Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & University of Toulouse - Jean Jaurès
ParCoLab Project (http://parcolab.univ-tlse2.fr/)

Wikimorph-sr: a morphosyntactic lexicon for Serbian based on the serbo-croatian Wiktionary

This is the documentation for the wikimorph-sr morphosyntactic lexicon for Serbian that can be used for POS-tagging, parsing and lemmatisation. The lexicon was developed as a part of the ParCoLab project (http://parcolab.univ-tlse2.fr/). It was mainly extracted from the serbo-croatian edition of the Wiktionary (sh.wiktionary.com).

Author

Aleksandra Miletic (UMR 5263 CLLE-ERSS, CNRS & University of Toulouse - Jean Jaurès, France)
Contact: aleksandra.miletic at univ-tlse2.fr

General description

Number of wordforms: 1 226 638
Number of lemmas: 117 445
Number of unique triples <*wordform, lemma, morphosyntactic description*>: 3 066 214

Format: [wordform][lemma][morphosyntactic description]
Field separator: tab (\t)
Character encoding: UTF-8
EOL character: LF (\n)

Source: Contents of the serbo-croatian Wiktionary (sh.wiktionary.com) in the form of the XML dump from 02/10/2015.

License

Referencing

If you are using this lexicon in your own research, please reference the following paper:

Miletic, A. (2017) *Building a morphosyntactic lexicon for Serbian from Wiktionary.* 6e édition des Journées d'étude toulousaines (JéTou 2017). Toulouse, France.

Acknowledgments

Many thanks to Franck Sajous (UMR 5263 CLLE-ERSS, CNRS & University of Toulouse – Jean Jaurès, France) for sharing his experience in working on the Wiktionary.

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & University of Toulouse - Jean Jaurès
ParCoLab Project (http://parcolab.univ-tlse2.fr/)

Content description

This lexicon was extracted from the serbo-croatian version of the Wiktionary. It was built with POS-tagging and parsing in mind, so it focuses on morphosyntactic information. Each line contains a unique triple <*wordform, lemma, morphosyntactic description*>. The morphosyntactic tags are the ones used in the ParCoLab project. A detailed description of the tag structure and possible feature values for each POS is given below.

Nouns

Examples:

strancem	stranac	N_com_ins_sg_m
Ivanom	Ivan	N_prop_ins_sg_0

Tag structure: [POS]_[subcategory]_[case]_[number]_[gender]
Possible values: N_(com|prop|col)_(nom|gen|dat|acc|voc|ins|loc)_(sg|pl)_(m|f|n)

/!\ The 0 value in the gender slot indicates that the gender information was absent from Wiktionary.

Adjectives

Examples:

najlepršavija	lepršav	A_qual_acc_pl_n_sup
mojega	moj	A_pos_gen_sg_m_-

Tag structure: [POS]_[subcategory]_[case]_[number]_[gender]_[degree of comparison]
Possible values: A_(qual|pos|dem|indef|inter|rel)_(nom|gen|dat|acc|voc|ins|loc)_(sg|pl)_(m|f|n)_(pos|comp|sup|-)

Only the qualificative adjectives are marked for degree of comparison.

Verbs

Examples:

abdiciraš	abdicirati	V_main_pres_2_sg_-_-
rabljeni	rabiti	V_main_partpass_-_pl_m_-

Tag structure: [POS]_[main or auxiliary]_[form]_[person]_[number]_[gender]_[negation]
Possible values: V_(main|aux)_(pres|aor|fut|imper|impf|inf|partact|partpass|partpast|partpres)_(1|2|3|-)_(sg|pl)_(m|f|n)_(neg|-)

Impersonal forms such as infinitive and active, passive, present and past participle are not marked for person. Personal forms (all other forms) are not marked for gender. Negation is indicated on synthetic forms such as *nemam* 'I don't have', *neću* 'I won't', *nisam* 'I am not', etc.

The active participle (*partact*) indicates the forms in *-o, -la, -lo, -li, -le, -la*, or *glagolski pridev radni.*
The passive participle (*partpass*) indicates the forms in *-n, -na, -no, -ni, -ne, -na* (and, alternatively, in *-t, -ta, -to, -ti, -te, -ta*), or *glagolski pridev trpni*.

The present participle (*partpres*) indicates the form in *-ći*, or *glagloski prilog sadašnji.*
The past participle (*partpast*) indicates the form in *-vši*, or *glagolski prilog prošli*.

Pronouns
Examples:

ga	on	P_pers_3_sg_m_gen
one	onaj	P_dem_-_pl_f_nom

Tag structure: [POS]_[subcategory]_[person]_[number]_[gender]_[case]
Possible values: P_(dem|indef|inter|pers|pos|rel)_(1|2|3|-)_(sg|pl)_(m|f|n)_(nom|gen|dat|acc|voc|ins|loc)

Numerals
Examples:

dvama	dva	Num_card_n_pl_ins
jednih	jedan	Num_card_n_pl_gen

Tag structure: [POS]_[subcategory]_[gender]_[number]_[case]
Possible values: Num_(card|ord|col)_(m|f|n)_(sg|pl)_(nom|gen|dat|acc|voc|ins|loc)

Adverbs
Examples:

agresivnije	agresivno	Adv_gen_comp
privatno	privatno	Adv_gen_pos

Tag structure: [POS]_[subcategory]_[degree of comparison]
Possible values: Adv_(gen|indef|rel|inter)_(pos|comp|sup|-)

Only the general adverbs are marked for degree of comparison.

Conjunctions
Examples:

jer	jer	C_sub
ali	ali	C_coor

Tag structure: [POS]_[subcategory]
Possible values: C_(sub|coor)

Prepositions
Examples:

ispod	ispod	Prep
prema	prema	Prep

Tag structure: Prep

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & University of Toulouse - Jean Jaurès
ParCoLab Project (http://parcolab.univ-tlse2.fr/)

Interjections
Examples:
ah ah I
hop hop I

Tag structure: I

Particles
No particles in the lexicon.

Tag structure: Part

Meaning of the feature values

Subcategories for nouns	
Feature value	**Meaning**
com	common
prop	proper
col	collective

Subcategories for adjectives	
Feature value	**Meaning**
qual	qualificative
pos	possessive
dem	demonstrative
indef	indefinite
inter	interrogative
rel	relative

Subcategories for verbs	
Feature value	**Meaning**
main	main
aux	auxiliary

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & University of Toulouse - Jean Jaurès
ParCoLab Project (http://parcolab.univ-tlse2.fr/)

Subcategories for pronouns	
Feature value	**Meaning**
pers	personal
pos	possessive
dem	demonstrative
indef	indefinite
inter	interrogative
rel	relative

Subcategories for numerals	
Feature value	**Meaning**
card	cardinal
ord	ordinal
col	collective

Subcategories for adverbs	
Feature value	**Meaning**
gen	general
indef	indefinite
rel	relative
inter	interrogative

Subcategories for conjunctions	
Feature value	**Meaning**
sub	subordinate
coor	coordinate

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & University of Toulouse - Jean Jaurès
ParCoLab Project (http://parcolab.univ-tlse2.fr/)

Case	
Feature value	**Meaning**
nom	nominative
gen	genitive
dat	dative
acc	accusative
voc	vocative
ins	instrumental
loc	locative

Gender	
Feature value	**Meaning**
m	masculine
f	feminine
n	neuter

Number	
Feature value	**Meaning**
sg	singular
pl	plural

Degree of comparison	
Feature value	**Meaning**
pos	positive
comp	comparative
sup	superlative

Person	
Feature value	**Meaning**
1	first
2	second
3	third

Verb form	
Feature value	**Meaning**
pres	present
aor	aorist
fut	future
imper	imperative
impf	imperfect
inf	infinitive
partact	active participle (*glagolski pridev radni*)
partpass	passive participle (*glagolski pridev trpni*)
partpast	past participle (*glagolski prilog prosli*)
partpres	present participle (*glagolski prilog sadasnji*)

Negation	
Feature value	**Meaning**
neg	negated

The presence of a dash (-) at a feature slot indicates that the feature does not apply to the wordform in question.

Caveat
Certain nouns (and especially proprer nouns) did not have a gender indication in their Wiktionary articles. The gender slot in their tags carries the 0 value, in order to indicate that the information was missing from the source rather than the trait doesn't apply to the noun.

Certain Wiktionary articles are generated through inflectional patterns that seem to be applied indiscriminately to all lemmas of the corresponding POS. As a consequence, there is a certain

amount of noise in the lexicon. This is especially true for the adjectives, for which the comparative and superlative forms are systematically present, even though the semantics of the adjective do not allow for comparison (cf. *alfabetskiji*, meaning 'more alphabetical', or *bakterijskiji*, meaning 'more bacterial').

The lexicon also contains an important amount of proper nouns: 355 178, which is more than 10% of the entries.

Given the fact that it was extracted from the serbo-croatian Wiktionary, it can contain forms with two reflections of the Old Slavic *yat*: those with '(i)je' as well as those with 'e'. For the same reason, it contains a certain amount of foreign proper names in their original spelling, in accordance with the croatian orthography.