Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Wikimorph-sr : un lexique morphosyntaxique pour le serbe basé sur l'édition serbo-croate du Wiktionary

Ceci est la documentation du lexique morphosyntaxique wikimorph-sr destiné à l'étiquetage morphosyntaxique, le parsing et la lemmatisation. Le lexique a été créé dans le cadre du projet ParCoLab (http://parcolab.univ-tlse2.fr/). Le contenu a été majoritairement dérivé de l'édition serbo-croate du Wiktionary (sh.wiktionary.com).

Auteur

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

Descriptif général

Nombre de formes fléchies : 1 226 638
Nombre de lemmes : 117 445
Nombre de triplets uniques <*forme fléchie, lemme, description morphosyntaxique*> : 3 066 214

Format: [forme fléchie][lemme][description morphosyntaxique]
Séoarateur de champs : tab (\t)
Encodage de caractères : UTF-8
Caractère de fin de ligne : LF (\n)

Source: Le contenu du Wiktionary pour le serbo-croate (sh.wiktionary.com) sous forme de l'XML dump datant du 02/10/2015.

Licence

Références

Si vous utilisez le lexique dans votre travail, vous êtes prié(e) de bien vouloir citer l'article suivant :

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

Remerciements

Je remercie chaleureusement Franck Sajous (UMR 5263 CLLE-ERSS, CNRS & Université de Toulouse – Jean Jaurès, France) d'avoir partagé avec moi ses expériences dans le travail sur le Wiktionary.

Descriptif du contenu

Ce lexique a été extrait à partir de la version serbo-croate du Wiktionary. Il est destiné à l'étiquetage morphosyntaxique, à la lemmatisation et au parsing, et par conséquent, il se concentre sur les

informations morphosyntaxiques. Chaque ligne contient un triplet unique *<forme fléchie, lemme, description morphosyntaxique>*. Les étiquettes morphosyntaxiques utilisées sont celles du projet ParCoLab. Une description détaillée de la structure des tags pour chaque partie du discours, ainsi que des valeurs possibles de différents traits, est donnée dans la suite.

Noms
Exemples:

strancem	stranac	N_com_ins_sg_m
Ivanom	Ivan	N_prop_ins_sg_0

Structure de l'étiquette : [POS]_[sous-catégorie]_[cas]_[nombre]_[genre]
Valeurs possibles : N_(com|prop|col)_(nom|gen|dat|acc|voc|ins|loc)_(sg|pl)_(m|f|n)

/!\ La valeur 0 dans la position du trait *genre* indique que l'information sur le genre n'était pas présent dans l'article du Wiktionary pour le nom donné.

Adjectifs
Exemples:

najlepršavija	lepršav	A_qual_acc_pl_n_sup
mojega	moj	A_pos_gen_sg_m_-

Structure de l'étiquette : [POS]_[sous-catégorie]_[cas]_[nombre]_[genre]_[degré de comparaison]
Valeurs possibles : A_(qual|pos|dem|indef|inter|rel)_(nom|gen|dat|acc|voc|ins|loc)_(sg|pl)_(m|f|n)_(pos|comp|sup|-)

Seuls les adjectifs qualificatifs portent les marques du degré de comparaison.

Verbes
Exemples:

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

Structure de l'étiquette : [POS]_[main or auxiliary]_[form]_[personne]_[nombre]_[genre]_[négation]
Valeurs possibles : V_(main|aux)_(pres|aor|fut|imper|impf|inf|partact|partpass|partpast|partpres)_(1|2|3|-)_(sg|pl)_(m|f|n)_(neg|-)

Les formes impersonnelles, comme l'infintif et les participes actif, passif, présent et passé ne portent pas de marque de personne, mais portent celles du genre. Les formes personnelles (toutes les autres) portent les marques de personne, mais pas celles du genre. La négation est indiqué sur les formes synthétiques comme *nemam* 'je n'ai pas', *neću* 'je ne veux pas', *nisam* 'je ne suis pas'.

Le terme de participe actif désigne les formes en *-o, -la, -lo, -li, -le, -la*, ou autrement dit *glagolski pridev radni.*
Le terme de participe passif désigne les formes en *-n, -na, -no, -ni, -ne, -na* (ou bien en *-t, -ta, -to, -ti, -te, -ta*), ou autrement dit *glagolski pridev trpni*.
Le terme de participe présent indique la forme en *-ći*, ou *glagloski prilog sadašnji.*
Le terme de participe passé indique la forme en *-vši*, ou *glagolski prilog prošli*.

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Pronoms
Exemples:

ga	on	P_pers_3_sg_m_gen
one	onaj	P_dem_-_pl_f_nom

Structure de l'étiquette : [POS]_[sous-catégorie]_[personne]_[nombre]_[genre]_[cas]
Valeurs possibles : P_(dem|indef|inter|pers|pos|rel)_(1|2|3|-)_(sg|pl)_(m|f|n)_(nom|gen|dat|acc|voc|ins|loc)

Numéraux
Exemples:

dvama	dva	Num_card_n_pl_ins
jednih	jedan	Num_card_n_pl_gen

Structure de l'étiquette : [POS]_[sous-catégorie]_[genre]_[nombre]_[cas]
Valeurs possibles : Num_(card|ord|col)_(m|f|n)_(sg|pl)_(nom|gen|dat|acc|voc|ins|loc)

Adverbes
Exemples:

agresivnije	agresivno	Adv_gen_comp
privatno	privatno	Adv_gen_pos

Structure de l'étiquette : [POS]_[sous-catégorie]_[degré de comparaison]
Valeurs possibles : Adv_(gen|indef|rel|inter)_(pos|comp|sup|-)

Seuls les adverbes généraux portent les marques de degré de comparaison.

Conjonctions
Exemples:

jer	jer	C_sub
ali	ali	C_coor

Structure de l'étiquette : [POS]_[sous-catégorie]
Valeurs possibles : C_(sub|coor)

Prépositions
Exemples:

ispod	ispod	Prep
prema	prema	Prep

Structure de l'étiquette : Prep

Interjections
Exemples:

ah	ah	I
hop	hop	I

Structure de l'étiquette : I

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Particules
Pas de particules dans le lexique.
Structure de l'étiquette : Part

Signifcation de valeurs de traits

Sous-catégories des noms	
Valeur du trait	**Signification**
com	commun
prop	propre
col	collectif

Sous-catégories des adjectifs	
Valeur du trait	**Signification**
qual	qualificatif
pos	possessif
dem	démonstratif
indef	indéfini
inter	interrogatif
rel	relatif

Sous-catégories des verbes	
Valeur du trait	**Signification**
main	principal
aux	auxiliaire

Sous-catégories des pronoms	
Valeur du trait	**Signification**
pers	personnel
pos	possessif
dem	démonstratif
indef	indéfini
inter	interrogatif
rel	relatif

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Sous-catégories des numéraux	
Valeur du trait	**Signification**
card	cardinal
ord	ordinal
col	collectif

Sous-catégories des adverbes	
Valeur du trait	**Signification**
gen	général
indef	indéfini
rel	relatif
inter	interrogatif

Sous-catégories des conjonctions	
Valeur du trait	**Signification**
sub	subordonné
coor	coordonné

Cas	
Valeur du trait	**Signification**
nom	nominatif
gen	genitif
dat	datif
acc	accusatif
voc	vocatif
ins	instrumental
loc	locatif

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Genre	
Valeur du trait	**Signification**
m	masculin
f	feminin
n	neutre

Nombre	
Valeur du trait	**Signification**
sg	singulier
pl	pluriel

Degré de comparaison	
Valeur du trait	**Signification**
pos	positif
comp	comparatif
sup	superlatif

Personne	
Valeur du trait	**Signification**
1	première
2	deuxième
3	troisième

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Forme verbale	
Valeur du trait	**Signification**
pres	présent
aor	aoriste
fut	futur
imper	impératif
impf	imparfait
inf	infinitif
partact	participe actif (*glagolski pridev radni*)
partpass	participe passif (*glagolski pridev trpni*)
partpast	participe passé (*glagolski prilog prosli*)
partpres	participe présent (*glagolski prilog sadasnji*)

Négation	
Valeur du trait	**Signification**
neg	négation présente

La présence d'un tiret (-) à la position d'un trait indique que le trait ne s'applique pas à la forme en question.

Caveat

Certains noms (notamment les noms propres) n'étaient pas accompagnés d'une indication de genre dans leur entrée dans Wiktionary. Dans leurs étiquette, à la place du genre se trouve la valeur 0, pour indiquer que l'information était manquante, et non pas qu'elle ne s'applique pas à la forme donnée.

Aleksandra Miletic
UMR 5263 CLLE-ERSS
CNRS & Université de Toulouse – Jean Jaurès
Projet ParCoLab (http://parcolab.univ-tlse2.fr/)

Certains articles du Wiktionary sont générés à travers des schémas de flexion qui semblent être appliqués à tous les lemmes de la partie du discours. Par conséquent, il existe un certain degré du bruit dans le lexique. Ceci est notamment le cas des adjectifs, pour lesquels les formes du comparatif et du superlatif sont systématiquement présentes, même pour les adjectifs relationnels (cf. a*lfabetskiji* 'plus alphabétique' ou *bakterijskiji* 'plus bactériel').

Le lexique contient également une part importante des noms propres : 355 178, ce qui repreésente plus de 10% d'entrées.

Etant donné que le lexique a été extrait de l'édition serbo-croate du Wiktionary, il peut contenir aussi bien des formes ékaviennes (celles en *e*) que des formes iékaviennes (elles en *(i)je*). Pour la même raison, il contient un certain nombre de noms propres étrangers en orthographe originale, suivant les règles de l'orthographe croate.