Construction du jeu d'étiquettes pour le parsing du serbe

Aleksandra Miletic, Cécile Fabre, Dejan Stosic
CLLE-ERSS, CNRS UMR 5263, Maison de la Recherche, Université de Toulouse-Jean Jaurès,
5, allées Antonio Machado, Toulouse Cedex 9
aleksandra.miletic@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr, dstosic@univ-tlse2.fr

Résumé. Cet article présente la démarche utilisée pour la construction d'un jeu d'étiquettes syntaxiques destiné à l'élaboration d'un corpus d'entraînement pour le parsing du serbe dans le but de doter le corpus ParCoLab (corpus parallèle serbe-français-anglais) d'une annotation syntaxique. Vu que le serbe ne dispose pas encore de treebank, il est nécessaire d'élaborer manuellement un corpus d'entraînement. Comme la structure et la taille du jeu d'étiquettes peuvent affecter les résultats du parsing, la définition du jeu est une étape cruciale. Dans le choix des étiquettes, nous avons été guidés par deux principes : réconcilier les traditions grammaticales serbe et française pour des raisons techniques et théoriques et maintenir la comparabilité avec les jeux d'étiquettes élaborés pour d'autres langues slaves. Cette démarche aboutit à un jeu de 28 étiquettes qui assurent la cohérence des traitements dans les différents volets du corpus et la possibilité d'exploiter les outils développés pour d'autres langues dans l'élaboration du corpus d'entraînement.

Abstract.

Constructing a syntactic tagset for the parsing of Serbian
This article presents the process of the construction of a syntactic tagset for Serbian. This tagset is intended for the constitution of a training corpus for the parsing of Serbian, in the global aim of linguistic annotation of the ParCoLab corpus, a parallel corpus of Serbian, French and English. Since there are still no treebanks for Serbian, a manually annotated training corpus must be created. As the parsing results can be affected by the structure and size of the tagset, its definition is a crucial stage. In the tag selection process, we were guided by two main goals: to reconcile the Serbian and the French grammar tradition for technical and linguistic reasons and to maintain comparability with existing tagsets for other Slavic languages. This strategy led us to 28 tags that ensure the coherence of annotation between different subcorpora and allow for the exploitation of tools developped for other languages in the manual annotation process.

Mots-clés : Jeu d'étiquettes, parsing, serbe, corpus parallèle
Keywords: Tagset, parsing, Serbian, parallel corpus

1. Introduction

Dans le domaine du TAL, le serbe reste une langue sous-dotée même si de nombreux outils et ressources existent déjà (pour un aperçu, voir Krstev 2008). En matière d'annotation morphosyntaxique, le seul corpus annoté qui soit librement disponible est celui développé dans le cadre du projet MULTEXT-east (Krstev et al., 2004), et le premier étiqueteur consacré au traitement de cette langue est BTagger, distribué en 2012 (Gesmundo, Samardžić, 2012). Aujourd'hui, il n'existe pas de treebank pour le serbe, et les premières expériences de parsing de cette langue sont très récentes (Jakovljević et al., 2014). Ce paysage a été enrichi récemment par ParCoLab, un corpus de textes littéraires en français, serbe et anglais (Stosic, 2015). Ce corpus parallèle a un intérêt important aussi bien pour la linguistique que pour le TAL car il est susceptible de favoriser des études contrastives théoriques sur ces trois langues, et de devenir une ressource importante pour le développement de différents outils TAL, notamment dans le domaine de la traduction automatique et de la traduction assistée par ordinateur. Or, pour que ce potentiel soit réalisé, il est nécessaire de doter ParCoLab d'annotations linguistiques de différents niveaux. Comme on dispose des outils nécessaires à l'annotation morphosyntaxique de ce corpus (Miletic, 2013), on se propose de développer les ressources pour le parsing et d'enrichir ParCoLab avec une annotation syntaxique.

Etant donné la tradition bien établie du parsing pour le français et l'anglais, reflétée dans le nombre et la diversité des ressources et outils disponibles pour le traitement de ces deux langues (Abeillé et al., 2003, Candito et al., 2010, Marcus et al., 1993, Petrov et al., 2006), l'annotation de ces deux volets du corpus n'est pas considérée comme problématique. Il n'en est pas de même pour le serbe : comme il n'existe pas encore de treebank pour cette langue, l'annotation du volet serbe de ParCoLab avec un parser statistique exige d'abord le

développement des ressources nécessaires à ce type de méthodes, notamment d'un corpus d'entraînement annoté manuellement. Pour pouvoir entamer l'annotation manuelle, il faut d'abord définir le jeu d'étiquettes à utiliser, autrement dit, déterminer quelles relations syntaxiques seront identifiées et codées dans le corpus. Vu que le nombre d'étiquettes, ainsi que leur définition, peuvent affecter l'exactitude du parsing, cette étape n'est pas anodine et exige une réflexion linguistique approfondie. Dans cet article, nous présentons le jeu d'étiquettes que nous avons établi à cette fin.

Plusieurs principes ont guidé le travail d'élaboration du jeu d'étiquettes. Tout d'abord, l'un des usages prévus de ParCoLab est de servir de support à des recherches linguistiques pour la communauté scientifique serbe aussi bien que française. Il fallait donc réconcilier deux traditions grammaticales différentes. La nécessité d'avoir un jeu d'étiquettes comparable vient aussi des contraintes techniques puisque nous allons annoter le volet français de ParCoLab avec le parser Talismane (Urieli, 2013) et que nous envisageons de tester le même outil sur le sous-corpus serbe. Comme le jeu d'étiquettes de Talismane est basé sur celui du *French Treebank* en dépendances (Candito et al., 2009) (dorénavant FTBDep), il diffère de manière importante des jeux élaborés pour les langues slaves (Hajič et al., 1988, Merkler et al., 2013). Les compromis qui sont à faire doivent donc assurer la cohérence des traitements linguistiques dans les deux volets du corpus et permettre une comparaison plus directe des résultats du parsing. Enfin, nous avons jugé intéressant de maintenir une comparabilité avec les jeux d'étiquettes déjà existants pour d'autres langues proches du serbe, notamment le croate. Ceci nous laisse la possibilité, suggérée par les travaux de (Agić et al., 2013), d'exploiter les outils développés pour cette langue dans le traitement du serbe (voir section 3).

Nous tenons à souligner qu'il s'agit ici d'une version préliminaire du jeu d'étiquettes, basée sur une réflexion théorique. Il est par conséquent fort probable que la couverture des phénomènes syntaxiques du serbe ne soit pas parfaite. Pour contrer ce problème, avant d'entamer l'élaboration du corpus d'entraînement, notre jeu d'étiquettes sera mis à l'épreuve d'un échantillon conséquent de texte authentique, ce qui nous permettra d'identifier et combler les éventuelles lacunes.

Dans la suite de cet article, nous présentons d'abord quelques spécificités du serbe et les travaux existants en parsing de cette langue (section 2). Nous définissons ensuite notre problématique et donnons une description brève du corpus ParCoLab (section 3). Dans la section 4, nous présentons le jeu d'étiquettes dans sa totalité, en justifiant quelques choix qui nous ont semblé importants. Nous clôturons enfin cet article en donnant une conclusion et des pistes pour les travaux à venir (section 5).

2. Parsing du serbe

L'annotation syntaxique automatique (ou *parsing*) repose aujourd'hui généralement sur l'utilisation des parsers (logiciels d'analyse syntaxique) statistiques, qui effectuent l'apprentissage des règles d'annotation à partir d'un corpus d'entraînement. Il s'agit d'un échantillon de texte annoté manuellement qui permet au parser de déterminer la probabilité de différentes analyses syntaxiques d'une phrase. Ainsi, une fois lancé sur un texte inconnu, le parser est capable de sélectionner l'analyse la plus probable et peut être utilisé à annoter la totalité d'un corpus de manière automatique (Kübler et al., 2009).

Pour pouvoir entamer l'élaboration d'un corpus d'entraînement, il est d'abord nécessaire de déterminer quelles fonctions syntaxiques seront annotées. Les appellations attribuées à ces fonctions constituent ce que l'on nomme *jeu d'étiquettes*. La structure du jeu d'étiquettes dépend avant tout du fonctionnement syntaxique de la langue en question, mais elle est conditionnée aussi par les choix théoriques retenus et par les spécificités de l'apprentissage automatique. La taille du jeu peut varier en fonction de l'usage envisagé du corpus : les applications TAL favorisent souvent des jeux restreints, alors que l'exploitation d'un corpus dans le domaine linguistique exige une granularité plus fine. Vu que la taille du jeu et la définition des étiquettes peuvent affecter les résultats du parsing, la définition d'un jeu d'étiquettes constitue une étape préalable cruciale.

2.1. Spécificités du serbe

Tout comme les autres langues slaves, le serbe dispose d'une morphologie flexionnelle riche : à titre d'illustration, les noms varient selon le genre (masculin, féminin ou neutre), le nombre (singulier, pluriel et paucal) et le cas (nominatif, génitif, datif, accusatif, vocatif, instrumental ou locatif). En plus de ces trois catégories, les adjectifs sont également affectés par celles du degré de comparaison (positif, comparatif ou superlatif) et de l'aspect (défini ou indéfini). Tous les pronoms se déclinent, et certains distinguent le genre, le nombre et la personne (première, deuxième ou troisième). L'indication de certaines fonctions syntaxiques étant portée par le marquage casuel, la structure de la phrase est très flexible. Même si l'ordre des constituants canonique est SVO, les ordres SOV, VOS, VSO, OVS et OSV sont non seulement grammaticaux, mais fréquents

(cf. par exemple Stanojčić, Popović, 2011, p.367, Ivić 2005). Il est également possible d'avoir des constituants discontinus, comme dans l'exemple 1.

[Lep-u	{ste	kuć-u]	kupi-li}.
beau-ACC.SG.F	être[PRS.2PL]	maison-ACC.SG.F	acheter-PTCP.PL.M

Vous avez acheté une belle maison / C'est une belle maison que vous avez achetée.

Exemple 1 : Constituant discontinu en serbe

Ici, l'un des constituants est délimité par des crochets (*Lepu kuću*), et l'autre par des accolades (*ste kupili*). On voit donc qu'il est possible d'insérer le verbe auxiliaire entre l'adjectif épithète et le nom auquel cet adjectif est rattaché.

Il est généralement considéré que le parsing de ce type de langues doit être basé sur l'analyse en dépendances, et non pas sur l'analyse en constituants, cette dernière ne disposant pas de mécanismes pour gérer la discontinuité des constituants (Buchholz, Marsi, 2006, Nivre et al., 2007). Or, à la différence d'autres langues slaves comme le tchèque et le russe, qui ont une tradition importante en syntaxe de dépendances (cf. Sgall et al., 1986, Мельчук, 1995), la description syntaxique du serbe repose traditionnellement sur l'analyse en constituants (Stanojčić, Popović, 2011, Ivić, 2005). Ceci signifie qu'il n'existe pas encore de formalisme pour l'annotation du serbe en syntaxe de dépendances[1], ce qui peut être l'une des causes du manque de travaux sur le parsing du serbe.

En effet, le premier travail sur le parsing de cette langue a été publié très récemment (Jakovljević et al., 2014). Ces expériences initiales ont été effectuées sur un treebank en cours de développement, basé sur le corpus *AlfaNum* (Sečujski, 2009). Pour la constitution du treebank, Jakovljević et ses collaborateurs reprennent le jeu d'étiquettes de *Prague Dependency Treebank* (dorénavant PDT) (Hajič, 1998). Ce jeu contient 28 étiquettes, dont 15 annotent les fonctions syntaxiques principales comme sujet, objet, prédicatif nominal etc., alors que les étiquettes restantes sont consacrées aux éléments considérés comme auxiliaires : verbes auxiliaires, mots emphatiques, différents types de ponctuation. L'annotation des fonctions syntaxiques est de faible granularité : à titre d'exemple, il existe une seule étiquette Obj (objet), alors que le tchèque distingue l'objet direct et l'objet indirect. (Jakovljević et al., 2014) utilisent ce jeu d'étiquettes en le modifiant de manière minimale pour optimiser le traitement des particules interrogatives, de l'auxiliaire *hteti* 'vouloir' et de certains cardinaux.

L'avantage principal de l'utilisation du jeu d'étiquettes de PDT pour l'annotation du serbe réside dans le fait que le guide d'annotation de PDT est disponible sur internet. En effet, afin d'assurer la cohérence des annotations manuelles, l'élaboration d'un corpus d'entraînement nécessite un ensemble de règles de traitement détaillées, appelé le guide d'annotation. L'accès à un guide existant permet d'entamer l'élaboration du corpus d'entraînement sans devoir consacrer un temps important à la définition des règles d'annotation. La même approche a été utilisée dans l'élaboration des premiers treebanks du croate (Hrvatska ovisnosna banka stabala ou HOBS, cf. (Tadić, 2007)) et du slovène (Slovene Dependency Treebank, cf. (Džeroski et al., 2006)). Vu la proximité typologique du tchèque avec le croate et le slovène, il a été possible d'adapter le jeu élaboré pour PDT à l'annotation syntaxique de ces deux langues. Cependant, cette stratégie a été remise en question dans des travaux plus récents : suite aux remarques des annotateurs humains selon lesquelles le jeu de PDT n'était pas intuitif dans son application au croate (Berović et al., 2012), un nouveau jeu de 15 étiquettes a été établi. Ce jeu reste fondé sur les principes de base de PDT, avec la réduction de taille obtenue par la simplification du traitement de certains éléments. La pertinence de ces choix a été justifiée par les résultats sur l'accord inter-annotateurs présentés dans (Agić, Merkler, 2013), qui montrent que l'utilisation du jeu d'étiquettes réduit apporte une hausse de 7,22 points pour le score LAS[2], et de 2,13 points pour le score UAS[3]. Le nouveau jeu a ensuite été utilisé dans l'élaboration d'un corpus de messages électroniques du croate (Merkler et al., 2013) et de *SETimes.hr*, un treebank du croate basé sur des textes journalistiques (Agić, Merkler, 2013). On peut remarquer la même tendance dans le parsing du slovène : lors de l'élaboration du JOS Corpus (Jezikoslovno označevanje slovenščine 'Annotation linguistique du slovène', (Erjavec et al., 2010)), un deuxième treebank du slovène, il a été noté que les annotateurs humains avaient des difficultés à maintenir la cohérence des annotations avec le jeu d'étiquettes basé sur le PDT. Par conséquent, le JOS Corpus a été élaboré en utilisant un jeu minimaliste de 10 étiquettes (*id.*).

[1] Un des relecteurs nous a signalé les travaux de P. Mrazović, que malheureusement nous n'avons pas réussi à nous procurer avant de terminer cet article.

[2] *Labeled attachment score* : pourcentage des tokens pour lesquels le parser a bien identifié le gouverneur et le type de la relation.

[3] *Unlabeled attachment score* : pourcentage des tokens pour lesquels le parser a bien identifié le gouverneur, sans tenir compte du type de relation attribuée.

Etant donné ces expériences, nous avons décidé de ne pas fonder notre jeu d'étiquettes sur celui du PDT, mais d'en concevoir un nouveau, tout en cherchant à respecter les différentes contraintes posées par la nature plurilingue de notre corpus et son usage envisagé.

3. Méthodologie et données

3.1. Méthodologie de construction du jeu d'étiquettes

Comme il a été mentionné ci-dessus, nous avons sélectionné le parser Talismane (Urieli, 2013) pour effectuer l'annotation du volet français du corpus. Ce parser, paramétré et testé sur le français, atteint une exactitude de 86,9 - 88,0% pour le score LAS et de 89,5 - 90,4% pour le score UAS (*id.*, p. 154), en fonction de la configuraiton utilisée. Ces résultats sont comparables à ceux obtenus par d'autres parsers disponibles pour le français, à savoir Berkeley (Petrov et al., 2006), MSTParser (McDonald et al., 2006), et MaltParser (Nivre et al., 2006), dont les performances sur le français sont présentées dans (Candito et al., 2010). Talismane a l'avantage de proposer également la possibilité de créer une approche hybride en intégrant des règles grammaticales afin d'améliorer la qualité de la sortie.

Nous envisageons également de paramétrer cet outil sur le corpus serbe pour tester sa capacité à s'adapter à une langue à ordre de constituants flexible. Si les résultats sont satisfaisants, il sera utilisé pour le parsing de la totalité du volet serbe de ParCoLab. Pour préserver l'intérêt scientifique de cette démarche et pouvoir comparer les résultats de l'outil sur ces deux langues, il est indispensable de maintenir un degré de comparabilité entre les deux jeux d'étiquettes. Cette démarche n'est pas simple, d'abord à cause des différences structurelles entre le français et le serbe, et ensuite à cause des différences entre les deux traditions grammaticales. Pour pouvoir faire les rapprochements nécessaires, nous avons été obligés d'adapter certaines analyses syntaxiques admises dans la tradition grammaticale serbe. Les plus importants de ces choix sont présentés et discutés dans la Section 4.

Un autre ensemble de compromis a été nécessaire pour respecter notre décision de garder notre jeu d'étiquettes proche du jeu croate utilisé dans SETimes.hr (Merkler et al., 2013), et ceci par souci d'accélérer l'élaboration du corpus d'entraînement. En effet, le croate dispose déjà des modèles de parsing : dans (Agić, Merkler, 2013), trois logiciels sont entraînés et testés sur le croate, et le modèle le plus performant, à savoir celui de MSTParser (McDonald et al., 2006), est mis à la disposition de la communauté scientifique. Etant donné la proximité syntaxique et morphologique du serbe et du croate, il est envisageable d'utiliser ce modèle pour effectuer une première annotation automatique du corpus d'entraînement, qui sera ensuite manuellement corrigée, à la fois pour éliminer les erreurs de parsing et pour rendre les annotations conformes à notre jeu d'étiquettes. Les résultats présentées dans (Agić et al., 2013) prouvent la pertinence de cette approche : dans ces expériences, MSTParser, entraîné exclusivement sur des données du croate, atteint le même niveau d'exactitude sur les échantillons du croate et du serbe. Afin de pouvoir tester cette possibilité, le jeu d'étiquettes ciblé doit garder un minimum de comparabilité avec le jeu intégré dans le modèle pour le croate de MSTParser. Sans cela, la correction de l'annotation de sortie exigerait trop d'ajustements et présenterait peu d'avantage par rapport à l'annotation manuelle pure.

En prenant en compte ces deux contraintes (comparabilité avec le jeu d'étiquettes de Talismane et celui de SETimes.hr), nous partons des fonctions syntaxiques traditionnellement utilisées en serbe (Stanojčić, Popović, 2011, Ivić, 2005) et établissons un jeu de 28 étiquettes (cf. § 4).

3.2. ParCoLab, corpus parallèle serbe-français-anglais

ParCoLab est un corpus parallèle trilingue. Il contient des textes originaux en serbe, français et anglais, ainsi que des traductions professionnelles de ces textes dans les deux autres langues du corpus. Aujourd'hui, il contient exclusivement des textes littéraires, mais une diversification en termes de genres est prévue pour un avenir proche, avec l'apport de textes provenant du web. Le corpus comporte à présent 1 650 501 tokens, dont 639 555 en serbe, 658 373 en français et 352 573 en anglais. Leur distribution par type de texte (original ou traduction) est donnée dans le Tableau 1.

Un échantillon de 150 000 tokens issu de la partie du corpus contenant des textes originaux en serbe a été transformé en corpus d'entraînement pour l'annotation morpho-syntaxique. Le jeu d'étiquettes morphosyntaxiques utilisé compte 47 étiquettes (Miletic, 2013), ce qui présente un compromis entre les deux jeux utilisés pour l'étiquetage du serbe dans la littérature : le premier est celui proposé dans le cadre du projet MULTEXT-East, qui encode toutes les informations morphosyntaxiques et compte plus de 900 tags (Krstev et al., 2004), et le deuxième est un jeu minimaliste de 15 tags, n'encodant que la partie du discours (Utvić, 2011)

Des 150 000 tokens que le sous-corpus comporte, 100 000 ont été annotés manuellement, et 50 000 ont fait l'objet d'un étiquetage automatique suivi d'une correction manuelle. Afin de profiter de la haute qualité des annotations morphosyntaxiques dans les expériences du parsing, c'est le même sous-corpus qui nous servira de base dans l'élaboration d'un corpus d'entraînement pour le parsing.[4]

	Volet français	**Volet serbe**	**Volet anglais**
	originaux	**originaux**	**originaux**
	113 973	459 708	229 044
	serbe → français	**français → serbe**	**français → anglais**
	388 326	104 310	123 529
	anglais → français	**anglais → serbe**	**serbe → anglais**
	156 074	75 537	-
TOTAL :	658 373	639 555	352 573

TABLEAU 1 : Nombre de tokens provenant des textes originaux et des traductions dans ParCoLab

Cependant, pour limiter le nombre d'étiquettes, le jeu de (Miletic 2013) encode seulement la partie du discours principale et sa sous-catégorie, en ajoutant le degré de comparaison pour les adjectifs et les adverbes. Comme une partie des fonctions syntaxiques en serbe sont indiquées par le marquage casuel, nous envisageons d'élaborer un module d'identification de cas qui nous permettra d'intégrer dans l'annotation morpho-syntaxique cette information cruciale pour le parsing.

4. Un nouveau jeu d'étiquettes pour le parsing du serbe

4.1. Présentation des étiquettes

Le jeu que nous proposons compte 28 étiquettes. Leur présentation accompagnée d'une brève définition de leur application est donnée dans le Tableau 2.

Etiquette		Définition	Exemple
1.	Pred	prédicat ; dans les temps composés, accordé au participe du verbe principal	*Filip **jede*** 'Filip **mange**' ; *Filip je **jeo*** 'Filip a **mangé**'
2.	AuxV	verbe auxiliaire dans les temps composés	*Filip **je** jeo* 'Filip **a** mangé'
3.	AuxVNeg	forme synthétique de verbe auxiliaire nié	*Filip **nije** jeo* 'Filip **n'a pas** mangé'
4.	Suj	sujet au nominatif	***Filip** jede* '**Filip** mange'
5.	SujLog	sujet logique exprimé au génitif, datif ou accusatif	cf. Exemple 4
6.	ObjDir	objet direct, exprimé à l'accusatif ou génitif	*Filip jede **jabuku*** 'Filip mange **une pomme**' ; *Filip pije **mleka*** 'Filip boit **du lait**'
7.	ObjIndir	objet indirect, exprimé au datif	*Filip daje jabuku **Milici*** 'Filip donne une pomme **à Milica**'
8.	ObjPrep	objet prépositionnel, réalisé sous forme d'un groupe prépositionnel régi par le verbe	*Filip misli **na porodicu*** 'Filip pense **à sa famille**'
9.	ComplInf	complément d'un verbe modal ou aspectuel sous forme d'un infinitif	cf. Exemple 11
10.	DepVAdv	dépendant adverbial d'un verbe	*Filip **vredno** radi* 'Filip travaille **assidumen**t'
11.	DepVCas	dépendant d'un verbe sous forme d'un GN fléchi	*Milica ide **kući*** 'Milica va **à la maison'**
12.	DepVPrep	dépendant d'un verbe sous forme d'un GP	*Milica radi **sa Filipom*** 'Milica travaille **avec Filip'**
13.	AttrSuj	complément d'un verbe attributif qui s'accorde avec le sujet	cf. Exemple 2
14.	AttrObj	complément d'un verbe attributif qui	cf. Exemple 3

[4] Davantage d'informations sur le corpus en question et l'étiquetage morphosyntaxique de ParCoLab peuvent être trouvées dans (Balvet et al., 2014) et (Miletic, 2013).

		s'accorde avec l'objet	
15.	Ep	épithète d'un nom sous forme d'un adjectif qui s'accorde avec le nom	*Alan kupuje **lepu** kuću* 'Alain achète une **belle** maison'
16.	DepNCas	dépendant d'un nom sous forme d'un GN	*kuća **moga strica*** 'maison de mon oncle'
17.	DepNPrep	dépendant d'un nom sous forme d'un GP	*kolač **sa višnjama*** 'gâteau **aux cerises**'
18.	DepAdjAdv	dépendant adverbial d'un adjectif	***vrlo** lepa kuća* '**très** belle maison'
19.	Ap	apposition	*Ivo Andrić, **pisac i nobelovac*** 'Ivo Andric, **écrivain et prix nobel**'
20.	EpDet	épithète occupant une place non canonique, en tête de phrase et/ou détaché du nom par des virgules	***Umoran**, Filip se vratio kući* '**Fatigué,** Filip est rentré'
21.	Sub	tout subordonnant sauf les relatifs	*Javiću se **kad** stignem* 'Je t'appellerai **quand** je serai arrivé'
22.	PredRel	prédicat de proposition relative	*Video sam čoveka koji se **doselio*** 'J'ai vu l'homme qui **vient d'emménager'**[5]
23..	Coord	conjonction de coordination	*Filip **i** Alan se smeju* 'Filip **et** Alain rient'
24.	DepCoord	éléments coordonnés sauf le premier (voir ci-dessous)	*Filip i **Alan** se smeju* 'Filip et **Alain** rient'
25.	CPrep	complément d'une préposition	*kolač sa **višnjama*** 'gâteau aux **cerises**'
26.	Elp	ellipse	*Stigao je **umoran*** '[~~il~~] est arrivé **fatigué**'[6]
27.	Neg	particule de négation	*Filip **ne** jede jabuku* 'Filip **ne** mange **pas** la pomme'
28.	Punc	ponctuation	*Filip se vratio**!*** 'Filip est rentré **!**'

TABLEAU 2 : Jeu d'étiquettes proposé

Ce jeu contient 13 étiquettes de plus que celui utilisé pour le croate dans SETimes.hr (Agić et al., 2013). Cependant, pour la majorité des étiquettes, il s'agit simplement d'une augmentation de granularité. Le Tableau 3 résume les différences principales.

SETimes.hr	**ParCoLab**
Sb (sujet)	Suj
	SujLog
Obj (objet)	ObjDir
	ObjIndir
	ObjPrep
Atr (modifieur nominal)	Ep
	DepNCas
	DepNPrep

TABLEAU 3: Correspondances d'étiquettes entre SETimes.hr et ParCoLab

Les distinctions faites dans notre jeu d'étiquettes sont basées sur des critères formels, notamment le cas et la catégorie du constituant (cf. Tableau 2). Par conséquent, la conversion de l'étiquette globale utilisée par SETimes.hr vers les étiquettes plus spécifiques de notre jeu peut se faire de manière automatique. Des différences plus fondamentales concernent le traitement de la fonction désignée dans SETimes.hr comme l'attribut verbal, des groupes prépositionnels et de la coordination, ce qui sera expliqué plus en détail dans la sous-section 4.2.

Quant à la comparabilité de notre jeu d'étiquettes avec celui de FTBDep, repris par Talismane, nous constatons d'abord que leurs tailles respectives sont proches. Le jeu de FTBDep compte 21 étiquettes de base, utilisées pour l'annotation automatique, et 8 étiquettes plus spécifiques, réservées à l'annotation manuelle (Candito et al., 2009). Nous prendrons ici en considération seulement les 21 étiquettes destinées au parsing.

[5] Le traitement des relatives est repris de FTBDep. Le prédicat de la relative est rattaché au prédicat de la proposition indépendante. Ceci permet d'annoter le relatif avec la fonction qu'il exerce au sein de la proposition relative. Sinon il serait annoté simplement en tant que subordonnant et son rôle syntaxique dans la relative serait perdu.

[6] Nous reprenons le traitement de l'ellipse du PDT. Si un constituant est absent de la phrase, les nœuds qui dépendraient de lui sont annotés avec l'étiquette Elp et rattachés à l'endroit où le nœud manquant se trouverait dans l'arbre. Dans l'exemple donné, c'est l'épithète détaché *umoran* 'fatigué' qui porterait l'étiquette Elp, vu que le sujet duquel il dépend n'est pas présent dans la phrase.

A la différence de FTBDep, qui dispose d'une étiquette pour le sujet, nous faisons la distinction entre le sujet et le sujet logique. Comme indiqué ci-dessus, cette opposition est basée sur des critères morphosyntaxiques et une correspondance directe peut être établie entre les étiquettes de notre corpus et celle de FTBDep. Les étiquettes pour l'attribut du sujet (*ats*) et l'attribut de l'objet (*ato*) de FTBDep ont des équivalents directs dans les étiquettes *AttrSuj* et *AttrObj*. Le traitement de la fonction objet est proche dans les deux jeux : l'étiquette *obj* de FTBDep (objet direct) correspond à celle de *ObjDir* dans notre jeu, *a_obj* et *de_obj* (objet indirect introduit par *à* ou *de*, respectivement) sont équivalentes de *ObjIndir*, et *p_obj* (objet introduit par une préposition autre que *à* et *de*) correspond à *ObjPrep*. La différence la plus importante entre ces deux jeux d'étiquettes concerne un aspect de l'application de l'étiquette *obj* dans FTBDep, et qui n'est pas repris dans ParCoLab. Ce point est discuté dans la sous-section suivante.

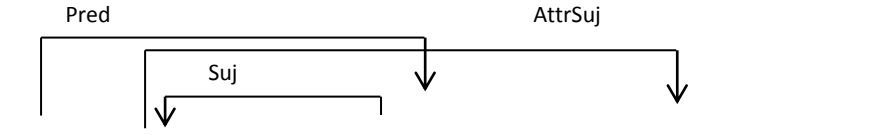
Un autre point commun que notre jeu d'étiquettes partage avec SETimes.hr et FTBDep est la décision de ne pas encoder la différence entre les compléments et les modifieurs. Comme cette distinction ne repose pas clairement sur des critères de surface accessibles à un parser, et qu'elle peut être problématique même pour les annotateurs humains, nous l'avons omise de notre jeu d'étiquettes. En revanche, nous considérons ces différentes fonctions comme des dépendants et précisons dans les étiquettes la catégorie de leur gouverneur et celle du dépendant lui-même. Ainsi, l'étiquette *DepNCas* désigne le dépendant d'un nom qui a la forme d'un GN fléchi, alors que *DepVPrep* indique le dépendant d'un verbe sous forme d'un GP. La structure des étiquettes permettra par la suite de faire des regroupements des fonctions s'il se prouve que la granularité actuelle est trop fine.

4.2. Justification des choix

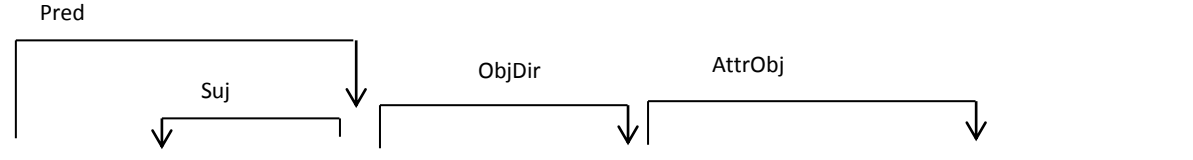
Cette partie de l'article est consacrée à la justification de plusieurs choix faits dans la construction du jeu d'étiquettes. Certains d'entre eux sont techniques, conditionnés par les caractéristiques intrinsèques des algorithmes de parsing, alors que d'autres concernent plutôt l'analyse linguistique de différentes fonctions. Dans chacun des points discutés, nous essaierons d'expliciter notre position par rapport à la tradition grammaticale serbe, ainsi que par rapport aux jeux d'étiquettes de SETimes.hr et de FTBDep.

Attribut du sujet et attribut de l'objet direct

Ces deux fonctions ne sont pas reconnues dans les travaux sur la syntaxe du serbe. Elles sont réparties entre les fonctions des prédicatifs nominal, complémentaire et optionnel, qui correspondent respectivement aux compléments du verbe *biti* 'être', ceux des autres verbes essentiellement attributifs (cf. (Riegel, 1981)) comme *proglasiti (se)* '(se) proclamer', *smatrati (se)* '(se) considérer', *prozvati (se)* '(se) nommer', et ceux des verbes occasionnellement attributifs (*id.*). Cependant, il n'y a pas de critères formels pour distinguer ces constituants : les trois favorisent la position post-verbale (quoiqu'ils puissent être placés devant le verbe), admettent des dépendants sous forme de groupe nominal ou adjectival et si le dépendant se réalise sous forme d'un groupe adjectival, il prend les marques du genre et du nombre soit du sujet soit de l'objet direct. Par conséquent, nous remplaçons la distinction traditionnelle citée ci-dessus par celle du jeu d'étiquettes de FTBDep et introduisons les étiquettes pour l'attribut de sujet (*AttrSuj*) et l'attribut d'objet (*AttrObj*) (cf. exemples 2 et 3).



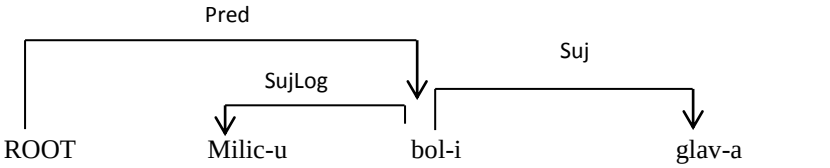
Exemple 2: Etiquette AttrSuj



Exemple 3: Etiquette AttrObj

Sujet logique

En serbe, le sujet est typiquement exprimé au nominatif et désigne l'agent du processus ou l'expérienceur de l'état décrit par le verbe (*Milica čita knjigu* 'Milica-NOM lit livre-ACC'). Cependant, un groupe de verbes exprimant un état physique ou mental exigent que leur expérienceur soit au datif ou à l'accusatif (cf. exemple 4). Ce constituant est désigné dans la littérature comme sujet logique (Ivic, 2005, Stanojčić, Popović, 2011).



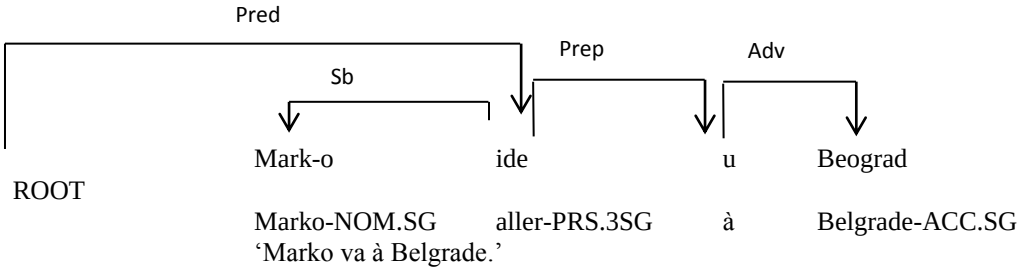
Exemple 4: Etiquette SujLog

Comme il existe un critère formel de distinction par rapport au sujet au nominatif, nous avons décidé d'introduire cette distinction dans notre jeu d'étiquettes : le sujet typique au nominatif sera annoté comme *Suj*, alors que pour le sujet logique on utilisera l'étiquette *SujLog*.

On peut remarquer que formellement ce constituant peut coïncider avec celui d'*ObjDir*, les deux étant des GN au génitif ou à l'accusatif. Ils montrent cependant des comportements différents quant à la linéarisation : le sujet logique est typiquement antéposé au verbe, avec une préférence pour la position initiale dans la phrase, alors que la position canonique de l'objet direct est à droite du verbe. Il est vrai que les deux sont mobiles et peuvent prendre la position typique de l'autre si la focalisation de la phrase l'exige. Il reste donc à voir si un parser sera capable d'opérer cette distinction.

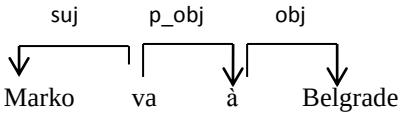
Groupes prépositionnels

Les jeux d'étiquettes de SETimes.hr et FTBDep adoptent deux approches différentes pour le traitement des groupes prépositionnels. Dans SETimes.hr, toutes les prépositions sont liées à leur gouverneur par la relation *Prep*, alors que c'est le complément de la préposition qui porte l'étiquette de la fonction exercée par le groupe prépositionnel dans la phrase (cf. Exemple 5).



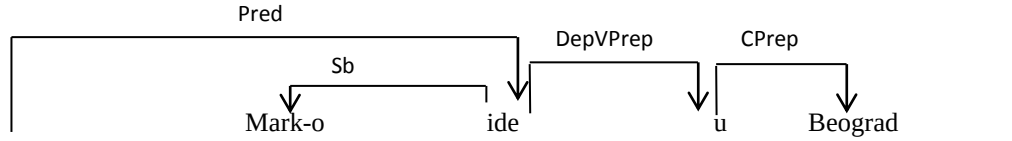
Exemple 5: Traitement des groupes prépositionnels dans SETimes.hr

En revanche, dans FTBDep, on annote la préposition avec la fonction du groupe prépositionnel, alors que le complément de la préposition est annoté en tant que *obj* (cf. exemple 6).



Exemple 6: Traitement des groupes prépositionnels dans FTBDep

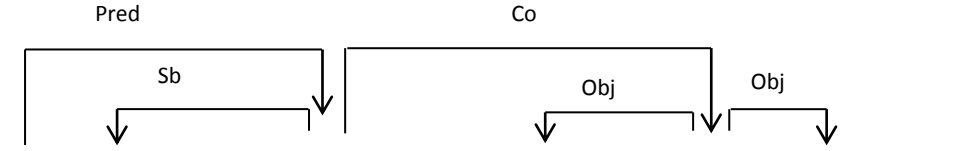
Nous reprenons l'approche du FTBDep avec une différence : au lieu d'étendre le domaine d'application de l'étiquette de l'objet direct aux compléments de préposition, nous définissions une étiquette spécialisée, *CPrep*. Ceci résulte dans le traitement présenté dans l'exemple 7.



Exemple 7: Traitement des groupes prépositionnels dans ParCoLab

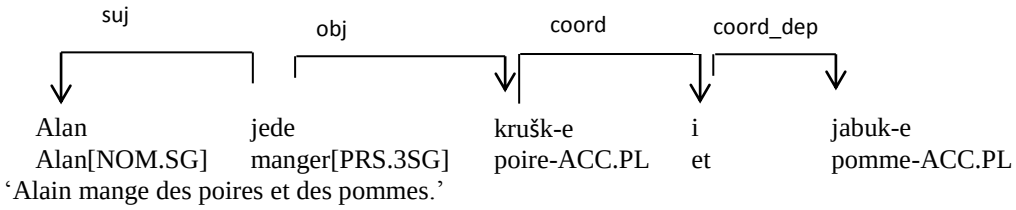
Coordination

Les formalismes d’annotation de SETimes.hr et de FTBDep diffèrent également dans leur traitement des constructions de coordination. Le premier reprend l’approche proposée par PDT, qui consiste à utiliser la fonction *Co* pour relier la conjonction de coordination au gouverneur des constituants coordonnés, et d’ensuite relier les coordonnés à la conjonction par l’étiquette de la fonction qu’ils exercent dans la phrase (cf. exemple 8).



Exemple 8: Traitement de coordination dans SETimes.hr

Bien qu’en accord avec l’intuition linguistique, ce traitement a des défauts du point de vue technique : ici, le parser est obligé de déterminer si la forme *kruške* fait partie d’une coordination avant d’identifier l’objet du verbe *jede*. Autrement dit, le parser doit reconnaître la coordination avant de pouvoir décider quel est le statut des coordonnés (Urieli, 2013). FTBDep propose une alternative que, tout en étant moins conforme à l’intuition linguistique, permet un traitement plus simple pour le parser (cf. exemple 9).

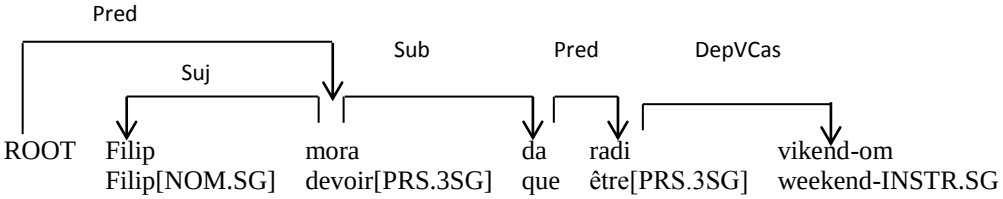


Exemple 9: Traitement de la coordination dans FTBDep

Avec ce traitement, le parser identifie d’abord la fonction du premier coordonné pour passer ensuite à l’analyse de la construction coordonnée. Même si la fonction des coordonnés autres que le premier n’est pas explicitement notée, elle peut être récupérée de l’étiquette du premier coordonné. Pour notre jeu d’étiquettes, nous adoptons cette approche et utilisons par conséquent l’étiquette *Coord* pour lier la conjonction de coordination au premier coordonné, et l’étiquette *DepCoord* pour lier tous les coordonnés sauf le premier à la conjonction.

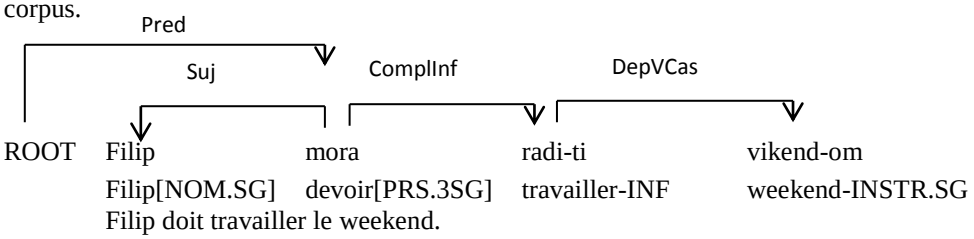
Traitement du prédicat complexe

La tradition grammaticale serbe considère comme « prédicats complexes » les constructions avec les verbes modaux et aspectuels (Stanojčić, Popović, 2011, p.269). Ces verbes sont le plus couramment complémentés par la construction *da + Vprezent* ‘que +V présent’, comme dans l’exemple 10.



Exemple 10: Prédicat complexe sous forme *da + Vprezent* ‘que + Vprésent’

Cependant, une complémentation en infinitif est également possible. Pour étiqueter cette construction alternative, deux traitements différents peuvent être envisagés. Il est possible de faire un rapprochement avec la construction équivalente *da + Vprezent* et d'annoter l'infinitif complément du verbe principal comme une subordonnée. Néanmoins, ceci veut dire que l'étiquette *Sub*, qui est dédiée à l'annotation des subordonnants, deviendrait également applicable aux verbes. De même, ses dépendants changeraient de manière importante : dans son emploi canonique, cette étiquette a un descendant *Pred* qui correspond au prédicat de la subordonnée, auquel sont ensuite rattachés les dépendants typiques du prédicat. Si on appliquait cette étiquette à l'infinitif, il n'y aurait plus de descendants *Pred*, et ce sont les dépendants du prédicat qui seraient attachés directement à l'étiquette *Sub*. Pour éviter la multiplication des contextes possibles pour cette étiquette, nous choisissons de considérer qu'il s'agit d'un complément verbal spécifique et introduisons une nouvelle étiquette, *ComplInf*. Cette approche nous permet également de maintenir la distinction linguistique entre ces deux constructions dans le corpus.



Exemple 11 : Utilisation de l'étiquette ComplInf

La discussion menée dans cette section montre que le jeu d'étiquettes que nous proposons respecte certaines distinctions traditionnellement admises dans la syntaxe du serbe (par exemple, pour le sujet et l'objet). Néanmoins, quelques adaptations ont également été nécessaires, notamment pour les fonctions regroupées sous le nom du prédicatif dans la grammaire serbe. Ces compromis sont justifiés par la contrainte de comparabilité entre notre jeu d'étiquettes, celui de FTBDep et celui de SETimes.hr. Grâce à un degré de comparabilité élevé, nous espérons à la fois maintenir une cohérence d'annotation entre les volets serbe et français de ParCoLab, tester les performances de Talismane sur le serbe, et exploiter les ressources existantes pour le croate dans l'élaboration du corpus d'entraînement.

5. Conclusion

Ce travail présente la première version d'un jeu d'étiquettes syntaxiques pour l'élaboration d'un corpus d'entraînement pour le parsing du serbe. Il s'inscrit dans le projet de doter le corpus parallèle serbe-français-anglais ParCoLab d'une couche d'annotation syntaxique.

En identifiant les fonctions syntaxiques qui doivent être représentées par le jeu, nous avons pris en compte les fonctions traditionnellement utilisées dans la syntaxe du serbe, mais nous avons également été guidés par le besoin de maintenir la comparabilité avec les jeux d'étiquettes d'un treebank du croate SETimes.hr et de French Treebank en dépendances. Cette approche a été choisie avec deux objectifs principaux : d'abord, nous souhaitons utiliser les modèles de parsing développés pour le croate sur notre corpus d'entraînement et accélérer ainsi son élaboration ; deuxièmement, une fois le corpus d'entraînement prêt, nous prévoyons de tester le parser Talismane (Urieli, 2013) et, si ses performances sont satisfaisantes, l'utiliser pour annoter la totalité du sous-corpus serbe de ParCoLab. Cette démarche aboutit à un jeu de 28 étiquettes. Cette taille dépasse celle des jeux de SETimes,hr et de FTBDep (15 et 21 étiquettes respectivement), mais la structure de notre jeu permet d'établir des correspondances nécessaires et a l'avantage de rendre possible la représentation des sous-types des fonctions syntaxiques principales (sujet, objet, dépendants nominaux, etc.).

Comme nous l'avons déjà indiqué, il s'agit d'une proposition initiale, fondée sur une réflexion théorique. Pour vérifier la pertinence de nos choix et combler d'éventuelles lacunes, nous testerons ce jeu d'étiquettes sur un échantillon du corpus. Une fois la version finale du jeu arrêtée, elle sera utilisée pour l'annotation manuelle du corpus d'entraînement.

Références

ABEILLE, A., CLEMENT, L., TOUSSENEL, F. (2003). Building a treebank for French. Dans: A. Abeillé, éd. *Treebanks.* Dordrecht : Kluwer, 165-187.

AGIC, Ž., MERKLER, D. (2013). Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. *LNCS* 8082, 560-567.

AGIC, Ž., MERKLER, D., BEROVIC, D. (2013). Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. Actes de *SPMRL* à *EMNLP*, 22-33.

BALVET, A., STOSIC, D., MILETIC, A. (2014). TALC-sef, A Manually-Revised POS-Tagged Litterary Corpus in Serbian, English, and French. Actes de *LREC 2014*, 4105-4110.

BEROVIC, D., AGIC, Ž.,TADIC, M. (2012). Croatian Dependency Treebank: Recent Development and Initial Experiments. Actes de *LREC 2012*, 1902-1906.

BUCHHOLZ, S., MARSI, E. (2006). CoNLL-X shared task on multilingual dependency parsing. Actes de *Tenth Conference on Computational Natural Language Learning*, 149-164.

CANDITO, M., CRABBE, B., FALCO, M. (2009). Dépendances syntaxiques de surface pour le français. *Rapport technique, Université Paris 7.*

CANDITO, M., NIVRE, J., DENIS, P., HENESTROZA ANGUIANO, E. (2010). Benchmarking of Statistical Dependency Parsers for French. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 108-116.

DZEROSKI, S., ERJAVEC, T., LEDINEK N., PAJAS, P., ŽABOKRTSKY, Z., ŽELE, A. (2006). Towards a Slovene Dependency Treebank. Actes de *LREC 2006*, 1388-1391.

ERJAVEC, T. (2004). *MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora.* Actes de *LREC 2004*, 1535-1538.

ERJAVEC, T., FISER , D., KREK, S., LEDINEK, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. Actes de *LREC2010*, 1806-1809.

GESMUNDO, A., SAMARDZIC, T. (2012). Lemmatising Serbian as a category tagging with bidirectional sequence classification. Actes de *LREC 2012*, 2103-2106.

HAJIC, J. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning.* Prague : Karolinum, 106-132.

HAJIC, J., HAJICOVA, E. (1997). Syntactic tagging in the Prague Treebank. *Proceedings of the Second European Seminar "Language Applications for a Multilingual Europe"*, 55-68.

HAJIC, J., PANEVOVA, J., BURANOVA, E., URESOVA, Z., BEMOVA, A. (1999). Annotations at analytical level: Instructions for annotators. *Rapport technique, UK MFF UFAL, Prague.*

IDE, N., VERONIS, J. (1994). MULTEXT (Multilingual text tools and corpora). *Proceedings of the 15th Conference on Computational Linguistics - Volume 1,* 588-592.

IVIC, M. éd., 2005. *Sintaksa savremenog srpskog jezika.* Beograd: Institut za srpski jezik SANU.

JAKOVLJEVIC, B., KOVACEVIC, A., SECUJSKI, M., MARKOVIC, M. (2014). A Dependency Treebank for Serbian: Initial Experiments. *Speech and Computer Lecture Notes in Computer Science* 8773, 42-49.

KRSTEV, C. (2008), *Processing of Serbian. Automata, Texts and Electonic Dictionaries,* Belgrade, Faculty of Philology, University of Belgrade.

KRSTEV, C., VITAS, D., ERJAVEC, T. (2004). MULTEXT-East resources for Serbian. Actes de *7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije*, 108-114.

KUBLER, S., MCDONALD, R., NIVRE, J. (2009). Dependency parsing. *Synthesis lectures on Human Language Technologies*, 1(1), 1-127.

LEDINEK, N., ŽELE, A. (2005). Building of the Slovene dependency treebank corpus according to the Prague dependency treebank corpus. *Proceedings of Grammar and Corpus*.

MARCUS, M. P., SANTORINI, B., MARCINKIEWICZ, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313-330.

MCDONALD, R., LERMAN, K., PEREIRA, F. (2006). Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Proceedings of the Tenth Conference on Computational Natural Laguage Learning*, 216-220.

MEL'CUK, I. (1988). *Dependency Syntax: Theory and Practice.*State University of New York Press.

MEL'CUK, I. (2011). Dependency in language. *Proceedings of DepLing 2011*, 1-16.

MERKLER, D., AGIC, Ž., AGIC, A. (2013). Babel Treebank of Public Messages in Croatian. *Procedia-Social and Behavioral Sciences* 95, 490-497.

MILETIC, A. (2013). Annotation semi-automatique en parties du discours d'un corpus littéraire serbe. *Mémoire de Master, Université Charles de Gaulle Lille 3*.

NIVRE, J., HALL, J., KUBLER, S., MCDONALD, R., NILSSON, J., RIEDEL, S., YURET, D. (2007). The CoNLL 2007 shared task on dependency parsing. *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, 915-932.

NIVRE, J., HALL, J., NILSSON, J. (2006). MaltParser A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of LREC 2006* 6, 2216-2219.

PETROV, S., BARRETT, L., THIBAUX, R., KLEIN, D. (2006). Learning accurate, compact, and interpretable tree annotation. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 433-440.

RIEGEL, M. (1981). Verbes essentiellement ou occasionnellement attributifs. *L'information grammaticale* 10, 23-27.

RIEGEL, M., PELLAT, J.-C., RIOUL, R. (1999) *Grammaire méthodique du français*, 5e éd. mise à jour. Paris : Presses Universitaires de France.

SECUJSKI , M. (2009). Automatic part-of-speech tagging of texts in the Serbian language. *Thèse de doctorat, Faculté des Sciences Techniques de Novi Sad*.

SGALL, P., HAJICOVA, E., PANEVOVA, J. (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspect.* Dordrecht : Kluwer.

STANOJCIC, Ž., POPOVIC, L. (2011). *Gramatika srpskog jezika*. 14 éd. Beograd: Zavod za udžbenike.

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

TADIC, M. (2000). Building the Croatian-English parallel corpus. *Proceedings of LREC 2000*, 523-530.

TADIC, M. (2007). Building the Croatian Dependency Treebank: the initial stages. *Suvremena Lingvistika* 33(63), 85-92.

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

UTVIC, M. (2011). Annotating the Corpus of contemporary Serbian. *Proceedings of INFOtheca '12*, 36-47.

VITAS, D., KRSTEV, C., OBRADOVIC I., POPOVIC, LJ., PAVLOVIC-LAZETIC, G. (2012). The Serbian Language in the Digital Age. *META-NET White Paper Series*. Springer. http://www.meta-net.eu/whitepapers

МЕЛЬЧУК, И. А. (1995). *Русский язык в модели «Смысл ↔ Текст».* Wiener Slawistischer Almanach/ Škola «Jazyki russkoj kul´tury»: Vienne/Moscou.