Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITE DE TOULOUSE

Délivré par :

Université Toulouse-Jean Jaurès

Présentée et soutenue par :
Aleksandra MILETIC

le mercredi 20 juin 2018

Titre :

Un treebank pour le serbe : constitution et exploitations

Tome 2 : Annexes

École doctorale et discipline ou spécialité :
ED CLESCO : Sciences du langage

Unité de recherche :
CLLE (UMR 5263)

Directeurs de Thèse :
Cécile FABRE, Université Toulouse-Jean Jaurès

Dejan STOSIC, Université Toulouse-Jean Jaurès

Jury :

SYLVAIN KAHANE	Professeur, Université Paris Nanterre	Rapporteur
PAOLA MERLO	Professeure, Université de Genève	Rapporteure
MARIE CANDITO	Maître de conférences, Université Paris Diderot	Examinatrice
VERAN STANOJEVIĆ	Professeur, Université de Belgrade	Examinateur
CECILE FABRE	Professeure, Université Toulouse-Jean Jaurès	Directrice
DEJAN STOSIC	Maître de conférences, Université Toulouse-Jean Jaurès	Directeur

Table des matières

Présentation du tome 2

Le deuxième tome de cette thèse regroupe les guides d'annotation utilisés dans la création du treebank ParCoTrain-Synt. Les principes d'annotation adoptés ayant déjà été présentés et discutés dans le tome 1 (chapitre 5), nous nous contentons ici de reprendre simplement les guides. Ils gardent donc chacun leur pagination, leur table des matières et leur bibliographie. L'annexe A contient donc le guide d'annotation morphosyntaxique, l'annexe B celui de la lemmatisation, et l'annexe **??** celui de l'annotation syntaxique.

Précisons encore qu'il s'agit de la version des guides la plus récente, établie à l'issu du travail sur ParCoTrain-Synt. Les guides contiennent donc toutes les modifications identifiées comme nécessaires dans le cadre des campagnes d'annotation manuelle.

Annexe A

Guide d'annotation morphosyntaxique

Guide d'annotation morphosyntaxique de ParCoLab, v2.0

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

17 avril 2018

Table des matières

1 Remarques introductives **5**

1.1 Comment lire ce guide 5

1.2 Quelques principes d'annotation morphosyntaxique de base 6

1.2.1 Correspondance tag-token 1 :1 6

1.2.2 Résolution des cas ambigus basée sur le contexte 7

2 Présentation du traitement par classe de mots **8**

2.1 Les noms 8

2.1.1 Les sous-catégories de noms 8

2.1.2 Noms du genre grammatical féminin désignant des êtres du sexe masculin 10

2.1.3 Noms composés contenant un trait d'union 10

2.2 Les adjectifs 11

2.2.1 Les sous-catégories d'adjectifs 11

2.2.2 Degré de comparaison de l'adjectif 13

2.2.3 Traits *genre* et *nombre* pour la sous-catégorie *prisvojni* 13

2.2.4 Adjectifs (pro)nominalisés 13

2.2.5 Adjectifs déverbaux dérivés des formes des participes passif et actif (*glagolski pridev trpni* et *glagolski pridev radni*) 14

2.2.6 Adjectifs composés 14

2.3 Les verbes 15

2.3.1 Trait *forme verbale* 15

2.3.2 Verbes auxiliaires dans les formes surcomposées 17

2.3.3 Trait *négation* 17

2.4 Les pronoms 19

2.4.1 Les sous-catégories des pronoms 20

2.4.2 Annotation du genre et du nombre 20

2.4.3 Pronoms indéfinis discontinus 20

2.4.4 Distinction entre le pronom personnel fléchi et l'auxiliaire 22

2.4.5 Le pronom *što* 22
2.5 Les numéraux 24
2.5.1 Le genre et le nombre des numéraux 24
2.5.2 Paucal 25
2.5.3 Formes en *-ak* 25
2.6 Les adverbes 26
2.6.1 Sous-catégories des adverbes 26
2.6.2 Degré de comparaison des adverbes 26
2.6.3 Forme *kad* ‘quand’ 27
2.7 Conjonctions 27
2.8 Les prépositions, les interjections, les particules, les mots étrangers et la ponctuation 28

3 Gestion des cas de figure spécifiques 29
3.1 Prépositions 29
3.2 Listes des particules 30
3.3 Autres cas de figure 32

Bibliographie 35

Liste des tableaux

1.1 Exemple d'annotation morphosyntaxique 5

2.1 Nom : traits utilisés et leurs valeurs possibles 8
2.2 Nom : distribution des traits en fonction de la sous-catégorie 9
2.3 Nom propre polylexical 9
2.4 Nom collectif 10
2.5 Nom de famille 10
2.6 Adjectifs : traits utilisés et leurs valeurs possibles 11
2.7 Adjectifs : distribution des traits en fonction de la sous-catégorie 11
2.8 Adjectif : exemples des sous-catégories différentes 12
2.9 Adjectif : degré de comparaison 13
2.10 Verbe : traits utilisés et leurs valeurs possibles 15
2.11 Verbe : distribution des traits en fonction de la forme verbale 16
2.12 Verbe auxiliaire : formes surcomposées 17
2.13 Pronom : traits utilisés et leurs valeurs possibles 19
2.14 Pronom : distribution des traits en fonction de la sous-catégorie 19
2.15 Pronom : genre et nombre 21
2.16 Pronom : discontinuité dans le GP 21
2.17 Pronom *vs* auxiliaire 22
2.18 Pronom *što* : exemple 1 23
2.19 Pronom *što* : exemple 2 23
2.20 Pronom *što* : exemple 3 23
2.21 Numéral : traits utilisés et leurs valeurs possibles 24
2.22 Numéral : distribution des traits en fonction de la sous-catégorie 24
2.23 Numéral *jedan* 24
2.24 Numéral *dva* : fléchi *vs* invariable 25
2.25 Adverbe : traits utilisés et leurs valeurs possibles 26
2.26 Adverbe : distribution des traits en fonction de la sous-catégorie 26
2.27 Conjonctions : traits utilisés et leurs valeurs possibles 27

2.28 Conjonctions : exemples des sous-catégories 27
2.29 Autres classes : étiquettes POS 28

1. Remarques introductives

Ce document s'articule comme suit : la première partie présente quelques principes généraux de l'annotation morphosyntaxique telle que définie dans le cadre du projet ParCoLab et propose une grille de lecture de ce document. La deuxième partie est la plus conséquente : elle contient la présentation du traitement adopté pour chaque classe de mots. Chaque sous-section donne les traits morphosyntaxiques traités pour la classe de mots donnée, les valeurs possibles des traits et les règles d'annotation pour les cas de figure nécessitant un traitement spécifique. La dernière partie du document contient l'indication des traitements adoptés pour certaines constructions problématiques. Dans cette partie aussi, les règles d'annotation sont organisées selon la classe de mots.

1.1 Comment lire ce guide

L'annotation morphosyntaxique de ParCoLab se fait en utilisant un tableur (un fichier Excel). Les fichiers à traiter se présentent dans un format verticalisé : une ligne correspond à un token du texte. Elle contient le token lui-même dans la première colonne, suivi de son annotation morphosyntaxique dans les colonnes suivantes.

Le traitement morphosyntaxique défini dans le cadre du projet ParCoLab met en place une annotation riche en informations : au-delà de l'indication de la classe et de la sous-catégorie distributionnelle (cf. nom commun, pronom démonstratif), les étiquettes indiquent également les traits morphosyntaxiques pertinents pour chaque classe de mots. À titre d'illustration, les informations accordées aux noms comprennent également l'indication du genre, du nombre et du cas, alors que le traitement des verbes fait appel à la personne, le nombre, le genre, la forme verbale et la présence ou l'absence de la négation. Dans le cadre de l'annotation manuelle, on utilise des indications explicites des valeurs de ces traits : elles sont épelées en toutes lettres (cf. tableau 1.1).

vidim	V	glavni	prezent	prvo lice	jednina	---	---
devojku	N	zajednicka	akuzativ	jednina	zenski rod		

Table 1.1 – Exemple d'annotation morphosyntaxique

Notons que ce ne sont pas les mêmes traits morphosyntaxiques qui sont indiqués pour les deux classes de mots. Par conséquent, les mêmes colonnes ne correspondent pas aux mêmes traits dans les deux cas. Or, pour que l'annotation manuelle soit exploitable par la suite, il est **essentiel** que les traits soient indiqués toujours dans le même ordre pour une classe de mots donnée. Dans la suite du document, nous présentons donc systématiquement les traits **dans l'ordre** dans lequel ils doivent apparaître dans l'annotation manuelle.

Pour faciliter ce travail, une macro VisualBasic a été implémentée dans les fichiers d'annotation manuelle. Cette macro permet la modification des en-têtes des colonnes en fonction de la valeur de la colonne contenant l'indication de la classe de mots. La modification s'active en passant d'une ligne du tableur à l'autre avec les flèches ↑ et ↓ sur le clavier. Si cette fonctionnalité ne démarre pas à l'ouverture du fichier, il est possible de la lancer en utilisant le raccourci **CTRL + q**, ou bien en allant dans l'onglet **Affichage/Macros/Affichage des macros** et en cliquant sur le bouton **Exécuter**.

Dans la partie 2 de ce document, nous définissons le traitement de chaque classe de mots. Nous présentons tous les traits morphosyntaxiques encodés, ainsi que leurs valeurs possibles. Dans le cadre de l'annotation manuelle, il est essentiel d'utiliser exactement les mêmes valeurs : attention à l'absence des diacritiques serbes.

Pour chaque classe de mots, dans un premier tableau nous présentons tous les traits utilisés, ainsi que toutes leurs valeurs possibles. Ici, les traits sont présentés dans le même ordre dans lequel ils doivent être utilisés en annotation manuelle. Ensuite, dans un deuxième tableau, nous définissons la distribution des traits en fonction de la sous-catégorie morphosyntaxique : souvent, certaines sous-catégories présentent des restrictions quant à certains traits.

Si une case du tableau est renseignée comme '[tous]', ceci signifie que toutes les valeurs du trait donné sont valides pour la sous-catégorie en question. Si la case contient '---', ceci veut dire que le trait donné n'est pas marqué dans la sous-catégorie donnée ; la même indication doit être reproduite dans l'annotation manuelle. Enfin, si la case contient une ou plusieurs valeurs concrètes (*mnozina* 'pluriel', *drugo lice* '2e personne'), ces valeurs seules sont valides pour la sous-catégorie donnée.

1.2 Quelques principes d'annotation morphosyntaxique de base

1.2.1 Correspondance tag-token 1 :1

L'identification des unités à annoter au niveau morphosyntaxique est basée sur le principe suivant : une forme orthographique ne peut porter qu'une seule étiquette, et

une étiquette ne peut être attachée qu'à une seule forme à la fois. Ce principe permet la réalisation d'un étiquetage du premier niveau (chaque forme orthographique se voit attribuer un tag), qui, du point de vue linguistique, n'est pas toujours correct : on dissocie ainsi les unités polylexicales et on leur impose une interprétation analytique. Pourtant, un étiquetage qui prendrait en compte les unités polylexicales serait problématique à ce stade, premièrement pour la question de la définition des unités polylexicales, mais aussi à cause de l'existence des unités discontinues en serbe (notamment les formes verbales complexes). L'approche adoptée permet d'éviter ces problèmes et facilite ainsi un maintien de cohérence de l'annotation. Cette première annotation pourra donner lieu ensuite à un traitement plus pertinent du point de vue linguistique.

1.2.2 Résolution des cas ambigus basée sur le contexte

Une partie de l'ambiguïté relevée dans le traitement d'un corpus peut provenir des formes capables d'assumer les rôles prototypiques de plusieurs classes de mots. Si un token, qui appartient typiquement à une classe de mots, prend dans le contexte le comportement d'une autre catégorie grammaticale, il est annoté selon sa fonction syntaxique. L'exemple typique est celui des adjectifs qui se trouvent nominalisés dans un contexte spécifique : le mot *mrtav* 'mort' est un adjectif ; cependant, dans l'exemple *razmišljati o mrtvima* 'penser aux morts', il est employé indépendamment d'un nom et fonctionne lui-même comme un groupe nominal. Par conséquent, il sera annoté comme un nom commun. D'une manière comparable, la forme *iza* 'derrière', qui est typiquement une préposition, n'introduit pas un groupe prépositionnel dans l'exemple *Parkirao sam se iza* 'Je me suis garé derrière'. Ici cette forme connaît un emploi adverbial, et peut être remplacée par un adverbe locatif typique comme *daleko* 'loin'. Par conséquent, dans ce contexte, *iza* sera annoté comme adverbe.

Les cas de figure de ce type qui sont relativement systématiques sont rappelés et illustrés dans les sections dédiées aux classes de mots.

2. Présentation du traitement par classe de mots

Cette partie du guide est dédiée à la présentation du traitement de chaque classe de mots. Dans chaque section, nous indiquons d'abord les traits moprhosyntaxiques pertinents et l'ordre dans lequel ils se présentent dans le cadre de l'annotation manuelle, ainsi que les valeurs valides pour chaque trait. Nous indiquons ensuite les spécificités du traitement si celui-ci ne coïncide pas avec la tradition grammaticale serbe. Si nécessaire, certains cas de figure particuliers sont également traités.

2.1 Les noms

Le traitement des noms fait appel aux traits morphosyntaxiques suivants : la classe de mots, la sous-catégorie, le cas, le nombre et le genre. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.1, alors que la distribution des traits en fonction de la sous-catégorie est présentée dans le tableau 2.2.

POS	**Sous-cat.**	**Cas**	**Nombre**	**Genre**
N	zajednicka vlastita zbirna	nominativ genitiv dativ akuzativ vokativ instrumental lokativ	jednina mnozina	muski rod zenski rod srednji rod

TABLE 2.1 – Nom : traits utilisés et leurs valeurs possibles

2.1.1 Les sous-catégories de noms

Stanojčić & Popović (2012) distinguent 6 sous-catégories de noms : noms propres (*vlastite*), communs (*zajedničke*), massifs (*gradivne*), abstraits (*apstraktne*) et déverbaux (*glagolske*) (équivalentes des mêmes sous-catégories en français), ainsi que les noms collectifs (*zbirne*), qui en serbe, à la différence du français et de l'anglais, comprennent les

Exemples	**POS**	**Sous-cat.**	**Cas**	**Nombre**	**Genre**
noms communs (cf. *konj* 'cheval', *stolica* 'chaise')	N	zajednicka	[tous]	[tous]	[tous]
noms propres (cf. *Beograd* 'Belgrade')	N	vlastita	[tous]	[tous]	[tous]
noms collectifs (cf. *granje* 'branches', *pilad* 'poussins')	N	zbirna	[tous]	jednina	[tous]

TABLE 2.2 – Nom : distribution des traits en fonction de la sous-catégorie

noms ayant la forme et le comportement d'un nom au singulier, mais le sémantisme du pluriel (cf. *pilad* 'poussins', qui désigne un ensemble de poussins, mais se décline comme le singulier du nom *noć* 'nuit').

Les noms massifs (*voda* 'eau', *brašno* 'farine'), abstraits (*ljubav* 'amour', *mržnja* 'haine') et déverbaux (*crtanje* 'dessin', *pevanje* 'chant') ne se distinguent pas par leur comportement morphologique des autres noms communss : ils suivent les mêmes modèles de déclinaison. Par conséquent, il a été décidé d'annoter les 4 sous-classes comme noms communs (valeur du trait sous-catégorie ***zajednicka***).

La distinction des noms propres a été gardée : dans la suite de l'enrichissement du corpus, cette propriété peut être exploitée dans la reconnaissance des entités nommées. La valeur du trait *sous-catégorie* pour ces noms est donc ***vlastita***. Cette étiquette s'applique aux noms des personnes et des animaux (cf. *Marko, Jovanović, Pera, Žuća* etc.), mais aussi aux noms des entités géographiques (cf. *Beograd, Italija*) ou des institutions (cf. *RTS*, *Kolarac*, etc.). Dans le cas d'un nom propre polylexical (cf. *Novi Sad*, *Fruška gora*), c'est la tête nominale qui est annotée comme nom propre, alors que le modifieur (dans ce cas, l'adjectif qualificatif) est traité en fonction de sa classe grammaticale. Un exemple est donné dans le tableau 2.3.

od	Prep					
Fruške	A	opsti	genitiv	jednina	zenski rod	pozitiv
gore	N	vlastita	genitiv	jednina	zenski rod	

TABLE 2.3 – Nom propre polylexical

Les noms collectifs sont distingués pour leur comportement spécifique dans l'accord. Ils sont par ailleurs identifiables par leurs terminaisons (cf. *-nje* dans *granje* 'ensemble de branches', *-ad* dans *pilad* 'ensemble de poussins', etc.) La valeur du trait sous-catégorie pour ces noms est donc ***zbirna*** (cf. tableau 2.4).

En ce qui concerne les autres traits :

po	Prep				
lišću	N	zbirna	lokativ	jednina	srednji rod

TABLE 2.4 – Nom collectif

— toutes les valeurs de tous les traits sont possibles pour la sous-catégorie ***zajednicka*** ;

— les noms collectifs (sous-catégorie ***zbirna***) sont systématiquement annotés comme singulier vu le patron de déclinaison qu'ils suivent ;

— pour les noms propres (sous-catégorie ***vlastita***), toutes les valeurs de tous les traits s'appliquent, à l'exception des noms de famille des femmes, qui sont invariables et dont il est impossible de déterminer le genre à partir des critères de surface (cf. *gospođa Petrović*, *gospođe Petrović* etc.). De manière arbitraire mais systématique, ces formes-là sont traitées comme un nominatif singulier, sans indication de genre.

gospođe	N	zajednicka	genitv	jednina	zenski rod
Petrović	N	vlastita	nominativ	jednina	---

TABLE 2.5 – Nom de famille

2.1.2 Noms du genre grammatical féminin désignant des êtres du sexe masculin

Le genre des noms tels *deda* 'grand-père', *tata* 'papa', *vladika* 'évèque' etc. est annoté selon leur genre naturel, vu que c'est souvent celui-ci qui impose l'accord aux dépendants du nom.

Dans le cas des noms tels *pijanica* 'ivrogne', *sudija* 'juge', etc., qui sont du genre grammatical féminin et qui peuvent désigner des êtres des deux sexes, le genre est annoté en fonction du contexte (par exemple, le genre imposé à un adjectif qui s'accorde avec le nom). Si aucun indice de ce type n'est disponible, le nom est annoté comme masculin.

2.1.3 Noms composés contenant un trait d'union

Dans la phase de la tokénisation du corpus, les noms composés contenant un trait d'union (cf. *žena-leptir* 'femme-papillon') ont été séparés. Dans la phase de l'annotation morphosyntaxique, il faut corriger ce fait en fusionnant les deux lignes de sorte à rétablir le token initial relié par un trait (celui de la touche 6 sur le clavier français). Les traits morpho-syntaxiques de ces formes sont ensuite annotés en considérant la forme complexe.

2.2 Les adjectifs

Le traitement des adjectifs fait appel aux traits suivants : la classe de mots, la sous-catégorie, le cas, le nombre, le genre et le degré de comparaison. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.6, alors que la distribution des traits en fonction de la sous-catégorie est présentée dans le tableau 2.7.

POS	**Sous-cat.**	**Cas**	**Nombre**	**Genre**	**Degré de comp.**
A	opisni pokazni prisvojni neodredjeni upitni odnosni	nominativ genitiv dativ akuzativ vokativ instrumental lokativ	jednina mnozina	muski rod zenski rod srednji rod	pozitiv komparativ superlativ

TABLE 2.6 – Adjectifs : traits utilisés et leurs valeurs possibles

Exemples	**POS**	**Sous-cat.**	**Cas**	**Nombre**	**Genre**	**Degré de comp.**
qualificatifs (cf. *lep* 'beau', *letnji* 'estival')	A	opisni	[tous]	[tous]	[tous]	[tous]
autres (*taj* 'ce', *neka* '(une) certaine', *kakvo* 'quel')	A	pokazni prisvojni neodredjeni upitni odnosni	[tous]	[tous]	[tous]	---

TABLE 2.7 – Adjectifs : distribution des traits en fonction de la sous-catégorie

2.2.1 Les sous-catégories d'adjectifs

Le traitement des adjectifs dans le cadre de ce projet diffère de celui proposé par les grammaires traditionnelles serbes. Par exemple, Stanojčić & Popović (2012) distinguent les adjectifs qualificatifs (*lep* 'beau', *hrabar* 'courageux'), possessifs (*Markov* 'qui appartient à Marko', *školski* 'qui appartient à/relatif à l'école'), massifs (*zlatan* 'en or', *drven* 'en bois'), temporels (*današnji* 'd'aujourd'hui', *godišnji* 'annuel') et spatiaux (*desni* 'de droite', *gornji* 'supérieur', 'd'en haut') (2012 : 93). Or, cette classification repose sur des critères sémantiques. Par ailleurs, tous les types d'adjectifs énumérés partagent le même

comportement morphosyntaxique. Il a donc été décidé de ne pas introduire ces distinctions sémantiques dans notre jeu d'étiquettes : tous ces sous-types d'adjectifs sont traités comme des adjectifs qualificatifs et leur trait *sous-catégorie* a la valeur ***opisni***.

Sont également traitées comme adjectifs les formes traditionnellement identifiées comme *pridevske zamenice* 'pronoms adjectivaux', quand elles figurent au sein d'un GN. Cette décision est motivée par le fait que ces formes ont le comportement adjectival typique : elles suivent les mêmes patrons de déclinaison que les adjectifs, et à l'intérieur d'un groupe nominal, elles occupent la position canonique d'un adjectif – à gauche du nom. Les valeurs possibles du trait sous-catégorie comprennent donc aussi les valeurs ***pokazni*** 'démonstratif', ***prisvojni*** 'possessif', ***neodredjeni*** 'indéfini', ***upitni*** 'interrogatif' et ***odnosni*** 'relatif'.

La valeur ***pokazni*** regroupe les formes des séries *taj, ovaj, onaj* ; ***prisvojni*** s'applique aux paradigmes de *moj, tvoj, njegov, njen, svoj*, etc., la valeur ***neodredjeni*** 'indéfini' regroupe les formes traditionnellement qualifiées comme pronoms généraux (*opšte*), négatifs (*odričnе*) et indéfinis (*neodređene*) et s'applique donc aux paradigmes de *svaki, svakakav, svačiji, neki, nekakav, nečiji, nikoji, nikakav, ničiji*, etc. ***Upitni*** correspond aux formes fléchies de *koji, kakav, koliki, čiji* etc. employées dans un contexte interrogatif, et ***odnosni*** concerne les mêmes formes, mais dans un contexte relatif. Le tableau 2.8 donne quelques illustrations.

pokazni	**taj** pas, **ova** sveska, **ona** pravila
prisvojni	**moj** pas, **tvoja** sveska, **njegova** pravila
neodredjeni	**svaki** pas, **nijedna** sveska, **neka** pravila
upitni	**Kojeg** psa si video ? **Kakvu** svesku si kupio ? **Čija** pravila slediš ?
odnosni	ljudi **čija** pravila sledim

TABLE 2.8 – Adjectif : exemples des sous-catégories différentes

NB1

Les formes des sous-catégories ***pokazni***, ***prisvojni*** et ***neodredjeni*** peuvent également connaître de véritables emplois pronominaux : *Hoću ovu, a ne njegovu* 'Je veux celle-ci et pas la sienne'. Dans ce cas, ces formes sont traitées comme des pronoms. Pour les traits qui s'appliquent dans ce cas, voir la section 2.4 Pronoms.

NB 2

Il est également à noter que le trait ***prisvojni*** ne s'applique pas aux adjectifs dénominatifs comme *Markov* 'qui appartient à Marko', *Marijin* 'qui appartient à Marija', *vojnički* 'militaire' ou *školski* 'scolaire'. Ces adjectifs sont traités comme ***opsini***, et la sous-catégorie ***prisvojni*** est limitée aux seules formes possessives considérées traditionnellement comme pronoms (cf. *moj*, *tvoj*, *njen* etc.).

2.2.2 Degré de comparaison de l'adjectif

Pour la sous-catégorie ***opisni*** 'qualificatif', le degré de comparaison par défaut est le positif, même pour les adjectifs dont le sémantisme se prête mal à la comparaison (cf. *sutrašnji* 'de demain', *letnji* 'estival', *školski* 'scolaire' etc.). En revanche, le degré de comparaison n'est pas marqué pour les autres sous-catégories et ce fait est indiqué en utilisant la valeur '---' dans la colonne correspondant au degré de comparaison.

moj	A	prisvojni	nominativ	jednina	muski rod	---
školski	A	opisni	nominativ	jednina	muski rod	pozitiv
pribor	N	zajednicka	nominativ	jednina	muski rod	

TABLE 2.9 – Adjectif : degré de comparaison

2.2.3 Traits *genre* et *nombre* pour la sous-catégorie *prisvojni*

Pour les adjectifs possessifs, le genre et le nombre indiqués sont celui du possédé. Le genre et le nombre du possesseur ne sont pas marqués. Il en est de même pour la personne du possesseur. Par conséquent, les possessifs dans les exemples ***njihov*** *stan* 'leur appartement', ***njen*** *stan* 'son appartement à elle', ***moj*** *stan* 'mon appartement' sont tous annotés comme singulier masculin.

2.2.4 Adjectifs (pro)nominalisés

En accord avec nos principes d'annotation de base, les adjectifs nominalisés sont traités en fonction de leur rôle syntaxique. De manière générale, les adjectifs qualificatifs nominalisés sont annotés comme noms, alors que les sous-catégories ***neodredjeni***, ***pokazni*** et ***prisvojni*** dans les emplois indépendants d'un nom sont considérés comme des pronoms (cf. NB1 ci-dessus). Par conséquent, la forme *mrtvih* dans l'exemple *enciklopedija mrtvih* 'encyclopédie des morts' est traitée comme un nom. Les formes *ovu* et *njegovu* dans

l'exemple *Hoću ovu, a ne njegovu* 'Je veux celle-ci et pas la sienne' sont traitées comme des pronoms.

2.2.5 Adjectifs déverbaux dérivés des formes des participes passif et actif (*glagolski pridev trpni* et *glagolski pridev radni*)

Il existe en serbe une sous-classe d'adjectifs qui coïncident formellement avec la forme du participe passif (cf. *otvoren* 'ouvert', *kupljen* 'acheté', *začaran* 'enchanté', etc) ou actif (cf. *zalutao* 'égaré', *procvetao* 'fleuri', etc.). Le traitement de ces formes est décrit en détail dans la section 2.3 Verbes.

2.2.6 Adjectifs composés

Dans la phase de la tokénisation du corpus, les adjectifs composés contenant un trait d'union (cf. *kabalističko-astrološki*, *jugo-zapadni*) ont été séparés. Dans la phase de l'annotation morphosyntaxique, il faut corriger ce fait en fusionnant les deux lignes de sorte à rétablir le token initial relié par un trait (celui de la touche 6 sur le clavier français). L'annotation des traits est faite en considérant l'unité dans sa totalité.

2.3 Les verbes

Le traitement des verbes comprend les traits morphosyntaxiques suivants : la classe de mots, la sous-catégorie, la forme verbale, la personne, le nombre, le genre et la négation. Précisons qu'ici la sous-catégorie correspond en effet au statut principal ou auxiliaire du verbe. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.10. Ici, la distribution des traits est définie en fonction de la forme verbale, et non pas selon la sous-catégorie, étant donné que cette première est plus pertinente pour la presence ou l'absence de certains traits (cf. tableau 2.11.

POS	**Sous-cat.**	**Forme verbale**	**Personne**	**Nombre**	**Genre**	**Négation**
V	glavni pomocni	aorist futur imperativ imperfekat infinitiv particip_pro particip_sad particip_radni particip_trpni prezent	prvo lice drugo lice trece lice	jednina mnozina	muski rod zenski rod srednji rod	negiran ---

TABLE 2.10 – Verbe : traits utilisés et leurs valeurs possibles

2.3.1 Trait *forme verbale*

L'utilisation des valeurs ***prezent***, ***futur***, ***aorist***, ***imperfekat***, ***imperativ*** et ***infinitiv*** correspondent à la définition de ces formes telles que présentées dans (Stanojčić & Popović, 2012). En revanche, les étiquettes utilisées pour les formes de différents participes ont été simplifiées : ***particip_radni*** correspond à *glagolski pridev radni*, ***particip_trpni*** à *glagolski pridev trpni*, ***particip_sad*** à *glagolski prilog sadašnji*, et ***particip_pro*** à *glagolski prilog prošli*.

Les formes dites du participe passif, telles que *otvoren* 'ouvert', *donet* 'apporté', et celles du participe actif comme *zalutao* 'égaré', *procvetao* 'fleuri', ont en serbe deux fonctionnements distincts : elles peuvent faire partie des formes verbales composées et être réellement des participes, ou avoir le rôle d'épithète ou d'attribut et être en effet des adjectifs, cf. *Ana je zalutala* 'Ana s'est égarée' *vs zalutala mačka* 'un chat égaré'.

La distinction entre ces deux fonctionnements devient problématique quand ces formes sont accompagnées de celles du verbe *biti* 'être' : cette combinaison des formes peut être interprétée soit comme une forme verbale composée, où la forme ambiguë correspondrait à

Exemples	**POS**	**Sous-cat.**	**Forme**	**Personne**	**Nombre**	**Genre**	**Nég.**
les temps (*radim* ‘je travaille’, *pričaće* ‘elle parlera’)	V	[tous]	prezent futur aorist imperfekat	[tous]	[tous]	---	[tous]
impératif (*radi !* ‘travaille !’)	V	glavni	imperativ	drugo lice	[tous]	---	[tous]
				prvo lice	mnozina	---	[tous]
infinitif (*raditi* ‘travailler’)	V	glavni	infinitiv	---	---	---	[tous]
participe actif (*radio* ‘travaillé’)	V	[tous]	particip_radni	---	[tous]	[tous]	[tous]
participe passif (*rađen* ‘travaillé’)	V	glavni	particip_trpni	---	[tous]	[tous]	---
participe présent (*radeći* ‘travaillant’)	V	glavni	particip_sad	---	---	---	[tous]
participe passé (*radivši* ‘ayant travaillé’)	V	glavni	particip_pro	---	---	---	[tous]

TABLE 2.11 – Verbe : distribution des traits en fonction de la forme verbale

un participe (et serait donc traitée comme un verbe), soit comme un emploi attributif du verbe *biti* ‘être’, où la forme ambiguë serait plutôt un adjectif utilisé comme attribut de sujet (ou *imenski predikativ* dans la terminologie serbe). C’est le cas des exemples comme *Hotel je otvoren* ‘L’hôtel est ouvert’, *Ključ je izgubljen* ‘La clé est perdue’, etc. Même si des critères sémantiques pourrait permettre à un annotateur humain de faire la distinction, elle est impossible à opérer sur base des critères de surface accessibles à un étiqueteur statistique. Par conséquent, le critère de distinction qui a été adopté est le suivant : **toute occurrence** d’une forme correspondant à un participe accompagnée du verbe *biti* ‘être’ est annotée comme participe, et la forme du verbe *biti* ‘être’ est considérée comme verbe auxiliaire. C’est donc le cas des formes soulignées dans les exemples *Hotel je otvoren* ‘L’hôtel est ouvert’ ou *Ključ je izgubljen* ‘La clé est perdue’. En revanche, les occurrences de ces formes dans la position de modifieur d’un nom (*atribut* en serbe), comme *otvoren hotel* ‘l’hôtel ouvert’, *izgubljen ključ* ‘la clé perdue’, etc., sont annotées comme adjectifs.

2.3.2 Verbes auxiliaires dans les formes surcomposées

Le serbe dispose de deux formes verbales surcomposées : le potentiel passé (équivalent dans un certain degré du conditionnel passé français) et le plus-que-parfait. Dans l'exemple *On je bio došao* 'Il était venu', le plus-que-parfait *je bio došao* est constitué de *je bio*, parfait du verbe *jesam* 'être', et de *došao*, participe actif du verbe *doći* 'venir'. Le parfait du verbe *jesam* lui-même consiste en *je*, présent du verbe *jesam*, qui est la forme de l'auxiliaire, et en *bio*, participe passé du verbe *biti*, le verbe principal.

Ici, le parfait du verbe *jesam* devrait être annoté en tant que verbe auxiliaire faisant partie du plus-que-parfait. Or, ceci serait contraire au principe de la correspondance 1 :1 entre les tokens et les étiquettes. On traite donc l'auxiliaire composé de manière suivante : on attribue l'étiquette du verbe auxiliaire à chacune des formes qui le constituent, alors que la forme du participe passé porte celle du verbe principal. Un exemple est donné dans le tableau 2.12.

je	V	pomocni	prezent	trece lice	jednina	---	---
bio	V	pomocni	particip_radni	---	jednina	muski rod	---
došao	V	glavni	particip_radni	---	jednina	muski rod	---

TABLE 2.12 – Verbe auxiliaire : formes surcomposées

NB3

Les verbes modaux ou aspectuels complémentés par un infinitif ne sont pas considérés comme temps composés : dans *trebati raditi* 'falloir travailler' ou *prestati raditi* lit. 'arrêter travailler', les deux verbes ont le statut d'un verbe principal.

2.3.3 Trait *négation*

Ce trait s'applique **seulement** aux formes niées **synthétiques** des verbes *biti* 'être', *imati* 'avoir' et *hteti* 'vouloir'. Autrement dit, il est marqué pour les formes comme *nisam* 'je ne suis pas', *nisi* 'tu n'es pas', *nije* 'il n'est pas', etc., *nemam* 'je n'ai pas', *nemaš* 'tu n'as pas', *nema* 'il n'a pas' etc., *neću* 'je ne veux pas', *nećeš* 'tu ne veux pas', *neće* 'il ne veut pas', etc. Il est également pertinent pour les formes du verbe défectif *nemoj* 'ne fais pas' (*nemoj* 'ne fais pas', *nemojte* 'ne faites pas', *nemojmo* 'ne faisons pas'). Dans le cas des formes niées **analytiques**, comme *ne radi* 'il ne travaille pas', la négation est marquée par la particule, et non pas par la forme verbale. Dans ce cas, la valeur du trait *négation*

pour le verbe est ‘---’.

2.4 Les pronoms

Le traitement des pronoms fait appel aux traits morphosyntaxiques suivants : la classe de mots, la sous-catégorie, le cas, le nombre, le genre et la personne. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.13, et la distribution des traits en fonction de la sous-catégorie est présentée dans le tableau 2.14.

POS	**Sous-cat.**	**Cas**	**Nombre**	**Genre**	**Personne**
P	licna pokazna neodredjena odnosna povratna upitna brojna	nominativ genitiv dativ akuzativ vokativ instrumental lokativ	jednina mnozina	muski rod zenski rod srednji rod	prvo lice drugo lice trece lice

Table 2.13 – Pronom : traits utilisés et leurs valeurs possibles

Exemples	**POS**	**Sous-cat.**	**Cas**	**Nombre**	**Genre**	**Personne**
personnels (*ja* 'je', *mi* 'nous')	P	licna	[tous]	[tous]	---	prvo lice drugo lice
		licna	[tous]	[tous]	[tous]	trece lice
démonstratifs, indéfinis, numéraux (*onaj* 'celui-là', *jedan* 'un (certain)')	P	pokazna neodredjena brojna prisvojna	[tous]	[tous]	[tous]	---
réflexif (*se*, *sebe* 'se, soi')	P	povratna	genitiv dativ akuzativ instrumental lokativ	---	---	---
intérrogatifs et relatifs (*ko* 'qui', *koji* 'qui')	P	upitna odnosna	[tous]	[tous]	[tous]	---

Table 2.14 – Pronom : distribution des traits en fonction de la sous-catégorie

2.4.1 Les sous-catégories des pronoms

Les formes traditionnellement appelées *pridevske zamenice* 'pronoms adjectivaux' ont un double traitement en fonction de leur contexte. Si elles figurent indépendemment d'un nom, elles sont traitées comme pronoms. Les valeurs ***pokazna*** 'démonstratif', ***neodredjena*** 'indéfini' et ***prisvojna*** 'possessif' s'appliquent donc aux mêmes formes que dans la section 2.2, mais dans les cas où elles sont utilisées en dehors d'un GN.

La valeur ***brojna*** 'numéral' s'applique aux emplois pronominaux des formes *jedan* 'un', *prvi* 'premier' et *drugi* 'deuxième, autre', comme dans l'exemple *Jedan je stigao, ali drugi kasni* 'L'un est arrivé, mais l'autre est en retard'.

2.4.2 Annotation du genre et du nombre

Le genre n'est pas annoté sur les pronoms personnels à la 1ère et à la 2e personne. Pour les pronoms personnels à la 3e personne, le genre est annoté, même si au niveau formel il n'est pas distingué à tous les cas (cf. la forme *ga* peut correspondre au masculin ou au neutre, alors que *ih* peut correspondre à chacun des trois genres). L'annotateur identifiera la bonne valeur à partir du contexte.

Pour les indéfinis et les relatifs, seuls les traits morphosyntaxiques formellement marqués seront encodés. Pour les pronoms tels que *neko* 'quelqu'un', *nešto* 'quelque chose', *svako* 'chacun', *svašta* 'toute chose', *niko* 'personne', *ništa* 'rien', *iko* 'personne/quelqu'un', *išta* 'rien/quelque chose', nous indiquons seul le type du pronom (indéfini) et le cas. Il en est de même pour les pronoms relatifs et interrogatifs *ko* 'qui', *šta/što* 'que'. En revanche, pour les relatifs *koji* 'qui', *kakav* 'quel', *koliki* 'de quelle taille', le genre et le nombre sont annotés aussi, et c'est également le cas des démonstratif *taj*, *ovaj* et *onaj* (cf. tableau 2.17).

Quant au traitement du pronom réflexif, sa forme clitique *se* est systématiquement annotée comme étant à l'accusatif. En revanche, pour la forme pleine, tous les cas obliques sont éligibles.

2.4.3 Pronoms indéfinis discontinus

Il existe une série des pronoms indéfinis en serbe dérivés des pronoms interrogatifs *ko* 'qui' et *šta* 'quoi' par préfixation en *ni-* et en *i-* (cf. *niko* 'personne', *ništa* 'rien', *iko* 'qui que ce soit', *išta* 'quoi que ce soit'). Si ces formes se trouvent dans un syntagme prépositionnel, elles deviennent discontinues : le préfixe se détache de la base et la préposition vient s'insérer entre les deux : *ni za koga* 'pour personne', *ni o čemu* 'de rien', *i prema kome* 'envers qui que ce soit', *i sa čim* 'avec quoi que ce soit'. Pour ces cas de figure, la solution suivante a été adoptée : le préfixe est annoté comme particule, la base du pronom porte

Pronom	**Sous-cat.**	**Cas**	**Nombre**	**Genre**	**Pers.**
brojne (*jedan*, *drugi*, *prvi*)	brojna	[tous]	[tous]	[tous]	---
niko	neodredjena	[tous]	---	---	---
ništa	neodredjena	[tous]	---	---	---
neko	neodredjena	[tous]	---	---	---
nešto	neodredjena	[tous]	---	---	---
iko	neodredjena	[tous]	---	---	---
išta	neodredjena	[tous]	---	---	---
***ko* (odnosna)**	odnosna	[tous]	---	---	---
***što* (odnosna)**	odnosna	[tous]	---	---	---
se	povratna	akuzativ	---	---	---
sebe	povratna	genitiv dativ akuzativ instrumental lokativ	---	---	
svi	neodredjena	[tous]	---	---	---
koji	odnosna	[tous]	---	---	---
pokazne (*taj, ta, to ; onaj, ona, ono ; ovaj, ova, ovo*)	pokazna	[tous]	---	---	---

Table 2.15 – Pronom : genre et nombre

l'étiquette du pronom indéfini, alors que la préposition est étiquetée de manière habituelle. Voir les exemples dans le tableau 2.16.

ni	Part					
za	Prep					
koga	P	neodredjena	akuzativ	---	---	---
ni	Part					
iz	Prep					
jednog	P	neodredjena	genitiv	jednina	muski rod	---

Table 2.16 – Pronom : discontinuité dans le GP

2.4.4 Distinction entre le pronom personnel fléchi et l'auxiliaire

Dans les exemples comme *Unosio je hladnoću u pregrštima, ili* ***je*** *gurao pred sobom*, la deuxième forme *je* est ambiguë : il n'est pas clair si elle correspond au pronom personnel *ona* 'elle' à l'accusatif singulier (qui serait alors une reprise de *hladnoća* 'le froid') ou bien à l'auxiliaire *jesam* 'être' à la troisième personne du singulier, qui ferait partie du parfait avec le participe *gurao* 'poussé'. Un test simple permet cependant de confirmer qu'il s'agit du pronom : il suffit de mettre la forme verbale à la première personne. Dans l'exemple *Unosio sam hladnoću u pregrštima, ili je gurao pred sobom*, la forme *je* ne peut représenter que le pronom, vu que l'auxiliaire à la première personne du singulier a la forme *sam*, comme dans la première proposition. Ceci montre que dans la deuxième proposition l'auxiliaire est élidé, et le pronom maintenu. Par conséquent, dans ce type d'exemples, la forme *je* doit être annotée comme pronom personnel.

Unosio	V	glavni	particip_radni	---	jednina	muski rod	---
je	V	pomocni	prezent	trece lice	jednina	---	---
hladnoću	N	zajednicka	akuzativ	jednina	zenski rod		
u	Prep						
pregrštima	N	zajednicka	instrumental	mnozina	zenski rod		
,	Z						
ili	C	koordinirani					
je	P	licna	akuzativ	jednina	zenski rod	trece lice	
gurao	V	glavni	particip_radni	---	jednina	muski rod	---
pred	Prep						
sobom	P	povratna	instrumental	---	---	---	

TABLE 2.17 – Pronom *vs* auxiliaire

2.4.5 Le pronom *što*

La forme *što* est ambiguë. Elle peut représenter un pronom relatif, comme dans les exemples *Pas što laje* (=*Pas koji laje*) 'le chien qui aboie', *Knjiga što čitam* (=*Knjiga koju čitam*) 'le livre que je lis'. Il se comporte ici comme un relatif typique : il a une double fonction syntaxique, celle de l'introducteur de la relative, mais aussi celle d'un élément syntaxique dans la relative elle-même (sujet dans le premier exemple, objet direct dans le deuxième). En revanche, sa nature dans l'exemple suivant est moins claire : *kuća što sam je kupio* lit. 'la maison que je l'ai achetée'. Ici, le rôle de l'objet direct dans la relative est assuré par le pronom personnel *je* 'la', et la forme *što* 'que' est en effet réduite à la fonction de subordination. Par conséquent, dans les exemples des relatives utilisant un pronom personnel au lieu d'un relatif à double emploi, la forme *što* sera traitée comme un subordonnant.

kuća	N	zajednicka	nominativ	jednina	zenski rod		
što	C	subordinirani					
sam	V	pomocni	prezent	prvo lice	jednina	---	---
je	P	licna	akuzativ	jednina	zenski rod	trece lice	
kupio	V	glavni	particip_radni	---	jednina	muski rod	---

TABLE 2.18 – Pronom *što* : exemple 1

Il faut souligner que ceci n'arrive jamais avec le sujet de la relative : les exemples comme **Pas što on laje* 'le chien qui il aboie' sont agrammaticaux. Dans l'exemple *pas što laje*, la forme *što* est considérée comme un pronom relatif au nominatif (cf. tableau 2.19).

pas	N	zajednicka	nominativ	jednina	muski rod		
što	P	odnosna	nominativ	---	---	---	
laje	V	glavni	prezent	trece lice	jednina	---	---

TABLE 2.19 – Pronom *što* : exemple 2

Il faut également faire la distinction entre les emplois relatifs de cette forme et de ses emplois en tant que complétif. Dans l'exemple *Raduje me to* ***što*** *mi je rekao* (au sens 'Ce qu'il m'a dit me réjouit'), il s'agit d'un relatif, dont l'antécédent est *to* et qui a la fonction du sujet dans la relative. En revanche, dans la phrase *Raduje me to što dolazi* (au sens 'Le fait qu'il vient me réjouit'), la forme *što* est un subordonnant introduisant une complétive qui exprime le contenu résumé par le pronom *to*. Il est utile de noter que *to* est remplaçable par *ono* dans le premier cas, mais pas dans le deuxième (sans changer le sens). Les deux traitements sont illustrés dans le tableau 2.20.

Raduje	V	glavni	prezent	trece lice	jednina	---	---
me	P	licna	akuzativ	jednina	---	prvo lice	
to	P	pokazna	nominativ	jednina	srednji rod	---	
što	P	odnosna	nominativ	---	---	---	
mi	P	licna	dativ	jednina	---	prvo lice	
je	V	pomocni	prezent	trece lice	jednina	---	---
rekao	V	glavni	particip_radni	---	jednina	muski rod	---
Raduje	V	glavni	prezent	trece lice	jednina	---	---
me	P	licna	akuzativ	jednina	---	prvo lice	
to	P	pokazna	nominativ	jednina	srednji rod	---	
što	C	subordinirani					
dolazi	V	glavni	prezent	trece lice	jednina	---	---

TABLE 2.20 – Pronom *što* : exemple 3

2.5 Les numéraux

Le traitement des numéraux fait appel aux traits morphosyntaxiques suivants : la classe de mots, la sous-catégorie, le cas, le nombre et le genre. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.21, alors que la distribution des traits en fonction de la sous-catégorie est présentée dans le tableau 2.22.

POS	**Sous-cat.**	**Cas**	**Nombre**	**Genre**
Num	opsti redni zbirni	nominativ genitiv dativ akuzativ vokativ instrumental lokativ	jednina mnozina	muski rod zenski rod srednji rod

TABLE 2.21 – Numéral : traits utilisés et leurs valeurs possibles

Exemples	**POS**	**Sous-cat.**	**Cas**	**Nombre**	**Genre**
jedan 'un', *dva* 'deux'	Num	opsti	[tous]	[tous]	[tous]
tri 'trois', *četiri* 'quatre'	Num	opsti	[tous]	---	---
à partir de *pet* 'cinq'	Num	opsti	---	---	---
ordinaux (*prvi* 'premier')	Num	redni	[tous]	[tous]	[tous]
collectifs (*dvoje* 'deux', *troje* 'trois')	Num	zbirni	[tous]	mnozina	[tous]

TABLE 2.22 – Numéral : distribution des traits en fonction de la sous-catégorie

2.5.1 Le genre et le nombre des numéraux

Pour les formes du numéral *jedan* 'un', nous annotons le cas, le nombre et le genre (cf. tableau 2.23).

jedne	Num	opsti	genitiv	jednina	zenski rod		
večeri	N	zajednicka	genitiv	jednina	zenski rod		

TABLE 2.23 – Numéral *jedan*

En ce qui concerne le numéral *dva* 'deux', deux cas de figure existent. Si le numéral est décliné en accord avec la fonction du groupe nominal dans la phrase, on annote le cas, le nombre et le genre (cf. *Priča sa dvama učenicima* lit. 'Il parle avec deux.INS élèves.INS).

Dans le cas contraire, où le numéral reste invariable et impose au nom la forme du paucal (cf. *Priča sa dva učenika* lit. 'Il parle avec deux élèves.PAUC), *dva* 'deux' est traité comme un numéral invariable (cf. le traitement de *pet* 'cinq') et par conséquent on n'annote que la sous-catégorie. Voir les exemples dans le tableau 2.24.

Un traitement parallèle est mis en place pour les numéraux *tri* 'trois' et *četiri* 'quatre' : s'ils sont déclinés au même cas que le nom, nous annotons le cas ; sinon, nous considérons qu'il s'agit d'un numéral invariable.

Priča	V	glavni	prezent	trece lice	jednina	---	---
sa	Prep						
dvama	Num	osnovni	instrumental	mnozina	muski rod		
učenicima	N	zajednicka	instrumental	mnozina	muski rod		
Priča	V	glavni	prezent	trece lice	jednina	---	---
sa	Prep						
dva	Num	osnovni	---	---	---		
učenika	N	zajednicka	genitiv	jednina	muski rod		

TABLE 2.24 – Numéral *dva* : fléchi *vs* invariable

2.5.2 Paucal

Les formes spéciales dites de *paukal* 'paucal', résidu du dual de l'ancien slave, imposées aux noms par les numéraux *dva* 'deux', *tri* 'trois' et *četiri* 'quatre', sont traitées comme **génitif singulier**. Ceci est fait dans un souci de limiter le nombre de valeurs possibles pour la catégorie du cas.

2.5.3 Formes en *-ak*

Les formes à sémantisme approximatif comme *dvadesetak* 'à peu près vingt' sont considérées comme des numéraux collectifs, mais leur genre et nombre ne sont pas annotés.

2.6 Les adverbes

Le traitement des adverbes fait appel aux traits morphosyntaxiques suivants : la classe de mots, la sous-catégorie et le degré de comparaison. L'ordre des traits et leurs valeurs possibles sont donnés dans le tableau 2.25, alors que la distribution des traits en fonction de la sous-catégorie est indiquée dans le tableau 2.26.

POS	**Sous-cat.**	**Degré de compar.**
Adv	opsti odnosni upitni neodredjeni	pozitiv komparativ superlativ

Table 2.25 – Adverbe : traits utilisés et leurs valeurs possibles

Exemples	**POS**	**Sous-cat.**	**Degré de compar.**
généraux (*sutra* 'demain', *mirno* 'tranquillement'	Adv	opsti	[tous]
autres (*kada* 'quand', *nikada* 'jamais')	Adv	odnosni upitni neodredjeni	---

Table 2.26 – Adverbe : distribution des traits en fonction de la sous-catégorie

2.6.1 Sous-catégories des adverbes

Nous abandonnons la classification sémantique proposée par Stanojčić & Popović (2012). Nous distinguons les adverbes relatifs (***odnosni***, cf. *Otišao je tamo* ***gde*** *sam mu rekla* 'Il est allé là où je lui ai dit'), interrogatifs (***upitni***, cf. ***Gde*** *je otišao ?* 'Où est-il allé ?') et indéfinis (***neodredjeni***, cf. *nikako* 'aucunement', *nikada* 'jamais', *nekako* 'd'une certaine façon', *nekada* 'autrefois' etc.). Tous les autres adverbes sont considérés comme ***opsti***.

2.6.2 Degré de comparaison des adverbes

Ce trait n'est pas marqué sur les sous-catégories ***odnosni***, ***upitni*** et ***neodredjeni***. Il a la valeur par défaut de ***pozitiv*** dans la sous-catégorie ***opsti***, même pour les adverbes qui ne se comparent à priori pas, comme *letos* 'l'été dernier' ou *levo* 'à gauche'.

2.6.3 Forme *kad* ‘quand’

Cette forme est traitée comme une conjonction quand elle introduit une subordonnée temporelle, cf. *Reći ću mu kad dođe* ‘Je le lui dirai quand il viendra’. En revanche, elle est traité comme adverbe interrogatif (***upitni***) dans les interrogatives directes et indirectes, cf. *Kad Marko dolazi ?* ‘Quand est-ce que Marko vient ?’ et *Pita kad Marko dolazi* ‘Il demande quand vient Marko’, et comme adverbe relatif (***odnosni***) dans les relatives, cf. *dan kad je stigao* ‘le jour où il est arrivé’. Il est utile de noter que les emplois relatifs de la forme *kad* peuvent toujours être remplacés par un relatif non ambigu : *dan u koji je stigao* lit. ‘le jour dans lequel il est venu’.

2.7 Conjonctions

Le traitement des conjonctions fait appel à la classe de mots et à la sous-catégorie. L’ordre des traits et leurs valeurs possibles sont donnés dans les tableaux 2.27 et 2.28.

POS	**Sous-catégorie**
C	subordinirani koordinirani

Table 2.27 – Conjonctions : traits utilisés et leurs valeurs possibles

Exemples	**POS**	**Sous-catégorie**
i ‘et’, *ili* ’ou’, *ali* ‘mais’	C	koordinirani
jer ‘parce que’, *da* ‘que’, *kad* ‘quand’	C	subordinirani

Table 2.28 – Conjonctions : exemples des sous-catégories

Certaines formes qui fonctionnent comme conjonctions de coordination peuvent également être employées comme des particules. Ceci est notamment le cas des formes *i* et *pa*.

I est traité comme conjonction dans les emplois où il établit réellement une coordination entre deux éléments syntaxiques, cf. *Petar i Marko* ‘Petar et Marko’, *Sedim i jedem* ‘Je suis assis et je mange’, *lep pas i divlja mačka* ‘un beau chien et un chat sauvage’. Il est considéré comme une particule quand il est employé au sens de ‘aussi’ : *I moja majka je tu* lit. ‘Et ma mère est là’, ‘Ma mère est là aussi’.

Pa est considéré comme conjonction de coordination dans les exemples comme *Oprao je zube pa je legao da spava* ‘Il s’est brossé les dents puis il s’est couché’. En revanche, il est traité comme particule dans les emplois comme *Pa rekla sam ti da to ne radiš* ‘Mais

je t'avais dit de ne pas le faire'. Pour plus d'informations, voir la section 3.2, dédiée au traitement des particules.

2.8 Les prépositions, les interjections, les particules, les mots étrangers et la ponctuation

Pour ces classes de mots nous n'indiquons que la classe de mots. Les étiquettes sont données dans le tableau 2.29. Des listes des prépositions et des particules ont été compilées à partir de (Mrazović, 2009) et sont disponibles dans la suite de ce document. Le traitement des particules y est présenté en détail.

Classe	**POS**
Prépositions	Prep
Particules	Part
Mots étrangers	X
Ponctuation	Z

TABLE 2.29 – Autres classes : étiquettes POS

3. Gestion des cas de figure spécifiques

3.1 Prépositions

Afin de faciliter leur identification dans le corpus, nous reprenons ici une liste de prépositions proposée par Mrazović (2009).

bez	mimo	pod	protiv	umesto
blagodareći	na	podno	put	unutar
blizu	nad	pokraj	putem	uoči
čelo	nadno	polovinom	radi	uprkos
dno	nadohvat	pomoću	sa	usled
do	nadomak	poput	saglasno	usred
dovrh	nakon	pored	saobrazno	usuprot
duž	nakraj	posle	sem	ususret
ispod	namesto	posred	shodno	uvrh
ispred	naspram	posredovanjem	silom	uz
iz	nasred	posredstvom	skraj	uzduž
iza	nasuprot	potkraj	sledstveno	van
između	navrh	poviše	slično	više
iznad	niz	povodom	sred	vrh
izuzev	niže	povrh	sredinom	za
izvan	o	pozadi	suprotno	zarad
ka	od	pre	tokom	zaslugom
kod	oko	pred	u	zbog
kraj	okolo	preko	udno	zahvaljujući
krajem	osim	prema	uinat	
kroz	po	pri	uključujući	
među	početkom	prilikom	ukraj	

NB4

Certaines de ces formes sont ambiguës : elles peuvent fonctionner comme prépositions, mais aussi comme d'autres classes de mots. Il s'agit typiquement d'adverbes, mais aussi de verbes et de noms. Par conséquent, il faut prendre en considération leur contexte syntaxique. Pour être considérées comme des prépositions, ces formes doivent introduire un complément nominal ou pronominal, dans un cas imposé par la forme, cf. *Izašao je* ***zahvaljujući*** *Ani* 'Il est sorti grâce à Ana', où *zahvaljujući* introduit le nom *Ani*, qui est au datif, imposé par la préposition. En revanche, dans *Izašao je* ***zahvaljujući*** *se Ani* 'Il est sorti en remerciant Ana', la forme *zahvaljujući* est en effet un participe. Cf. également les oppositions *Ostavio ga je* ***nadohvat*** *ruke* 'Il l'a laissé à portée de main' (préposition) *vs Pobeda je bila* ***nadohvat*** 'La victoire était proche' (adverbe) ; *Kupio je sve,* ***uključujući televizor*** 'Il a tout acheté, y compris la télé' (préposition) *vs Zaćutao je* ***uključujući televizor*** 'Il s'est tu en allumant la télé' (participe) ; *Obratio se novinarima* ***prilikom*** *konferencije* 'Il s'est adressé aux journalistes lors de la conférence de presse' (préposition) *vs Obraća se novinarima svakom* ***prilikom*** 'Il s'adresse aux journalistes à toutes occasion' (nom).

3.2 Listes des particules

Notre traitement des particules est basé sur celui de Mrazović (2009, p. 466-512). L'auteure introduit 4 types de particules, basés largement sur des distinctions sémantiques ; nous ne les reprenons pas ici. L'auteure souligne également que certaines particules coïncident formellement avec des adverbes et indique que la distinction entre ces deux classes de mots se base sur le fait que les adverbes sont des dépendants verbaux qui servent à préciser ou situer le contenu verbal, alors que les particules réfèrent à la phrase entière et servent à exprimer l'attitude de l'énonciateur. Cependant, cette différentiation est délicate à opérer en contexte. Pour faciliter le travail des annotateurs, nous donnons d'abord la liste des particules censées être non ambiguës (par rapport aux adverbes). Ensuite, nous reprenons celles marquées comme ambiguës par Mrazović (2009) et donnons des indices pour opérer la désambiguïsation.

Liste des particules non ambiguës par rapport aux adverbes

a	doduše	međutim	obavezno	uglavnom
ako	dosta	možda	očito	umalo
ala	e	naime	odveć	uopšte
ama	eno	najzad	odviše	uostalom
bar	eto	naprotiv	oko	upravo
barem	evo	naravno	ono	vala
baš	gle	naročito	otprilike	valjda
besumnje	hajde	navodno	pa	veoma
bezmalo	i	nažalost	samo	verovatno
bogme	inače	ne	srećom	vrlo
čak	ionako	neka	štaviše	zaista
čas	ipak	nemoguće	što	zamalo
časkom	izgleda	neosporno	suviše	zapavo
da	jedino	nepotrebno	sve	zar
dabogda	jedva	nešto	ta	zbilja
dabogme	još	nesumnjivo	takođe	
dakako	kao	ni	taman	
dakle	li	nipošto	tek	
de	ma	no	to	

Liste des particules **ambiguës** par rapport aux adverbes

badava	konačno	onako	sigurno	teško
daleko	lično	onda	skoro	već
dobro	malo	potpuno	slobodno	više
dosta	mnogo	praktično	slučajno	
gotovo	najmanje	prilično	svakako	
jednostavno	nikako	prosto	tačno	
još	obično	sad	tako	

En ce qui concerne les formes ambiguës, les critères suivants peuvent être utilisés pour identifier de quelle classe de mots il s'agit :

1. construction **Particule + da** : des particules en serbe ont la capacité d'introduire cette construction spécifique, cf. *Skoro/gotovo/prosto da ne poveruješ*, *Teško da će stići*, *Nikako da stigne*. Si une forme ambiguë se trouve dans cette construction, elle sera traitée comme particule.

2. reformulation du type **Particule + je + da/što** : si une forme ambiguë permet ce type de reformulation, cela montre qu'elle porte sur toute la phrase, et non seulement sur le processus verbal (cf. *Sigurno si umorna* 'Tu es certainement fatiguée' => *Sigurno je da si umorna* 'Il est certain que tu es fatiguée'). Par conséquent, elle sera traitée comme particule .
3. présence des virgules : étant donné qu'elles portent sur toute la phrase, les particules peuvent se trouver en dehors de sa structure syntaxique proprement dite. Dans ce cas, elles sont séparées par des virgules du reste de la phrase, cf. *Jednostavno, treba se odlučiti*. Ce type d'utilisation des formes ambiguës sera traité comme particule.
4. questions **kako ?** 'comment ?', **kada ?** 'quand ?', **gde ?** 'où ?', **koliko ?** 'combien ?' : si la forme ambiguë répond à l'une de ces questions dans la phrase, elle sera traitée comme adverbe (cf. *Govori jednostavno* 'Il parle simplement' => **kako govori ?** 'comment parle-t-il ?' ***jednostavno*** 'simplement').

Des exemples des oppositions adverbe - particule :

1. *Sigurno si umorna* (particule [critère 2]) *vs Kako je govorio ? - Sasvim sigurno* (adverbe [critère 4])
2. *Teško da će doći* (particule [critère 1]) *vs Kako si uspeo da ga ubediš ? -Teško.* (adverbe [critère 4])
3. *Badava se trudiš* (particule [critère 2]) *vs Jabuke su badava* (adverbe [critère 4])
4. *Jednostavno, treba se odlučiti* (particule [critère 3]) *vs Oblači se jednostavno* (adverbe [critère 4])

3.3 Autres cas de figure

1. Adjectifs
   (a) **ma koliki** : *ma* = particule, *koliki* = adjectif indéfini. Voir le traitement des pronoms discontinus (cf. section 2.4.3).
2. Pronoms
   (a) **sve to** je doneo : *sve* = adjectif indéfini, *to* = pronom démonstratif
   (b) **svi mi** mu verujemo : *svi* = adjectif indéfini, *mi* = pronom personnel
   (c) **onaj novi** je došao : *onaj* = adjectif démonstratif, *novi* = nom commun
   (d) **onaj koji** je došao : *onaj* = pronom démonstratif, *koji* = pronom relatif
3. Numéraux

(a) **drugi** : traité comme numéral ordinal seulement dans le sens de ‘deuxième’ cf. *Ovo mi je druga knjiga ove nedelje* ‘C’est le deuxième livre que je lis cette semaine’. Dans le sens de ‘autre, différent’, il est annoté comme adjectif qualificatif, cf. *Nađi neki drugi način* ‘Trouve un moyen différent’. S’il est utilisé indépendamment d’un nom, il est traité comme pronom numéral, cf. *Stigao je i drugi* ‘Le deuxième est arrivé aussi’.

4. Noms
   (a) L’instrumental de certains noms s’est grammaticalisé et fonctionne également comme un mot invariable. Des exemples :
      1. **većinom** : *Većinom su zadovoljni* ‘Ils sont majoritairement contents’ (adverbe) *vs Zadovoljan je većinom odgovora* ‘Il est content de la majorité des réponses’ (nom)
      2. **početkom** : *Došao je početkom popodneva* ‘Il est venu en début de l’après-midi’ (préposition) *vs Razočaran je početkom filma* ‘Il est déçu du début du film’ (nom)
      3. **tokom** : *Upoznala sam ga tokom leta* ‘Je l’ai rencontré durant l’été’ (préposition) *vs Nezadovoljan je tokom stvari* ‘Il n’est pas content du déroulement des choses’ (nom)
      4. **mahom** : *Nalazili su se mahom u gradu* ‘Ils se retrouvaient majoritairement en ville’ (adverbe)
5. Mots invariables
   (a) **kao**
      — conjonction si utilisé au sens comparatif : *Voli ga kao brata* ‘Elle l’aime comme un frère’
      — particule quand il est possible de le remplacé par *tobože* : *On je kao bolestan* ‘Il est soi-disant malade’
      — adverbe quand il est utilisé au sens de ‘en tant que’ : *Radi kao novinar* ‘Il travaille comme journaliste’
   (b) **stoga**, **zato** : des adverbes, même quand ils se trouvent en tête de proposition, cf. *Stoga/Zato je došao da proveri* ‘Ainsi il est venu s’assurer’. Bien qu’ils semblent introduire ici une subordonnée (utilisée indépendamment d’une proposition principale), ces formes sont en effet mobiles : *Došao je stoga/zato da proveri.* Il ne s’agit donc pas de véritables subordonnants. La situation est différente s’ils font partie des unités polylexicales *stoga što* et *zato što* : ces unités sont effectivement des subordonnants et elles introduisent des subordonnées.

Néanmoins, même à l'intérieur de ces expressions, les formes *stoga* et *zato* sont annotées comme adverbes, et c'est la forme *što* qui porte l'étiquette de conjonction subordonnée.

(c) **tako da** : *tako* = adverbe, *da* = conjonction de subordination

(d) **kao da** : *kao* = conjonction de subordination, *da* = conjonction de subordination

(e) **kao i** : *kao* = conjonction de subordination, *i* = particule

Bibliographie

Pavica Mrazović. *Gramatika srpskog jezika za strance*. Izdavačka knjižarnica Zorana Stojanovića, 2009.

Živojin Stanojčić and Ljubomir Popović. *Gramatika srpskog jezika*. Zavod za udžbenike, 2012.

Annexe B

Guide de lemmatisation

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 **2**
1.1 Principes généraux adoptés 2

2 Règles de lemmatisation **3**
2.1 Les noms 3
2.2 Les adjectifs 3
2.3 Les verbes 4
2.3.1 Traitement des verbes *jesam* et *biti* 5
2.3.2 Traitement des verbes à négation synthétique 5
2.3.3 Traitement des participes 5
2.3.4 Lemmes doublons 6
2.4 Les pronoms 6
2.5 Autres catégories 7

3 Traitement de cas de figure spécifiques **8**
3.1 Liste des lemmes adjectivaux problématiques 8
3.2 Liste des lemmes verbaux problématiques 9

Bibliographie **10**

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.

Annexe C

Guide d'annotation syntaxique

Guide d'annotation syntaxique de ParCoLab, v2.0

Aleksandra Miletic, Dejan Stosic, Cécile Fabre
CLLE-ERSS, Université Toulouse - Jean Jaurès

26 septembre 2018

Table des matières

Remarques préliminaires **3**

1 Annotation syntaxique et le TAL **4**
1.1 Analyse en constituants et analyse en dépendances 4
1.2 Déterminer l'existence de la dépendance et son orientation 7

2 Règles d'annotation syntaxique **9**
2.1 Présentation globale du jeu d'étiquettes 9
2.2 Élément racine 13
2.2.1 Verbe principal 13
2.2.2 Autres types de tête 14
2.2.3 Caractère unique de la tête de phrase 15
2.3 Dépendants du verbe 16
2.3.1 Verbe auxiliaire 17
2.3.2 Sujet 17
2.3.3 Sujet logique 18
2.3.4 Objet direct 20
2.3.5 Objet indirect 21
2.3.6 Prédicatif nominal 22
2.3.7 Prédicatif adverbial 24
2.3.8 Prédicatif complémentaire 25
2.3.9 Prédicatif optionnel 27
2.3.10 Dépendant sous forme d'un adverbe 28
2.3.11 Dépendant sous forme d'un nom fléchi 29
2.3.12 Dépendant sous forme d'une préposition 29
2.3.13 Propositions participiales 30
2.3.14 Prédicat complexe : verbe modal ou aspectuel introduisant un infinitif 30
2.3.15 Traitement des enchaînements des dépendants 31
2.4 Dépendants du nom 32
2.4.1 Dépendant sous forme d'un nom fléchi 32
2.4.2 Dépendant sous forme de préposition 32
2.4.3 Dépendant sous forme d'adjectif 33
2.4.4 Apposition 34
2.5 Dépendants de l'adjectif 36
2.5.1 Dépendant sous forme d'un adverbe 36
2.5.2 Dépendant sous forme d'un nom fléchi 37
2.5.3 Dépendant prépositionnel 38
2.5.4 Construction *sav* 'tout' + Adjectif 38
2.6 Dépendants de l'adverbe 39
2.6.1 Dépendant sous forme d'adverbe 39
2.6.2 Dépendant sous forme de nom fléchi 39

2.6.3 Dépendant sous forme de préposition . . . 40
2.7 Relations diverses . . . 40
2.7.1 Complément de préposition . . . 40
2.7.2 Complément de numéral . . . 41
2.7.3 Négation . . . 41
2.7.4 Interrogation . . . 42
2.7.5 Réflexif . . . 44
2.7.6 Éléments extra-prédicatifs . . . 46
2.7.7 Emphase . . . 46
2.7.8 Eléments polylexicaux . . . 49
2.7.9 Citations et emplois métalinguistiques . . . 53
2.7.10 Ponctuation . . . 54
2.8 Subordination . . . 54
2.8.1 Subordonnées adverbiales à subordonnant mono-fonctionel . . . 54
2.8.2 Subordonnées complétives . . . 55
2.8.3 Subordonnées interrogatives indirectes . . . 58
2.8.4 Subordonnées relatives . . . 60
2.8.5 Subordonnées corrélatives . . . 62
2.8.6 Ambiguïté des subordonnées en *da* . . . 63
2.9 Discours indirect . . . 64
2.10 Coordination . . . 65
2.11 Juxtaposition . . . 68
2.12 Ellipse . . . 69

Index des exemples par étiquette **73**

Bibliographie **75**

Remarques préliminaires

Ce guide est destiné aux annotateurs qui travaillent sur l'annotation syntaxique du volet serbe du corpus *ParCoLab*. Il s'organise de manière suivante : le chapitre 1 présente le cadre théorique dans lequel s'inscrit l'annotation syntaxique de *ParCoLab* et les règles d'annotation générales adoptées dans le cadre du projet, alors que le chapitre 2 contient le jeu d'étiquettes et les règles d'annotation concrètes relatives à chaque étiquette. À la fin du document, un index des exemples disponibles pour chaque étiquette syntaxique est proposé (Index des exemples par étiquette).

1. Annotation syntaxique et le TAL

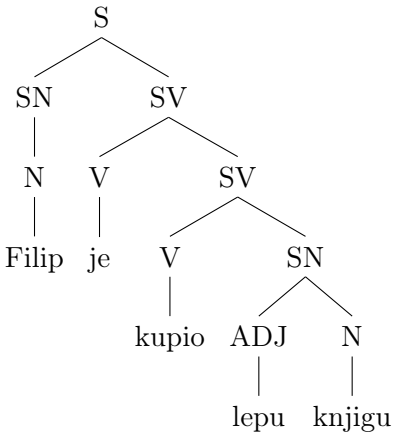
Dans l'analyse syntaxique, aussi bien théorique qu'appliquée au TAL, deux approches principales existent : l'analyse en constituants et l'analyse en dépendances. Ce chapitre est dédié à une brève présentation des deux cadres, à l'argumentation du choix qui a été fait dans le cadre de ce projet, ainsi qu'à la présentation des notions de base de l'approche sélectionnée.

1.1 Analyse en constituants et analyse en dépendances

L'analyse en constituants aussi bien que l'analyse en dépendances ont donné naissance à de nombreuses théories syntaxiques. Parmi les représentant de la grammaire en constituants, on trouve *Government and Binding Theory* (Chomsky, 1993, 1982), *Generalized phrase structure grammar* (GPSG) (Gazdar et al., 1985) et *Head-driven phrase structure grammar* (HPSG) (Pollard and Sag, 1994). Du côté de la grammaire en dépendances on peut citer les théories suivantes : *Word Grammar* (WG) (Hudson, 1984), *Functional Generative Description* (FGD) (Sgall et al., 1986), *Dependency Unification Grammar* (DUG) (Hellwig, 1986), *Meaning-Text Theory* (MTT) (Mel'čuk, 1988), ou *Functional Dependency Grammar* (FDG) (Tapanainen and Järvinen, 1997). Le principal critère de distinction entre ces deux approches théoriques est leur vision de la structure syntaxique et des représentations syntaxiques dont elles se servent. Comme les deux visions entraînent des implications importantes aussi bien pour la linguistique que pour le TAL, nous les présentons dans la suite.

L'analyse en constituants est en général plus répandue et mieux connue que l'analyse en dépendances. Dans ce cadre théorique, on considère que l'unité de base de la structure syntaxique sont les constituants (syntagmes, groupes). Analyser une phrase consiste à la décomposer en constituants niveau par niveau, jusqu'à arriver aux mots mêmes. Les arbres syntaxiques qui en résultent se présentent comme dans l'exemple (i).

(i)



Nous pouvons remarquer plusieurs caractéristiques de l'arbre :

1. l'arbre représente une structure à **plusieurs niveaux** :
2. la **racine** de l'arbre (le nœud supérieur) est un nœud représentant **la phrase** ;

3. il n'y a **pas de marquage explicite des fonctions syntaxiques** : on indique les catégories morphosyntaxiques des syntagmes, alors que la fonction syntaxique est dérivée de la structure de l'arbre.

L'application de cette approche dans le cadre du TAL consiste souvent à indiquer la structure en constituants de la phrase en ajoutant des crochets délimitant les syntagmes. L'arbre montré ci-dessus serait donc représenté sous la forme suivante :

[S [SN [N Filip]] [SV [V je] [SV [V kupio] [SN [ADJ lepu] [N knjigu]]]]]

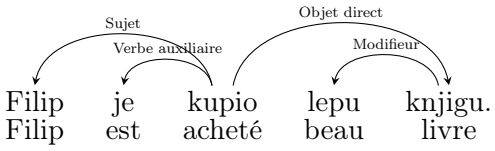
Cette représentation (ainsi que les fondements théoriques de l'analyse en constituants) pose cependant un problème important : l'analyse est basée sur l'ordre linéaire des mots dans la phrase. Ceci pose des difficultés dans le traitement des langues à ordre des mots flexible, tel le serbe, qui permet par ailleurs des constituants discontinus. Il est donc tout à fait possible d'avoir des phrases comme *Lepu je knjigu Filip kupio*, présentée dans l'exemple (ii).

(ii) Lep-u je knjig-u Filip kupi-o
beau-ACC.SG.F VAUX-3SG.PRES livre-ACC.SG Filip.NOM.SG acheté
'C'est un beau livre que Filip a acheté'

Ici, la focalisation spécifique de la phrase entraîne un ré-ordonnancement des mots, et la décomposition de la phrase en constituants devient ardue : le sujet *Filip* se trouve au milieu du syntagme verbal, l'auxiliaire *je* est séparé du verbe principal *kupio* par la tête de l'objet direct *knjigu* et le sujet *Filip*, et par ailleurs, l'adjectif modifieur de l'objet direct *lepu* est séparé de sa tête *knjigu* par l'auxiliaire. Il est donc impossible d'effectuer une analyse en constituants en respectant l'ordre linéaire des mots dans la phrase. C'est précisément pour cette raison que l'analyse syntaxique des langues comme le serbe se fait de préférence dans le cadre de l'analyse en dépendances.

Cette deuxième approche considère que la structure syntaxique d'une phrase est un ensemble de relations de dépendance qui s'établissent entre les mots individuels d'une phrase : chaque mot de la phrase a un gouverneur et peut à son tour en gouverner d'autres. Les arbres syntaxiques de cette approche se présentent sous la forme indiquée dans l'exemple (iii).

(iii)



'Filip a acheté un beau livre.'

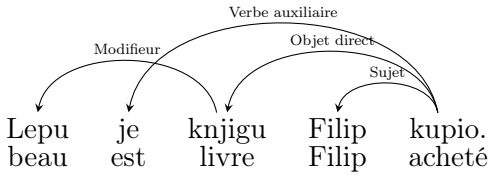
Les principales différences par rapport à l'arbre en constituants sont les suivantes :

1. l'arbre a **un seul niveau** : les relations s'établissent directement entre les formes fléchies ;
2. **la racine** de l'arbre est représentée par **le verbe principal** de la phrase ;
3. les **fonctions syntaxiques** sont **annotées explicitement**, mais les syntagmes ne le sont pas [1].

L'avantage de cette approche est que l'ordre linéaire des mots dans la phrase a peu d'impact sur l'analyse. Ainsi, les mêmes relations présentes dans l'exemple (iii) se mettent en place dans l'exemple (iv), malgré le fait que les mots sont distribués différemment.

1. Ils peuvent néanmoins être récupérés en considérant les différents sous-arbres.

(iv)



'C'est un beau livre que Filip a acheté.'

En fonction du courant de la grammaire en dépendances dans lequel on s'inscrit, les relations peuvent s'établir entre différents types d'unités : formes fléchies, lemmes, ou unités polylexicales. Également, les relations peuvent exprimer différents types de dépendances : il peut s'agir des rôles thématiques ou des relations syntaxiques (souvent dites "de surface"). Dans le cadre du projet *ParCoLab*, l'annotation portera sur des relations syntaxiques de surface, et les relations s'établiront entre les formes fléchies individuelles (les tokens) constituant la phrase.

En ce qui concerne la forme de l'arbre syntaxique créé par l'analyse en dépendances, elle est soumise à plusieurs contraintes :

1. **complétude** : tous les mots de la phrase doivent être inclus dans l'arbre (chaque mot doit avoir un gouverneur) ;
2. **acyclicité** : un mot ne peut pas gouverner son gouverneur, ni un mot hiérarchiquement supérieur à son gouverneur (il ne peut pas y avoir des cycles fermés dans l'arbre) ;
3. **nature unique du gouverneur** : un mot ne peut avoir qu'un seul gouverneur (un mot peut être la cible d'une seule dépendance).

Dans certains cadres applicatifs en TAL, on impose une dernière contrainte aux arbres créés dans l'annotation : celle de la **projectivité**. Un arbre syntaxique est considéré comme non projectif si au moins une dépendance qui le constitue est elle-même non projective. Une dépendance entre un gouverneur G et un dépendant D est considérée comme non projective s'il existe au moins un token entre G et D dans l'ordre linéaire de la phrase qui n'est pas dominé par G (autrement dit, que G n'est ni son gouverneur immédiat, ni son ancêtre). Si l'on revient encore une fois à l'exemple (iv), on remarque que la relation entre le gouverneur *knjigu* et le dépendant *lepu* est non projective, vu que l'auxiliaire *je*, situé entre ces deux formes, n'est pas dominé par *knjigu*. En revanche, la dépendance entre le verbe principal *kupio* et l'auxiliaire *je* est projective, car les deux tokens intervenants *knjigu* et *Filip* sont gouvernés par *kupio*. La non-projectivité d'une relation (et, par conséquent, d'un arbre) se traduit typiquement par le croisement des arcs désignant les dépendances dans la représentation graphique de l'arbre syntaxique.

Vu que le traitement de ce type de relations reste encore problématique dans le cadre du parsing, certains projets (et certains outils de parsing) exigent que les arbres analysés soient projectifs et imposent un traitement projectif artificiel aux constructions syntaxiques non-projectives par nature. Or, la non-projectivité est le reflet de la nature discontinue des constituants dans une phrase. Le taux de non-projectivité dans les corpus de différentes langues est souvent utilisée comme un indicateur du degré de liberté de l'ordre des constituants. Alors que le taux des relations non-projectives est relativement bas pour toutes les langues (en général, 1%-2%), le pourcentage d'arbres syntaxiques contenant au moins une relation non-projective peut atteindre (voire dépasser) 20% dans les langues à ordre de constituants flexible telles les langues slaves. Ceci signifie qu'un schéma d'annotation strictement projectif appliqué à une telle langue produirait des analyses fausses pour une phrase sur cinq dans le corpus. Par conséquent, dans le cadre de ce projet, la non-projectivité est autorisée dans la phase de l'annotation manuelle. Ceci permettra d'estimer le taux de non-projectivité

en serbe, et également de mettre à l'épreuve les méthodes de parsing proposant une solution pour ce type de relation.

1.2 Déterminer l'existence de la dépendance et son orientation

L'une des questions centrales dans l'analyse en dépendances est de savoir déterminer si deux formes sont reliées par une relation ou non. Pour ce faire, la Théorie Sens-Texte (Mel'čuk, 1988) propose les critères suivants :

— le critère de linéarité : la position dans la phrase de l'une des deux formes est déterminée par rapport à la position de l'autre si une relation de dépendance existe entre elles, et

— le critère d'unité prosodique : deux formes liées par une dépendance font une unité prosodique, ou bien l'une des formes peut être liée prosodiquement avec une unité prosodique dont l'autre forme est la tête. (Mel'čuk, 1988, pp. 129-132).

Il existe également plusieurs critères qui permettent de déterminer laquelle entre les deux formes reliées par une dépendance est le gouverneur, et laquelle est le dépendant. Certains d'entre eux sont donnés dans la suite. Ils font appel aux notions de gouverneur, dépendant et construction, cette dernière désignant la construction syntaxique créée par le gouverneur et le dépendant.

1. C'est le **gouverneur** qui **détermine la catégorie distributionnelle de la construction** et **peut** souvent **la remplacer**.
   Par exemple, si l'on reprend la phrase *Filip je kupio lepu knjigu* et qu'on analyse la construction *lepu knjigu*, nous constatons que la construction occupe la fonction de l'objet direct dans la phrase, qui est une fonction typique du nom. Par conséquent, nous pouvons conclure que c'est *knjigu* qui est le gouverneur, et *lepu* le dépendant. Par ailleurs, on peut également remplacer la construction par le gouverneur sans compromettre la gramaticalité de la phrase : *Filip je kupio knjigu*.
2. **Le dépendant** d'une relation **peut être optionnel** (il peut être omis de la phrase), mais **le gouverneur est obligatoire** dans la phrase.
   Par exemple, en considérant encore la construction *lepu knjigu*, on peut ommetre le modifieur *lepu* de la phrase *Filip je kupio lepu knjigu* et obtenir une phrase grammaticale *Filip je kupio knjigu*, mais ceci n'est pas le cas du gouverneur *knjigu* : ? ?*Filip je kupio lepu*.
3. **La réalisation morphosyntaxique du dépendant est déterminée par le gouverneur** (à travers les règles d'accord ou de rection).
   La forme *lepu*, qui correspond à l'accusatif singulier féminin de l'adjectif *lep* 'beau', est déterminée par le fait que son gouverneur, la forme *knjigu* est un nom féminin, à l'accusatif singulier. Le cas du nom est déterminé à son tour par son gouverneur *kupio* (participe passé du verbe *kupiti* 'acheter'), dont la structure argumentale exige un objet direct à l'accusatif.
4. **La position du dépendant** dans la phrase est **déterminée par rapport au gouverneur**.
   Ce critère est reflété, par exemple, dans le fait que l'adjectif *lepu* se trouve à gauche du nom *knjigu*, ce qui est la position de préférence des modifieurs adjectivaux d'un nom. En revanche, ce critère est beaucoup plus pertinent pour des langues à ordre de constituants rigide que pour le serbe.

On remarque que ces règles sont très générales et qu'elles ne font pas mention des relations syntaxiques spécifiques. Effectivement, les règles citées ci-dessus sont indépendantes de langue et servent d'outil de base dans l'analyse en dépendances de toute langue. En revanche, chaque

langue individuelle dispose d'un inventaire de relations syntaxiques de surface qui lui sont spécifiques. Certaines langues (cf. le russe, le tchèque) bénéficient de travaux théoriques décrivant en détail leur fonctionnement syntaxique dans le cadre de l'analyse en dépendants. Il est donc possible de se baser sur ces ouvrages pour dresser l'inventaire des relations syntaxiques qui seront encodées en corpus. Or, la syntaxe théorique du serbe repose traditionnellement sur l'analyse en constituants. Il n'y a donc pas de formalisme prêt à être exploité dans le cadre d'une application en corpus. Par conséquent, les relations syntaxiques proposées dans le jeu d'étiquettes décrit dans ce document ont été définies spécifiquement pour ce projet, avant le démarrage du travail sur le corpus. Elles ont été mises à l'épreuve dans le cadre d'une annotation manuelle initiale, mais le jeu d'étiquettes reste certainement perfectible. Les fonctions retenues et leurs domaines d'application respectifs sont présentés dans la suite de ce document.

2. Règles d'annotation syntaxique

Avertissement

Afin de pouvoir exploiter ce guide d'annotation, il est nécessaire de maîtriser les notions suivantes :

- dépendance syntaxique ;
- arbre de dépendances syntaxiques ;
- racine de l'arbre de dépendances syntaxiques ;
- dépendances projectives et non-projectives ;
- structure argumentale d'un verbe.

Ce chapitre est dédié à la présentation du jeu d'étiquettes syntaxiques[1] et du schéma d'annotation associé de *ParCoLab*. Nous proposons d'abord une vue d'ensemble du jeu d'étiquettes pour présenter ensuite en détail les règles d'application de chaque étiquette.

2.1 Présentation globale du jeu d'étiquettes

Le jeu d'étiquettes que nous présenterons dans la suite a été construit sur base de plusieurs principes. D'abord, étant donné la nature parallèle de *ParCoLab*, le corpus est destiné à être utilisé par la communauté serbophone aussi bien que francophone. Par conséquent, le noyau des relations retenues coïncide avec les fonctions syntaxiques traditionnellement reconnues dans la grammaire serbe (cf. sujet grammatical, sujet logique, objet direct, objet indirect, prédicatifs nominal et adverbial, etc.). En revanche, ces fonctions traditionnelles ont été soumises à une analyse basée sur des critères formels de surface dans l'objectif de garantir que la distinction des fonctions syntaxiques encodées en corpus se base sur des propriétés accessibles à un parser. Il s'agit notamment des critères suivants :

- réalisations morphosyntaxiques possibles du gouverneur et du dépendant (POS, lemme) ;
- flexion du dépendant et du gouverneur si des traits morphosyntaxiques spécifiques sont liés à la fonction en question (par exemple, un cas ou une forme verbale spécifique) ;
- pour les dépendants verbaux considérés comme objets, la possibilité de pronominalisation avec un clitique et le type de clitique utilisé ;
- accord : les constituants et les traits concernés ;
- règles de linéarisation :
  - ordre canonique gouverneur - dépendant ;
  - caractère flexible ou rigide de l'ordre gouverneur - dépendant ;

1. Le terme *jeu d'étiquettes* désigne l'ensemble des étiquettes ou des labels utilisés à un niveau d'annotation dans un corpus. Il existe donc des jeux d'étiquettes morphosyntaxiques, qui expriment typiquement les parties du discours, des jeux d'étiquettes syntaxiques, exprimant des relations syntaxiques, des jeux d'étiquettes discursifs, servant à l'analyse du discours, etc.

— caractère obligatoire ou non de l'adjacence du gouverneur et du dépendant (possibilité que le dépendant soit séparé du gouverneur par un autre constituant) ;
— non-projectivité possible ou non de la relation (possibilité que le dépendant soit séparé du gouverneur par un mot n'ayant pas le même gouverneur, ce qui entraîne un croisement des arcs dans l'arbre syntaxique).

Pour certains phénomènes plus complexes, moins fréquemment abordés dans les travaux théoriques, des traitements déjà existants dans d'autres corpus ont été repris ou adaptés.

Cette démarche a abouti à un jeu de 48 relations syntaxiques de base, et un traitement spécifique pour l'ellipse, qui sera détaillé dans la suite. Le jeu est présenté dans le tableau 2.1. Chaque étiquette est accompagnée d'une brève définition de la relation et d'un exemple. Dans l'exemple, le **gouverneur** est indiqué en gras, et le dépendant en souligné.

NB

Avant de continuer, rappelons encore que dans le cadre de la syntaxe en dépendances, les relations s'établissent entre les mots individuels. Par conséquent, dans les exemples donnés ci-dessous, le dépendant correspond toujours à un seul mot. Dans le cadre de la syntaxe en constituant, ce mot correspondrait à la tête du constituant exerçant la fonction syntaxique en question. C'est pourquoi, dans la première ligne du tableau, nous annotons seulement la forme predsednik 'président' en tant qu'apposition, et non pas tout le syntagme predsednik Francuske 'président de la France'.

TABLE 2.1: Jeu d'étiquettes syntaxiques ParCoLab

Etiquette	Définition	Exemple
Ap	apposition	***Oland***, *predsednik Francuske* '**Hollande**, président de la France'
AuxV	verbe auxiliaire dans une forme verbale composée	*Filip je* ***stigao*** 'Filip est **arrivé**'
Cit	élément métalinguistique	***misao*** *« budan sam »* '**idée** « je suis réveillé »'
ComplNum	complément d'un numéral sous forme du paucal[2]	***dva*** *čoveka* '**deux** *hommes*'
ComplPrep	complément de préposition	*kolač* ***od*** *višanja* 'gâteau **aux** cerises'
ConjCoord[3]	relation entre le coordonné précédant immédiatement la conjonction de coordination et la conjonction elle-même	*Filip je* ***vredan*** *i pametan* 'Filip est **travailleur** et intelligent'
Coord	relation entre la conjonction de coordination et le dernier coordonné	*Filip je vredan* ***i*** *pametan* 'Filip est travailleur **et** intelligent'
Correl	relation entre deux éléments d'une structure corrélative	***tako*** *vruće da peče* '**si** chaud que ça brûle'

2. Le paucal est une forme casuelle spécifique, imposée aux mots qui se déclinent par les numéraux *dva* 'deux', *tri* 'trois' et *četiri* 'quatre'. Il s'agit d'une trace de l'ancien dual.

TABLE 2.1: Jeu d'étiquettes syntaxiques ParCoLab

Etiquette	Définition	Exemple
DepAdjAdj	dépendant d'un adjectif sous forme d'un autre adjectif	*sav* ***zadihan*** 'tout **essouflé**'
DepAdjAdv	dépendant d'un adjectif sous forme d'un adverbe	*Jedva* ***vidljiv*** *i sasvim* ***tih*** 'A peine **visible** et complètement **silencieux**'
DepAdjCas	dépendant d'un adjectif sous forme d'un nom fléchi	***sličan*** *ocu* '**semblable** père.DAT'
DepAdjPrep	dépendant d'un adjectif sous forme d'un groupe prépositionnel	*On je* ***zaljubljen*** *u Milicu* 'Il est **amoureux** de Milica'
DepAdvAdv	dépendant d'un adverbe sous forme d'un adverbe[4]	*još* ***dugo*** 'encore **longtemps**'
DepAdvCas	dépendant d'un adverbe sous forme d'un nom fléchi	***mnogo*** *ljudi* '**beaucoup** gens.GEN'
DepAdvPrep	dépendant d'un adverbe sous forme d'une préposition	***više*** *od njega* '**plus** *que* lui'
DepEx_	ellipse : préfixe ajouté à l'étiquette de l'élément dont le gouverneur est élidé [5]	(cf. section 2.12)
DepNAdj	dépendant de nom sous forme d'un adjectif	*dugo* ***pismo*** ' longue **lettre**'
DepNCas	dépendant de nom sous forme d'un nom fléchi	***zrak*** *sunca* '**rayon** soleil.GEN'
DepNPrep	dépendant de nom sous forme d'un groupe prépositionnel	***pismo*** *od Filipa* '**lettre** de Filip'
DepVAdv	dépendant d'un verbe sous forme d'un adverbe	*Filip lepo* ***peva*** 'Filip **chante** bien'
DepVCas	dépendant d'un verbe sous forme d'un nom fléchi et qui n'est pas un ObjDir, ObjIndir ou prédicatif	***Plaši*** *se grmljavine* 'Il **a peur** tonnerre.GEN'
DepVInf	dépendant infinitif d'un verbe	***prestati*** *plakati* '**arrêter** pleurer'
DepVPart	dépendant d'un verbe introduit par un participe présent ou passé	***Vratio*** *se pevajući* 'Il est **rentré** en chantant'
DepVPrep	dépendant d'un verbe sous forme d'un groupe prépositionnel qui n'est pas un ObjIndir ou un prédicatif	***Učestvovao*** *je u organizaciji* 'Il a **participé** dans l'organisation'
Emph	dépendant de la racine de la proposition exprimant l'emphase, privé de fonction syntaxique	*To* ***dolazi*** *zima* lit. 'Ça **vient** hiver'
ExtraPred	dépendant de la racine de la proposition sous forme d'un adverbe extra-prédicatif	*On zapravo* ***kasni*** 'Il **est** en fait en retard'
Interrog	dépendant de la racine de la proposition sous forme d'un marqueur d'interrogation	***Dolazi*** *li Filip ?* 'Est-ce que Filip **vient** ?'

4. Il s'agit typiquement d'intensifieurs.
5. Le traitement de l'ellipse a été repris de Prague Dependency Treebank (Hajič et al., 1999).

TABLE 2.1: Jeu d'étiquettes syntaxiques ParCoLab

Etiquette	Définition	Exemple
Juxt	juxtaposition de deux éléments de haut niveau où aucune autre relation ne s'applique	*Posledice su se brzo* ***osetile*** *: vrtoglavica i mučnina* 'Les conséquences se sont vite fait **sentir** : vertige et nausée'
Neg	négation (verbale ou nominale)	*Ne* ***dolazi*** 'Il ne **vient** pas'
ObjDir	objet direct, à l'accusatif ou au génitif	*Filip* ***jede*** *jabuku* 'Filip **mange** pomme.ACC'
ObjIndirCas	objet indirect au datif	*Filip* ***daje*** *jabuku Ani* 'Filip **donne** pomme.ACC *Ana.DAT*
ObjIndirPrep	objet indirect réalisé comme *o* 'de' + N_locatif	*Filip* ***misli*** *o putovanju* 'Filip **pense** de voyage.LOC'
Polylex	relation réunissant les éléments d'une locution prépositionnelle ou adverbiale ou d'une conjonction complexe	*Dolazi* ***zato*** *što mora* 'Il vient **parce** qu'il est obligé'
PredCompletive	relation entre le subordonnant et le prédicat de la complétive	*Zna* ***da*** *dolazim* 'Il sait **que** je viens'
PredPercont	relation entre le prédicat de la principale et le prédicat de la percontative	***Pitao*** *je zašto dolaze* 'Il a **demandé** pourquoi ils venaient'
PredRap	relation entre le verbe introductif et le verbe principal du discours rapporté	*« Dolazim »,* ***kaže****.* « J'arrive », **dit**-il.
PredRel	relation entre l'antécédent d'une relative et son prédicat	***čovek*** *koji je došao* '**l'homme** qui est venu'
PredSub	relation entre le subordonnant et le prédicat de la subordonnée	*Dolazi* ***kad*** *završi* 'Il vient **quand** il finit'
PredicAdv	prédicatif adverbial : complément adverbial du verbe *biti* 'être'	*Filip* ***je*** *u Beogradu* 'Filip **est** à Belgrade'
PredicComplObj	prédicatif complémentaire lié à l'objet direct : complément nominal, adjectival ou prépositionnel d'un verbe obligatoirement attributif autre que le verbe *biti* 'être'	***Proglasili*** *su ga kraljem* 'Ils l'ont **proclamé** roi.INS'
PredicComplSuj	prédicatif complémentaire lié au sujet : complément nominal, adjectival ou prépositionnel d'un verbe obligatoirement attributif autre que le verbe *biti* 'être'	***Proglasio*** *se kraljem* 'Il s'est **proclamé** roi.INS'
PredicNom	prédicatif nominal : complément nominal du verbe *biti* 'être'	*Filip* ***je*** *profesor* 'Filip **est** professeur'
PredicOpt	prédicatif optionnel : complément nominal ou adjectival d'un verbe optionnellement attributif	*Filip se* ***vratio*** *umoran* 'Filip est **rentré** fatigué'
Ref	relie le verbe au pronom réflexif	***Osvežio*** *se* 'Il s'est **rafraîchi**'
Root	relie la racine externe et la tête de la phrase	***ROOT*** *Dolazi sutra* '**ROOT** Il vient demain'

TABLE 2.1: Jeu d'étiquettes syntaxiques ParCoLab

Etiquette	Définition	Exemple
Sub	relation entre le verbe principal et le subordonnant introduisant une proposition subordonnée	***Dolazi*** *kad* *završi* 'Il **vient** quand il finit'
Suj	sujet grammatical, exprimé par le nominatif	*Filip* *je* ***stigao*** 'Filip est **arrivé**'
SujLog	sujet logique, exprimé par le datif, génitif ou accusatif	*Filipu* ***je*** *dosadno* 'Filip s'**ennuie**' (lit. 'Filip.DAT est ennuyeux')
Ponct	relie la ponctuation au premier token précédent qui n'en est pas une	***Razočaran*** , *vratio* *se* ***kući*** . '**Déçu** , il est rentré à la **maison** .'

Les modifications les plus importantes par rapport à la tradition grammaticale serbe concernent le traitement des dépendants du nom et des dépendants du verbe autre que les objets et les prédicatifs. Comme la distinction entre les ajouts et les arguments est souvent basée sur des critères sémantiques plutôt que syntaxiques, elle est difficile à opérer pour un annotateur humain, et d'autant plus pour un parser. Par conséquent, nous adoptons une autre approche pour le traitement de ces éléments : nous mettons en place une série d'étiquettes sous-spécifiées, qui commencent par Dep, désignant le terme neutre de dépendant, suivi des indications de la nature morphosyntaxique du gouverneur et du dépendant. Ainsi, l'étiquette DepVPrep indique un dépendant d'un verbe sous forme d'une préposition (qui n'est pas un prédicatif ou un objet indirect casuel), tandis que DepNCas désigne un dépendant d'un nom sous forme d'un nom fléchi (cf. *padežni atribut*). Par analogie, le même type d'étiquettes a été mis en place pour les dépendants d'un adjectif et d'un adverbe. Chacune de ces étiquettes est présentée en détail dans la suite de ce document (cf. sections 2.3.10, 2.3.11, 2.3.12, 2.3.13, 2.3.14, 2.4.1, 2.4.2, 2.4.3, 2.5.1, 2.5.2, 2.5.3, 2.5.4, 2.6.1, 2.6.2 et 2.6.3).

2.2 Élément racine

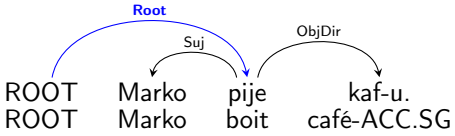
Cette section décrit la relation syntaxique Root, qui s'établit entre la racine externe de la phrase et la phrase elle-même. En effet, dans la section 1.1, nous avons mentionné que l'une des contraintes que doit respecter un arbre syntaxique réalisé dans le cadre de l'analyse en dépendances était la complétude : chaque mot de la phrase doit avoir un gouverneur. Nous avons également vu que la racine de la phrase était le verbe principal. Se trouvant au sommet de l'arbre, il n'était gouverné par aucun autre mot de la phrase. Pour contraindre la complétude de l'arbre et permettre au verbe principal aussi d'avoir un gouverneur, on introduit un nœud artificiel, représentant la racine "externe". Dans le cadre de ce projet, ce nœud se trouve à gauche de la phrase et il est marqué comme *ROOT*. Selon notre schéma d'annotation, *ROOT* a un descendant unique dans la phrase. La liste des cas de figure possibles est donnée dans la suite.

2.2.1 Verbe principal

Le dépendant prototypique de la racine externe est le **verbe principal** de la phrase.

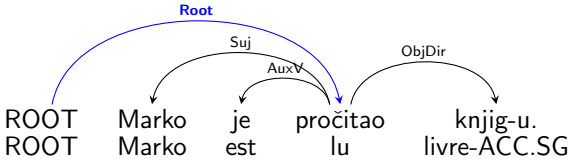
3. Le traitement de la coordination sera présenté en détail dans la section 2.10.

(1)



'Marko boit un café'

(2)



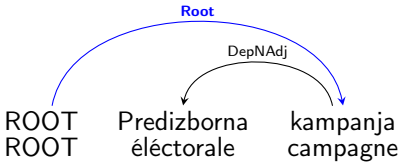
'Marko a lu un livre.'

2.2.2 Autres types de tête

Phrases averbales

Une phrase peut ne pas contenir de verbe fini (par exemple, en cas d'ellipse ou s'il s'agit d'un titre). Dans ce cas, on considère comme descendant de la racine externe la **tête du segment averbal**.

(3)

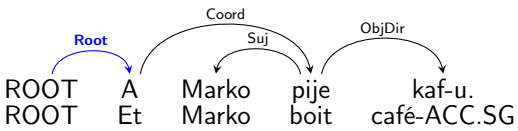


'Campagne électorale'

Phrases introduites par un connecteur

Dans le cas où une phrase est introduite par une conjonction en position initiale (ayant donc une foncton trans-phrastique, ou, plus précisément, le rôle d'un connecteur discursif), c'est la **conjonction** qui est annotée comme descendant de la racine externe. La conjonction gouverne, quant à elle, la tête de la proposition. La relation qui relie la conjonction et le verbe dépend de la nature de la conjonction : dans le cas des conjonctions de coordination (exemple 4), il s'agit de la relation Coord (cf. section 2.10), alors que pour les conjonctions de subordination (exemple 5) c'est PredSub (cf. section 2.8).

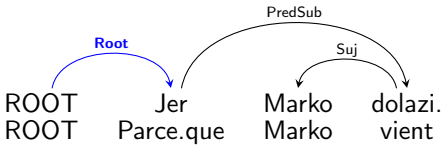
(4)



'Et Marko boit du café.'

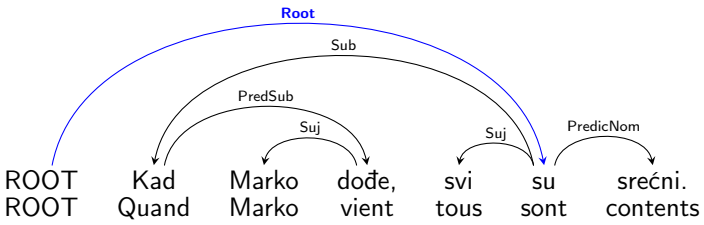
Il ne faut pas confondre ce cas de figure avec celui où la subordonnée est simplement antéposée à la principale, cf. l'exemple 6 :

(5)



'Parce que Marko vient.'

(6)



'Quand Marko vient, tout le monde est content.'

Malgré le fait qu'ici aussi la phrase commence par une conjonction de subordination, la phrase contient également la proposition principale. La racine correspond donc au verbe principal de la proposition principale, qui gouverne à son tour la subordonnée.

2.2.3 Caractère unique de la tête de phrase

Dans le cadre du projet *ParCoLab*, nous considérons qu'il n'existe qu'un descendant de la racine externe dans une phrase. Autrement dit, nous adoptons le principe de la racine unique. Ceci est notamment important dans le traitement des cas de figure suivants.

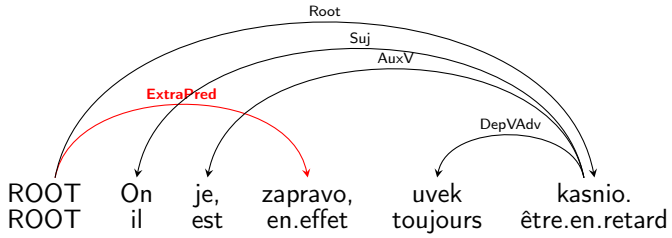
Rattachement des modifieurs phrastiques

La linguistique théorique considère que certains éléments de la phrase se trouvent au même niveau de la structure syntaxique que le contenu propositionnel lui-même : il s'agit notamment des adverbes phrastiques. Logiquement, ces éléments devraient se positionner au même niveau de l'arbre syntaxique que la tête de la proposition qu'ils modifient. Autrement dit, ils devraient être considérés comme descendants de la racine externe. Or, dans le cadre de ce travail, ces éléments ne sont pas rattachés à la racine externe, mais à la tête interne de la proposition (typiquement au verbe principal, cf. 2.7.6). Ceci est fait dans un effort d'éviter la sur-production des relations non-projectives dans l'arbre syntaxique : les modifieurs phrastiques pouvant intervenir au milieu de la phrase aussi bien qu'en début, on risquerait d'avoir des arcs qui se croisent dans la représentation de la structure syntaxique, comme dans l'exemple 7.

Pour éviter cette situation, nous adoptons le traitement dans lequel le modifieur phrastique est gouverné par le verbe principal de la proposition (cf. exemple 8).

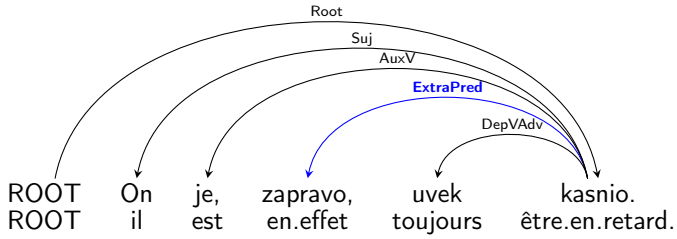
Le point problématique qui se pose ici est que l'extra-prédicatif paraît modifier le verbe au même titre qu'un ajout adverbial classique, plutôt que de modifier la phrase. Pour remédier à ce problème, nous introduisons une étiquette spécialisée ExtraPred pour les éléments extra-prédicatifs, ce qui permet d'identifier facilement ce type de dépendant en corpus (cf. section 2.7.6).

(7)



'Il était, en effet, toujours en retard.'

(8)

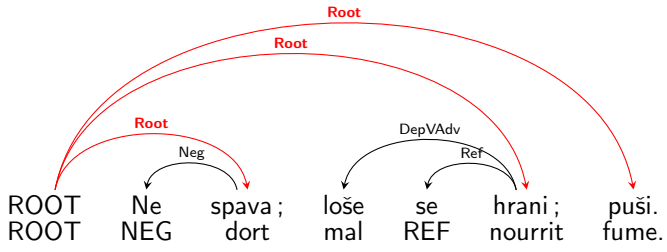


'Il était, en effet, toujours en retard.'

Juxtaposition de plusieurs propositions

Un autre cas de figure où l'on peut considérer le rattachement de plusieurs éléments à la racine externe est celui où plusieurs propositions indépendantes sont enchaînées en une phrase sans être en coordination. Dans ce cas, il peut paraître intuitif de rattacher chacun des verbes principaux directement à la racine externe (cf. exemple 9).

(9)



'Il ne dort pas ; il se nourrit mal ; il fume.'

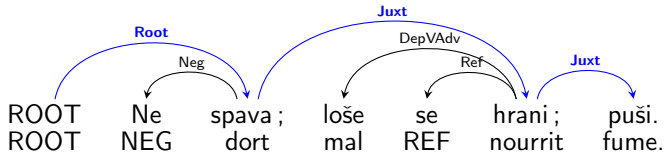
Cependant, pour maintenir la cohérence du traitement, nous considérons que le seul descendant de la racine externe est la tête de la première proposition et que les autres ont une relation de juxtaposition par rapport à elle. Le traitement adopté est donc celui montré dans l'exemple 10.

La juxtaposition est décrite en détail dans la section 2.11.

2.3 Dépendants du verbe

Cette section est destinée à la description du traitement des fonctions régies par le verbe.

(10)

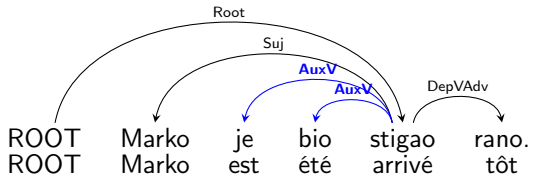


'Il ne dort pas ; il se nourrit mal ; il fume.'

2.3.1 Verbe auxiliaire

La relation AuxV a pour gouverneur le verbe principal, et pour dépendant le verbe auxiliaire d'un temps composé. Dans le cas des temps surcomposés, tous les verbes auxiliaires sont reliés directement au verbe principal par cette relation, indépendemment les uns des autres.

(11)



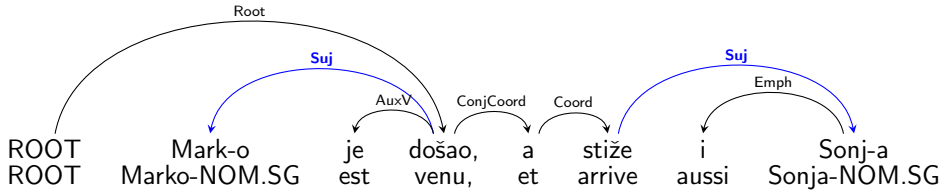
'Marko était arrivé tôt.'

Ce traitement est parallèle à celui adopté au niveau morphosyntaxique, où tous les éléments du verbe auxiliaire d'une forme verbale surcomposée sont traités comme des auxiliaires.

2.3.2 Sujet

La relation Suj correspond à la notion de *sujet grammatical* dans la tradition syntaxique serbe. Cet élément se réalise typiquement sous forme d'un **élément nominal au nominatif** qui répond à la question *Ko ?* 'qui-sujet-humain' ou *Šta ?* 'quoi-sujet-humain'. Le gouverneur de cette relation est donc le verbe principal de la proposition, alors que le dépendant peut avoir la forme d'un nom, pronom ou numéral au nominatif.

(12)

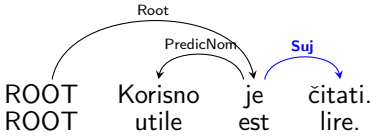


'Marko est venu, et Sonja arrive aussi.'

Il est également possible d'avoir un **infinitif** dans cette fonction (exemple 13).

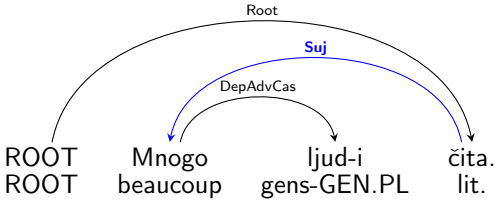
Le sujet peut également être représenté par un syntagme partitif, dont la tête est l'**adverbe de quantité** (exemple 14).

(13)



'Lire est utile.'

(14)

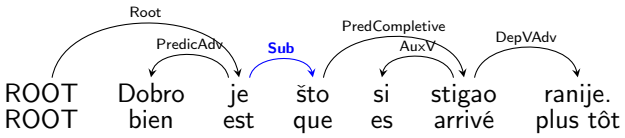


'Beaucoup de gens lisent.'

Complétives en *da* et *što*

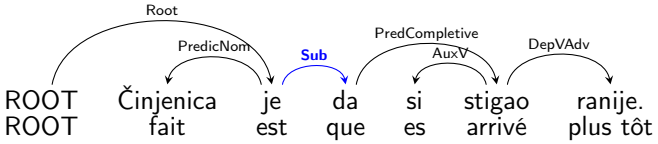
À la différence de la tradition grammaticale serbe, qui considère que les propositions en *da* et *što* dans une phrase de type *Dobro je da si tu* 'C'est bien que tu sois là' ont la fonction du sujet, nous considérons en effet qu'il s'agit des phrases impersonnelles, sans sujet, et que les propositions sont en effet des subordonnées complétives qui dépendent du verbe. Pour ces cas de figure, nous proposons donc les traitements illustrés dans les exemples 15 et 16.

(15)



'C'est bien que tu sois arrivé plus tôt.'

(16)



'Le fait est que tu es arrivé plus tôt.'

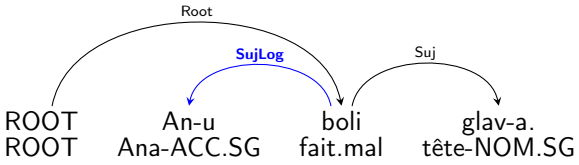
Le traitement des complétives est expliqué en détail dans la section 2.8.2.

2.3.3 Sujet logique

L'étiquette SujLog correspond à la fonction du sujet logique de la syntaxe serbe (cf. Stanojčić and Popović, 2012; Ivić, 2005). Cette fonction représente de manière générale l'**expérienceur d'un processus verbal** ou le **sujet d'un énoncé existentiel**. Il peut se réaliser **au datif, au génitif ou à l'accusatif**. Le sujet logique fait partie de la structure argumentale du verbe dont il dépend, et le remplacement par un sujet grammatical n'est pas possible. En revanche, certains verbes ouvrent deux places pour les deux types de sujet (cf.

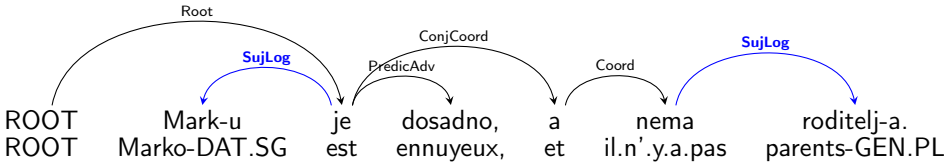
exemple 17). Ces faits confirment que le sujet grammatical et le sujet logique représentent deux fonctions syntaxiques distinctes. D'autres constructions qui contiennent le sujet logique sont listées dans les exemples 18 à 23.

(17)



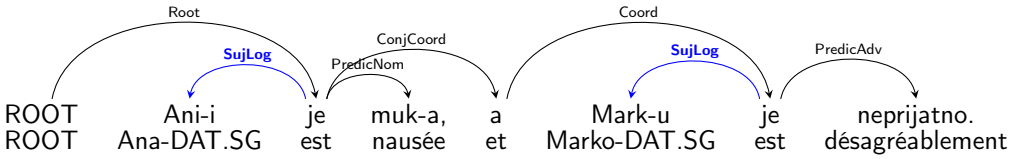
'Ana a mal à la tête.'

(18)



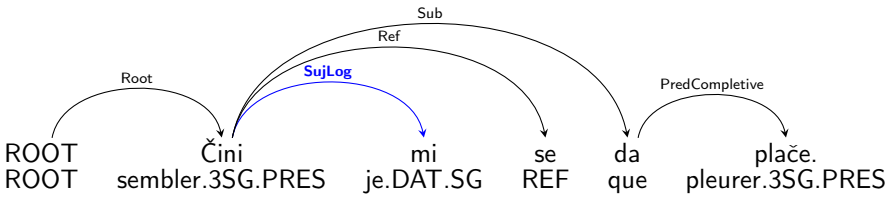
'Marko s'ennuie, et ses parents ne sont pas là'

(19)



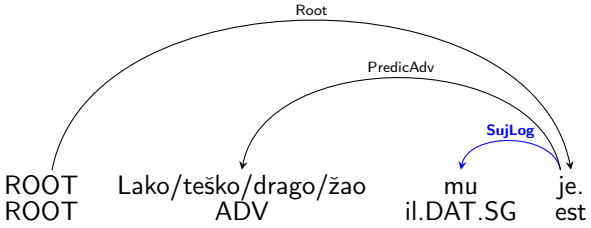
'Ana a la nausée, et Marko est mal à l'aise.'

(20)



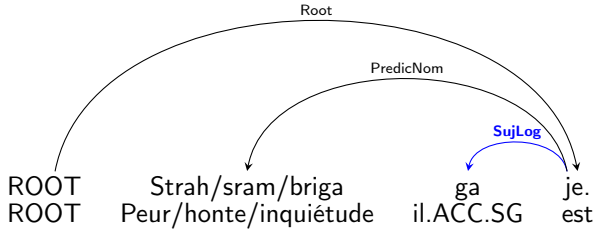
'Il me semble qu'il pleure.'

(21)



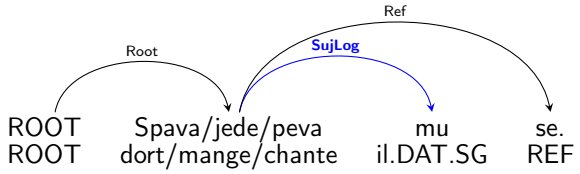
'Il trouve ça facile/difficile ; cela lui fait plaisir ; il est désolé'

(22)



'Il a peur ; il a honte ; il s'en fiche'

(23)



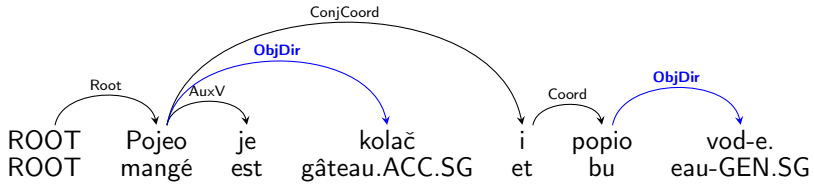
'Il a envie de dormir/manger/chanter'

À la différence de (Stanojčić and Popović, 2012), nous ne considérons pas le datif dit possessif (cf. *Popeo se ocu na kolena*, lit. 'Il est monté père.DAT sur les genoux' = 'Il est monté sur les genoux de son père') comme un sujet logique. Ce datif, tout comme le datif éthique, est traité comme un objet indirect (cf. section 2.3.5).

2.3.4 Objet direct

Cette fonction est représentée par l'étiquette ObjDir. L'objet direct en serbe a typiquement la forme d'un **élément nominal** (nom, pronom ou numéral) **à l'accusatif**. Il existe également l'objet direct **au génitif**, le génitif en question étant un génitif partitif, ou bien un génitif dit 'slave' (utilisé après la négation). Les deux réalisations (celle à l'accusatif et celle au génitif) sont annotées à l'aide de l'étiquette ObjDir (cf. exemple 24).

(24)



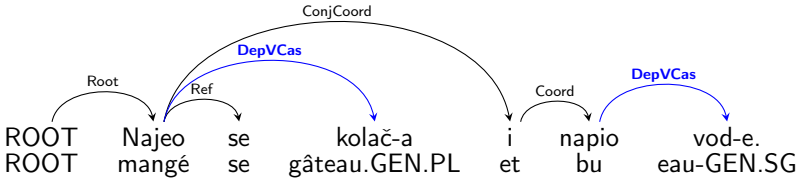
'Il a mangé le gâteau et bu de l'eau'

Pour décider si un génitif est un objet direct, nous utilisons le **test** suivant : si le génitif peut être remplacé par l'accusatif, il s'agit effectivement d'un objet direct (cf. *Popio je* ***vode*** => *Popio je* ***vodu***). En revanche, le dépendant au génitif qui ne peut pas être remplacé par un accusatif n'est pas considéré comme objet direct. Il est annoté comme dépendant verbal sous forme d'un nom fléchi (étiquette DepVCas, cf. exemple 25 [6]). Ici, il n'est pas possible de remplacer le génitif par l'accusatif : *Najeo se* ***kolača*** => **Najeo se* ***kolači***, *Napio se* ***vode*** => **Napio se* ***vodu***.

L'utilisation de l'étiquette DepVCas est présentée en détail dans la section 2.3.11.

6. Les verbes *najesti se* et *napiti se* ont une interprétation aspectuelle spécifique : ils ont une lecture perfective associée à l'idée d'une action effectuée *à satiété*.

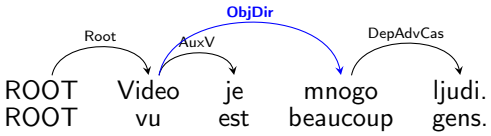
(25)



'Il a mangé assez de gâteau et bu assez d'eau.'

Tout comme dans le cas du sujet, l'objet direct peut également être représentée par un adverbe de quantité introduisant un syntagme partitif (cf. exemple 26).

(26)



'Il a vu beaucoup de gens.'

2.3.5 Objet indirect

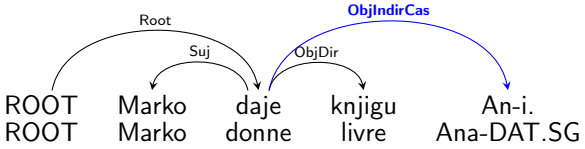
Cette étiquette est consacrée à un sous-ensemble des relations regroupées sous le nom de l'objet indirect dans (Stanojčić and Popović, 2012). En effet, ces auteurs considèrent comme objet indirect tout argument verbal qui n'est pas un objet direct et qui n'a pas un sémantisme adverbial net. Les caractéristiques de surface de ces éléments sont très disparates : il peut s'agir aussi bien des groupes nominaux à des cas différents que des groupes prépositionnels introduits par un nombre élevé de prépositions différentes (les auteurs en citent 9). Pour éviter de rassembler cette grande diversité de dépendants sous une étiquette dénotant une fonction bien précise, nous avons préféré réserver la fonction de l'objet indirect à deux cas de figure certes spécifiques mais prototypiques et de traiter les autres avec des étiquettes sous-spécifiées en DepV (cf. section 2.3).

Quant à l'objet indirect, nous introduisons deux étiquettes : ObjIndirCas, qui correspond à l'objet indirect casuel, exprimé sous forme d'un élément nominal fléchi au datif, et ObjIndirPrep, objet indirect prépositionnel, qui correspond à l'objet indirect d'un verbe de parole ou de processus mental introduit par la préposition *o* 'de', complétée par un nom au locatif.

Objet indirect casuel

ObjIndirCas correspond à la réalisation prototypique de l'objet indirect : il s'agit d'un **élément nominal au datif** exprimant typiquement le bénéficiaire ou le destinataire d'un processus verbal.

(27)

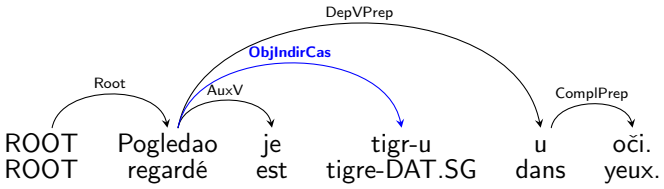


'Marko donne le livre à Ana.'

Cette étiquette est également accordée à d'autres types des dépendants verbaux au datif, comme le **datif possessif** (exemple 28) et le **datif éthique** (exemple 29). Même s'il ne s'agit

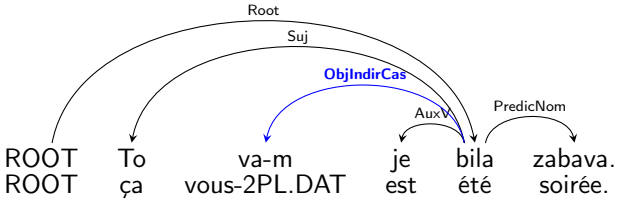
pas de véritables objets indirects dans ces deux cas, ces datifs ont le même comportement syntaxique que l'objet indirect, et le seul paramètre de distinction est le sémantisme du verbe. Ceci dépasse donc le cadre d'une annotation syntaxique de surface et sera traité dans une étape ultérieure du développement du corpus.

(28)



'Il a regardé le tigre dans les yeux.'

(29)

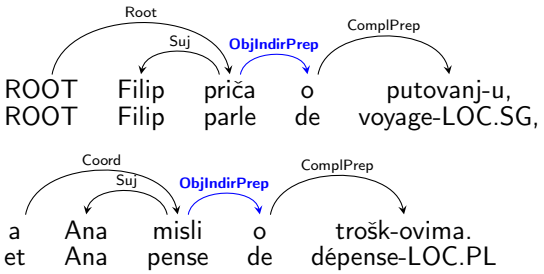


'Ça a été une vraie soirée.'

Objet indirect prépositionnel

Nous considérons comme objet indirect prépositionnel (ObjIndirPrep) les dépendants des verbes de parole et de processus mentaux ayant la forme d'un **groupe prépositionnel introduit par la préposition *o* 'de' complétée par un nom au locatif**. Tout autre dépendant prépositionnel d'un verbe (qui n'est pas un prédicatif) est traité comme DepVPrep (cf. section 2.3.12).

(30)



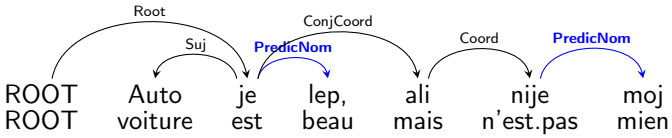
'Filip parle du voyage et Ana pense aux dépenses.'

2.3.6 Prédicatif nominal

Dans la tradition syntaxique serbe (et, plus généralement, slave), le prédicatif désigne un dépendant nominal non-objet dans l'expression d'un prédicat. La notion du prédicatif nominal (PredicNom) correspond à un élément à nature nominale introduit par le verbe *biti* 'être', qui exprime une caractéristique du sujet du même verbe. Cette fonction est donc équivalente de l'attribut du sujet dans la tradition grammaticale française, mais limitée à un

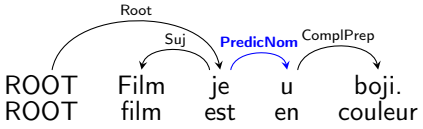
seul gouverneur (le verbe *être*). Le dépendant de cette relation peut avoir la forme d'un **nom, pronom, adjectif ou numéral au nominatif**, ou bien d'une **préposition** introduisant un groupe prépositionnel.

(31)



'La voiture est belle, mais elle n'est pas à moi.'

(32)



'Le film est en couleur.'

Les phrases comme *To je bila moja majka* 'C'était ma mère'

En serbe, le sujet impose au verbe l'accord en nombre, genre et personne. Si l'on considère la phrase dans l'exemple 33, nous constatons que le verbe porte les marques d'accord avec le nom, qui est du genre féminin, mais pas avec le pronom démonstratif, qui est du genre neutre.

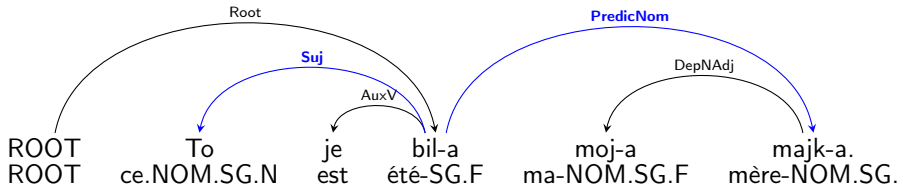
(33)

To	je	bil-a	moj-a	majk-a.
ce.NOM.SG.N	est	été-SG.F	ma-NOM.SG.F	mère-NOM.SG.

'C'était ma mère.'

Cet indice morphosyntaxique pourrait nous amener à considérer que c'est le nom qui a le rôle du sujet, et que le pronom est en effet le prédicatif nominal. Cependant, le pronom démonstratif est fonctionnellement équivalent ici d'un groupe nominal ou d'un pronom personnel, réalisations type du sujet en serbe : *Profesorka matematike/ona/to je bila moja majka* 'La professeure de maths/elle/c'était ma mère'. Nous adoptons donc l'analyse suivante : le pronom démonstratif est traité comme sujet, alors que le nom (ou l'élément à nature nominale) a la fonction du prédicatif nominal.

(34)

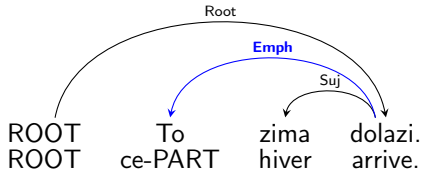


'C'était ma mère.'

NB

Ce cas de figure n'est pas à confondre avec les phrases dans lesquelles la forme *to* 'ce' n'est pas un pronom, mais une particule, et dans lesquelles elle n'a pas de rôle syntaxique. Dans la phrase *To zima dolazi* 'C'est l'hiver qui arrive', le verbe *dolaziti* 'arriver, venir' ouvre une seule place dans sa structure argumentale : celle du sujet, qui est dans cet exemple occupée par le nom *zima* 'hiver'. La particule *to* 'ce' n'a donc pas de place dans la structure syntaxique de la phrase. Son rôle est de l'ordre emphatique : elle met en valeur le contenu propositionnel. Par conséquent, ce type d'occurrences de la forme *to* est annoté avec la relation Emph (cf. exemple 35). L'utilisation de cette étiquette est décrite en détail dans la section 2.7.7.

(35)

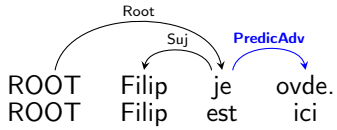


'L'hiver arrive.'

2.3.7 Prédicatif adverbial

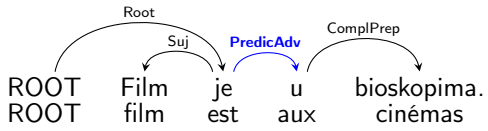
Le prédicatif adverbial (PredicAdv) est également un dépendant du verbe *biti* 'être', mais à sens adverbial. Il peut se réaliser comme un **adverbe** ou comme une **préposition** introduisant un groupe prépositionnel à sens adverbial.

(36)



'Filip est ici.'

(37)



'Le film est à l'affiche.'

NB

Comme mentionné dans la section 2.3.6, un groupe prépositionnel gouverné par le verbe *biti* 'être' peut également avoir le rôle d'un prédicatif nominal. Il faut donc veiller à faire la distinction entre l'exemple 37, où l'on trouve un prédicatif adverbial, et l'exemple 32, qui illustre un prédicatif nominal sous forme d'un groupe prépositionnel. Pour identifier la bonne étiquette, le test suivant peut être utilisé : si le groupe prépositionnel peut être remplacé par un adjectif, il s'agit d'un prédicatif nominal, et s'il peut être remplacé par un adverbe, il s'agit d'un prédicatif adverbial.

2.3.8 Prédicatif complémentaire

La fonction appelée *dopunski predikativ* 'prédicatif complémentaire' en serbe correspond dans la syntaxe française à l'attribut du sujet ou de l'objet direct avec les verbes obligatoirement attributifs autres que le verbe *être*. Il s'agit donc des dépendants obligatoires des verbes comme *zvati (se)* '(s')appeler', *proglasiti (se)* '(se) proclamer', *smatrati (se)* '(se) considérer', etc. En syntaxe serbe (cf. Stanojčić and Popović, 2012), le prédicatif complémentaire regroupe aussi bien les prédicatifs qui définissent le sujet que ceux qui sont liées à l'objet direct. Dans le cadre de ce projet, nous utilisons deux étiquettes distinctes selon ce critère, à savoir PredicComplSuj pour le prédicatif complémentaire lié au sujet et PredicComplObj pour le prédicatif complémentaire lié à l'objet direct.

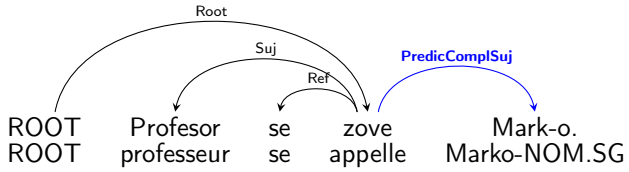
Prédicatif complémentaire lié au sujet

S'il détermine le sujet, le prédicatif complémentaire peut se réaliser sous les formes suivantes :

— un *nom*, *pronom*, *numéral* ou *adjectif* au **nominatif** ou à l'**instrumental**, ou
— la **préposition *za* 'pour'** complétée par un **accusatif**.

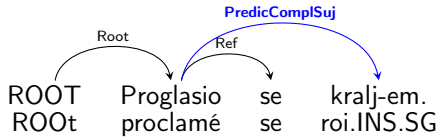
Dans le corpus, il est annoté en utilisant l'étiquette PredicComplSuj.

(38)



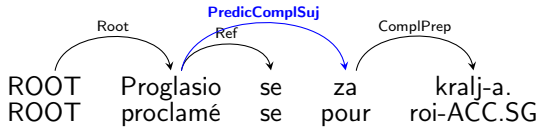
'Le professeur s'appelle Marko.'

(39)



'Il s'est proclamé roi.'

(40)



'Il s'est proclamé roi.'

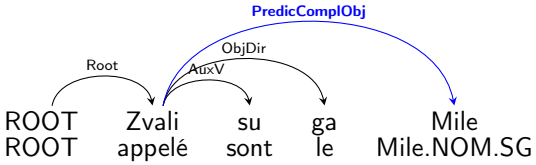
Prédicatif complémentaire lié à l'objet direct

Le prédicatif lié à l'objet direct peut avoir les mêmes formes que celui lié au sujet :

— *nom*, *pronom*, *numéral* ou *adjectif* au **nominatif** ou à l'**instrumental**, ou
— un *groupe prépositionnel* introduit par la **préposition *za* 'pour'** complétée par un **accusatif**.

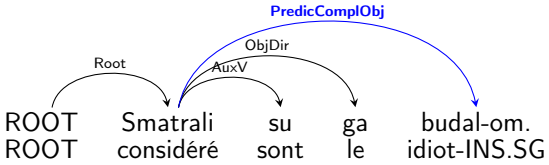
Ce qui distingue cette construction de la précédente est la présence de l'objet direct dans la phrase. Dans le corpus, il est annoté avec l'étiquette PredicComplObj.

(41)



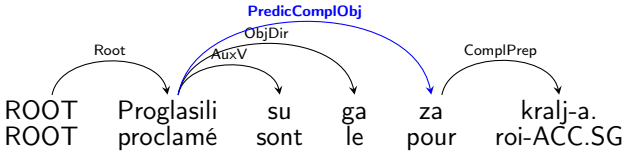
'Ils l'appelaient Mile.'

(42)



'On le considérait comme un idiot.'

(43)



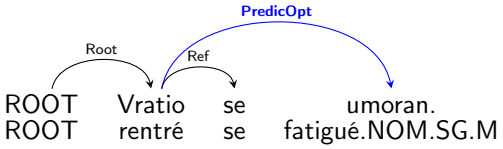
'Ils l'ont proclamé roi.'

2.3.9 Prédicatif optionnel

Le prédicatif optionnel (PredicOpt), désigné typiquement par le terme de *aktuelni kvalifikativ* dans la tradition grammaticale serbe, correspond à un dépendant verbal qui qualifie le sujet ou l'objet direct dans le cadre du processus verbal . En français, il correspond à la fonction de l'attribut du sujet ou de l'objet direct avec les verbes ocasionnellement attributifs ou à l'épithète détachée. C'est un dépendant optionnel qui peut s'associer à différents verbes, comme *zateći* 'retrouver', *stići* 'arriver', *krenuti* 'partir', etc. Il a trois réalisations principales :

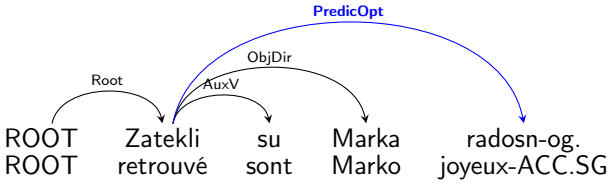
— sous forme d'un *adjectif*, il peut être au **nominatif** (s'il modifie le sujet) ou à l'**accusatif** (s'il modifie l'objet direct) ;
— sous forme d'un *nom*, *pronom* ou *numéral*, il est au **génitif** (peu importe l'élément qu'il modifie) ;
— sous forme d'un **groupe prépositionnel**, indépendemment de la fonction à laquelle il est lié.

(44)



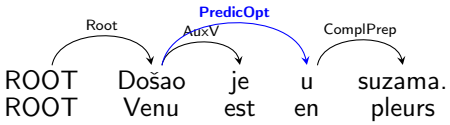
'Il est rentré fatigué.'

(45)



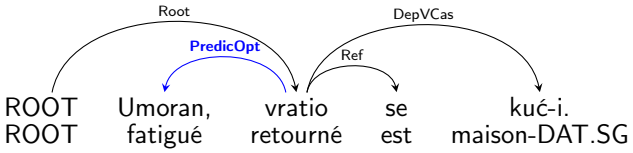
'Ils ont retrouvé Marko joyeux.'

(46)



'Il est venu en larmes.'

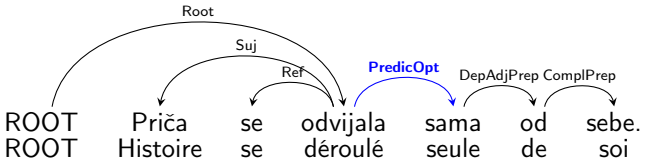
(47)



'Fatigué, il est retourné à la maison.'

Le prédicatif optionnel peut également se trouver en tête de la phrase (cf. exemple 47). La construction *sam od sebe* 'tout seul, de son gré' relève souvent du prédicatif optionnel :

(48)

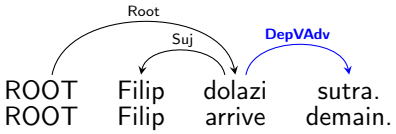


'L'histoire se déroulait toute seule.'

2.3.10 Dépendant sous forme d'un adverbe

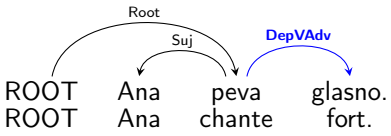
Comme il a déjà été indiqué, dans le cadre du projet *ParCoLab*, nous ne faisons pas la distinction entre les ajouts et les arguments verbaux. Par conséquent, tout **dépendant adverbial d'un verbe** qui n'est pas un prédicatif adverbial est annoté comme DepVAdv.

(49)



'Filip arrive demain.'

(50)

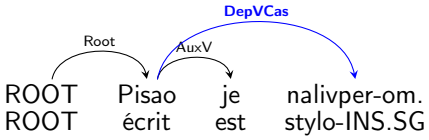


'Ana chante fort.'

2.3.11 Dépendant sous forme d'un nom fléchi

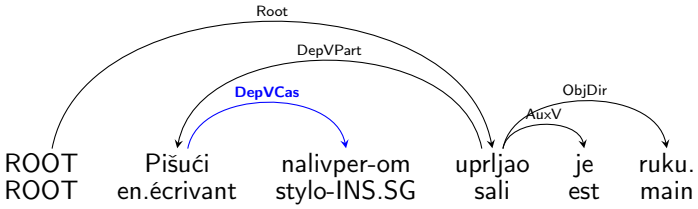
L'étiquette DepVCas regroupe tous les **dépendants casuels d'un verbe** qui ne représentent pas un sujet, sujet logique, objet direct, objet indirect ou prédicatif. Il s'agit notamment des groupes nominaux exprimant des valeurs comme celle de l'instrument, l'accompagnement, le temps, etc. Le gouverneur de cette relation peut être le verbe principal d'une proposition (cf. exemple 51) ou bien un participe présent ou passé (cf. exemple 52).

(51)



'Il écrivait au stylo.'

(52)

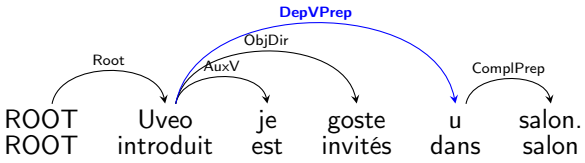


'En écrivant au stylo il s'est sali la main.'

2.3.12 Dépendant sous forme d'une préposition

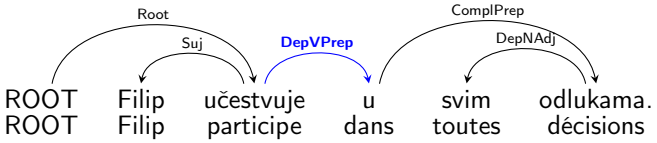
L'étiquette DepVPrep regroupe tous les **dépendants prépositionnels d'un verbe** qui ne représentent pas un objet indirect prépositionnel ou un prédicatif. Elle s'applique donc aux cas de figure considérés par Stanojčić and Popović (2012) comme des objets indirects (*strahovati od nečega* 'avoir peur *de quelque chose*'), aux ajouts à sens adverbial (*sedeti u sobi* 'être assis *dans la pièce*'), ou aux compléments adverbiaux (*ići u školu* 'aller *à l'école*').

(53)



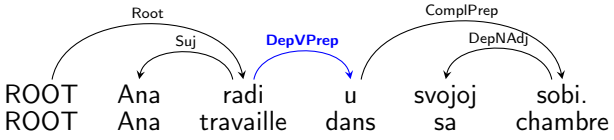
'Il a introduit les invités dans le salon.'

(54)



'Filip participe dans toutes les décisions.'

(55)

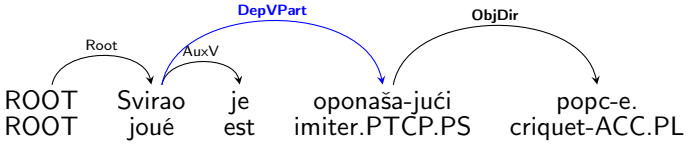


'Ana travaille dans sa chambre.'

2.3.13 Propositions participiales

Outre les deux participes utilisés dans la construction des formes verbales complexes [7], le serbe dispose de deux autres formes appelées des participes : le participe présent (*glagolski prilog sadašnji* 'adverbe déverbal présent', par ex. *radeći* 'en travaillant') et le participe passé (*glagolski prilog prošli* 'adverbe déverbal passé', par ex. *uradivši* 'ayant fait'). Malgré leur sémantisme qui pourrait être qualifié d'adverbial, ces participes ont un comportement fondamentalement différent de celui des adverbes. Plus particulièrement, ils gardent la structure argumentale du verbe dont ils sont la réalisation et peuvent avoir des dépendants verbaux typiques, notamment des objets directs et indirects. Vu cette capacité, on pourrait les rapprocher des subordonnées, mais dans la tradition grammaticale serbe on considère comme propositions seulement les constructions ayant une forme verbale finie pour noyau (cf. Stanojčić and Popović, 2012; Mrazović, 2009; Ivić, 2005). Par conséquent, nous introduisons une étiquette spéciale pour le traitement de ces formes : DepVPart. Le gouverneur de la relation est le verbe principal, et le dépendant est obligatoirement un autre verbe sous forme d'un **participe présent ou passé**. Les dépendants éventuels des participes sont, quant à eux, traités comme les dépendants d'un verbe principal.

(56)



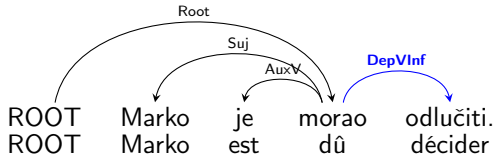
'Il jouait en imitant les criquets.'

2.3.14 Prédicat complexe : verbe modal ou aspectuel introduisant un infinitif

En serbe, les verbes modaux et aspectuels peuvent introduire deux types de constituants : un infinitif (cf. *mogu raditi* lit. 'je.peux travailler') ou une complétive en *da* 'que' introduisant un verbe au présent (cf. *mogu da radim* lit. 'je.peux que je.travaille'). Le premier type de construction est traité avec l'étiquette DepVInf, cf. les exemples 57 et 58.

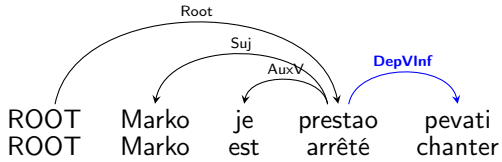
7. Dans le cadre du projet *ParCoLab*, nous dénotons comme *participe actif* la forme connue comme *glagolski pridev radni* 'adjectif verbal actif' dans la tradition grammaticale serbe, alors que le terme *glagolski pridev trpni* 'adjectif verbal passif' est traduit comme *participe passif*

(57)



'Marko devait se décider.'

(58)



'Marko a arrêté de chanter.'

Le cas de figure où ces verbes introduisent une complétive est présenté dans la section dédiée à ce type de subordonnée (cf. section 2.8.2).

2.3.15 Traitement des enchaînements des dépendants

Il est possible qu'une phrase présente plusieurs éléments qui semblent avoir le même rôle syntaxique par rapport au verbe : *Ana, Filip i Alan su došli* 'Ana, Filip et Alain sont venus', ou *Filip je putovao u Francusku, u Italiju, u Liban.* 'Filip est allé en France, en Italie, au Liban'. Dans ce cas, deux traitements sont possibles, en fonction du type de dépendant en question.

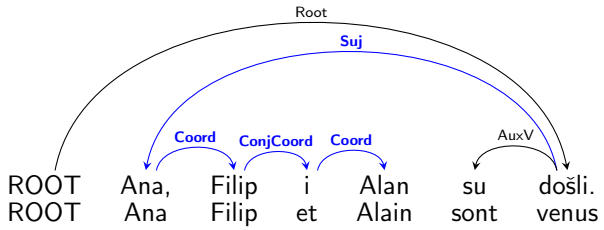
Pour les dépendants suivants :
— Suj,
— SujLog,
— ObjDir,
— ObjIndirCas,
— ObjIndirPrep,
— PredicNom,
— PredicAdv,
— PredicComplSuj,
— PredicComplObj et
— ComplPrep,

nous considérons qu'il s'agit *d'un seul dépendant complexe*, composé de plusieurs dépendants en coordination entre eux. Autrement dit, dans la phrase *Ana, Filip i Alan su došli* 'Ana, Filip et Alain sont venus', il y aura une seule relation *Suj*, qui reliera le premier élément de l'enchaînement au verbe, et les suivants seront considérés comme des dépendants coordonnés. Ce traitement se justifie par la structure argumentale des verbes : le verbe n'ouvre qu'une position pour un sujet ou un objet [8], et par conséquent, le fait d'annoter plusieurs dépendants avec l'une de ces fonctions fausserait la représentation de la structure argumentale des verbes en corpus. L'annotation correcte du premier exemple est montrée dans l'exemple 59.

Il faut noter que s'il n'y avait pas de conjonction de coordination dans la phrase, on considérerait tout de même qu'il s'agit d'une coordination. (v. section 2.10).

8. Avec quelques exceptions possibles en serbe, notamment *pitati* 'demander', qui exige un double accusatif (*pitati nekoga nešto* lit. 'demander quelqu'un quelque chose').

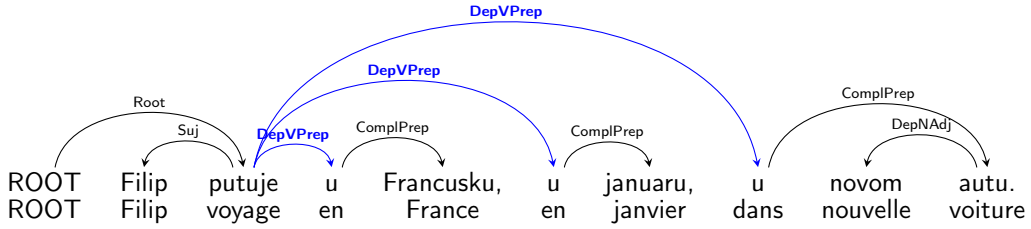
(59)



'Ana, Filip et Alain sont venus.'

En revanche, pour tout autre dépendant verbal, nous adoptons le rattachement direct de chaque élément au verbe gouverneur.

(60)



'Filip voyage en France, en janvier, dans sa nouvelle voiture.'

Ce traitement reflète le fait que les verbes peuvent admettre plusieurs réalisations des dépendants qui n'appartiennent pas à leur structure argumentale.

Nous avons ainsi passé en revue l'ensemble des dépendants du verbe. Dans la section suivante, nous présentons les dépendants du nom.

2.4 Dépendants du nom

Cette section regroupe les relations dont le gouverneur est un nom (ou une forme à comportement nominal, comme un pronom ou un numéral). Il s'agit principalement des dépendants adjectivaux, casuels et prépositionnels. Précisons que nous abandonnons la classification traditionnelle de (Stanojčić and Popović, 2012) en *kongruentni atribut*, *padežni atribut* et *atributiv* : les critères de distinction de ces fonctions étaient trop disparates pour permettre une identification fiable en corpus. À leur place, nous introduisons des étiquettes sous-spécifiées à l'instar de celles utilisées pour les dépendants verbaux. Leur utilisation est expliquée dans la suite.

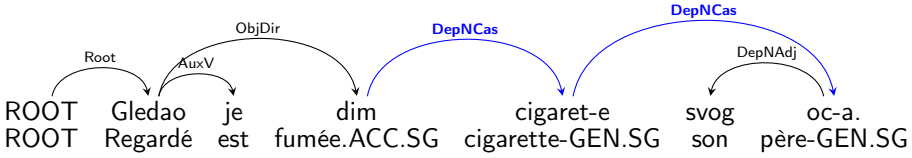
2.4.1 Dépendant sous forme d'un nom fléchi

La relation DepNCas s'applique aux dépendants du nom sous forme d'un **élément nominal à un cas oblique** (génitif, datif, accusatif ou instrumental) (cf. exemple 61).

2.4.2 Dépendant sous forme de préposition

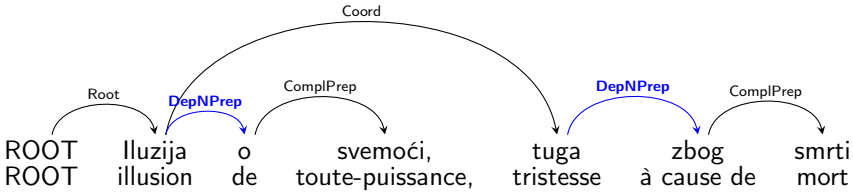
L'étiquette DepNPrep regroupe les dépendants prépositionnels d'un nom, peu importe leur sémantisme (cf. exemple 62).

(61)



'Il regardait la fumée de la cigarette de son père.'

(62)



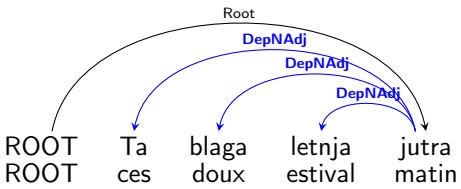
'Illusion de toute-puissance, tristesse à cause de la mort'

2.4.3 Dépendant sous forme d'adjectif

L'étiquette DepNAdj regroupe tous les dépendants adjectivaux du nom. Il s'agit ici aussi bien des adjectifs qualificatifs (***lep*** *auto* '**belle** voiture') que des autres sous-catégories (cf. ***naš*** *auto* '**notre** voiture'). Elle s'applique également aux **adjectifs dérivés des participes**.

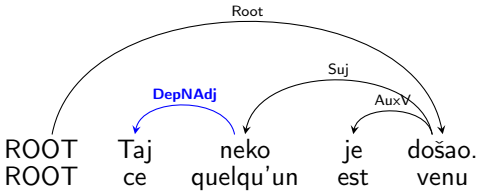
Le gouverneur de la relation est un nom ou un pronom, alors que le dépendant prend la forme d'**un adjectif ou d'un numéral qui s'accorde avec le nom**.

(63)



'Ces doux matins d'été'

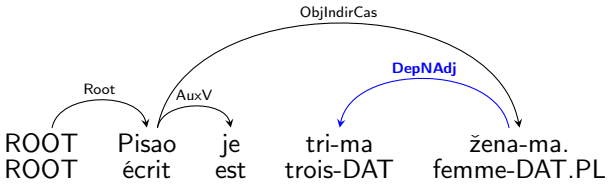
(64)



'Ce quelqu'un est venu.'

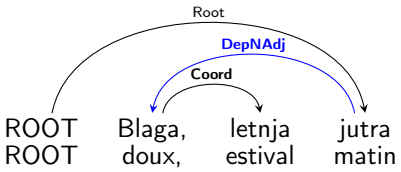
Il est à noter que lorsque plusieurs adjectifs sont placés au même niveau (sans virgule intervenante, cf. exemple 63), ils sont rattachés au nom gouverneur indépendemment les uns des autres. En revanche, si les adjectifs sont séparés par une virgule, on considère qu'il s'agit d'un seul dépendant complexe, contenant des adjectifs coordonnés (cf. exemple 66).

(65)



'Il écrivait à trois femmes.'

(66)

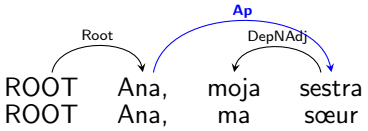


'Doux matins d'été'

2.4.4 Apposition

L'apposition est marquée par l'étiquette Ap. Cette étiquette s'applique, bien évidemment, aux cas prototypiques de l'apposition. Le gouverneur et le dépendant de la relation sont des noms, qui doivent être au même cas. Le dépendant apporte une précision par rapport au gouverneur, et il est le plus souvent séparé du contexte par des virgules (cf. exemple 67).

(67)

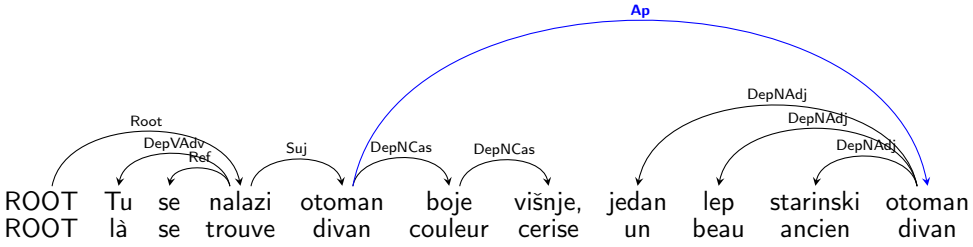


'Ana, ma sœur'

Cependant, l'étiquette Ap est également utilisée pour plusieurs cas de figure qui ne relèvent pas traditionnellement de l'apposition. Il s'agit notamment de la **répétition** du même élément dans la phrase, des **noms des personnes**, des **honorifiques** antéposés aux noms et des **enchaînements des constituants suivis d'un pronom de reprise**. Tous les cas de figure sont détaillés dans la suite.

L'exemple 68 illustre la **répétition** du même élément dans la phrase.

(68)

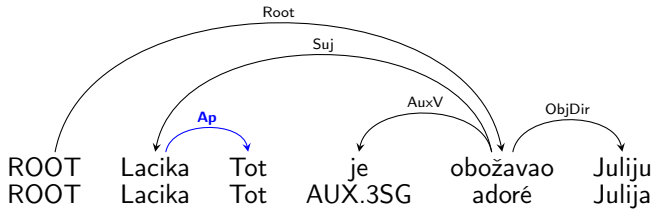


'Ici se trouve un divan couleur cerise, un beau divan à l'ancienne'

Dans le cas des **noms de personne** composés du prénom et du nom de famille, le premier élément (typiquement le prénom) porte l'étiquette de la fonction exercée par le groupe

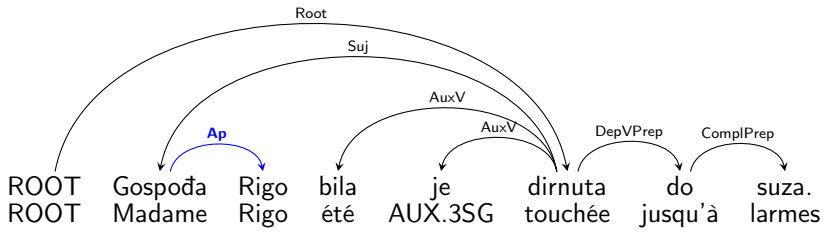
dans la phrase, et il gouverne à son tour le deuxième élément, qui porte l'étiquette Ap. Le même principe s'applique dans le cas des honorifiques devant un nom de famille, et d'autres cas relevant de la relation syntaxique connue sous le nom de *atributiv* dans la tradition grammaticale serbe. Dans ce type d'application, nous considérons systématiquement que la tête du syntagme est le nom à l'extrême gauche du groupe, et l'apposition s'établit de gauche à droite, en cascade si plusieurs éléments entretenant le même type de rapport s'enchaînent dans la phrase (cf. exemples 69, 70, 71).

(69)



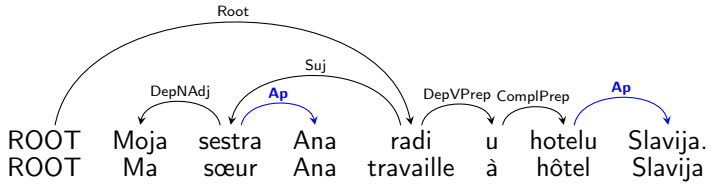
'Lacika Tot adorait Julija.'

(70)



'Madame Rigo était émue aux larmes.'

(71)



'Ma sœur Ana travaille à l'hôtel Slavija.'

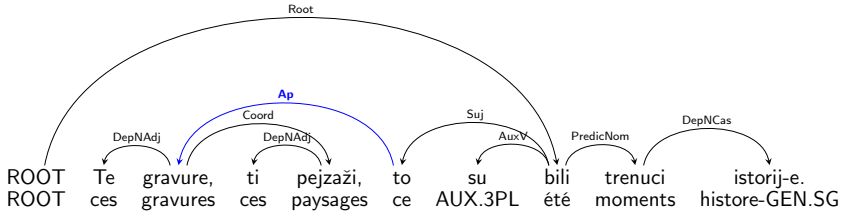
L'étiquette Ap est utilisée aussi dans le traitement d'un phénomène syntaxique plus complexe. Il s'agit des cas où un élément de la structure syntaxique est représenté par une sorte d'énumération, reprise ensuite par le pronom démonstratif *to*. Dans les cas rencontrés jusqu'à présent en corpus, il s'agit typiquement du sujet.

Comme les éléments de l'énumération ont le plus souvent des traits morphosyntaxiques disparates, alors que le prédicat s'accorde (le plus souvent) avec le pronom *to*, nous annotons le pronom comme sujet, et considérons qu'il gouverne le premier élément de l'énumération *via* la relation Ap. Les éléments de l'énumération sont reliés entre eux par la relation Coord, en cascade (cf. l'exemple 72).

Il peut parfois être délicat de déterminer s'il s'agit d'une apposition ou d'une coordination, cf. example 73.

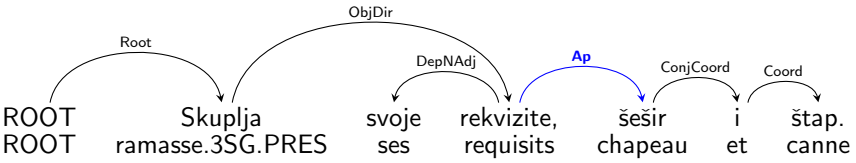
Ici, le seul critère disponible est de l'ordre référentiel : si les éléments enchaînés ont le même référent (autrement dit, s'ils désignent la même entité), il s'agit d'une apposition ;

(72)



'Ces gravures, ces paysages, c'était des moments de l'histoire.'

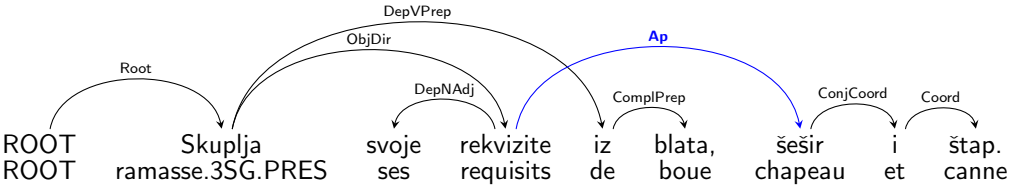
(73)



'Il ramasse ses requisits, le chapeau et la canne.'

sinon, il s'agit d'une coordination. Un autre critère qui peut être utile est l'insertion d'un autre élément entre le premier nom et le reste du groupe nominal (cf. example 74).

(74)



'Il ramasse ses requisits, le chapeau et la canne.'

Le positionnement du groupe prépositionnel indique qu'il s'agit plutôt d'une apposition, vu que ce type de clivage d'une coordination est moins probable.

Cet aperçu des dépendants du nom montre que leur inventaire est beaucoup plus restreint que celui des dépendants du verbe. Dans la section qui suit, nous présentons les dépendants de l'adjectif.

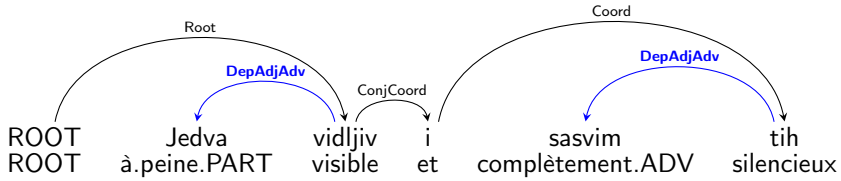
2.5 Dépendants de l'adjectif

Cette section regroupe les relations gouvernées par l'adjectif. En effet, en serbe l'adjectif peut avoir des dépendants adverbiaux (typiquement des intensifieurs comme *vrlo* 'très'), mais aussi des dépendants casuels et prépositionnels, et il existe également un cas de figure dans lequel un adjectif en gouverne un autre. Tous ces types de dépendants sont présentés ci-dessous.

2.5.1 Dépendant sous forme d'un adverbe

L'étiquette DepAdjAdv s'applique aux **dépendants adverbiaux** d'un adjectif. Cette relation peut également être portée par une particule à sens adverbial gouvernée par un adjectif (cf. exemple 75).

(75)



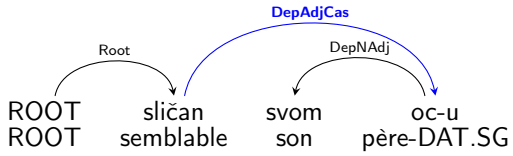
'À peine visible et complètement silencieux'

2.5.2 Dépendant sous forme d'un nom fléchi

L'étiquette DepAdjCas est appliquée aux **dépendants casuels** d'un adjectif. Il s'agit des éléments nominaux à un cas oblique (génitif, datif, accusatif ou instrumental).

Il peut s'agir d'un dépendant exigé par l'adjectif, comme dans l'exemple 76.

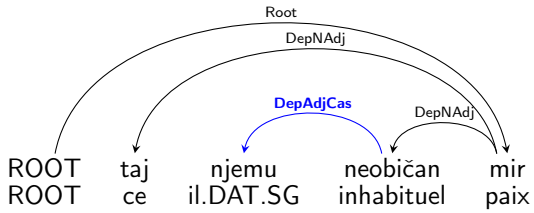
(76)



'semblable à son père'

Il peut également s'agir des compléments qui ne sont pas exigés par l'adjectif (cf. exemple 77).

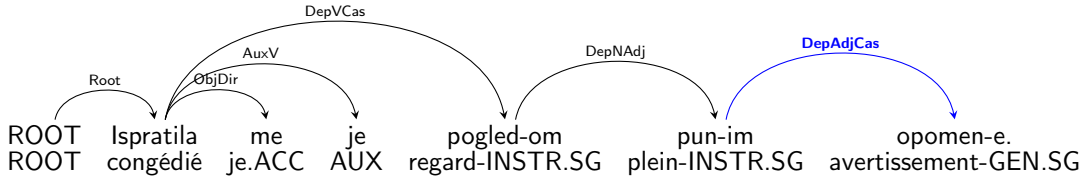
(77)



'cette paix qu'il trouvait inhabituelle'

Il faut remarquer le fait que les adjectifs ayant un dépendant sous forme d'un nom fléchi sont souvent post-posés au nom (cf. exemple 78).

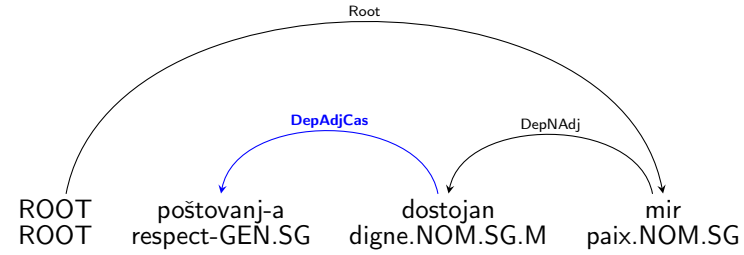
(78)



'Elle m'a congédié avec un regard plein d'avertissements.'

En revanche, si un tel adjectif est antéposé au nom, son dépendant sous forme de nom fléchi se trouve antéposé à lui (cf. exemple 79).

(79)

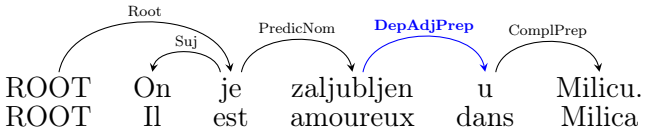


'Un calme digne de respect'

2.5.3 Dépendant prépositionnel

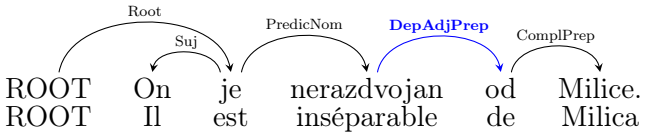
L'étiquette DepAdjPrep est dédiée aux **dépendants prépositionnels** d'un adjectif. Tout comme dans le cas des dépendants casuels, il peut s'agir d'un dépendant exigé par l'adjectif (cf. exemples 80 et 81).

(80)



'Il est amoureux de Milica'

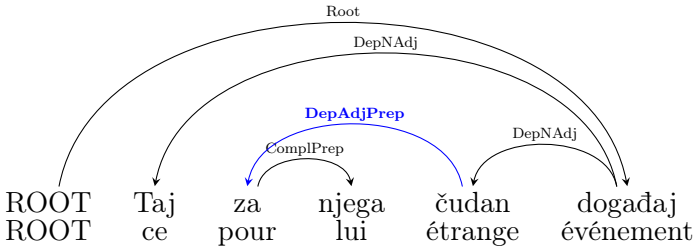
(81)



'Il est inséparable de Milica'

Cependant, il peut également s'agir des dépendants qui ne sont pas exigés par l'adjectif (cf. exemple 82).

(82)



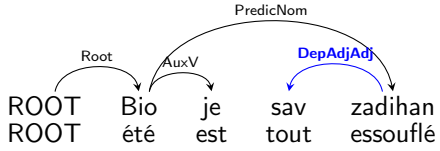
'cet événement qui lui était étrange'

Nous remarquons le même patron de linéarisation que pour les adjectifs dotés d'un dépendant casuel : typiquement, un adjectif doté d'un dépendant prépositionnel se trouve à droite du nom. S'il est tout de même antéposé au nom, son dépendant prépositionnel se trouve à sa gauche (cf. exemple 82).

2.5.4 Construction *sav* 'tout' + Adjectif

L'étiquette DepAdjAdj est utilisée pour la construction dans laquelle un adjectif (typiquement qualificatif) gouverne l'adjectif indéfini *sav* 'tout' (cf. exemple 83).

(83)



'Il était tout essouflé'

À la différence du français, où cette fonction est exercée par un adverbe, en serbe il s'agit d'un adjectif, ce qui est confirmé par sa capacité d'être décliné, ainsi que par la variation en nombre et en genre : *Zatekli su ga* ***sveg*** *zadihanog* 'Ils l'ont retrouvé tout.ACC essouflé'. Nous déterminons que l'adjectif qualificatif gouverne l'indéfini grâce au fait que ce dernier peut être omis de la phrase sans compromettre sa grammaticalité (*Bio je zadihan*, *Zatekli su ga zadihanog*), alors que le qualificatif non (**Bio je sav*, *Zatekli su ga sveg*).

Nous avons ainsi parcouru les dépendants possibles d'un adjectif. Dans la suite, nous présentons les dépendants de l'adverbe.

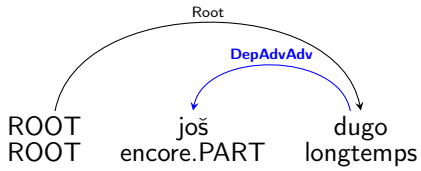
2.6 Dépendants de l'adverbe

Cette section est consacrée aux relations destinées au traitement des dépendants de l'adverbe. En serbe, un adverbe peut avoir des dépendants adverbiaux, prépositionnels ou casuels. Les trois types sont présentés dans la suite.

2.6.1 Dépendant sous forme d'adverbe

La relation DepAdvAdv est consacrée à l'annotation des **dépendants adverbiaux** d'un adverbe sous forme d'un autre adverbe ou de particule (cf. exemple 84).

(84)



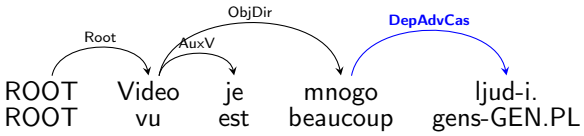
'encore longtemps'

2.6.2 Dépendant sous forme de nom fléchi

La relation DepAdvCas est destinée à l'annotation des dépendants casuels d'un adverbe. Il s'agit en fait des syntagmes dits *partitifs* dans la tradition grammaticale du serbe : la tête du syntagme est un adverbe de quantité comme *malo* 'peu' ou *mnogo* 'beaucoup', qui exige un complément au génitif. Dans ce cas, nous considérons que la tête de la relation est l'adverbe, et le complément au génitif est annoté comme DepAdvCas.[9] Notons que l'adverbe lui-même porte la fonction de l'ensemble du syntagme (cf. exemple 85).

9. Les syntagmes partitifs peuvent également avoir une tête nominale, notamment sous forme d'un nom exprimant la quantité comme *čaša* 'verre' ou *gomila* 'tas', qui exigent le même type de complément au génitif. Ce cas de figure est traité avec l'étiquette DepNCas (cf. la section 2.4.1).

(85)

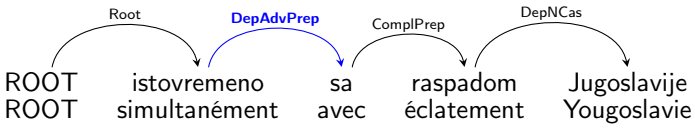


'Il a vu beaucoup de gens'

2.6.3 Dépendant sous forme de préposition

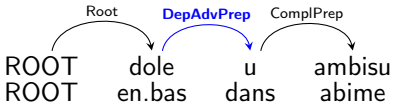
L'étiquette DepAdvPrep sert à annoter les **dépendants prépositionnels** d'un adverbe (cf. les exemples 86 à 89).

(86)



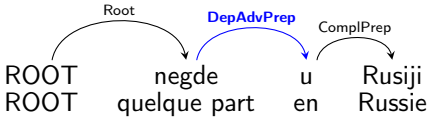
'simultanément avec l'éclatement de la Yougoslavie'

(87)



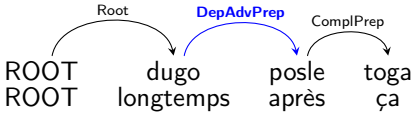
'en bas dans l'abîme'

(88)



'quelque part en Russie'

(89)



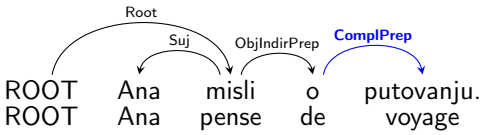
'longtemps après ça'

2.7 Relations diverses

2.7.1 Complément de préposition

L'étiquette ComplPrep est utilisée pour relier la préposition avec son dépendant. Pour rappel, la préposition elle-même est annotée avec l'étiquette de la fonction exercée par le groupe prépositionnel dans la phrase.

(90)



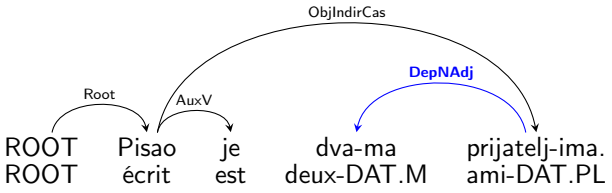
'Ana pense au voyage'

2.7.2 Complément de numéral

Les numéraux cardinaux *dva* 'deux', *tri* 'trois' et *četiri* 'quatre' en serbe se déclinent. Ils prennent donc par défaut le même cas que le nom, et le numéral *dva* s'accorde également au genre (cf. exemple 91). Par ailleurs, dans ce type de construction, le numéral est omissible, ce qui n'est pas le cas du nom (cf. *Pisao je prijateljima* 'Il a écrit à des amis' *vs* **Pisao je dvama* lit. 'Il a écrit à deux'). Par conséquent, nous considérons que c'est le nom qui est le gouverneur de la relation, et comme le numéral se comporte dans ce contexte comme tout adjectif antéposé au nom, on lui accorde l'étiquette DepNAdj (cf. section 2.4.3).

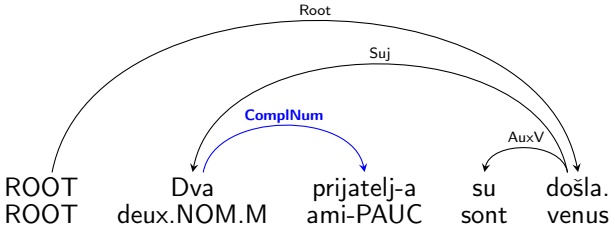
Cependant, dans certains cas, le numéral peut rester invariable et imposer par ailleurs une forme spécifique aux noms masculins et neutres (cf. exemple 92). La forme en question s'appelle *paucal* et elle est un résidu du dual de l'ancien slave, dont l'apparition est limitée à ce contexte spécifique. Dans cette configuration, c'est le numéral qui impose la réalisation morphosyntaxique du nom, ce qui implique que c'est le numéral qui gouverne la relation. Le critère d'omissibilité corrobore également cette hypothèse : *Dva su došla* 'Deux sont venus' reste grammatical, ce qui n'est pas le cas de **Prijatelja su došla* 'Amis sont venus'. Par conséquent, nous annotons le numéral avec la fonction exercée par le groupe nominal dans la phrase, et le nom est ratacché au numéral via l'étiquette ComplNum (cf. exemple 92).

(91)



'Il a écrit à deux amis.'

(92)



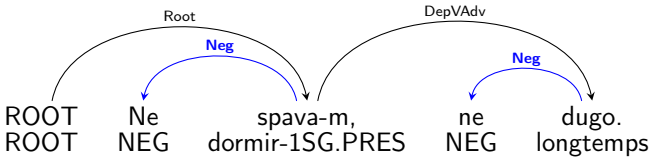
'Deux amis sont venus.'

2.7.3 Négation

L'étiquette Neg est accordée à la particule de négation *ne* 'non, pas' quel que soit son gouverneur. En effet, elle est porteuse de la négation verbale, mais peut également être asso-

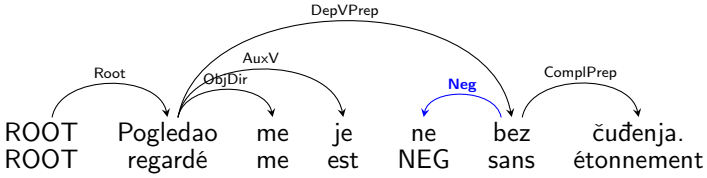
ciée à d'autres parties du discours, par exemple aux noms ou aux adverbes (cf. exemples 93 et 94).

(93)



'Je ne dors pas, pas pour longtemps.'

(94)



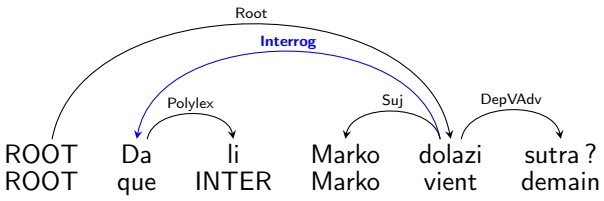
'Il m'a regardé, pas sans étonnement.'

2.7.4 Interrogation

La relation Interrog est utilisée pour annoter les **marqueurs de la modalité interrogative** comme *da li* (cf. exemple 95) et *li* (cf. exemple 96). Le gouverneur de la relation est le verbe principal de la proposition interrogative.

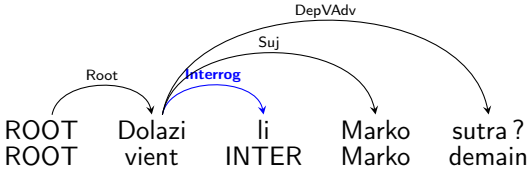
Le marqueur complexe *da li* est considéré comme une forme polylexicale ; par conséquent, ses éléments sont reliés entre eux par l'étiquette Polylex (cf. section 2.7.8).

(95)



'Est-ce que Marko vient demain ?'

(96)

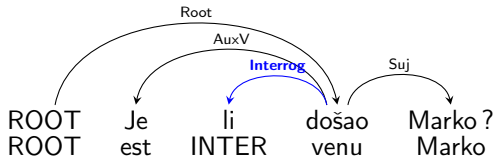


'Marko vient-il demain ?'

Il faut faire attention à distinguer les cas dans lesquels la forme *je li* est un marqueur d'interrogation de ceux où il s'agit d'une combinaison de la forme fléchie du verbe *jesam* 'être' et du marqueur simple *li* : dans l'exemple *Je li došao Marko ?* 'Marko est-il venu ?',

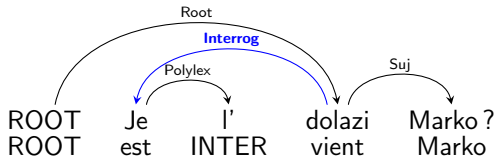
la forme *je* correspond effectivement à une forme de l'auxiliaire *jesam* 'être' faisant partie de la forme complexe du parfait *je došao* 'est venu'. Ceci peut être démontré en substituant un sujet au pluriel, qui entraîne également la modification de l'auxiliaire à cause des règles d'accord : *Jesu li došli prijatelji ?* 'Les amis sont-ils venus ?', et non **Je li došli prijatelji ?*. Dans ce cas, *je* doit être traité en accord avec sa nature et sa fonction dans la proposition (cf. exemple 97). En revanche, dans le langage parlé, *je li* (et sa forme abréviée *je l'*) peut avoir la fonction d'un véritable marqueur d'interrogation : *Je l' dolazi Marko ?*. Ici, la forme *je* ne peut pas être interprétée comme une forme fléchie du verbe *jesam* 'être', et *je li* bénéficie du même traitement que *da li* (cf. exemple 98).

(97)



'Marko est-il venu ?'

(98)



'Est-ce que Marko vient ?'

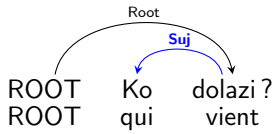
Traitement des phrases interrogatives

Le traitement de l'interrogation totale indiquée par un marqueur d'interrogation a été montré ci-dessus. L'interrogation partielle bénéficie, en revanche, d'un traitement différent.

Dans le cas des questions comme ***Ko*** *dolazi ?* '**Qui** vient ?' ou ***Kada*** *počinje film ?* '**Quand** est-ce que commence le film ?', les formes interrogatives en gras ne sont pas de purs marqueurs de modalité. Il s'agit en effet d'un pronom interrogatif dans le cas de *ko* 'qui', et d'un adverbe interrogatif dans celui de *kada*. Les deux formes exercent des fonctions par rapport au verbe : dans le premier cas, le pronom a le rôle du sujet, et dans le deuxième, l'adverbe a le rôle d'un dépendant adverbial du verbe. Pour montrer l'équivalence de *ko* et d'un sujet typique, il suffit de comparer les phrases *Marko dolazi* 'Marko vient' et *Ko dolazi ?* 'Qui vient ?' : *Marko* et *ko* sont tous les deux au nominatif, et expriment l'agent du processus verbal. Le pronom interrogatif porte donc l'étiquette du sujet (cf. exemple 99). Un procédé comparable peut être utilisé pour l'analyse du deuxième exemple : *Kada dolazi Marko ?* 'Quand est-ce que Marko vient ?' peut être comparé à *Sutra dolazi Marko*, lit. 'Demain vient Marko'. Dans cette phrase canonique, il est facile d'identifier le rôle de l'adverbe *sutra* : il s'agit d'un dépendant du verbe *dolaziti* sous forme d'un adverbe. Il porterait donc l'étiquette DepVAdv. L'adverbe interrogatif *kada* dans la phrase de départ exerce exactement le même rôle par rapport à son verbe ; par conséquent, il porte la même étiquette (cf. exemple 100).

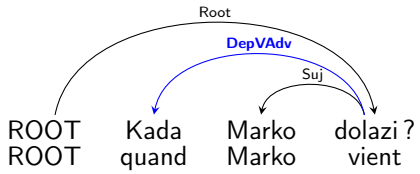
Dans le traitement des phrases interrogatives, il est donc utile de rétablir l'ordre canonique de mots et de remplacer la forme interrogative par un mot équivalent (nom, pronom ou adverbe, en fonction du cas traité). Par exemple, la phrase *S* ***kime*** *Marko dolazi ?* 'Avec qui vient Marko ?' peut se transformer en *Marko dolazi s* ***kime*** 'Marko vient avec qui', puis en

(99)



'Qui vient ?'

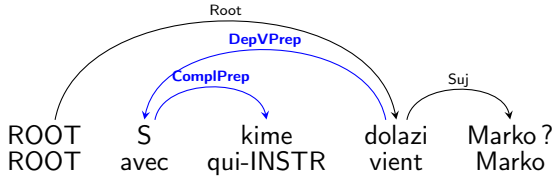
(100)



'Quand est-ce que Marko vient ?'

Marko dolazi s ***Anom*** 'Marko vient avec Ana'. En analysant cette phrase à ordre de mots canonique, nous voyons que la préposition *s* 'avec' a la fonction DepVPrep par rapport au verbe *dolaziti* 'venir', et que la forme à l'instrumental *Anom* est le complément de cette préposition. Le même traitement peut être transposé sur la phrase de départ : la proposition *s* 'avec' est un DepVPrep du verbe *dolaziti* 'venir', alors que l'instrumental *kime* est le ComplPrep de la préposition (cf. exemple 101).

(101)

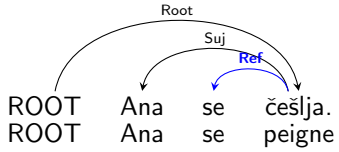


'Avec qui est-ce que Marko vient ?'

2.7.5 Réflexif

L'étiquette Ref est utilisé pour annoter le pronom réflexif sous forme clitique *se* 'se'. [10] Le gouverneur de la relation est le verbe pronominal.

(102)



'Ana se peigne.'

Bien qu'on puisse distinguer plusieurs types de pronoms réfléxifs, ils seront tous annotés avec l'étiquette unique Ref. Ce choix est dû au fait qu'il n'y a pas de critères syntaxiques suffisamment fiables qui permettent de faire cette distinction. Dans la phrase 103, il semble

10. Remarque : en serbe, il n'existe qu'une forme du pronom réflexif pour tous les nombres et toutes les personnes.

évident que le pronom réflexif a le rôle de l'objet direct par rapport au verbe *osvežiti* 'rafraîchir' ; il peut par ailleurs être remplacé par un autre type d'objet (cf. *Krenuo sam da* ***ga*** *osvežim* 'Je partis le rafraîchir'). Mais le statut du réflexif dans la phrase 104 est moins clair. À la base, *prekrstiti* 'signer' est un verbe transitif et le pronom réflexif devrait donc avoir le rôle de l'objet direct. Cependant, dans la langue contemporaine, ce verbe semble réduit à sa forme pronominale, et un remplacement du pronom réflexif par un autre objet n'est pas possible : **Majka* ***me*** *prekrsti* 'Ma mère me signa'. Il est donc difficile d'établir les critères de différenciation même dans le cadre d'une annotation manuelle, alors que pour un parser les différences seront presque certainement inaccessibles et très difficiles à généraliser. Par conséquent, nous retenons une seule étiquette pour tous les types du pronom réflexif.

(103)

Krenuh da se osveži-m.
partis que REF rafraîchir-1SG.PRES.

'Je partis me rafraîchir.'

(104)

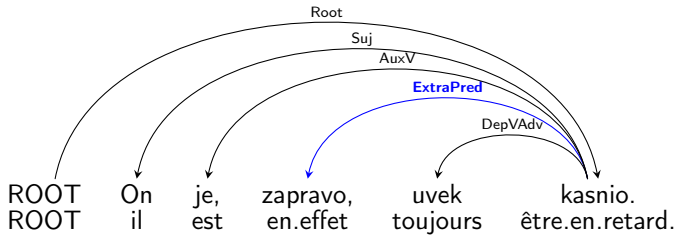
Moja majka se prekrsti.
ma mère REF signa.

'Ma mère se signa.'

2.7.6 Éléments extra-prédicatifs

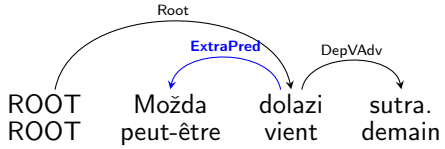
Les **éléments extra-prédicatifs** sont traités à l'aide de l'étiquette ExtraPred. Comme il a été mentionné dans la section 2.2.3, on peut considérer que ces éléments, étant des modifieurs phrastiques, doivent se trouver au même niveau de la structure syntaxique que le contenu propositionnel. En revanche, dans le cadre de ce projet, ils seront annotés comme étant gouvernés par la tête de la proposition qu'ils modifient. Cette approche permet de minimiser la création des arcs non-projectifs, étant donné que les éléments extra-prédicatifs peuvent aparaître à l'intérieur de la proposition aussi.

(105)



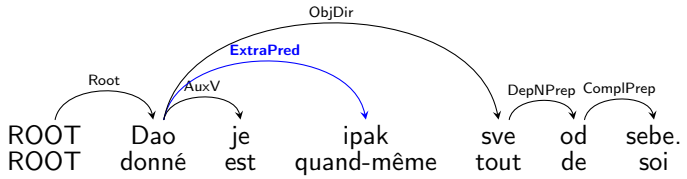
'Il était, en effet, toujours en retard.'

(106)



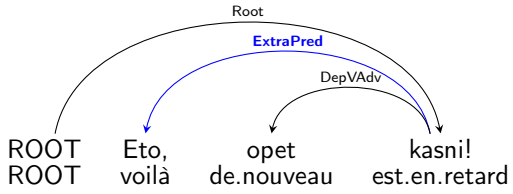
'Peut-être il vient demain.'

(107)



'Il a quand-même fait de son mieux.'

(108)

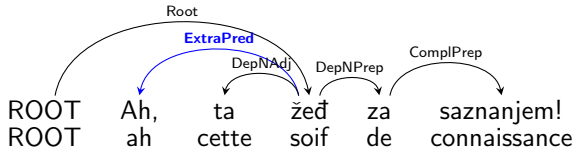


'Voilà, il est encore en retard !'

2.7.7 Emphase

L'étiquette Emph est destinée aux éléments n'ayant pas de rôle syntaxique net dans la phrase et dont la fonction sémantique est de souligner une partie du contenu phrastique. Il s'agit typiquement de particules et d'adverbes. Le gouverneur doit être déterminé au cas par cas. À la différence de la fonction ExtraPred, Emph concerne des éléments infra-phrastiques.

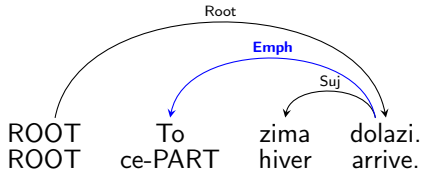
(109)



'Ah, cette soif du savoir !'

Un emploi particulier concerne la particule *to* soulignant la totalité du contenu d'une proposition (exemple 110). À ne pas confondre avec le sujet sous forme du pronom démonstratif *to* : dans le cas où il s'agit d'une particule, la position du sujet est déjà occupée par un autre élément de la phrase, et la forme *to* ne correspond en effet à aucune fonction syntaxique auprès du verbe.

(110)

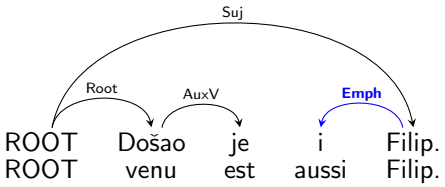


'L'hiver arrive.'

Les formes qui ont typiquement le rôle des conjonctions de coordination, tel que *i* 'et' et *ni* 'ni', peuvent se trouver dans le rôle d'emphase (cf. exemples 111, 112). Le critère pour distinguer ces cas de figure de la coordination se trouve dans le fait qu'avec l'emphase il n'est pas possible d'identifier l'autre élément de la coordination. L'emphase peut également porter sur d'autres types d'éléments, par exemple sur des noms ou des groupes prépositionnels (cf. exemples 115 et 116).

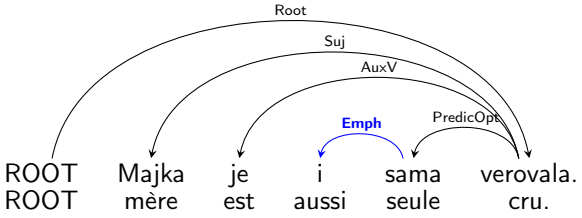
Il est à noter que cette étiquette ne doit être utilisée que si aucune autre ne s'applique. Autrement dit, si une particule dépend d'un verbe, elle sera préférablement traitée comme DepVAdv ou comme ExtraPred, et non pas comme Emph.

(111)



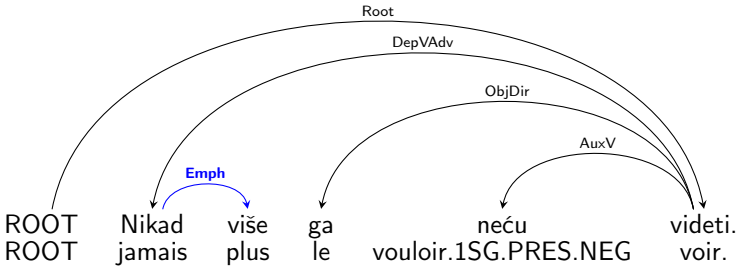
'Filip est venu aussi.'

(112)



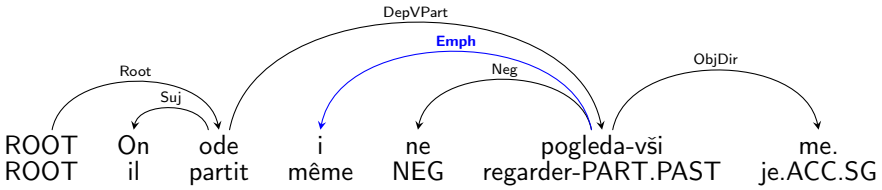
'Ma mère y croyait elle-même.'

(113)



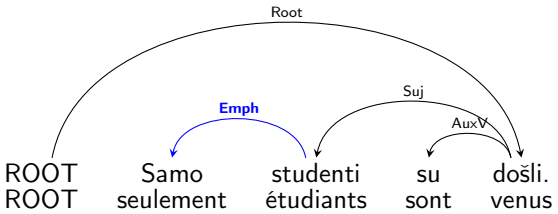
'Je ne le verrai plus jamais.'

(114)



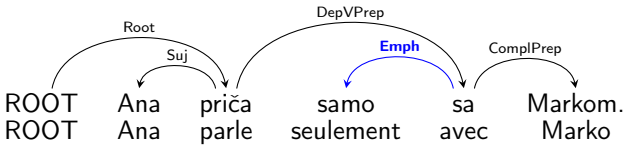
'Il partit sans même me regarder.'

(115)



'Seuls les étudiants sont venus.'

(116)



'Ana parle seulement avec Marko.'

2.7.8 Eléments polylexicaux

L'étiquette Polylex est une abréviation de *polylexical* et on l'utilise pour relier les éléments d'une unité polylexicale entre eux. Elle est utilisée seulement pour les véritables figements grammaticaux : elle concerne en premier lieu les locutions conjonctives et adverbiales, et n'est appliquée ni aux phraséologismes, ni aux collocations. En revanche, elle est utilisée pour l'annotation des numéraux cardinaux et ordinaux complexes, ainsi que pour le traitement des pronoms indéfinis polylexicaux (*ma ko* 'toute personne', *bilo ko* 'n'importe qui'), ainsi que lors du clivange d'un pronom indéfini clivé par une préposition (cf. *ni za šta* 'pour rien'). Elle est également utilisée pour relier les formes des mots étrangers entre elles quand ces formes-là constituent des suites de tokens. Des exemples de ces différents cas de figure sont donnés dans la suite.

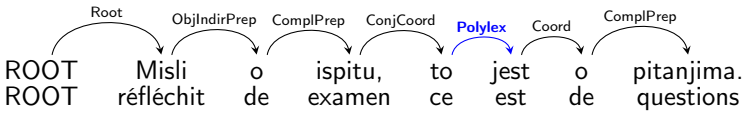
Dans le traitement de ces formes, nous retenons deux principes de base : le premier élément de l'unité polylexicale porte l'étiquette de la fonction exercée par l'unité entière, et les éléments de l'unité polylexicale sont reliés entre eux par la relation Polylex de gauche à droite, en cascade.

NB

Les exemples donnés dans la suite recensent les structures traitées comme polylexicales dans le corpus à présent. Ce recensement n'est sans doute pas exhaustif (de nouvelles structures polylexicales peuvent être rencontrées en corpus), mais les constructions qui y figurent doivent être systématiquement traitées en accord avec le Guide.

Il existe aussi bien des conjonctions de coordination que des conjonctions de subordination polylexicales. [11]

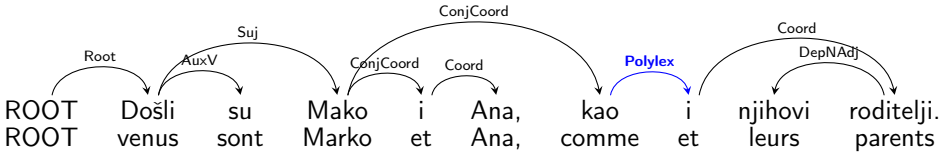
(117)



'Il réfléchit à l'examen, c'est-à-dire aux questions.'

Strictement parlant, il ne s'agit pas ici d'une coordination, mais d'une reprise d'un élément de la phrase avec le but d'apporter une précision. Ce rapport relève cependant des relations de discours, et non pas des relations syntaxiques. Comme les deux éléments sont mis en parallèle, cette construction se rapproche au niveau de surface d'une coordination. C'est pourquoi nous décidons de la traiter comme telle.

(118)



'Marko et Ana sont venus, ainsi que leurs parents.'

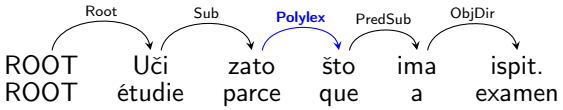
La situation est comparable dans l'exemple 118 : il s'agit, au niveau discursif, d'une précision plutôt que d'une coordination. Or, ce dernier élément est mis en parallèle avec

11. Pour le traitement de la coordination de base, voir section 2.10

la première partie de la coordination, et la forme *kao* semble avoir le rôle de souligner la conjonction de coordination *i*. Par conséquent, nous choisissons de considérer qu'il s'agit ici d'une forme de coordination.

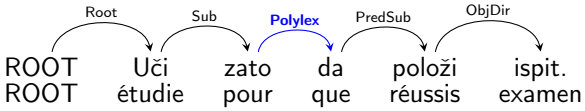
Dans la suite, nous listons d'autres locutions conjonctives polylexicales et indiquons leur traitement.

(119)



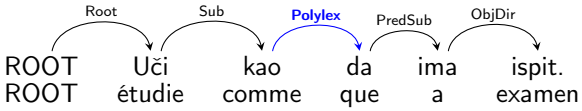
'Il travaille parce qu'il a un examen.'

(120)



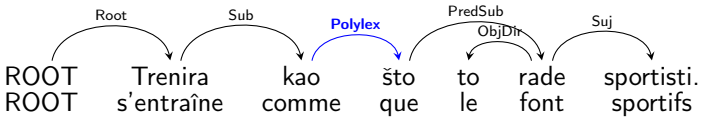
'Il travaille pour réussir son examen.'

(121)



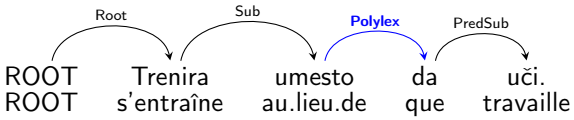
'Il travaille comme s'il avait un examen.'

(122)



'Il s'entraîne comme le font les sportifs.'

(123)



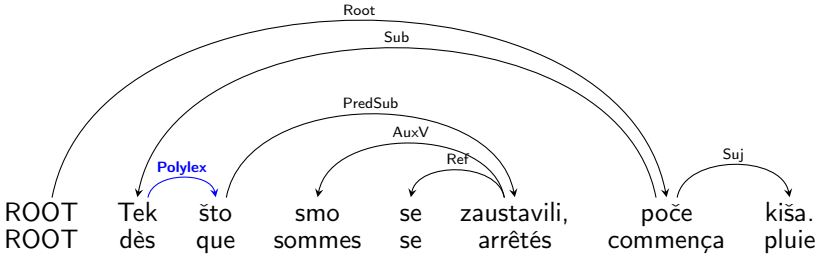
'Il s'entraîne au lieu de travailler.'

Les propositions comparatives exprimant une comparaison d'inégalité introduite par ***nego što*** relèvent des structures corrélatives (cf. exemple 127). Pour plus de détails sur ce sujet, voir la section 2.8.5.

Une attention spéciale doit être accordée à la structure ***tako da*** 'de sorte que', qui peut avoir deux rôles, et par conséquent, deux analyses syntaxiques différentes.

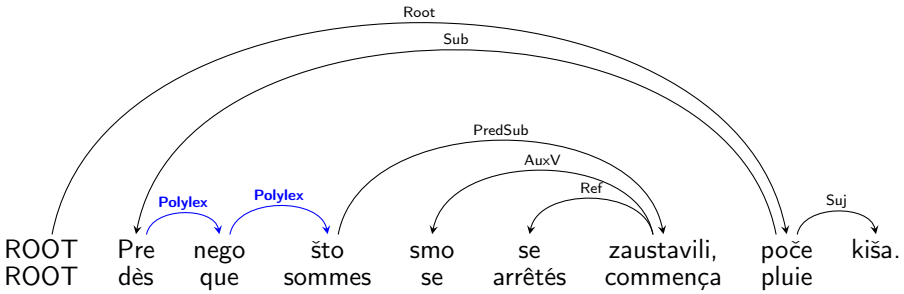
Dans le premier cas possible, *tako* 'ainsi' est un véritable adverbe de manière qui introduit une **structure corrélative** (cf. exemple 129).

(124)



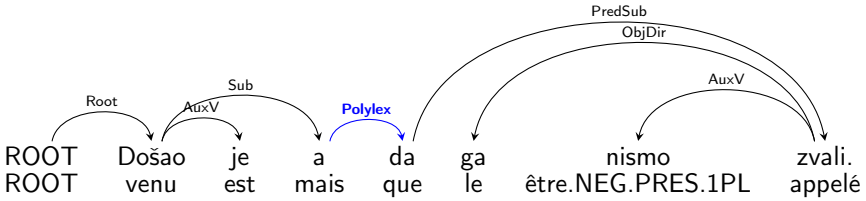
'Dès que nous nous sommes arrêtés, la pluie commença.'

(125)



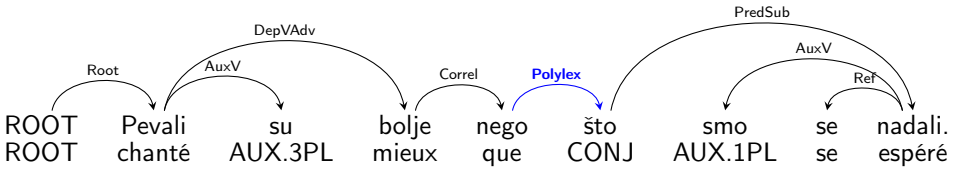
'Avant qu'on ne se soit arrêtés, la pluie commença.'

(126)



'Il est venu sans qu'on l'appelle.'

(127)



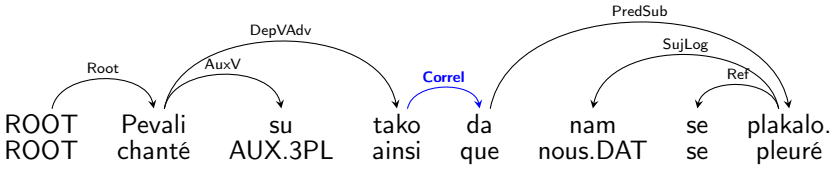
'Ils ont chanté mieux qu'on n'avait espéré.'

Dans le deuxième cas de figure, il s'agit d'une **conjonction polylexicale**, et la forme *tako* n'est plus un dépendant verbal sous forme d'adverbe (cf. exemple 129).

Les **numéraux cardinaux et ordinaux complexes** (*dvadest dva* 'vingt-deux', *dvadeset drugi* 'vingt-deuxième') comme des éléments polylexicaux, qu'ils contiennent la conjonction *i* ou non.

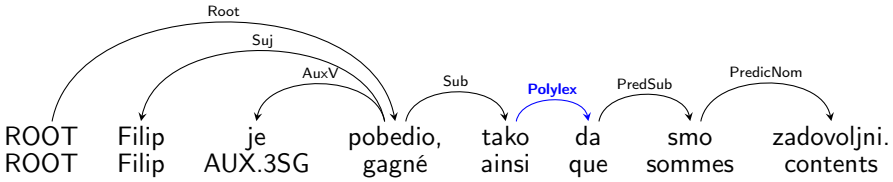
Cet emploi concerne également l'expression de l'heure :

(128)



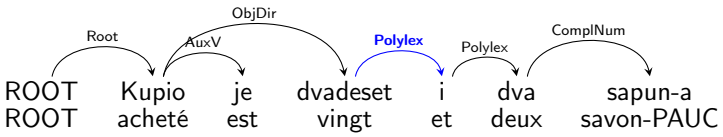
'Ils sont chanté de sorte (de telle manière) qu'on avait envie de pleurer.'

(129)

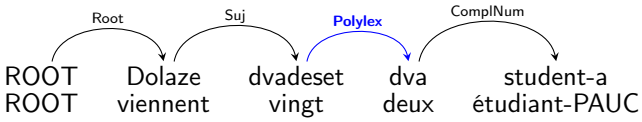


'Filip a gagné si bien qu'on est contents.'

(130)

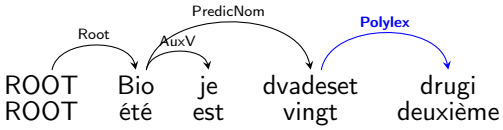


'Il a acheté vingt-deux savons.'



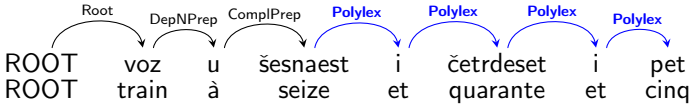
'Vingt-deux étudiants viennent.'

(131)



'Il était le vingt-deuxième.'

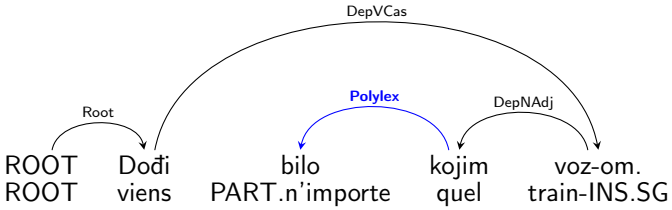
(132)



'le train de seize heures quarante-cinq'

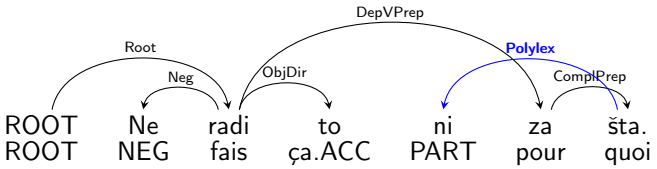
Cette relation est également utilisée pour relier les parties des **pronoms et adjectifs indéfinis polylexicaux**, comme *bilo koji* 'n'importe lequel', *ma ko* 'n'importe qui', mais aussi dans les cas où un **pronom indéfini simple** comme *ništa* 'rien' ou *iko* 'd'aucuns' a le rôle du complément d'une préposition et se voit **clivé** par cette préposition (cf. *ni za šta* 'pour rien', *i za koga* 'pour qui que ce soit').

(133)



'Prends n'importe quel train.'

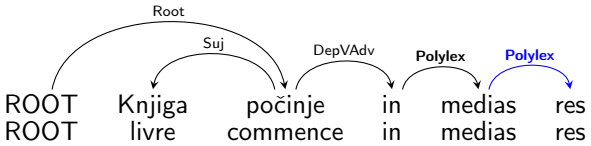
(134)



'Ne le fais pas pour quoi que ce soit.'

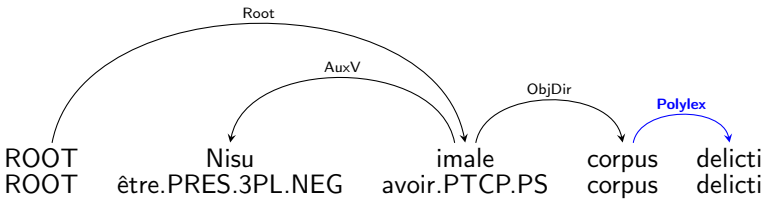
En ce qui concerne les **suites de plusieurs tokens provenants d'une langue étrangère**, le pemier est annoté avec la fonction que le groupe exerce dans la phrase, alors que les autres sont reliés en cascade par l'étiquette Polylex.

(135)



'Le livre commence *in medias res*.'

(136)



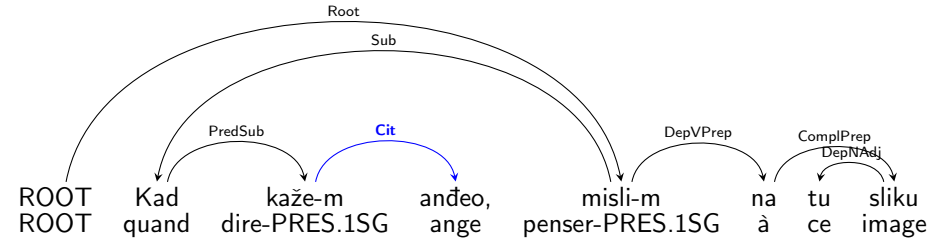
'Elles n'avaient pas de *corpus delicti*.'

2.7.9 Citations et emplois métalinguistiques

L'étiquette Cit est utilisée pour marquer les emplois métaliniguistiques de mots individuels ainsi que pour les segments du discours rapporté averbaux.

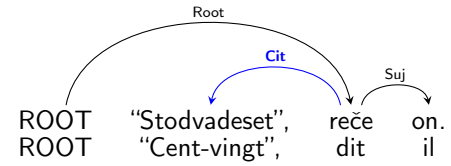
Ces élements pourraient être considérés comme des objets directs des verbes de parole qui les introduisent. En revanche, pour le premier exemple, cette approche signifierait qu'une forme au nominatif peut exercer la fonction de l'objet direct, ce qui semble peu justifié. Pour ce qui est du deuxième cas de figure, le verbe introducteur peut facilement ne pas être un verbe transitif direct : *Sto dvadeset, uvredi se on* 'Cent-vingt, s'offusqua-t-il.' Pour simplifier le traitement, nous choisissons de considérer que ces formes ont un lien moins direct avec le reste de l'arbre syntaxique et les annotons simplement comme morceaux de discours intervenus dans la phrase.

(137)



'Quand je dis ange, je pense à cette image.'

(138)



' "Cent-vingt", dit-il.'

2.7.10 Ponctuation

Toutes les ponctuations sont traitées par l'étiquette Ponct. L'approche adoptée est celle signalée comme la plus performante par A. Urieli : chaque signe de ponctuation est rattaché au premier token précédent qui n'en est pas une. Cette approche permet d'avoir un traitement systématique des ponctuations, facile à effectuer de manière automatique. Par conséquent, il n'est pas nécessaire d'annoter la ponctuation dans le cadre du traitement manuel.

2.8 Subordination

Dans le cadre du projet *ParCoLab*, nous ne reprenons pas la typologie traditionnelle des subordonnées en serbe. Nous introduisons un traitement plus global, qui distingue 5 types de subordonnées basé sur leurs propriétés syntaxiques : celui des subordonnées adverbiales à subordonnant simple (section 2.8.1), celui des complétives (section 2.8.2), celui des interrogatives indirectes (section 2.8.3), celui des relatives (section 2.8.4) et celui des subordonnées corrélatives (les comparatives et les consécutives - section 2.8.5).

Le traitement de base de la subordination est celui des subordonnées adverbiales : ces propositions sont introduites par un subordonnant dont la seule fonction syntaxique est d'assurer l'inclusion de la subordonnée dans la proposition principale. Les quatre traitements restants représentent des variations de ce traitement de base. Nous distinguons les complétives du fait de leur statut spécifique par rapport au verbe qui les introduit (elles font souvent partie de la structure argumentale du verbe). Les relatives et les interrogatives indirectes ont la spécificité d'être introduites par un subordonnant à double fonction syntaxique, qui, outre son rôle de subordination, effectue aussi une fonction à l'intérieur de la subordonnée. Enfin, les propositions corrélatives ont une structure syntaxique spécifique qui mérite un traitement à part. Le traitement de chacun de ces cas de figure est présenté dans la suite.

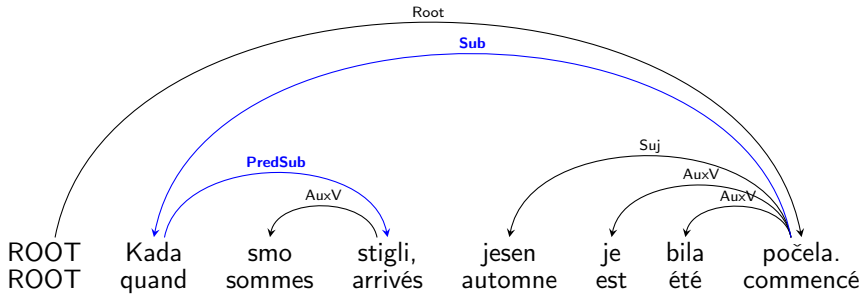
2.8.1 Subordonnées adverbiales à subordonnant mono-fonctionel

Nous ne faisons pas la distinction entre les différents types sémantiques de subordonnées adverbiales. Le seul critère pour appliquer le traitement décrit dans cette section à une proposition subordonnée est qu'elle soit dotée d'un subordonnant à fonction syntaxique unique, servant seulement à établir le lien avec la proposition principale. À titre d'exemple, ceci est

le cas des conjonctions comme *kada* ‘quand’, *pošto* ‘après que, puisque’, *jer* ‘parce que’, *iako* ‘bien que’, *ako* ‘si’, etc.

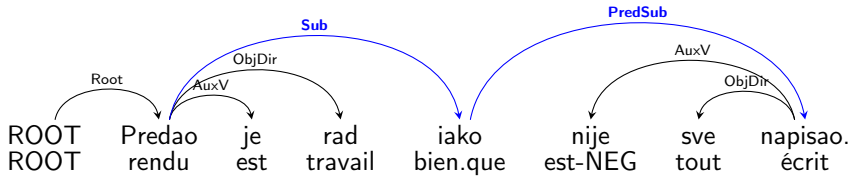
Deux étiquettes sont liées à ce traitement : Sub est l’étiquette qui relie le verbe de la principale et le **subordonnant**, et PredSub est utilisée pour établir le lien entre le subordonnant et le **verbe principal de la subordonnée**.

(139)



‘Quand nous sommes arrivés, l’automne avait commencé.’

(140)



‘Il a rendu son travail bien qu’il n’ait pas tout écrit.’

S’il s’agit d’un subordonnant polylexical, comme *nakon što* ou *zato što*, il faut faire appel à l’étiquette Polylex, dont l’utilisation est présentée dans la section 2.7.8.

NB

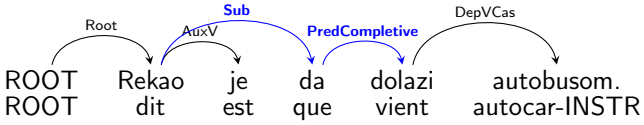
Ce traitement *ne concerne pas* les subordonnées en *da* ‘que’ complétant les verbes aspectuels ou modaux, ni les propositions déclaratives. Voir la section 2.8.2.

2.8.2 Subordonnées complétives

Nous traitons comme complétives les **propositions** introduites par des **verbes de parole** (*govoriti* ‘parler’, *reći* ‘dire’, *pričati* ‘raconter’, etc.) ou **de processus mentaux** (*misliti* ‘penser’, *smatrati* ‘considérer’, *verovati* ‘croire’, etc.), introduites par la **conjonction *da* ‘que’ ou *kako* ‘comment’**, ainsi que les propositions introduites par des **verbes de perception** (cf. *gledati* ‘regarder’, *slušati* ‘écouter’) avec la conjonction ***kako* ‘comment’**. La relation entre le verbe de la principale et le **subordonnant** est la même que pour les propositions adverbiales : Sub. En revanche, pour indiquer la nature spécifique des complétives, la relation allant du subordonnant vers le **verbe principal de la subordonnée** est PredCompletive (prédicat de la complétive).

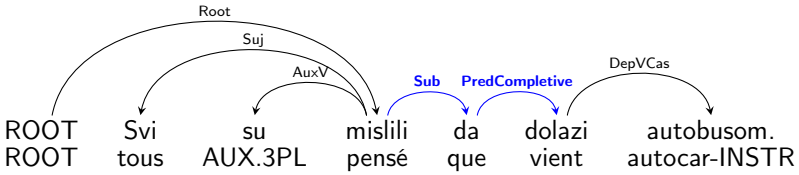
Les complétives dans les exemples 153, 142 et 143 représentent l’objet direct du verbe qui les introduit. Nous avons cependant écarté la possibilité de les traiter comme tel pour le fait que toutes les complétives en serbe ne sont pas des objets directs. Par exemple, les verbes *verovati* ‘croire’, *spremati se* ‘se préparer’ et *uspeti* ‘réussir’ admettent tous des compléments

(141)



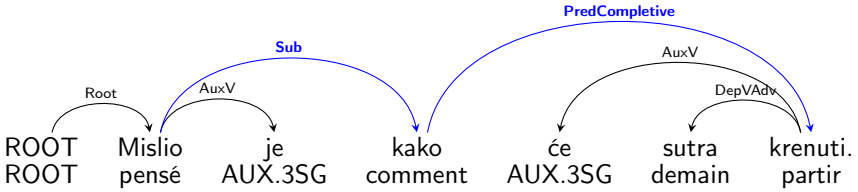
'Il a dit qu'il venait en autocar.'

(142)



'Tous pensaient qu'il venait en autocar.'

(143)

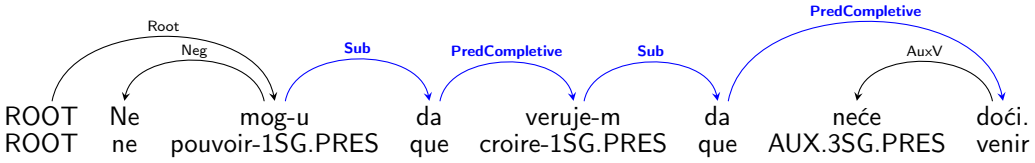


'Il pensait qu'il allait partir le lendemain.'

sous forme d'une proposition en *da*. Cependant, tous les trois ont des constructions de base différentes : *verovati u nešto* 'croire en quelque chose', *spremati se na nešto* lit. 'se préparer à quelque chose', *uspeti u nečemu* lit.'réussir en quelque chose'. On peut donc difficilement affirmer que les complétives en *da* 'que' introduites par ces verbes soient des objets directs, vu que les verbes eux-mêmes n'ouvrent pas cette position dans leur structure argumentale. Ainsi, toutes ces propositions sont simplement traitées comme des complétives, alors que leur statut par rapport au verbe gouverneur reste sous-spécifié.

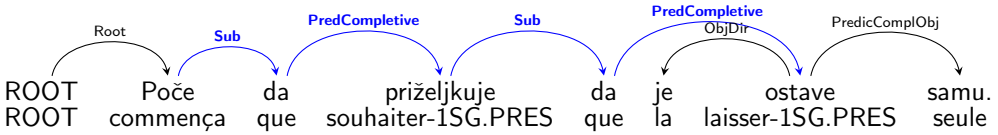
Nous considérons également comme complétives les constructions traitées dans la tradition grammaticale serbe comme **prédicat complexe**. Autrement dit, les compléments des verbes modaux (*morati* 'devoir', *moći* 'pouvoir', etc.) et aspectuels (*početi* 'commencer', *prestati* 'arrêter', etc.) sous forme de la construction *da* 'que' + $V_{present}$ sont annotés comme complétives.

(144)



'Je ne peux pas croire qu'il ne viendra pas.'

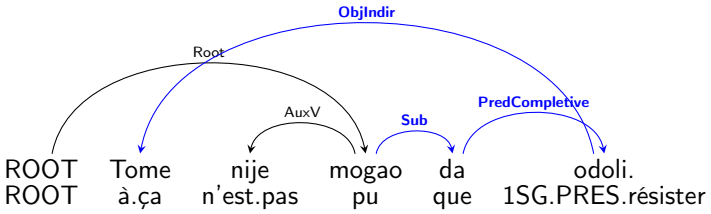
(145)



'Elle commença à souhaiter que l'on la laisse seule.'

Dans ces constructions, il faut veiller à identifier correctement le verbe qui gouverne les dépendants nominaux : un objet direct ou indirect du verbe de la complétive peut se trouver en dehors de sa proposition. Il doit tout de même être rattaché au verbe qui l'introduit, et non pas au verbe introducteur de la relative (cf. exemple 146).

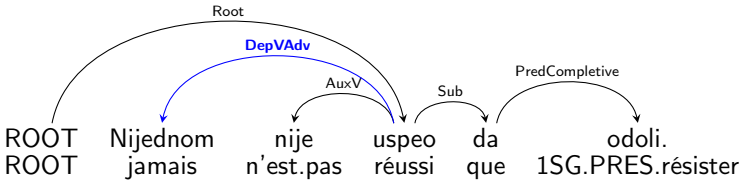
(146)



'A ça, il ne pouvait pas résister.'

Or, dans le cas des dépendants adverbiaux, il n'est pas possible d'identifier lequel des deux verbes est la tête en utilisant des critères de surface. Par conséquent, ce type de dépendant sera rattaché au verbe introducteur de la complétive.

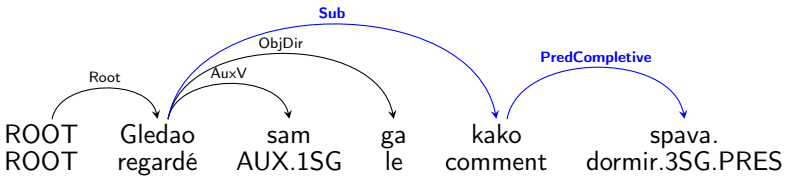
(147)



'Jamais il n'a pu résister.'

Une autre sous-classe des complétives concerne les verbes de perception introduisant une proposition en *kako* 'comment'. Ces propositions sont équivalentes de la constructions infinitivales en français du type *Je le regardais dormir*, d'autant plus que la proposition principale en serbe contient également un objet direct à l'accusatif.

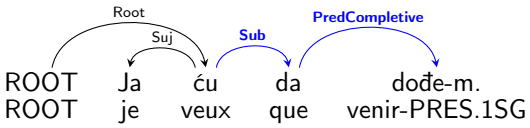
(148)



'Je le regardais dormir.'

Nous annotons de la même manière la **forme irrégulière du futur** comme dans *ja ću da dođem* 'je vais venir', construite de l'auxiliaire *hteti* 'vouloir' qui n'est pas suivi d'un infinitif (comme c'est le cas pour les formes régulières du futur), mais d'une proposition en *da* 'que' contenant le verbe principal au présent. Dans ce cas de figure, étant donné que la forme du verbe *hteti* 'vouloir' n'est pas suivie d'un participe, nous considérons qu'il s'agit d'un verbe principal, et la construction *da* 'que' + $V_{present}$ est traitée comme une complétive.

(149)

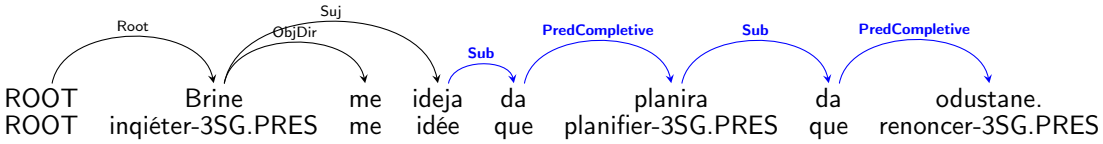


'Je vais venir.'

Complétives introduites par des parties du discours autres qu'un verbe

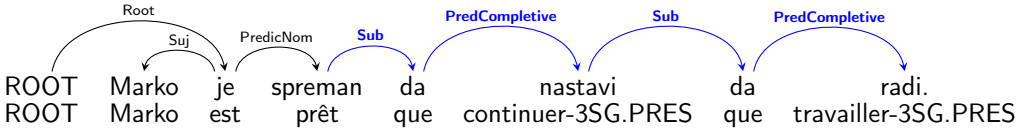
Les complétives peuvent également être introduites par certains noms et adjectifs, voire par des particules. Elles expriment alors le contenu encapsulé par le mot introducteur : *ideja da će Marko doći*, *rešen da ne odustane*, *jedva da diše*.

(150)



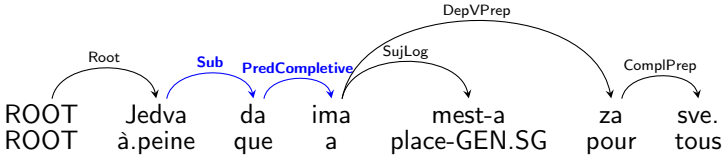
'L'idée qu'il envisage de renoncer m'inquiète.'

(151)



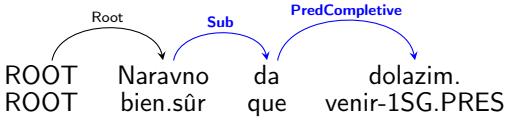
'Marko est prêt à continuer à travailler.'

(152)



'À peine s'il y a de la place pour tous.'

(153)



'Bien sûr que je viens.'

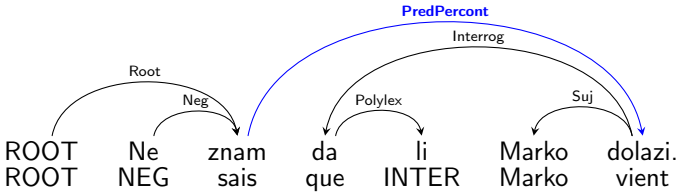
2.8.3 Subordonnées interrogatives indirectes

Les subordonnées interrogatives indirectes diffèrent des subordonnées adverbiales et des complétives par le fait que leur subordonnant a une double fonction syntaxique : il établit le lien entre la principale et la subordonnée, mais il a également une fonction à l'intérieur de la subordonnée. [12] Le comportement syntaxique des formes interrogatives a déjà été expliqué dans la section 2.7.4 et il a été constaté qu'un interrogatif peut avoir la fonction d'un sujet,

12. Ceci est également le cas des relatives : voir la section 2.8.4.

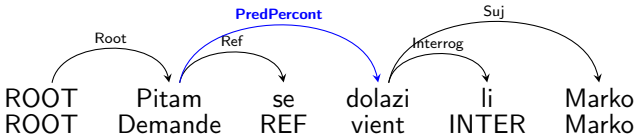
objet direct ou indirect, dépendant verbal ou dépendant nominal au sein de la proposition dans laquelle il figure. Il en est de même des interrogatifs dans les interrogatives indirectes subordonnées. En revanche, comme il a déjà été mentionné, dans ce cas de figure les interrogatifs ont également une deuxième fonction : celle de la subordination. On souhaiterait donc marquer les deux dépendances de ces formes : celle par rapport à la proposition subordonnée elle-même, mais aussi celle par rapport au verbe de la principale. Or, ceci est impossible : comme il a été expliqué dans les règles générales de l'analyse en dépendances, un élément de l'arbre ne peut avoir qu'un seul gouverneur. Il faut donc choisir entre les deux dépendances. Comme l'interrogatif occupe souvent une position syntaxique qui ne peut pas rester vide dans la proposition interrogative, nous préférons encoder celle-ci. Les interrogatives indirectes sont donc traitées comme suit : c'est le verbe de la principale qui est relié directement au verbe de la subordonnée par la relation PredPercont (= prédicat de percontative), et le contenu de la subordonnée est traité en accord avec les principes décrits dans la section 2.7.4.

(154)



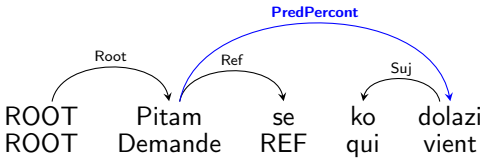
'Je ne sais pas si Marko vient.'

(155)



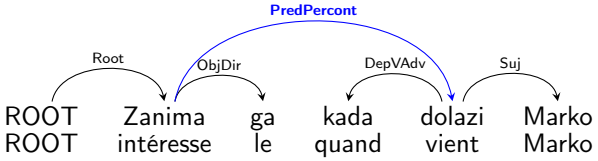
'Je me demande si Marko vient.'

(156)



'Je me demande qui vient.'

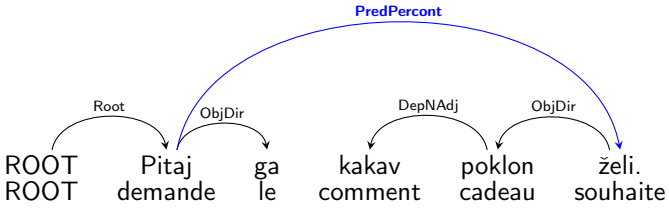
(157)



'Il veut savoir quand vient Marko.'

Pour une méthode qui facilite l'identification de la fonction syntaxique de l'interrogatif au sein de la subordonnée, voir la section 2.7.4.

(158)



'Demande-lui comment il veut son cadeau.'

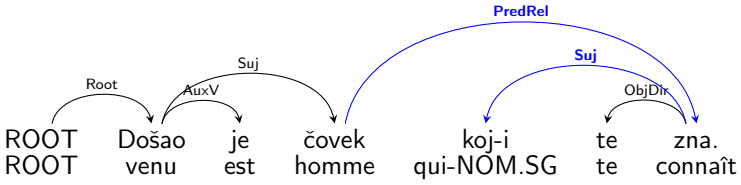
Rappel

Dans le cas des interrogatives exprimant une interrogation totale, le marqueur d'interrogation n'a pas de rôle spécifique par rapport au prédicat de la subordonnée. Par conséquent, il est relié au verbe de la subordonnée interrogative par la relation Interrog.

2.8.4 Subordonnées relatives

Les relatifs exhibent un comportement syntaxique proche de celui des interrogatifs : au-delà de leur fonction de subordination, ils exercent également une fonction syntaxique à l'intérieur de la proposition subordonnée qu'ils introduisent. La même question se pose donc ici : il faut décider si le relatif doit être gouverné par son antécédent en tant que subordonnant ou bien par l'élément dont il dépend à l'intérieur de la relative. Ici aussi, nous optons pour la fonction du relatif exercée à l'intérieure de la subordonnée. Ceci est notamment fait pour maintenir la représentation de la structure argumentale des verbes dans les relatives. Par conséquent, le lien entre la proposition principale et la relative s'établit sans passer par la forme relative. À la différence du traitement des interrogatives indirectes, le lien ne s'établit pas entre les prédicats de la principale et de la subordonnée : la relative est plutôt introduite par l'antécédent du relatif, ce qui permet d'indiquer le rôle de cet élément dans la structure relative. L'étiquette utilisée pour relier l'**antécédent au prédicat de la relative** est PredRel (= prédicat de relative). Le relatif est annoté avec l'étiquette qui exprime le mieux sa fonction à l'intérieur de la relative (tout comme les interrogatifs) (cf. exemples ci-dessous).

(159)

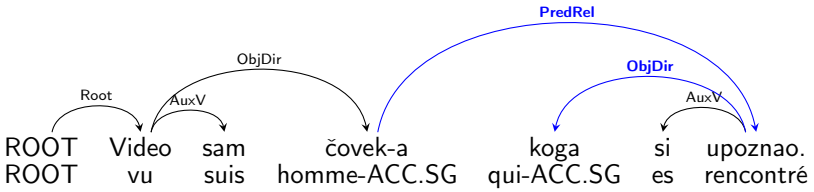


'L'homme qui te connaît est venu.'

Il faut souligner que l'antécédent peut avoir des focntions différentes dans la proposition principale, et il en est de même pour le relatif au sein de la relative.

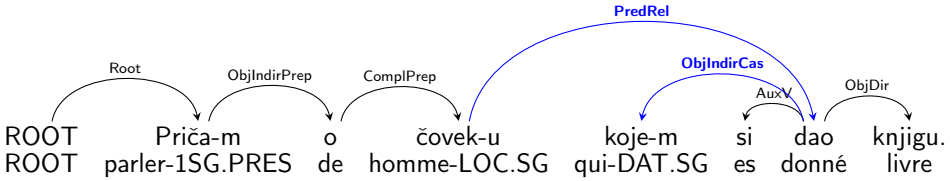
Le relatif peut également ne pas dépendre directement du verbe de la relative, mais faire partie d'une construction prépositionnelle introduite par le verbe (cf. exemple 162). Dans l'identification de la fonction du relatif, il peut être utile de remplacer le pronom relatif par un pronom personnel ou démonstratif et de rétablir l'ordre des mots canonique dans la relative : *o kojem sam ti pričao* peut se transformer ainsi en *o njemu sam ti pričao*, et ensuite

(160)



'J'ai vu l'homme que tu as rencontré.'

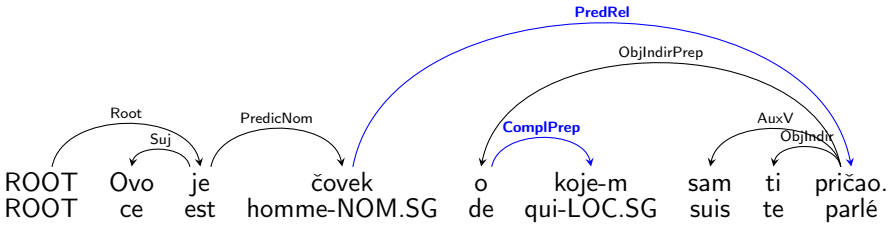
(161)



'Je parle de l'homme à qui tu as donné le livre.'

en *pričao sam ti o njemu*. Cette manipulation peut aider à voir plus nettement que le verbe *pričao* introduit la préposition *o*, qui, à son tour, introduit une forme au locatif (*njemu* dans la proposition transformée, ou bien *kojem* dans la proposition de départ). La proposition correspond donc à la relation ObjIndirPrep, et le pronom (dans les deux cas) a le rôle du complément de la préposition.

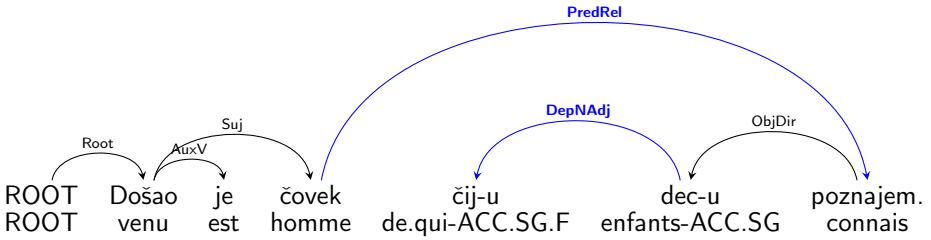
(162)



'C'est l'homme duquel je t'ai parlé.'

L'adjectif relatif ***čiji*** 'de qui' fonctionne comme tout autre adjectif : il s'accorde pleinement avec le nom dont il dépend, et il lui est antéposé. Par conséquent, il est traité en utilisant la relation DepNAdj.

(163)

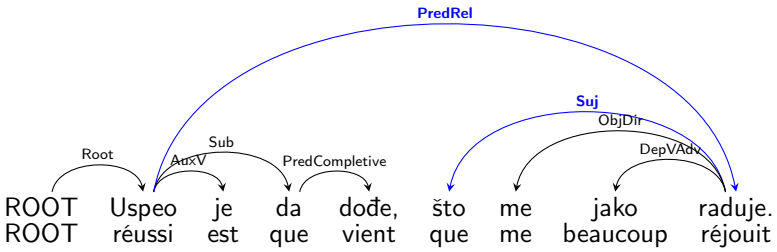


'L'homme dont je connais les enfants est venu.'

Il existe également des **relatives qui ont pour antécédent toute la proposition principale**, comme dans *Uspeo je da dođe, što me jako raduje* 'Il a pu venir, ce qui me

réjouit beaucoup'. Comme le prédicat de la relative est introduit toujours par la tête de l'antécédent, dans ce cas de figure, c'est le prédicat de la principale qui introduit le prédicat de la relative (cf. exemple 164).

(164)

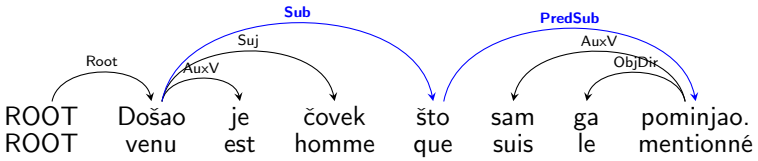


'Il a pu venir, ce qui me réjouit beaucoup.'

NB

Les relatives introduites par *što* 'que' dans lesquelles la fonction censée être exercée par le relatif est reprise par le pronom personnel ne sont pas traitées de la même manière. En effet, leur traitement est celui des subordonnées à subordonnant simple (cf. exemple 165). Ceci se justifie par le fait que le relatif n'a plus de rôle syntaxique par rapport au verbe de la relative, celui-ci étant déjà accompagné d'un objet direct (le pronom personnel *ga* 'le').

(165)



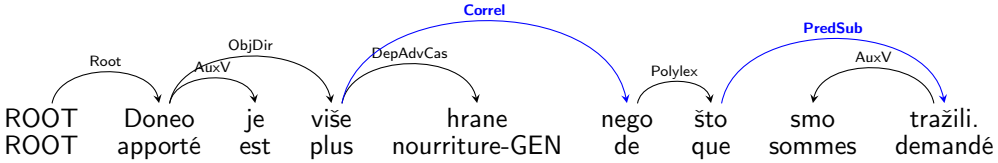
'L'homme que j'ai mentionné est venu.'

2.8.5 Subordonnées corrélatives

Cette section est consacrée aux subordonnées introduites par un corrélatif : il s'agit d'un élément qui a une fonction syntaxique dans la principale, et qui appelle un autre élément dans la subordonnée : la présence du premier terme étant subordonnée à la présence du second et réciproquement, on parle de corrélation. Il s'agit notamment de propositions consécutives et comparatives.

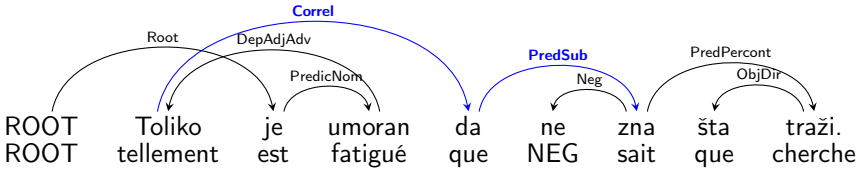
La relation de subordination s'établit donc entre le corrélatif et le subordonnant via la relation Correl. Le subordonnant gouverne à son tour le verbe de la subordonnée à travers la relation PredSub.

(166)



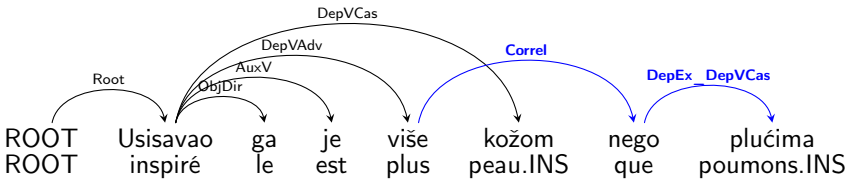
'Il a apporté plus de nourriture que nous avions demandé.'

(167)



'Il est tellement fatigué qu'il ne sait pas ce qu'il cherche.'

(168)



'Il l'inspirait plus avec sa peau qu'avec ses poumons.'

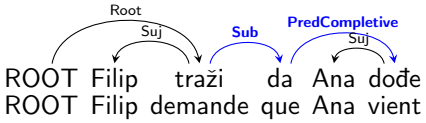
Dans l'exemple 168, nous trouvons également une proposition averbale, dont le verbe a été élidé. Le traitement de l'ellipse est présenté en détail dans la section 2.12.

2.8.6 Ambiguïté des subordonnées en *da*

Une attention particulière doit être accordée au traitement des subordonnées introduites par la conjonction *da*. En effet, cette conjonction est polysémique et peut introduire différents types de subordonnées.

Premièrement, la conjonction *da* est la conjonction prototypique pour les **complétives** (cf. section 2.8.2). Notons que dans ce cas la subordonnée est introduite par un verbe de parole ou de processus mental et correspond souvent à un argument du verbe. Par conséquent, elle est difficilement omissible (cf. exemple 169).

(169)

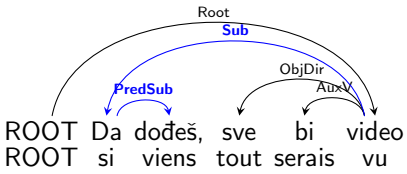


'Filip demande qu'Ana vienne.'

Deuxièmement, la conjonction *da* peut introduire une proposition hypothétique irréelle (cf. exemples 170 et 171). Notons que dans ce cas la conjonction introduit le présent d'un verbe imperfectif (cf. exemple 170) ou le parfait (cf. exemple 171).

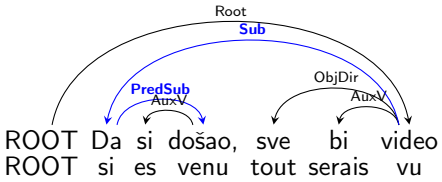
Troisièmement, la conjonction *da* peut également introduire une proposition finale (cf. exemple 172). Dans ce cas, le verbe dans la subordonnée est au présent ou au potentiel, et ces

(170)



'Si tu venais, tu verrais tout.'

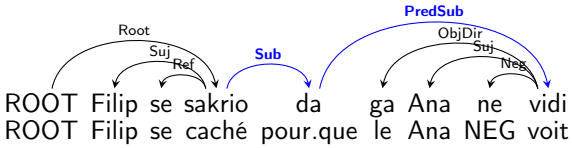
(171)



'Si tu étais venu, tu aurais tout vu.'

deux formes verbales sont intérchangeables (cf. *Filip se sakrio da ga Ana ne vidi => Filip se sakrio da ga Ana ne bi videla*).

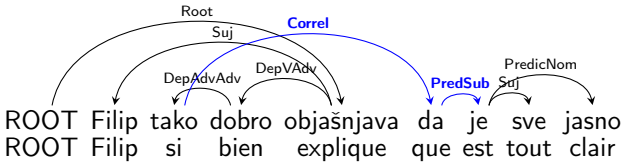
(172)



'Filip s'est caché pour qu'Ana ne le voie pas.'

Enfin, elle peut également introduire une consécutive, mais dans ce cas, elle se combine avec un corrélatif dans la principale (cf. exemple 173). Le traitement des structures corrélative est décrit en détail dans la section 2.8.5.

(173)

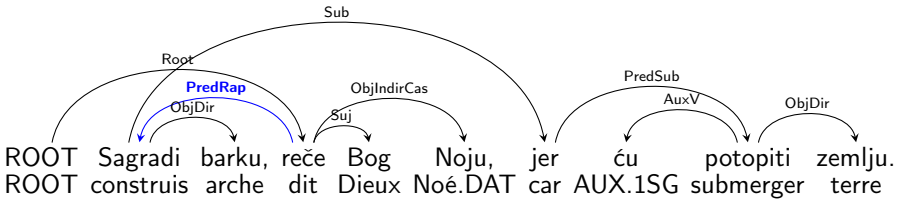


'Filip explique tellement bien que tout est clair.'

2.9 Discours indirect

Si un élément de **discours rapporté** contient un verbe principal, autrement dit, s'il a la structure d'une **proposition indépendante**, il est traité par la étiquette PredRap. Cette relation est gouvernée par le verbe introducteur du discours rapporté. Si le verbe introducteur se trouve au milieu de la phrase rapportée, les éléments séparés sont reliés par les mêmes relations qui s'appliqueraient dans une phrase typique (cf. exemple 174). En revanche, si le discours rapporté ne contient pas de verbe, il relève plutôt de la relation Cit (cf. section 2.7.9).

(174)

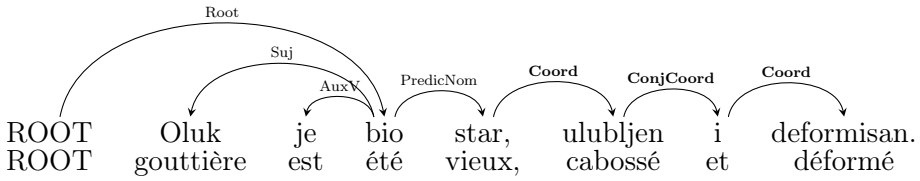


'Construis une arche, dit Dieu à Noé, car je submergerai la terre.'

2.10 Coordination

Pour les structures coordonnées, nous adoptons un traitement en cascade. Le premier conjoint porte l'étiquette qui exprime la fonction syntaxique de la structure coordonnée dans la phrase. Tous les coordonées entre le premier et la conjonction de coordination sont annotés en cascade comme Coord. La conjonction est gouvernée par le conjoint immédiatement précédent à l'aide de l'étiquette ConjCoord. Le dernier conjoint dépend de la conjonction à l'aide de l'étiquette Coord (cf. exemple 175).

(175)

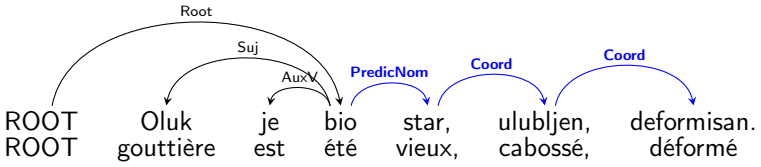


'La gouttière était vieille, cabossée et déformée.'

NB

Nous considérons également comme coordination les enchaînements de plusieurs éléments séparés par des virgules, sans conjonction de coordination entre les deux derniers éléments (cf. exemple 176). Ce phénomène est considéré comme juxtaposition dans la littérature linguistique, mais il est traité comme coordination dans plusieurs autres corpus (FTBDep, PDT). Comme ce traitement permet de simplifier le traitement de ce phénomène, nous l'adoptons ici.

(176)

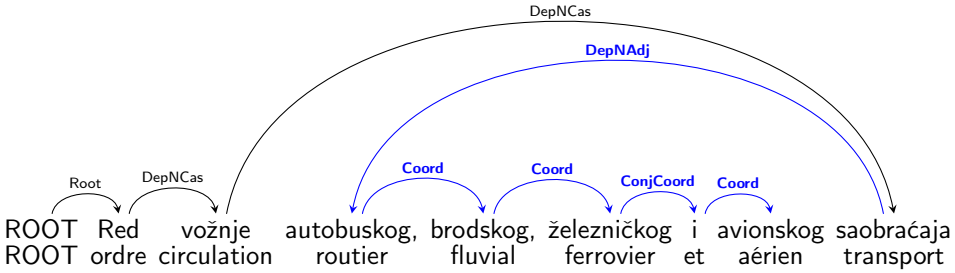


'La gouttière était vieille, cabossée, déformée.'

Les adjectifs antéposés au nom peuvent également être coordonnés (cf. exemple 177).

La relation de coordination s'établit entre les **têtes** des éléments coordonnés. Par conséquent, il faut veiller à bien identifier les structures mises en parallèle par la coordination et à sélectionner les bons gouverneurs pour la relation Coord. Comparez l'exemple 178, où ce sont

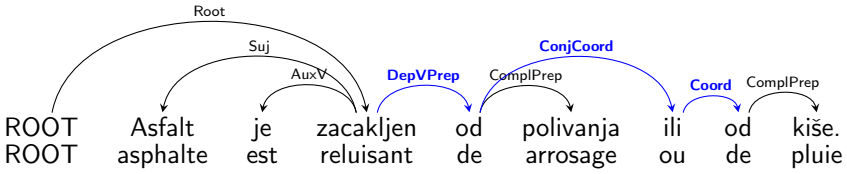
(177)



'Les horaires des transports en bus, bateau, train et avion'

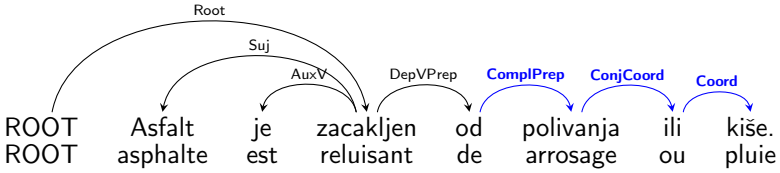
des groupes prépositionnels qui sont coordonnés, avec l'exemple 179, où ce sont les compléments de la préposition. Dans l'exemple 180, la coordination porte sur les deux complément du nom *tama* 'obscurité', alors que dans l'exmple 181, elle met en parallèle les deux objets directs *tamu* 'obscurité' et *daljinu* 'distance'.

(178)



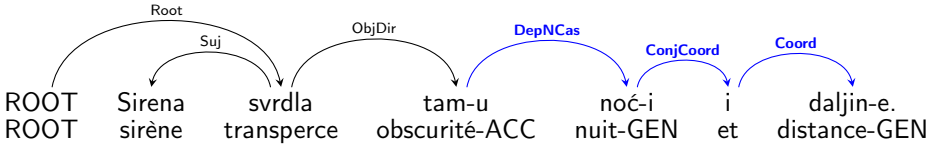
'L'asphalte reluit de l'arrosage ou de la pluie.'

(179)



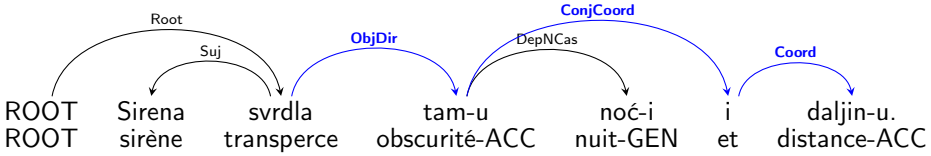
'L'asphalte reluit de l'arrosage ou de la pluie.'

(180)



'La sirène transperce l'obscurité de la nuit et de la distance.'

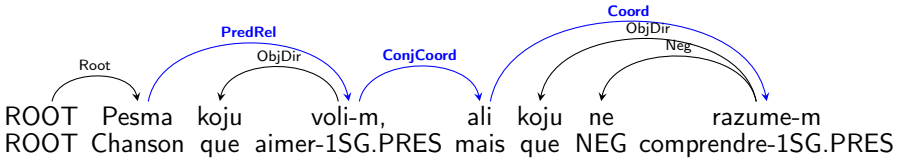
(181)



'La sirène transperce l'obscurité de la nuit et la distance.'

Dans le cas de la **coordination des relatives**, c'est entre les **verbes principaux des relatives** que la coordination s'établit. En cas d'absence de conjonction de coordination, la relation Coord relie directement les deux verbes, et sinon, la relation ConjCoord part du verbe de la première relative vers la conjonction de coordination, et la relation Coord s'établit entre la conjonction et le verbe de la deuxième relative (cf. exemple 182).

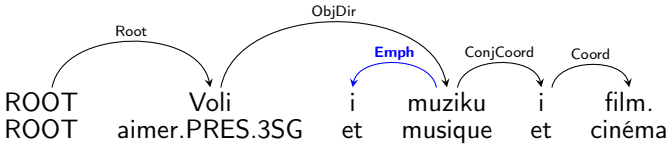
(182)



'la chanson que j'aime, mais que je ne comprends pas'

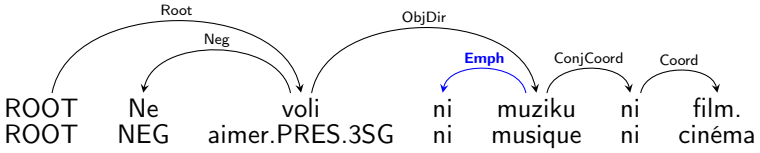
Les structures en ***i - i*** comme *Voli i muziku i film* 'Il aime et la musique et le cinéma' et en ***ni - ni*** comme *Ne voli ni muziku, ni film* 'Il n'aime ni la musique, ni le film' bénéficient d'un traitement spécifique. En effet, étant donné que la première des deux formes répétées est optionnelle, (cf. *Voli muziku i film*, *Ne voli muziku ni film*), on considère que ce n'est que la deuxième qui a le rôle d'un coordonnant. La première est interprétée comme un élément d'emphase. Voir les exemples ci-dessous.

(183)



'Il aime aussi bien la musique que le cinéma.'

(184)

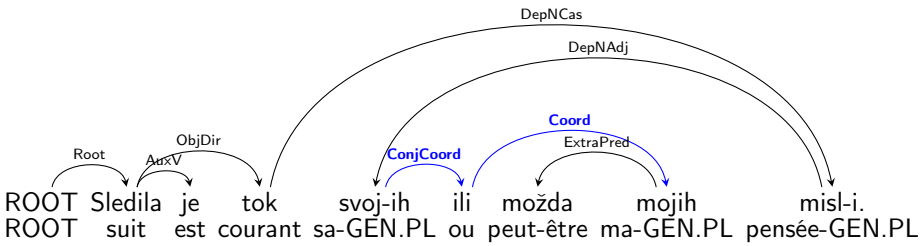


'Il n'aime ni la musique ni le cinéma.'

Constituants incomplets

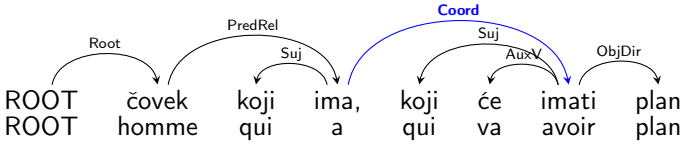
Dans le cas des constituants incomplets, si pertinent, la coordination peut être appliquée.

(185)



'Elle suivait le trajet de ses pensées, ou peut-être des miennes.'

(186)

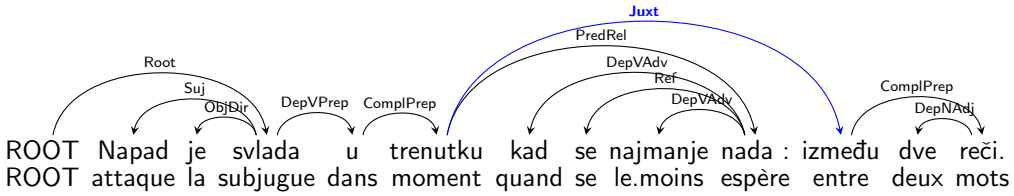


'homme qui a, qui va avoir un plan'

2.11 Juxtaposition

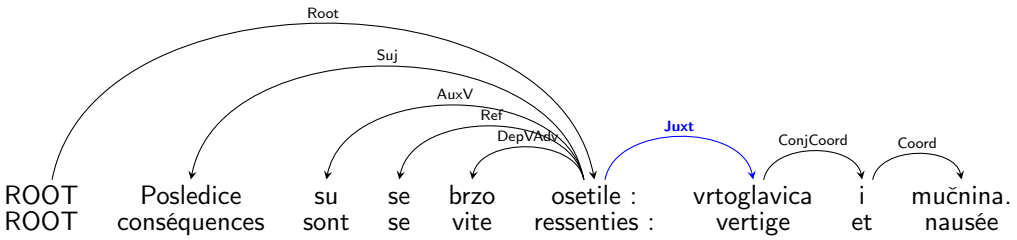
Cette relation est réservée aux éléments de haut niveau (deux propositions, une proposition et un syntagme) qui sont liés au niveau discursif, mais qui paraissent indépendants l'un de l'autre au niveau syntaxique et entre lesquels aucune autre étiquette syntaxique ne s'applique. Ces éléments sont typiquement séparés par deux points ou par des points-virgules. Le dépendant, qui est l'élément qui se trouve à la périphérie de la structure de la phrase, est gouverné par l'élément avec lequel il est mis en parallèle (cf. exemple 187). S'il n'est pas possible d'identifier cet élément, on utilise comme gouverneur la tête du segment à gauche (cf. exemple 188).

(187)



'L'attaque la subjuge dans le moment où elle s'y attend le moins : entre deux mots.'

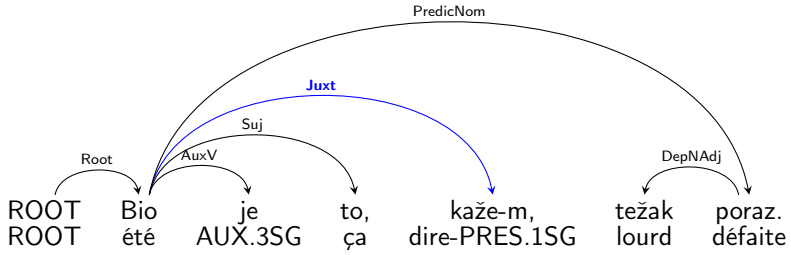
(188)



'Les conséquences se sont vite fait sentir : le vertige et la nausée.'

Cette relation est également utilisée pour annoter les verbes de la parole (typiquement des verbes pouvant servir d'introducteur du discours indirect) insérés dans la phrase sans qu'il s'agisse du discours rapporté. Ils sont gouvernés par le verbe principal de la proposition dans laquelle ils se trouvent (cf. exemple 189).

(189)

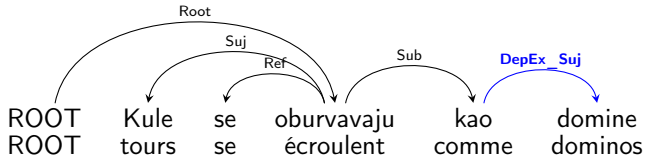


'Ce fut, je dis, une lourde défaite.'

2.12 Ellipse

Le traitement de l'ellipse reste une question ouverte aussi bien en linguistique théorique qu'en TAL. Pour cette première annotation de notre corpus, nous retenons l'approche mise en place dans le treebank tchèque PDT : une forme dépendante d'un élément phrastique élidé est gouvernée par le gouverneur de l'élément élidé, en utilisant l'étiquette de la relation que la forme concernée a par rapport à son gouverneur élidé préfixé de DepEx_ (pour 'dépendance externe'). À priori, n'importe quelle étiquette du jeu peut être modifiée de cette manière pour permettre le traitement de l'ellipse. De nombreux cas de figure sont possibles ; quelques-uns des cas les plus fréquents sont illustrés dans la suite.

(190)

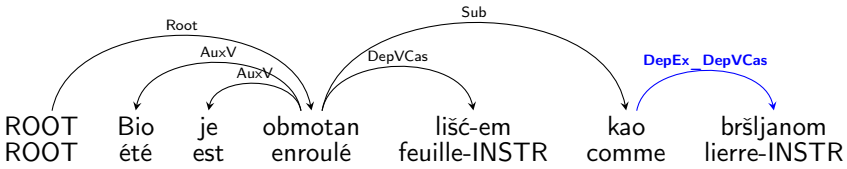


'Les tours s'écroulent comme des dominos'

Pour identifier la fonction syntaxique que la forme porterait par rapport à son gouverneur élidé, il convient de reconstituer la proposition élidée. Dans l'exemple 190, la proposition reconstituée correspond à *Kule se oburvavaju kao što se oburvavaju domine* 'Les tours s'écroulents comme s'écroulent des dominos'. À l'intérieur de cette proposition, la forme *domine* a la fonction du sujet. Par conséquent, on lui accorde l'étiquette DepEx_Suj dans la phrase de départ. Quant à son gouverneur, c'est le subordonnant : c'est lui qui serait le gouverneur du verbe de la subordonnée si celui-ci était présent dans la phrase.

Une approche comparable peut être transposée à l'exemple 191. Ici, la proposition complète serait *kao da je obmotan bršljanom* 'comme s'il était enroulé du lierre', mais le verbe, qui gouvernerait normalement la forme à l'instrumental, est ellidé. Par conséquent, c'est la conjonction qui est le gouverneur de la forme à l'instrumental, et comme cette forme aurait la fonction DepVCas par rapport au verbe ellidé, c'est l'étiquette DepEx_DepVCas qui unit la conjonction à la forme nominale.

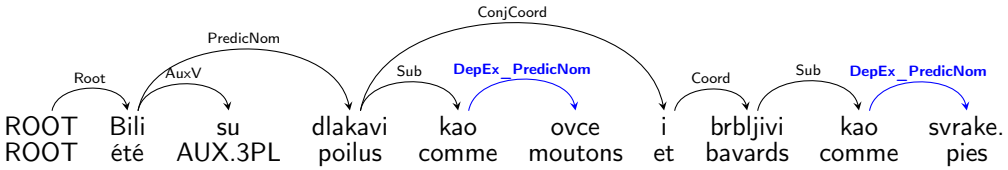
(191)



'Enroulé avec des feuilles comme avec du lierre'

Dans l'exemple 192, les propositions rétablies sont les suivantes : *Bili su dlakavi kao što su ovce dlakave i brbljivi kao što su svrake brbljive* 'Ils étaient poilus comme sont poilus les moutons et bavards comme sont bavardes les pies'. Par conséquent, les adjectifs introduits par *kao* 'comme' ont le rôle du prédicatif nominal par rapport à un verbe *être* élidé.

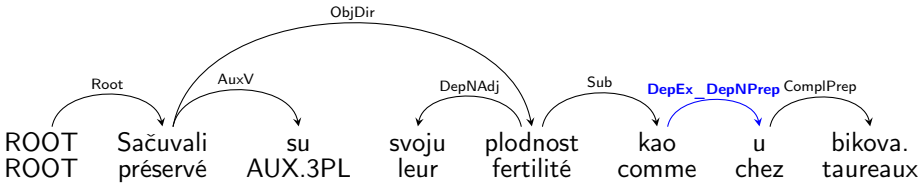
(192)



'Ils étaient poilus comme des moutons et bavards comme des pies.'

Ce type de construction peut également être gouverné par d'autres parties du discours, si la comparaison porte sur un dépendant nominal (cf. exemple 193).

(193)



'Ils avaient préservé leur fertilité comme celle des taureaux.'

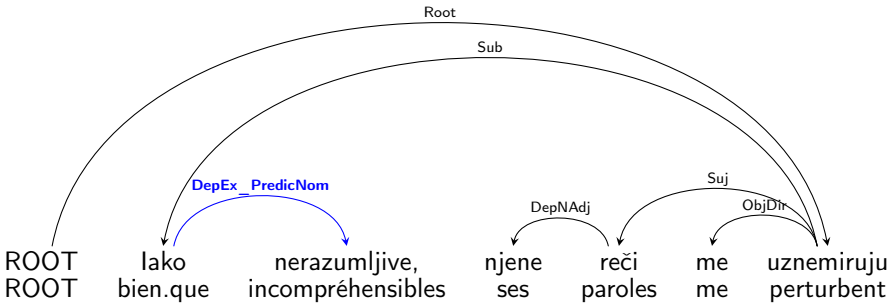
Ici, il s'agit en effet d'une ellipse à plusieurs niveaux : le groupe prépositionnel dépend d'un nom ellidé, qui dépendrait, lui, du verbe, qui est ellidé aussi(cf. *sačuvali su svoju plodnost koja je bila kao plodnost u bikova* 'ils avaient préservé leur fertilité qui était comme la fertilité des tauraux'). Nous encodons ici un seul niveau d'ellipse et marquons la dépendance de ce groupe prépositionnel par rapport à un nom absent de la phrase.

La même approche est appliqué aux **conjonctions de subordination introduisant des syntagmes** : dans ce cas, nous considérons qu'il s'agit d'une proposition subordonnée avec le verbe élidé et nous appliquons le traitement illustré dans l'exemple 194, qui correspond à un prédicatif nominal.

Un cas spécifique de l'ellipse concerne la construction en ***sve* 'tout'** comme *sve mangup do mangupa* lit. 'tous coquin à côté de coquin', 'que des coquins'. Elle est traitée comme dans l'exemple 195.

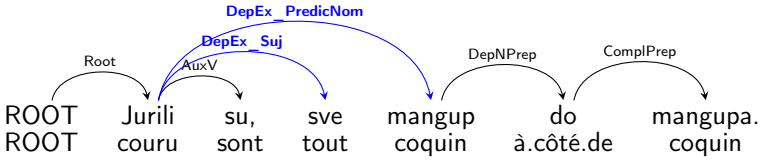
Il faut préciser aussi qu'il peut y avoir des situations dans lesquelles le **gouverneur de l'ellipse** est **ambigu**. Par exemple, dans la phrase *Kroz kosu joj proviruje pupak kao kiklopsko oko* 'Son nombril transperce ses cheveux tel un œil de Cyclope', la conjonction *kao* 'comme' peut être introduite par le verbe (et dans ce cas, la phrase s'interpréterait comme *Pupak*

(194)



'Bien qu'incompréhensibles, ses paroles me perturbent'

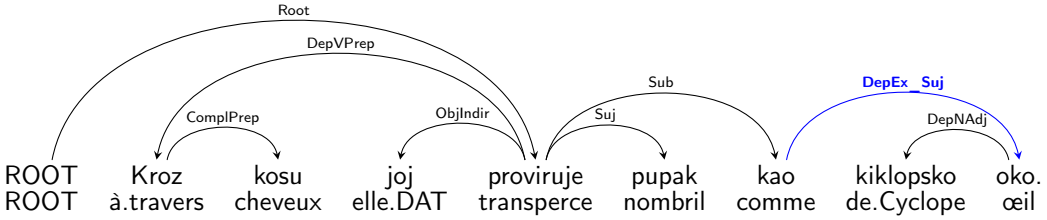
(195)



'Ils couraient, tous des coquins.'

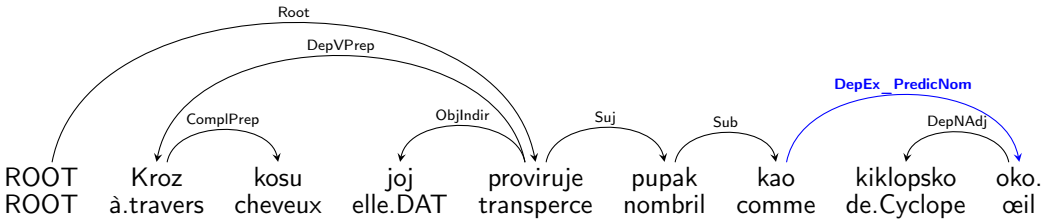
joj proviruje kroz kosu kao što proviruje kiklopsko oko 'Son nombril transperce ses cheveux comme le ferait un œil de Cyclope'), ou bien par le nom *pupak* 'nombril' (et alors le phrase s'interprète comme *Kroz kosu joj proviruje pupak, koji je kako kiklopsko oko* 'Ses cheveux sont trasnpercés par son nombril, qui est comme un œil de Cyclope'). Le premier cas correspond à l'analyse montrée dans l'exemple 196, alors que le deuxième est illustré dans l'exemple 197.

(196)



'Son nombril transperce ses cheveux tel un œil de Cyclope.'

(197)



'Son nombril transperce ses cheveux tel un œil de Cyclope.'

Dans ce type d'exemples, la décision du traitement à adopter revient à la discrétion de l'annotateur. Il est cependant invité à signaler tout exemple ambigu à l'annotateur expérimenté pour que celui-ci puisse l'analyser et l'inventorier si nécessaire.

Bilan

Le guide recense les 48 relations de dépendance utilisées dans l'annotation syntaxique du corpus ParCoTrain-Synt, ainsi que le traitement de l'ellipse. Les étiquettes se répartissent comme suit :

— 17 dépendants du verbe,
— 4 dépendants du nom,
— 4 dépendants de l'adjectif,
— 3 dépendants de l'adverbe,
— 7 étiquettes pour le traitement de la subordination,
— 2 étiquettes pour le traitement de la coordination,
— 10 étiquettes pour d'autres types de dépendants.

Dans la suite, nous proposons un index des exemples par étiquette syntaxique.

Index des exemples par étiquette

Ap : 67, 68, 69, 70, 71, 72, 73, 74.

AuxV : 2, 7, 8, 12, 15, 16, 24, 26, 28, 34, 41, 42, 43, 45, 46, 51, 52, 53, 56, 57, 58, 59, 61, 64, 65, 69, 70, 72, 78, 83, 85, 91, 92, 94, 97, 105, 107, 111, 112, 113, 115, 118, 124, 125, 126, 129, 131, 136, 140, 153, 142, 143, 146, 147, 149, 159, 160, 161, 163, 164, 165, 167, 168, 170, 171, 174, 175, 176, 178, 179, 185, 186, 188, 189, 191, 192, 193, 195.

Cit : 138.

ComplNum : 92.

ComplPrep : 28, 30, 32, 37, 40, 43, 46, 48, 53, 54, 55, 60, 62, 70, 71, 74, 80, 81, 82, 86, 87, 88, ??, 90, 94, 101, 107, 109, 116, 117, 132, 134, 138, 153, 161, 178, 179, 187, 193, 195, 196, 197.

ConjCoord : 12, 18, 19, 24, 25, 31, 59, 73, 74, 75, 117, 118, 175, 177, 178, 179, 180, 181, 182, 183, 184, 185, 188, 192.

Coord : 4, 12, 18, 19, 24, 25, 30, 31, 59, 62, 66, 72, 73, 74, 75, 117, 118, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 188, 192.

Correl : 127, 129, 167, 168, 173.

DepAdjAdj : 83.

DepAdjAdv : 75, 167.

DepAdjCas : 76, 77, 78, 79.

DepAdjPrep : 48, 80, 81, 82.

DepAdvAdv : 84, 173.

DepAdvCas : 13, 26, 85, 167.

DepAdvPrep : 86, 87, 88, ??.

DepEx_DepNPrep : 193.

DepEx_DepVCas : 168, 191.

DepEx_PredicNom : 192, 194, 195, 197.

DepEx_Suj : 190, 195, 196.

DepNAdj : 3, 34, 54, 55, 61, 63, 64, 65, 66, 67, 68, 71, 72, 73, 74, 76, 77, 78, 79, 82, 91, 109, 118, 133, 138, 158, 163, 177, 185, 187, 189, 193, 194, 196, 197.

DepNCas : 61, 68, 72, 86, 177, 180, 181, 185.

DepNPrep : 62, 107, 109, 132, 195.

DepVAdv : 7, 8, 9, 10, 15, 16, 49, 50, 68, 93, 95, 96, 100, 105, 106, 108, 113, 127, 129, 135, 143, 147, 157, 164, 168, 173, 187, 188.

DepVCas : 25, 47, 51, 52, 78, 133, 153, 142, 168, 191.

DepVInf : 57, 58.

DepVPart : 52, 56, 114.

DepVPrep : 28, 53, 54, 55, 60, 70, 71, 74, 94, 101, 116, 134, 138, 153, 178, 179, 187, 196, 197.

Emph : 12, 35, 108, 109, 110, 112, 113, 114, 115, 116, 183, 184.

ExtraPred : 7, 8, 105, 106, 107, 185.

Interrog : 95, 96, 97, 98, 154, 155.

Juxt : 10, 187, 188, 189.

Neg : 9, 10, 93, 94, 114, 134, 153, 154, 167, 169, 182, 184.

ObjDir : 1, 2, 4, 24, 26, 27, 41, 42, 43, 45, 52, 53, 56, 61, 69, 73, 74, 78, 85, 94, 107, 113, 114, 119, 120, 121, 122, 126, 134, 136, 140, 153, 149, 157, 158, 159, 160, 161, 163, 164, 165, 167, 168, 170, 171, 169, 174, 180, 181, 182, 183, 184, 185, 186, 187, 193, 194.
ObjIndir : 146, 162, 196, 197.
ObjIndirCas : 27, 28, 65, 91, 161, 174.
ObjIndirPrep : 30, 90, 117, 161.
Polylex : 95, 98, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 129, 131, 132, 133, 134, 135, 136, 154, 167.
PredCompletive : 15, 16, 20, 153, 142, 143, 146, 147, 149, 164, 169.
PredPercont : 154, 155, 156, 157, 158, 167.
PredRap : 174.
PredRel : 159, 160, 161, 163, 164, 182, 186, 187.
PredSub : 5, 6, 119, 120, 121, 122, 123, 124, 125, 126, 129, 138, 140, 165, 167, 170, 171, 169, 173, 174.
PredicAdv : 15, 18, 19, 21, 36, 37.
PredicComplObj : 41, 42, 43, 153.
PredicComplSuj : 38, 39, 40.
PredicNom : 6, 13, 16, 19, 22, 28, 31, 32, 34, 72, 80, 81, 83, 129, 131, 153, 162, 167, 173, 175, 176, 189, 192.
PredicOpt : 44, 45, 46, 47, 48, 112.
Ref : 9, 10, 20, 23, 25, 38, 39, 40, 44, 47, 48, 68, 102, 124, 125, 129, 155, 156, 169, 187, 188, 190.
Root : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, **??**, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 105, 106, 107, 108, 109, 110, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 129, 131, 132, 133, 134, 135, 136, 138, 140, 153, 142, 143, 146, 147, 149, 154, 155, 156, 157, 158, 159, 160, 161, 163, 164, 165, 167, 168, 169, 170, 171, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197.
Sub : 6, 15, 16, 20, 119, 120, 121, 122, 123, 124, 125, 126, 129, 138, 140, 153, 142, 143, 146, 147, 149, 164, 165, 169, 170, 171, 174, 190, 191, 192, 193, 194, 196, 197.
Suj : 1, 2, 4, 5, 6, 7, 8, 12, 13, 17, 27, 28, 30, 31, 32, 34, 35, 36, 37, 38, 48, 49, 50, 54, 55, 57, 58, 59, 60, 64, 68, 69, 70, 71, 72, 80, 81, 90, 92, 95, 96, 97, 98, 99, 100, 101, 102, 105, 110, 111, 112, 114, 115, 116, 118, 122, 124, 125, 129, 135, 138, 140, 142, 149, 153, 154, 155, 156, 157, 159, 162, 163, 164, 165, 169, 173, 174, 175, 176, 178, 179, 180, 181, 186, 187, 188, 189, 190, 194, 196, 197.
SujLog : 17, 18, 19, 20, 21, 22, 23, 129, 153.

Bibliographie

Noam Chomsky. *Some concepts and consequences of the theory of government and binding*. MIT press, 1982.

Noam Chomsky. *Lectures on government and binding : The Pisa lectures*. Walter de Gruyter, 1993.

Gerald Gazdar, Klein Ewald, Geoffrey Pullum, and Ivan Sag. *Generalized phrase structure grammar*. Harvard University Press, 1985.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. Annotations at analytical level. Instructions for annotators. *UK MFF ÚFAL, Praha, Czech Republic. URL http ://ufal. mff. cuni. cz/pdt2. 0/doc/manuals/en/a-layer/pdf/a-man-en. pdf (2012-03-18)*, 1999.

Peter Hellwig. Dependency unification grammar. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, pages 195–198, Stroudsburg, PA, USA, 1986. Association for Computational Linguistics. doi: 10.3115/991365.991423. URL `http://dx.doi.org/10.3115/991365.991423`.

Richard A Hudson. *Word grammar*. Blackwell Oxford, 1984.

Milka Ivić, editor. *Sintaksa savremenog srpskog jezika*. Institut za srpski jezik SANU, Beograd, 2005.

Igor Mel'čuk. *Dependency syntax : Theory and practice*. State University Press of New York, 1988.

Pavica Mrazović. *Gramatika srpskog jezika za strance*. Izdavačka knjižarnica Zorana Stojanovića, 2009.

Carl Pollard and Ivan A Sag. *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

Petr Sgall, Eva Hajicová, and Jarmila Panevová. *The meaning of the sentence in its semantic and pragmatic aspects*. Springer Science & Business Media, 1986.

Živojin Stanojčić and Ljubomir Popović. *Gramatika srpskog jezika*. Zavod za udžbenike, 2012.

Pasi Tapanainen and Timo Järvinen. A non-projective dependency parser. In *Proceedings of the fifth conference on Applied natural language processing*, pages 64–71. Association for Computational Linguistics, 1997.

Résumé. Au début de cette thèse, aucun corpus annoté syntaxiquement (treebank) n'était disponible pour le serbe. Or, les treebanks annotés manuellement sont une condition *sine qua non* du développement (entraînement et évaluation) d'outils statistiques dédiés à l'annotation syntaxique automatique (parsers). L'existence des parsers performants permet à son tour l'annotation syntaxique de corpus plus larges, qui peuvent ensuite alimenter des recherches en linguistique théorique. De fait, l'absence de ces ressources pour le serbe freine le développement des recherches sur cette langue dans ces deux directions, et plus généralement les efforts visant l'informatisation et la valorisation du serbe.

Afin de combler cette lacune, nous avons constitué un ensemble de ressources pour le traitement automatique du serbe. Il s'agit en premier lieu du treebank ParCoTrain-Synt, qui contient 101 000 tokens annotés en morphosyntaxe, en lemmes et en syntaxe de dépendances. Nous avons également confectionné le lexique ParCoLex, doté de 7 millions d'entrées provenant de 157 000 lemmes différents. En exploitant ces deux ressources, nous avons développé des modèles pour le parsing, pour l'étiquetage et pour la lemmatisation. Toutes les ressources citées sont librement diffusées à l'adresse suivante : `https://github.com/aleksandra-miletic/serbian-nlp-resources`. Les ressources constituées ont également été exploitées dans le cadre de deux études linguistiques, montrant ainsi que le corpus ParCoTrain-Synt ouvre la porte aux études empiriques basées sur des analyses quantitatives dans le domaine de la linguistique serbe.

Abstract. At the beginning of this PhD, no treebank for Serbian was available. However, manually annotated treebanks are an essential resource for developing (training and evaluating) statistical tools for syntactic analysis (parsers). Efficient parsers, in turn, facilitate the annotation of large corpora, which can be used as a basis for research in theoretical linguistics. The lack of these resources for Serbian slows down the research in these two directions. It also hinders the creation of digital resources for Serbian in general.

In order to address this issue, we created a suite of NLP resources for Serbian. Firstly, we created the ParCoTrain-Synt treebank, a 101 000 token corpus, complete with morphosyntactic annotation, lemmatisation and syntactic dependency annotation. We also built the ParCoLex lexicon, containing 7 million entries for 157 000 different lemmas. Using these two resources, we trained models for parsing, morphosyntactic tagging and lemmatisation. All of the above resources are available at the following address : `https://github.com/aleksandra-miletic/serbian-nlp-resources`. We also used these resources in two experiments in Serbian linguistics, demonstrating that the ParCoTrain-Synt treebank is well suited to empirical studies based on quantitative data analysis.