Guide d'annotation pour la lemmatisation de ParCoLab, v2.0

Aleksandra Miletic
CLLE-ERSS, Université de Toulouse - Jean Jaurès

17 avril 2018

Table des matières

1. Remarques introductives

Ce document s'articule comme suit : la première partie présente quelques principes généraux de lemmatisation adoptés dans le cadre du projet ParCoLab et propose une grille de lecture de ce guide. La deuxième partie définit les traitements mis en place pour différentes parties du discours. Enfin, la troisième contient des listes des solutions adoptées pour certains cas de figure problématiques.

1.1 Principes généraux adoptés

À la différence de l'étiquetage morphosyntaxique et de l'annotation syntaxique, qui font appel à des jeux d'étiquettes et des règles d'annotation complexes, la lemmatisation consiste simplement à identifier la forme canonique de chaque forme fléchie dans un corpus. Un exemple de la tâche est donné dans le tableau 1.1.

Token	**Lemme**
Filip	Filip
studira	studirati
lingvistiku	lingvistika
u	u
Italiji	Italija

Table 1.1 – Exemple de lemmatisation

L'ouvrage de référence pour ce travail sera le dictionnaire électronique du serbe de Simić (2005). Les annotateurs sont invités à s'en servir pour vérifier les lemmes dont ils ne sont pas certains.

Les règles de lemmatisation adoptées pour chaque partie du discours seront présentées dans la partie 2. Pour différentes parties du discours, nous définissons la forme qui est considérée comme lemme et donnons ensuite les règles de traitement de certains cas particuliers ou problématiques.

2. Règles de lemmatisation

2.1 Les noms

Le lemme d'un nom correspond à son **nominatif singulier** (cf. tableau 2.1).

Token	**Lemme**
knjigama 'livre.INS.PL'	knjiga 'livre.NOM.SG'
pevačima 'chanteurs.DAT.PL'	pevač 'chanteur.NOM.SG'
sela 'village.GEN.SG'	selo 'village.NOM.SG'

TABLE 2.1 – Nom : exemples de lemmatisation

Pour les noms féminins qui désignent un métier ou une fonction dérivés d'un nom masculin (cf. *pevač* 'chanteur' *vs pevačica* 'chanteuse'), le lemme est le nominatif singulier **féminin**. Donc, *pevačicom* 'chanteuse.INS.SG' doit être lemmatisé comme *pevačica*.

2.2 Les adjectifs

Le lemme d'un adjectif correspond à son **nominatif singulier masculin du positif** (cf. tableau 2.2).

Token	**Lemme**
lepom 'beau.INS.SG.F'	lep 'beau.NOM.SG.M'
takva 'ce.ACC.PL.F'	takav 'ce.NOM.SG.M'
mojoj 'mon.DAT.SG.F'	moj 'mon.NOM.SG.M'
najlepšoj 'beau.DAT.SG.F.SUP'	lep 'beau.DAT.SG.F.POS'

TABLE 2.2 – Adjectif : exemples de lemmatisation

Comme le montrent les exemples du tableau 2.2, la même règle s'applique aux adjectifs qualificatifs, ainsi qu'à toute autre sous-catégorie (démonstratifs, possessifs, indéfinis, relatifs, interrogatifs).

Par ailleurs, c'est typiquement le nominatif singulier du masculin de l'aspect **indéfini** qui est utilisé (cf. *lep*, et non pas *lepi*). Cependant, certains adjectifs – et notamment les

adjectifs relationnels, dits *prisvojni pridevi* en serbe – n'ont pas de formes de l'indéfini et sont cités donc au nominatif singulier du masculin du **défini** (cf. *seoski* 'villageois', *alfabetski* 'alphabétique').

La frontière entre ces deux ensembles d'adjectifs n'est pas clairement déterminée : pour certains adjectifs massifs (*gradivni pridevi*), les deux lemmes sont possibles (cf. *mermeran/mermerni* 'en marbre', *papiran/papirni* 'en papier', *kristalan/kristalni* 'en cristal'). Il en est de même pour certains adjectifs qualificatifs, rarement utilisés à l'indéfini pour des raisons sémantiques (cf. *davan/davni* 'ancien', *divalj/divlji* 'sauvage'), etc. En rencontrant un cas de figure de ce type, les annotateurs sont invités à consulter le dictionnaire de Simić (2005). Si l'indéfini est reconnu comme possible, c'est cette forme-là qui doit être utilisée.

Par ailleurs, une liste des formes adjectivales déjà vérifiées est proposée dans la section 3.1. Si les annotateurs rencontrent un adjectif qui ne figure pas dans cette liste, ils le signaleront à l'annotateur expérimenté de sorte qu'il puisse l'intégrer dans la prochaine version du guide.

2.3 Les verbes

Le lemme d'un verbe correspond à son **infinitif** (cf. tableau 2.3).

Token	**Lemme**
jedemo 'mangeons'	jesti 'manger'
pojedemo 'mangeons.PERF'	pojesti 'manger.PERF'
ješćemo 'mangerons'	jesti 'manger'

Table 2.3 – Verbe : exemples de lemmatisation

Une attention spéciale doit être accordée à la question de l'aspect : il faut veiller à choisir le lemme approprié, notamment dans les cas des verbes qui ont des séries aspectuelles bien développées. Par exemple :

— sedim 'je.suis.assis' => **sedeti** 'être.assis'

— sednem 'je.m'assois.PERF' => **sesti** 's'asseoir.PERF'

— sedam 'je.m'assois.IMPERF' => **sedati** 's'asseoir.IMPERF'

— ležim 'je.suis.couché' => **ležati** 'être.couché'

— legnem 'je.me.couche.PERF' => **leći** 'se.coucher.PERF'

— ležem 'je.me.couche.IMPERF' => **legati** 'se.coucher.IMPERF'

2.3.1 Traitement des verbes *jesam* et *biti*

Il existe en serbe deux équivalents du verbe 'être' : *jesam* 'je suis' (et sa forme négative *nisam* 'je ne suis pas') et *biti* 'être'. Le verbe *jesam* est un verbe défectif : il existe seulement au présent et ne dispose pas d'un infinitif, et c'est lui qui exprime le présent indicatif. Le verbe *biti* est un verbe régulier, qui dispose d'un paradigme complet. À partir de ce critère morphosyntaxique, on considère traditionnellement qu'il s'agit de deux lemmes différents. Nous préservons cette distinction dans la lemmatisation. Par conséquent, les formes de la série *jesam, jesi, jeste...*, ainsi que les formes clitiques correspondantes (*sam, si, je...*) sont annotées comme ***jesam***. En revanche, les formes des séries *budem, budeš, bude...*, *bio, bila, bilo...*, *bih, bi, bi...*, *biću, bićeš, biće...*, etc., sont annotées comme ***biti***.

2.3.2 Traitement des verbes à négation synthétique

Certains verbes en serbe présentent des formes qui intègrent la négation de manière synthétique : *neću* 'je ne veux pas', *nisam* 'je ne suis pas', *nemam* 'je n'ai pas' et *nemoj* 'ne fais pas'.

Les formes de la série ***nemam, nemaš, nema...*** disposent également d'un infinitif indépendant (cf. *nemati* 'ne pas avoir' *vs imati* 'avoir'). Par conséquent, elles sont lemmatisées comme ***nemati***.

Les formes de la série ***neću, nećeš, neće...*** ne disposent pas d'un infinitif indépendant ; elles sont par conséquent lemmatisées comme ***hteti*** 'vouloir'.

Les formes de la série ***nisam, nisi, nije...*** n'en disposent pas non plus ; par conséquent, elles sont lemmatisées comme ***jesam*** (voir la section 2.3.1 pour une explication de la lemmatisation du verbe *jesam*).

Quant aux formes de la série ***nemoj, nemojmo, nemojte***, ce verbe est défectif : ces trois formes sont les seules dont il dispose. Par conséquent, elles sont lemmatisées comme ***nemoj***.

2.3.3 Traitement des participes

Les participes actif (*glagolski pridev radni*) et passif (*glagolski pridev trpni*) sont lemmatisés comme verbes quand ils sont accompagnés du verbe *jesam* 'être' ; autrement dit, quand ils font partie d'une forme verbale composée. Quand ils se trouvent à l'intérieur d'un groupe nominal, nous considérons qu'il s'agit des adjectifs ; leur lemme ne correspond donc pas à leur infinitif, mais au nominatif singulier masculin indéfini. Quant au participe présent (*glagolski prilog sadašnji*), il est annoté comme verbe s'il dépend d'un autre verbe, et comme adjectif s'il dépend d'un nom. En revanche, cette forme ne dispose pas de l'indéfini ; elle est donc lemmatisée au nominatif singulier masculin défini. Des

Token	Lemme
pas	pas
je	jesam
vezan	**vezati**
pred	pred
ulazom	ulaz
stoji	stajati
vezan	**vezan**
pas	pas

Token	Lemme
staze	staza
su	jesam
zarasle	**zarasti**
u	u
korov	korov
idu	ići
stazama	staza
zaraslilm	**zarastao**
u	u
korov	korov

Token	Lemme
izašao	izaći
je	jesam
trčeći	**trčati**
izašao	izaći
je	jesam
trčećim	**trčeći**
korakom	korak

TABLE 2.4 – Verbe : lemmatisation des participes

exemples sont donnés dans le tableau 2.4.

2.3.4 Lemmes doublons

Pour certains paradigmes, il existe deux formes de l'infinitif reconnues par la norme : la forme *brojim* 'je compte' peut correspondre à l'infinitif *brojati* ou à *brojiti* 'compter', *podignem* 'je soulève' peut correspondre à *podići* ou *podignuti* 'soulever', et *stojim* 'je me tiens debout' peut avoir comme infinitif *stojati* ou *stajati* 'se tenir debout', etc.

Dans le cas des verbes où on a le choix entre l'infinitif en *-ći* et celui en *-ti* (cf. *podići* et *podignuti*), nous choisissons systématiquement celui en *-ći*. Pour les autres cas, les infinitifs retenus sont notés dans une liste dans la section 3.2. Si un nouveau verbe avec un lemme doublon est rencontré, les annotateurs sont invités à le signaler à l'auteure du guide pour qu'elle détermine le lemme qui sera utilisé et l'intègre à la liste.

2.4 Les pronoms

Pour les **pronoms personnels**, le lemme correspond au **nominatif** du pronom en question. La personne et le nombre correspondent à celui de la forme fléchie. En ce qui concerne les formes de la **troisième personne**, on utilise systématiquement le **nominatif masculin**.

Pour les autres sous-catégories des pronoms (possessifs, relatifs, indéfinis, interrogatifs, démonstratifs), le **nominatif singulier masculin** est utilisé systématiquement. Des exemples sont donnés ci-dessous.

— ***nama*** 'nous.DAT' => ***mi*** 'nous.NOM' (et non pas *ja* 'je.NOM')

— ***njoj*** 'elle.DAT' => ***on*** 'il.NOM' (et non pas *ona* 'elle.NOM')

— ***njima*** ‘elles/ils.DAT’ => ***oni*** ‘ils.NOM’ (et non pas *one* ‘elles.NOM’)

— ***tim*** ‘ces.DAT’ => ***taj*** ‘ce.NOM.SG.M’

— ***čijih*** ‘de.qui.GEN.PL’ => ***čiji*** ‘de.qui.NOM.SG.M’

2.5 Autres catégories

Les numéraux

Pour les **numéraux variables**, le lemme correspond au **nominatif singulier masculin** : *dvema* ‘deux.DAT/INS.F’ => *dva* ‘deux.NOM.M’. Pour les **numéraux invariables**, le lemme correspond à la forme du token : *pet* ‘cinq’ => *pet*.

Si un numéral est écrit en chiffres, on reprend le token en tant que lemme.

Les adverbes

La majorité des adverbes étant invariables, le lemme d’un adverbe correspond typiquement au token trouvé dans le texte : *lako* ‘facilement’ => *lako*. La seule exception concerne les adverbes qui se comparent : le lemme d’un adverbe au **comparatif** ou au **superlatif** correspond à la forme du **positif** de cet adverbe : *lakše* ‘plus facilement’ => *lako* ‘facilement’, *najlakše* ‘le plus facilement’ => *lako* ‘facilement’.

Les prépositions

Certaines prépositions comme *ka* ‘vers’ et *sa* ‘avec’ disposent également des formes allomorphes courtes *k* et *s*. Pour la lemmatisation, nous utilisons systématiquement les formes longues.

Le pronom réflexif

Le pronom réflexif clitique *se* ‘se’ dispose d’une forme pleine *sebe*. Comme cette forme pleine est beaucoup moins fréquente en corpus, nous utilisons la forme brève pour la lemmatisation.

3. Traitement de cas de figure spécifiques

3.1 Liste des lemmes adjectivaux problématiques

La liste ci-dessous indique si c'est la forme de l'indéfini ou du défini qui a été retenue comme lemme. C'est la forme en **gras** qu'il faut utiliser.

davan *vs* **davni**
divalj *vs* **divlji**
dokazan *vs* dokazni (materijal)
drevan *vs* drevni
duhovan *vs* duhovni
dvojan *vs* dvojni
istočan *vs* **istočni**
izazovan *vs* izazovni
javan *vs* javni
južan *vs* **južni**
kasan *vs* **kasni**
kaznen *vs* **kazneni**
kristalan *vs* kristalni
ljubavan *vs* **ljubavni**
mermeran *vs* mermerni
minijaturan *vs* minijaturni
narodan *vs* **narodni**
nedostojan *vs* nedostojni
okolan *vs* **okolni**
okrutan *vs* okrutni
paran *vs* parni (kao u *parni brod*)
popodnevan *vs* **popodnevni**
preostao *vs* preostali
prvobitan *vs* prvobitni
ručan *vs* **ručni** (kao u *ručni sat*)
srebrn *vs* srebrni
starozavetan *vs* **starozavetni**
susedan *vs* susedni
svet *vs* sveti
vekovan *vs* **vekovni**
žarak *vs* **žarki**
zabrinjavajuć *vs* **zabrinjavajući**
zaslepljujuć *vs* **zaslepljujući**

3.2 Liste des lemmes verbaux problématiques

La liste ci-dessous indique le lemme verbal retenu pour les verbes qui ont des infinitifs doublons. C'est la forme en **gras** qu'il faut utiliser.

brojati *vs* brojiti
dići *vs* dignuti
dostići *vs* dostignuti
izbeći *vs* izbegnuti
izdići *vs* izdignuti
navići *vs* naviknuti
nići *vs* niknuti
odbeći *vs* odbegnuti
podići *vs* podignuti
postići *vs* postignuti
prepući *vs* prepuknuti
promaći *vs* promaknuti
razleći *vs* razlegnuti
razmaći *vs* razmaknuti
stajati *vs* stajati
steći *vs* steknuti
stići *vs* stignuti
uzdići *vs* uzdignuti
zadići *vs* zadignuti

Bibliographie

Milorad Simić. Srpski elektronski rečnik. http ://www.rasprog.com, 2005.