Parsing des textes journalistiques en serbe à l'aide du logiciel Talismane

Dusica Terzic

Parsing des textes journalistiques en serbe par le logiciel *Talismane*

Dusica Terzic[1]
(1) Faculté de philologie, Université de Belgrade, Studentski trg 3, 11 000 Belgrade, Serbie
duterzic@gmail.com

RÉSUMÉ

Cet article présente la création d'un treebank journalistique serbe, *ParCoJour*. Il est composé de 30K tokens et doté de trois couches d'annotation : étiquetage morphosyntaxique, lemmatisation et annotation syntaxique. Une fois construit, *ParCoJour* a été utilisé dans trois expériences afin d'évaluer l'impact du domaine textuel sur le parsing du serbe en comparant les performances de *Talismane*, un système par apprentissage automatique, sur deux types de corpus, journalistique et littéraire : 1) parsing du corpus journalistique avec un modèle entraîné sur le corpus journalistique ; 2) parsing du corpus journalistique avec un modèle entraîné sur le corpus littéraire ; 3) parsing du corpus littéraire avec un modèle entraîné sur le corpus journalistique. Les résultats sont comparés à ceux où les deux corpus relevaient du domaine littéraire. Le changement de domaine textuel dans la deuxième et la troisième expérience entraîne une baisse des performances, mais les résultats de parsing restent satisfaisants.

ABSTRACT

Parsing of newspaper texts in Serbian using *Talismane*

This article presents the creation of a newspaper treebank for Serbian. The corpus contains 30K tokens and it is tagged, lemmatised and parsed. After its creation, the corpus was used in three experiments the goal of which was to evaluate the impact of text domain on the automatic processing of Serbian by comparing the performance of *Talismane*, a supervised machine learning system, on two types of corpora, journalistic and literary: 1) parsing of the newspaper corpus with the model trained on the same corpus; 2) parsing of the newspaper corpus with the model trained on the literary corpus; 3) parsing of the literary corpus with the model trained on the newspaper corpus. The results are compared with the results of the experiment in which both corpora were literary. The change of domain in the second and third experiment resulted in a drop of the parsing results, but they remained satisfactory.

MOTS-CLÉS : Parsing, corpus d'entraînement, serbe, adaptation de domaine.

KEYWORDS: Parsing, training corpus, Serbian, domain adaptation.

1 Introduction

Le serbe, la langue de notre étude, fait partie des langues slaves du sud. Il est parlé en Serbie et dans les pays voisines par presque 9 millions de personnes (Fagard et al., 2017). C'est une langue à ordre de mots flexible et à morphologie flexionnelle riche. Il figure dans le groupe des langues

relativement peu dotées en ressources du TAL. Certaines de ces ressources sont dotées de trois couches d'annotation : l'annotation morphosyntaxique, la lemmatisation, l'annotation syntaxique. Néanmoins, les ressources analysées au niveau syntaxique sont les moins nombreuses. Pour contribuer au développement des ressources serbes dotées d'annotation syntaxique, nous décidons de mettre accent sur le parsing dans notre recherche. L'état de l'art en parsing du serbe est atteint avec *Talismane,* un logiciel statistique à base de transitions qui traite les arbres en dépendances. Pour cette raison, dans cet article, l'accent est mis sur les systèmes par apprentissage automatique bien que d'autres techniques de traitement automatique des langues naturelles existent, et qu'elles donnent parfois de bons résultats en étiquetage morphosyntaxique et en lemmatisation du serbe (voir Gesmundo, Samardžić, 2012).

Les ressources existantes et librement disponibles relèvent d'un domaine textuel restreint, à savoir littéraire (*1984* (Krstev et al., 2004) ; *ParCoTrain-Synt* (Miletic, 2018)) et journalistique (*SETimes.SR* (Samardžić et al., 2017)). Ce manque de diversité de corpus disponibles affecte la qualité des résultats obtenus dans le cas d'apprentissage supervisé. Afin de contribuer à la diversification des corpus disponibles, nous avons défini notre premier objectif : constituer une nouvelle ressource du TAL pour le serbe, enrichie d'annotations linguistiques à plusieurs niveaux et provenant d'un domaine textuel sous-représenté pour le moment, à savoir le domaine journalistique. Or, la création des ressources du TAL exige des investissements humains et matériels considérables. Pour faciliter la création de la nouvelle ressource, nous nous sommes servi des ressources existantes. Le point de départ dans la constitution du corpus *ParCoJour* est le treebank littéraire *ParCoTrain-Synt* (Miletic, 2018). *ParCoTrain-Synt* est doté de trois couches d'annotation, il est librement disponible et l'état de l'art de parsing du serbe est atteint en l'utilisant en tant que corpus d'entraînement (LAS[1]=87,48, UAS[2]=91,22). Nous exploitons donc les modèles de traitement entraînés sur ce corpus pour effectuer un prétraitement de nos données, suivi d'une correction manuelle.

Notre deuxième objectif est d'effectuer une première évaluation des effets des domaines textuels différents sur les résultats du parsing du serbe. Malgré les performances élevées qu'il permet d'atteindre en parsing, *ParCoTrain-Synt* reste limité à un seul domaine textuel – la littérature. Cela remet en question son utilité lors du traitement de textes provenant d'autres domaines, comme des textes journalistiques ou scientifiques. En effet, des expériences en adaptation de domaine montrent que les modèles de parsing entraînés sur des corpus mono-genre sont particulièrement peu robustes lors du passage à un nouveau type de texte (Nivre et al. 2007a, Gildea 2001). Les expériences d'Agić et al. (2013) sur l'étiquetage et la lemmatisation du serbe et du croate montrent des pertes plus importantes lors du changement de domaine (journalistique *vs* encyclopédique) que lors du changement de langue (croate *vs* serbe). Le travail d'Agić & Ljubešić (2015) sur le parsing de ces deux langues aboutit aux mêmes observations.

Il faut cependant noter que dans toutes les expériences citées ci-dessus le corpus d'entraînement relevait du domaine journalistique. Le choix de ce domaine pour le nouveau corpus du serbe est, entre autres, motivé par le but de mener les expériences dans le sens inverse. Dans nos expériences, nous effectuons donc des évaluations dans les deux sens (entraînement sur le littéraire et évaluation sur le journalistique et *vice versa*). Notre intuition linguistique nous mène à considérer les textes littéraires comme syntaxiquement plus complexes que les textes journalistiques. Par conséquent, nous nous demandons si la baisse lors de l'adaptation de domaine serait plus importante dans le cas où le corpus d'entraînement serait du domaine littéraire et le corpus d'évaluation du domaine

1 Le score de LAS (score de rattachement labellisé) représente le pourcentage des tokens pour lesquels le parser a bien déterminé le gouverneur aussi bien que la fonction.

2 Le score de UAS (score de rattachement non labellisé) équivaut au pourcentage des gouverneurs bien identifiés. Pour calculer le score de UAS, l'identification de la fonction n'est pas prise en compte.

journalistique[3]. Autrement dit, l'hypothèse de notre recherche serait la suivante : un système d'apprentissage automatique entraîné sur un corpus littéraire atteint de meilleurs scores dans le parsing des textes journalistiques que l'inverse.

Dans la suite de l'article, nous présentons l'état de l'art du TAL pour le serbe dans la partie 2. La méthodologie de la constitution du nouveau corpus ainsi que le corpus créé dans le cadre de cette recherche sont décrits dans la section 3. Les expériences de l'adaptation de domaine que nous avons menées et leurs résultats sont détaillés dans la section 4.

2 Le serbe en traitement automatique des langues

Pour justifier notre choix du treebank de départ, nous présentons d'abord les ressources qui existent déjà pour le traitement automatique du serbe. Nous nous focalisons sur la disponibilité des ressources. Dans un deuxième temps, nous présentons l'état de l'art dans le parsing du serbe pour avoir un point de référence dans l'analyse des résultats de nos expériences.

2.1 Disponibilité des ressources

Ressource	Type de ressource	Annotation	Taille (en tokens)
corpus *NETK* et *SrpKor2003*	corpus de la langue serbe contemporaine	non annotés	22 millions
SrpKor2013	corpus du serbe contemporain de domaines diversifiés	bibliographique et morphologique	122 millions
SrpLemKor[4]	sous-corpus journalistique du corpus *SrpKor2013*	lemmatisation	3,7 millions
Corpus du serbo-croate de Henning Moerk de l'Université d'Aarhus	corpus des textes de prose publiés entre 1955 et 1990		
Crise éléctorale de l'année 2000	corpus journalistique		
	corpus des proverbes de Vuk Karadžić		

TABLE 1 : Ressources du TAL pour le serbe – indisponibles

Parmi les ressources auxquelles nous n'avons pas réussi à accéder figurent les bases textuelles citées sur le site[5] de la Faculté des mathématiques de l'Université de Belgrade. Elles sont énumérées dans le tableau 1 et accompagnées d'informations disponibles sur le site. Nous n'avons pas trouvé le lien pour télécharger les ressources, même si certaines sont censées être disponibles sous la licence

3 Il faut mentionner que des techniques d'adaptation des outils à un nouveau domaine autre que l'entraînement d'un outil sur un corpus des textes de ce domaine existent. Nous avons cependant choisi la technique décrite pour rendre les résultats de nos expériences comparables aux résultats des expériences antérieures et pour essayer de répondre à la question posée.

4 D'après les informations sur le site, *SrpLemKor* peut être téléchargé depuis l'adresse suivante : http://www.korpus.matf.bg.ac.rs/SrpLemKor/. Dernier accès : le 6 janvier 2019. Pourtant, nous n'avons pas réussi à le faire.

5 http://korpus.matf.bg.ac.rs/prezentacija/korpusi.html. Dernier accès : le 6 janvier 2019.

CC_BY-NC. D'après les informations sur le site, il est possible d'effectuer la recherche en ligne dans certains corpus, mais il faut avoir le compte utilisateur que nous n'avons pas réussi à créer.

Ressource	Type	Annotation	Taille
Lexiques			
lexique de MULTEXT-East (Krstev et al., 2004)	NA	NA	16 907 formes fléchies
wikimorph-sr (Miletic, 2017)	NA	NA	1 226 638 formes fléchies
Corpus			
1984 (Krstev et al., 2004) [6]	traduction du roman *1984*	morphosyntaxique	environ 100 000 mots
ParCoTrain (Miletic, 2013)	corpus de domaines divers	morphosyntaxique et lemmatisation	environ 150 000 tokens
ParCoTrain-Synt, (Miletic, 2018)	treebank littéraire	morphosyntaxique, syntaxique, lemmatisation	la première version de 81 000 tokens
SETimes.SR[7] (Samardžić et al., 2017)	treebank journalistique	morphosyntaxique, syntaxique, lemmatisation[8]	86 726 tokens

TABLE 2 : Corpus du TAL pour le serbe – disponibles

Le tableau 2 regroupe les ressources serbes du TAL que nous avons réussi à télécharger. Les premières ressources du TAL pour le serbe ont été développées à l'issu du projet MULTEXT-East (Dimitrova et al., 1998). Il s'agit du corpus *1984* et du lexique composé à partir des mots de ce corpus (Krstev et al., 2004) qui contient la traduction serbe du roman *1984* de George Orwell. Toutefois, dans le développement des ressources du TAL pour le serbe, le projet *ParCoLab*[9] (Miletic et al., 2017) représente un point crucial surtout concernant la disponibilité. *ParCoLab* est un corpus parallèle de textes en serbe, anglais et français. Chaque langue du corpus est représentée aussi bien à travers des traductions que des originaux. Les textes qui y figurent proviennent majoritairement du domaine littéraire. Néanmoins, les domaines se diversifient au fur et à mesure que le corpus s'enrichit. Il est désormais composé de 11 359 761 mots et il est accompagné de ressources pour le traitement automatique du serbe.

En plus des ressources développées directement pour le serbe, il est important de mentionner celles développées pour le croate et, dans la plupart des cas, adaptées au serbe par une équipe de chercheurs de l'Université de Zagreb (Agić et al., 2013 ; Ljubešić et al., 2016 ; Agić, Ljubešić, 2015). Ils énumèrent les corpus suivants : *Croatia Weekly*, corpus parallèle journalistique anglo-croate (100 000 tokens) ; *hrWaC* (1,9 milliard de tokens) et *srWaC* (894 millions de tokens), corpus collectés sur Internet ; *hr500k* (500 000 tokens), corpus provenant de domaines diversifiés ; *Setimes.HR* (90 000 tokens), corpus journalistique. Certains de ces corpus sont dotés d'annotation morphosyntaxique, de lemmatisation et d'annotation syntaxique. Quant aux lexiques

6 CC BY-NC-SA 4.0 (https://creativecommons.org/licenses/by-nc-sa/4.0/

7 Téléchargeable depuis l'adresse du projet Universal Dependencies : https://universaldependencies.org/. Dernier accès : le 18 février 2019.

8 Les annotations ont été effectuées conformément à la méthode du projet Universal Dependencies (McDonald et al., 2013).

9 http://parcolab.univ-tlse2.fr/en/about/resources/ Dernier accès : le 15 octobre 2018.

morphologiques, ils présentent *Hrvatski morfološki leksikon* (lexique morphologique croate) de 3,8 millions de formes fléchies et les lexiques *hrLex* et *srLex* construits à partir d'un lexique serbo-croato-bosniaque d'Apertium[10], un système de traduction automatique de langues proches. Toutes ces ressources développées par l'équipe de l'Université de Zagreb sont librement disponibles[11].

Ce survol bref des ressources pour le traitement automatique du serbe montre qu'elles ne sont pas nombreuses. La plupart de ces ressources sont citées dans la littérature et énumérées sur les sites, mais elles ne sont pas téléchargeable. Autrement dit, bien qu'elles soient créées, elles ne peuvent pas être effectivement utilisées. D'autres défauts des ressources serbes du TAL existent. Par exemple, la plupart des corpus serbes sont des bases textuelles non annotées. Par conséquent, ils ne peuvent pas servir de corpus d'apprentissage pour des systèmes d'apprentissage automatique. En revanche, des corpus serbes qui sont annotés ne contiennent pas de textes de domaines diversifiés et se limitent le plus souvent au domaine littéraire. En conclusion, il existe un grand besoin de développement de nouvelles ressources annotées, provenant d'autres domaines textuels, afin de favoriser le progrès du TAL pour le serbe.

2.2 État de l'art

	Étiquetage morphosyntaxique fin (%)		**Lemmatisation (%)**	
	Croate	Serbe	Croate	Serbe
Agić et al. (2013), corpus d'entraînement - *SETimes*, lemmatiseur - *CST*, étiqueteur - *HunPOS*	87,72	85,56	97,78	96,30
Ljubešić et al. (2016), corpus d'entraînement - *hr500k*, *CRF*, lexique - hrLex	92,53	92,33	NA	NA
Gesmundo & Samardžić (2012), corpus d'entraînement *1984*, *BTagger*	NA	86,65	NA	97,72
Miletic (2018), corpus d'entraînement *ParCoTrain-Synt*, lemmatiseur – *CST*, étiqueteur – *HunPOS*, lexique *ParCoLex*	NA	85,00	NA	96,50

TABLE 3 : État de l'art de l'étiquetage morphosyntaxique fin et de la lemmatisation du serbe

Le tableau 3 contient les taux d'exactitude les plus élevés atteints dans la lemmatisation et dans l'étiquetage du serbe et du croate que nous avons trouvés dans la littérature. Agić et al. (2013)[12] effectuent les expériences qui montrent que *BTagger* est beaucoup plus chronophage que *HunPOS* (Halácsy et al., 2007) en tant qu'étiqueteur et que *CST* (Jongejan, Dalianis, 2009) en tant que lemmatiseur même si l'étiqueteur et lemmatiseur *BTagger* donne les résultats les plus élevés (Gesmundo, Samardžić, 2012). L'état de l'art actuel en parsing du serbe et du croate est présenté dans le tableau 4. Étant donné que les schémas d'annotation dans le cadre de ces deux travaux diffèrent, les résultats ne sont pas directement comparables. Néanmoins, ce sont les résultats de Miletic (2018) qui dépassent largement l'état de l'art antérieur.

10 L'interface se trouve à l'adresse suivante https://www.apertium.org/index.eng.html?dir=cat-arg#translation. Dernier accès : le 22 septembre 2018.

11 Depuis les adresses suivantes : https://github.com/ffnlp/sethr et https://www.clarin.si/repository/xmlui/. Dernier accès : le 18 février 2019.

12 L'apprentissage sur un même corpus prend 6 heures avec *BTagger*, alors qu'avec *HunPOS* il ne prend qu'une seconde.

	Croate		Serbe	
	UAS	LAS	UAS	LAS
Agić & Ljubešić (2015), parser *Mate*, corpus *Setimes.hr*	86,9	81,5	86,0	81,5
Miletic (2018), parser *Talismane*, corpus *ParCoTrain-Synt*	NA	NA	91,22	87,48

TABLE 4 : État de l'art dans le parsing du serbe et du croate

Une des conclusions de Nivre et al. (2007a) est que les schémas d'annotation et les principes adoptés lors de l'adaptation de domaine doivent être pareils si l'on veut profiter de cette démarche. Nous avons ainsi décidé d'adopter la méthodologie de Miletic (2018) dans la création de notre ressource. Nous nous servons donc de *ParCoTrain-Synt* pour faciliter ce processus et pour rendre les résultats de nos expériences comparables à ceux de Miletic (2018).

3 Création d'un treebank littéraire serbe, *ParCoJour*

Pour réaliser notre premier objectif, nous avons procédé à la création d'un treebank journalistique, *ParCoJour*. Il est composé de 37 articles journalistiques de 30 566 tokens au total. Les articles ont été tirés de deux journaux nationaux, *Danas*[13] (quotidien) et *NIN*[14] (hebdomadaire). Pour essayer d'éviter tout biais, nous avons choisi les textes de manière aléatoire. La longueur des textes varie de 251 à 15 291 tokens. Les articles de *NIN* couvrent la période de 2003 à 2017 et ceux de *Danas* de 2007 à 2017. Les articles ont été écrits par 28 auteurs différents. Après les avoir collectés, nous avons procédé à l'annotation en suivant la méthode de Miletic (2018). Nous présentons d'abord cette méthode. Ensuite, nous présentons le travail pratique effectué.

3.1 Méthodologie de création de *ParCoJour*

La méthode globale préconisée par Miletic (2018) est la suivante : nous effectuons l'étiquetage morphosyntaxique détaillé (incluant des traits morphosyntaxiques fins, comme le cas, le nombre, le genre, etc.), la lemmatisation et le parsing. Ces niveaux de traitement sont effectués en cascade. Pour chaque niveau, nous nous servons des outils entraînés dans le cadre de (Miletic 2018) afin d'effectuer une préannotation automatique de nos données. Plus précisément, nous utilisons *HunPOS* (Halácsy et al., 2007), pour l'étiquetage morphosyntaxique, *CST* (Jongejan, Dalianis, 2009) pour la lemmatisation et *Talismane* (Urieli, 2013) pour le parsing. Nous adoptons également les mêmes jeux d'étiquettes et schémas d'annotation. La préannotation automatique est ensuite corrigée manuellement. Dans la correction manuelle des annotations automatiques, nous avons suivi les instructions des trois guides d'annotation qui figurent dans les annexes de Miletic (2018). Il s'agit de *Guide d'annotation morphosyntaxique, Guide de lemmatisation* et *Guide d'annotation syntaxique.* Lorsque le *Guide de lemmatisation* n'offrait pas la solution aux problèmes rencontrés, nous avons consulté le dictionnaire de référence de (Miletic 2018), à savoir *Srpski elektronski rečnik* de Milorad Simić. Cette approche de préannotation automatique qui est ensuite corrigée manuellement s'est montrée comme moins chronophage que l'annotation manuelle (Miletic et al., 2016 ; Fort, Sagot, 2010).

13 https://www.danas.rs/ Dernier accès : le 17 janvier 2019.
14 http://www.nin.co.rs/ Dernier accès : le 17 janvier 2019.

Le choix d'effectuer une analyse morphosyntaxique détaillée est motivé par les observations que les scores de parsing d'un même texte en serbe aussi bien qu'en croate sont meilleurs si ce texte est doté au préalable d'une annotation morphosyntaxique détaillée (voir Agić, Ljubešić, 2015 ; Miletic, 2018). Cela est dû au fait que l'ordre des constituants en serbe est flexible et que ce sont les indices morphosyntaxiques qui nous informent sur les fonctions syntaxiques.

En ce qui concerne le parsing, nous adoptons l'analyse syntaxique en dépendances. *Talismane* traite les arbres en dépendances et le treebank que nous utilisons dans le processus est composé d'arbres syntaxiques de ce type. Ce choix a été motivé par la nature du serbe, qui est une langue à ordre de mots flexible.

3.2 Travail pratique de création du corpus *ParCoJour*

Le corpus textuel avant annotation a été encodé au format XML. Chaque élément <text> contient un article du corpus. Différentes métadonnées sont stockées au niveau de cet élément en tant qu'attributs : l'id (le numéro de l'article dans le corpus), la source (soit « nin » soit « danas »), la date de publication de l'article, l'auteur et l'adresse URL à partir de laquelle il a été récupéré. Afin d'être exploitable dans les tâches d'étiquetage morphosyntaxique, de lemmatisation et d'annotation syntaxique, le corpus a dû être transformé en d'autres formats, accessibles aux outils automatiques sélectionnés. Les formats exploités par des outils que nous avons utilisés sont présenté dans le tableau 5.

Outil	*HunPOS* (Halácsy et al., 2007)	*CST* (Jongejan, Dalianis, 2009)	*Talismane* (Urieli, 2013)
Format d'entrée	document XML segmenté et tokénisé	[token]\t[tag]	CoNLL-X
Format de sortie	[token]\t[tag]	[token]\t[tag]\t[lemme]	CoNLL-X
Format de correction manuelle	CSV	CSV	*standoff* de *Brat* (Stenetorp et al., 2012)
Corpus d'entraînement	*ParCoTrain-Synt*[15] (Miletic, 2018)	Le modèle[16] créé par Miletic (2018)	Le modèle le plus performant de Miletic (2018)[17]

TABLE 5 : Formats exploités par outils choisis

Il n'est pas facile de corriger manuellement les annotations dans les formats exploités par les outils. Tableau 5 contient également les informations sur les formats de corpus sous lesquels la correction manuelle a été effectuée. Pour faciliter la correction de l'annotation morphosyntaxique et de la lemmatisation pour un annotateur humain, nous avons converti la sortie de *HunPOS* et de *CST* en format CSV. Figure 2 illustre une phrase lemmatisée sous format CSV. Le token se trouve dans la première colonne, l'étiquette morphosyntaxique dans la deuxième et le lemme dans la troisième.

Seule l'auteure a participé dans la correction manuelle des préannotations automatiques. La correction manuelle de l'annotation syntaxique automatique a cependant été vérifiée par l'auteure de *ParCoTrain-Synt.* La vitesse moyenne de correction de l'annotation morphosyntaxique était de 430

15 Il est disponible sous la licence Creative Commons BY-NC-SA 3.0 et il peut être téléchargé à partir de l'adresse suivante : http://parcolab.univ-tlse2.fr/en/about/resources/. Dernier accès : le 15 octobre 2018.

16 Librement disponible à l'adresse : https://github.com/aleksandra-miletic/serbian-nlp-resources/tree/master/Models/Lemmatisation. Dernier accès : le 15 octobre 2015.

17 Ce modèle peut être téléchargé à partir de l'adresse suivante : https://github.com/aleksandra-miletic/serbian-nlp-resources/tree/master/Models/Parsing/*Talismane*. Dernier accès : le 15 octobre 2018.

tokens par heure, ce qui est moins que la vitesse de l'annotateur expérimenté dans le cadre de Miletic (2018) (800 tokens par heure), mais plus que la vitesse de l'annotateur novice (325 tokens par heure). La comparaison de notre vitesse avec celle de l'annotation manuelle sans pré-annotation automatique d'un annotateur novice dans le cadre de Miletic (2018) (80 tokens) justifie notre décision d'effectuer la correction d'une annotation automatique au lieu d'annoter tout le corpus manuellement. Notre vitesse moyenne de la correction manuelle de lemmatisation était de 1500 tokens par heure, ce qui est de nouveau plus vite que l'annotation manuelle sans pré-annotation automatique dans le cadre de Miletic (2018) (825 tokens par heure).

52	Lideri	N_com_nom_pl_m	Lider
53	EU	N_prop_gen_sg_f	eu
54	pristali	V_main_partact_-_pl_m_-	pristati
55	na	Prep	na
56	kompromis	N_com_acc_sg_m	kompromis
57	,	Z	,
58	ali	C_coord	ali
59	ne	Part	ne
60	odustaju	V_main_pres_3_pl_-_-	odustajati
61	od	Prep	od
62	sopstvenih	A_qual_gen_pl_m_pos	sopstven
63	vojnih	A_qual_gen_pl_m_pos	vojni
64	planova	N_com_gen_pl_m	plan
65			

FIGURE 1: Extrait de *ParCoJour* étiqueté et lemmatisé en format CSV

Pour corriger manuellement le parsing, nous nous sommes servi de l'outil *Brat*[18] de Stenetorp et al. (2012). L'interface graphique de cet outil (voir Figure 2) facilite la correction de l'annotation syntaxique étant donné que la représentation graphique dans cet outil permet de visualiser l'arbre. Les phrases sont présentées linéairement avec l'étiquette morphosyntaxique détaillée en dessous de chaque token.

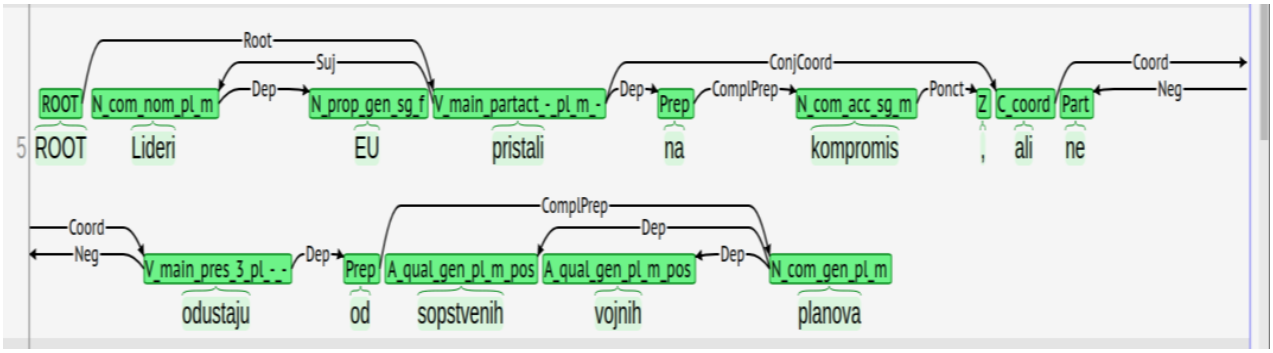

FIGURE 2: Interface graphique de *Brat*

18 Il peut être téléchargé à partir de l'adresse suivante : http://brat.nlplab.org/. Dernier accès : le 15 octobre 2018.

Ces trois processus effectués, nous avons atteint notre premier but : nous avons obtenu un corpus de textes journalistiques, *ParCoJour*, composé de 30 566 tokens et doté d'annotations syntaxique, morphosyntaxique et de lemmatisation. La suite de l'article porte sur notre objectif méthodologique. La méthode adoptée dans les expériences, l'application de cette méthode et les réponses à nos questions qui en découlent sont détaillées dans la section suivante.

4 Adaptation au domaine

Les corpus à notre disposition (*ParCoJour* et *ParCoTrain-Synt*) relevant des domaines restreints (journalistique et littéraire respectivement), nous pouvons les utiliser dans des expériences dont le but est d'évaluer l'impact des domaines textuels des corpus d'entraînement et d'évaluation sur le parsing. Pour ce faire, nous avons mené trois expériences. Nous les présentons dans cette section aussi bien que les résultats de ces expériences. Ils sont commentés et confrontés aux scores des deux études antérieures du parsing du serbe que nous avons présentées (Miletic, 2018 ; Agić, Ljubešić, 2015).

4.1 Expériences d'adaptation au domaine avec *ParCoJour*

Pour effectuer des évaluations sur notre corpus, nous l'avons divisé en trois parties : *dev* (1 596 tokens), *train* (30 063 tokens), *test* (3 036 tokens). La section *test* est parsée automatiquement et comparée à l'annotation *gold* lors de l'évaluation. Lorsque nous parlons des corpus *ParCoJour* et *ParCoTrain-Synt* dans l'analyse des résultats des expériences, nous nous référons aux sections test des corpus respectifs.

Talismane est doté d'un sytème d'évaluation et nous l'exploitons dans nos expériences. Lors de l'évaluation, *Talismane* crée deux fichiers, un fichier TXT et un fichier CSV. Le fichier TXT contient l'annotation que *Talismane* a produite sous le format conforme aux propositions des campagnes CoNLL-X. Le fichier CSV contient les résultats quantitatifs de la comparaison sous forme d'une matrice de confusion accompagnée de scores standard de parsing. Il s'agit du nombre total d'occurrences d'une étiquette dans le corpus *test*, ainsi que la précision, le rappel et la f-mesure liés à cette étiquette. Ce fichier contient également le score LAS, alors que le score UAS doit être calculé ultérieurement.

Pour comparer les performances de *Talismane* lorsqu'il est entraîné sur un corpus littéraire à ses performances lorsque le corpus d'entraînement est un corpus journalistique nous avons effectué trois expériences : 1) parsing de *ParCoJour* avec le modèle entraîné sur *ParCoJour* ; 2) parsing de *ParCoJour* avec le modèle entraîné sur *ParCoTrain-Synt* ; 3) parsing de *ParCoTrain-Synt* avec le modèle entraîné sur *ParCoJour*. Dans chaque expérience, l'évaluation a été effectuée sur la section *test* des corpus correspondant. Les résultats obtenus dans les expériences menées sont ensuite comparés aux résultats des expériences de Miletic (2018) où le modèle entraîné sur *ParCoTrain-Synt* a été évalué sur la section *test* du même corpus.

La configuration optimale proposée par Miletic (2018) comprend le paramètre *cutoff* de 3 et le paramètre *beam width* de 5. L'augmentation de la valeur de *beam width* entraîne l'augmentation du temps nécessaire d'effectuer l'apprentissage. Ici, nous nous éloignons de la configuration optimale : le temps étant un facteur important dans notre étude, nous effectuons nos expériences en utilisant le *beam width* de 1. En conséquence, nous comparons nos résultats avec les résultats obtenus par Miletic (2018) avec la même valeur de ce paramètre (voir Table 7). La comparaison des résultats dans les conditions optimales reste une piste à suivre dans les travaux futurs.

Miletic (2018) évalue aussi l'effet du type d'annotation morphosyntaxique qui précède le parsing. Les résultats montrent qu'une annotation manuelle donne de meilleurs résultats qu'une annotation automatique non corrigée. Le but de notre recherche étant l'évaluation de l'effet du genre sur le parsing, nous n'analysons pas les effets du type d'annotation qui précède, mais nous utilisons les conditions d'apprentissage concernant l'étiquetage qui se sont montrées comme les meilleures dans le cadre de Miletic (2018). L'analyse des effets de type de l'étiquetage morphosyntaxique sur le parsing du corpus journalistique *ParCoJour* peut être envisagée dans des travaux ultérieurs.

4.2 Résultats des expériences d'adaptation au domaine

Les tableaux 6 et 7 contiennent les résultats quantitatifs globaux des expériences menées dans le cadre de notre travail. Dans le tableau 6 figurent les résultats des deux premières expériences où le corpus d'évaluation était la section test du corpus *ParCoJour* et le corpus d'entraînement changeait. Le tableau 7 donne les résultats du parsing avec le corpus *ParCoTrain-Synt* comme le corpus d'évaluation. Pour le cas où le modèle entraîné sur *ParCoTrain-Synt* est évalué sur le corpus *test* de *ParCoTrain-Synt* nous reprenons les résultats de Miletic (2018) avec le paramètre *beam width* de 1 et où l'étiquetage moprhosyntaxique était manuel (voir Table 7).

		Corpus d'évaluation *ParCoJour*	
	Corpus d'entraînement	UAS	LAS
Expérience 1	*ParCoJour*	89,72	85,08
Expérience 2	*ParCoTrain-Synt*[19]	87,15	80,86
	différentiel de performance dans l'expérience avec changement de domaine de corpus d'entraînement	-2,57	-4,22

TABLE 6 : Résultats du parsing dans les expériences avec le corpus *ParCoJour* comme corpus d'évaluation

		Corpus d'évaluation *ParCoTrain-Synt*	
	Corpus d'entraînement	UAS	LAS
Expérience de Miletic (2018)	*ParCoTrain-Synt*	91,19	87,83
Expérience 3	*ParCoJour*	85,75	81,24
	différentiel de performances dans l'expérience avec changement de domaine de corpus d'entraînement	-5,44	-6,59

TABLE 7 : Résultats du parsing dans les expériences avec le corpus *ParCoTrain-Synt* comme corpus d'évaluation

Dans les cas où le domaine des corpus d'entraînement et d'évaluation ne change pas (expérience 1 et expérience de Miletic (2018)), les meilleurs scores sont atteints dans le parsing du corpus littéraire

19 Résultats du parsing de *ParCoTrain-Synt* si beam width = 1 (Miletic, 2018, p. 215).

(voir Table 6 et 7). Étant donné que les conditions d'apprentissage, le schéma d'annotation et le jeu d'étiquettes ne diffèrent pas, nous supposons que les scores moins bons dans le parsing des textes journalistiques d'après le modèle entraîné sur le corpus journalistique viennent de la différence des tailles des deux corpus analysés. En effet, le corpus *ParCoJour* contient environ 30 000 tokens, alors que la version de *ParCoTrain-Synt* que nous utilisons ici en contient environ 81 000. Pour tester cette hypothèse, nous proposons d'enrichir *ParCoJour* de nouveaux textes journalistiques ou de limiter la taille de *ParCoTrain-Synt* à 30 000 tokens dans les travaux à suivre.

Pour tester les résultats du changement de domaine, nous confrontons ensuite les résultats de la deuxième et de la troisième expérience. Les deux expériences confirment une chute des résultats lors de ce processus. Les dernières lignes des tableaux 6 et 7 contiennent le différentiel des performances dans ces deux expériences. En analysant ces baisses, nous essayons de déterminer s'il est plus utile d'avoir à disposition un corpus d'entraînement littéraire ou un corpus journalistique si l'on doit traiter les deux genres textuels. Les résultats ont baissé de 2,57 en UAS et de 4,22 en LAS si le corpus d'évaluation était *ParCoJour* et de 5,44 en UAS et de 6,59 en LAS si le corpus d'évaluation était *ParCoTrain-Synt* (voir Table 6 et 7), ce qui est moins que ce qui a été observé auparavant dans le cadre de (Nivre et al., 2007a). Néanmoins, les résultats ne sont pas directement comparables étant donné que les schémas d'annotation et les tailles de corpus diffèrent.

Dans le cadre de notre recherche, la baisse des scores est plus importante si le corpus à parser est le corpus littéraire. Cette observation correspond aux résultats de Gildea (2001). Un parser est, donc, plus performant sur les textes journalistiques que sur les textes littéraires. Nous pouvons conclure que le corpus d'entraînement littéraire est plus utile dans le parsing des textes journalistique que le corpus journalistique ne l'est dans le parsing des textes littéraires. Il faut, toutefois, prendre en considération les tailles différentes des corpus d'entraînement dans nos expériences. D'autre part, le score de UAS est meilleur de 1,4 dans l'expérience où le corpus d'entraînement était le corpus *ParCoTrain-Synt* et le corpus d'évaluation *ParCoJour*. Le score de LAS est, en revanche, meilleur de 0,38 dans la troisième expérience où le corpus d'entraînement était *ParCoJour* et le corpus d'évaluation *ParCoTrain-Synt*. Le fait que les résultats ne sont pas toujours meilleurs avec un corpus littéraire comme corpus d'entraînement, et cela dans le cas où ce corpus est plus petit, confirment l'importance d'avoir les corpus de genres diversifiés.

Il serait intéressant d'entrer dans une analyse plus profonde des différences linguistiques entre ces deux types de corpus. Cette analyse pourrait expliquer la cause des différences dans les performances d'un parser sur ces deux types de corpus. Faute de place, nous sommes obligée de laisser cette analyse pour un prochain travail.

Dans l'introduction, nous avons présenté les scores du parsing du serbe et du croate dans le cadre du travail d'Agić et Ljubešić (2015). Avant Miletic (2018), c'était Agić & Ljubešić (2015) qui avaient atteint l'état de l'art dans le parsing du serbe. Le score de LAS était de 81,5 et le score de UAS de 86,0. Les scores dans le cadre de notre étude sont inférieurs à ceux de Miletic (2018), mais dépassent les scores d'Agić et Ljubešić (2015). Ceci est vrai même dans les scénarios avec changement de domaine. Il faut, cependant, noter que le schéma d'annotation d'Agić et Ljubešić (2015) diffère du schéma dans le cadre de notre recherche. Par conséquent, les résultats ne sont pas directement comparables. Néanmoins, les scores atteints dans le cadre de notre étude justifient le parti pris dans notre travail de suivre la méthodologie adoptée par Miletic (2018) et illustrent également l'utilité de la diversification des genres dans la constitution des corpus annotés.

5 Conclusion et pistes

Deux grands objectifs de notre travail étaient de créer un treebank journalistique pour le serbe et de faire des expériences en adaptation du domaine sur cette langue. Par le premier objectif, nous avons envisagé de contribuer à la diversification de *ParCoTrain-Synt* (Miletic, 2018), un treebank littéraire pour le serbe. L'objectif méthodologique était censé contribuer à la compréhension du mécanisme d'adaptation de domaine. Le corpus *ParCoJour* sera librement disponible sur le site du projet *ParCoLab.*

Lors de la création de notre treebank, nous nous sommes servi de *ParCoTrain-Synt,* un treebank existant. Le corpus *ParCoJour* que nous avons créé est composé de 37 articles journalistiques de 30 566 tokens au total. *ParCoJour* est doté de l'annotation morphosyntaxique, de la lemmatisation et de l'annotation syntaxique en dépendances suivant les principes et les schémas d'annotation présentés dans (Miletic, 2018).

Nous utilisons le corpus *ParCoTrain-Synt* pour mener des expériences censées donner des résultats qui répondraient à quel point la baisse des performances est due au domain concret des corpus d'entraînement et si la baisse serait importante au cas où le corpus d'entraînement relevait du domaine littéraire. Les limites de ce travail ne nous permettant pas d'analyser le traitement à tous les niveaux possibles, nous avons décidé de mener les expériences concernant le parsing.

Dans le parsing du corpus journalistique, *Talismane*, l'outil que nous avons utilisé, atteint le score de 85,08 en LAS et de 89,72 en UAS si entraîné sur le corpus du même domaine. Ces scores sont inférieurs à ceux de Miletic (2018), mais dépassent les scores d'Agić et Ljubešić (2015). Ceci est vrai même pour les scores dans les scénarios avec changement de domaine. Même si le schéma d'annotation d'Agić et Ljubešić (2015) diffère du schéma dans le cadre de notre recherche et même si les résultats ne sont pas directement comparables, les scores atteints dans le cadre de notre étude justifient le parti pris dans notre travail de suivre la méthodologie adoptée par Miletic (2018). Ils confirment aussi que l'utilisation d'un corpus littéraire, quoique du domaine restreint, donne des résultats satisfaisants dans le parsing des textes journalistiques.

D'autre part, le score de UAS est meilleur de 1,4 dans l'expérience où le corpus d'entraînement était le corpus *ParCoTrain-Synt* et le corpus d'évaluation *ParCoJour*. Le score de LAS est, en revanche, meilleur de 0,38 dans l'expérience où le corpus d'entraînement était *ParCoJour* et le corpus d'évaluation *ParCoTrain-Synt.* Le fait que les résultats ne sont pas toujours meilleurs avec un corpus littéraire comme corpus d'entraînement confirment l'importance d'avoir des corpus de genres diversifiés et justifie le choix de créer un corpus journalistique pour le serbe. Ainsi, nous avons contribué à l'enrichissement des ressources du TAL pour la langue serbe.

Le travail fait dans le cadre de cette recherche peut être poursuivi dans le but d'améliorer les résultats obtenus et d'approfondir l'analyse de l'adaptation de domaine. D'abord, la piste la plus importante serait d'augmenter la taille du corpus *ParCoJour* pour rendre les résultats des expériences plus immédiatement comparables à ceux des expériences menées sur *ParCoTrain-Synt.* L'autre possibilité est d'entraîner l'outils sur seulement une partie du corpus *ParCoTrain-Synt* dont la taille serait comparable à celle de *ParCoJour.* Ensuite, il serait bien d'évaluer l'exactitude de la lemmatisation et de l'étiquetage morphosyntaxique lors du changement de domaine. Un autre scénario d'évaluation possible comprendrait un entraînement de l'outil effectué dans des conditions définies comme optimales dans le cadre de Miletic (2018), à savoir, *cutoff* de 3, *beam width* de 5. Une autre piste possible serait d'évaluer l'effet de l'étiquetage automatique au lieu de l'étiquetage *gold* sur le parsing du corpus journalistique. Enfin, il serait également utile d'évaluer les performances de l'outil si entraîné sur un corpus mixte où une moitié serait composée des textes d'un genre, et l'autre moitié des textes de l'autre. Ceci nous permettrait d'évaluer pleinement l'intérêt de la diversification de treebank et l'apport du contenu journalistique au parsing du serbe.

Références

Agić Ž., Ljubešić N., Merkler D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. Actes de *The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, 48-57.

Agić Ž., Ljubešić N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). Actes de *5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, 1-8.

Ljubešić N., Klubička F., Agić Ž, Jazbec I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. Actes de *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Dimitrova L., Erjavec T., Ide N., Kaalep H. J., Petkevič V., Tufiş D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Actes de *COLING-ACL '98.*

Fagard B., Stosic D., Cerruti M. (2017). Within-type variation in Satellite-framed languages: The case of Serbian. *STUF, Akademie Verlag*, 70 (4), 637-660.

Fort K., Sagot B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. *The Fourth ACL Linguistic Annotation Workshop*, 56-63.

Gesmundo A., Samardžić T. (2012). Lemmatising Serbian as category tagging with bidirectional sequence classification. Acte de *The 8th Language Resources and Evaluation Conference (LREC 2012)*, 2103-2106.

Gildea D. (2001). Corpus Variation and Parser Performance. Actes de *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Halácsy P., Kornai A., Oravecz C. (2007). *HunPOS* : an open source trigram tagger. Actes de *The 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 209-212.

Jongejan B., Dalianis H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. Actes de *The Joint Conference of the 47th Annual Meeting of ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.*,145-153.

Krstev C., Vitas D., Erjavec T. (2004). MULTEXT-East resources for Serbian. Actes B de *The 7th international multiconference information society: Language technologies,*108-114.

Kübler S., McDonald R., Nivre J. (2009). *Dependency parsing*. Morgan & Claypool Publishers.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Bartomeu Castelló N., Lee J. (2013). Universal dependency annotation for multilingual parsing. Actes de *The 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92-97.

Miletic A. (2017). Building a morphosyntactic lexicon for Serbian using Wiktionary. Actes de *Sixièmes Journées d'études Toulousaines (JéTou2017)*, 30-34.

Miletic A., Stosic D., Marjanović S. (2017). ParCoLab : A Parallel Corpus for Serbian, French and English. *Text, Speech and Dialogue 2017, LNAI 10415,* 156-164.

Miletic A., Fabre C., Stosic D. (2016). Mise au point d'une méthode d'annotation morphosyntaxique fine du serbe. Actes de *Traitement Automatique des Langues Naturelles(TALN 2016)*, 506-514.

Miletic A. (2013). *Annotation morphosyntaxique semi-automatique d'un corpus littéraire serbe*. Master's thesis. Université Charles de Gaulle - Lille 3.

Miletic A. (2018). *Un treebank pour le serbe : constitution et exploitation.* PhD thesis. Université Toulouse-Jean Jaurès.

Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D. (2007). The CoNLL 2007 shared task on dependency parsing. Actes de *CoNLL shared task session of EMNLP-CoNLL*, 915-932.

Samardžić T., Agić Ž., Starović M., Ljubešić N. (2017). Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages. Actes de *The 6th Workshop on Balto-Slavic Natural Language Processing* 39-44.

Simić M. (2005). *Srpski elektronski rečnik*.

Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J. (2012). BRAT : a web-based tool for NLP-assisted text annotation. Actes de *The Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102-107.

Urieli A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université Toulouse le Mirail - Toulouse II.