De la constitution d'un corpus arboré à l'analyse syntaxique du serbe

Aleksandra Miletic, Cécile Fabre, Dejan Stosic

De la constitution d'un corpus arboré à l'analyse syntaxique du serbe

Aleksandra Miletic* — **Cécile Fabre*** — **Dejan Stosic***

* *CLLE, Université de Toulouse, CNRS, UT2J, France*
aleksandra.miletic@univ-tlse2.fr

RÉSUMÉ. *Cet article retrace une expérience de constitution d'un corpus arboré pour le serbe, conçu dans le but de doter cette langue des instruments nécessaires à l'analyse syntaxique et, plus généralement, de favoriser des recherches plus systématiques aussi bien en TAL (traitement automatique des langues) qu'en linguistique serbe. Au-delà de la description des résultats de ce projet, nous présentons une méthode de confection d'un corpus arboré qui vise à optimiser les ressources, par définition rares, dont on dispose dans le cas d'une langue peu dotée, qu'il s'agisse de moyens matériels (corpus et outils) ou humains. Nous montrons comment tirer au mieux parti de l'existant pour faciliter le travail des annotateurs humains et accélérer l'enrichissement du corpus, tout en garantissant la validité de l'annotation produite. Cette méthode, basée sur des principes transposables à d'autres langues, a vocation à faciliter la création des corpus arborés pour les langues sous-dotées en général.*

ABSTRACT. *In this paper we describe our work on a treebank for Serbian, which aims to provide this language with tools and resources needed for parsing and, more globally, to encourage research on this language both in NLP (natural language processing) and in theoretical linguistics. Beyond the results of this resource-building project, we also provide a description of a treebank-building method that optimizes the limited resources available for an under-resourced language, both from the technical point of view (tools and corpora) and from that of human resources (annotation process). We show how best to take advantage of what is available in order to facilitate the manual work and accelerate the corpus enrichment process, all the while maintaining a high-quality annotation. Being based on language-independent principles, this method should help forward the creation of treebanks for other under-resourced languages.*

MOTS-CLÉS : *corpus arboré, serbe, méthode d'annotation, optimisation, langues à morphologie flexionnelle riche, langues sous-dotées.*

KEYWORDS: *treebank, Serbian, annotation method, optimization, morphologically rich languages, under-resourced languages.*

1. Introduction

Cet article retrace une expérience de constitution d'un corpus arboré pour le serbe, conçue dans le but de doter cette langue des instruments nécessaires à l'analyse syntaxique et, plus généralement, de favoriser des recherches plus systématiques aussi bien en TAL (traitement automatique des langues) qu'en linguistique serbe. Le serbe n'est pas à proprement parler une langue dépourvue de ressources et d'outils (Krstev *et al.*, 2004 ; Vitas et Krstev, 2004 ; Pavlović-Lažetić *et al.*, 2004), et des efforts récents tendent à rattraper le retard pris (Gesmundo et Samardžić, 2012 ; Jakovljević *et al.*, 2014 ; Samardžić *et al.*, 2017). Néanmoins, une partie importante des ressources existantes n'est pas diffusée ou l'est sous des licences restrictives (section 2.2). Par ailleurs, les seules expériences en analyse syntaxique de cette langue sur un corpus serbe (Jakovljević *et al.*, 2014) ont été basées sur un corpus d'apprentissage minimal (7 000 *tokens*) qui n'a pas été distribué. Le corpus arboré présenté ici, publié alors qu'un corpus serbe issu du projet *Universal Dependencies* (dorénavant UD)[1] vient d'être diffusé (Samardžić *et al.*, 2017), est assorti d'un ensemble de ressources (lexiques et modèles d'analyse syntaxique, de lemmatisation et d'étiquetage morphosyntaxique). Quant aux annotations linguistiques du corpus arboré, outre les indications des fonctions syntaxiques et des catégories grammaticales, le corpus est également doté d'une lemmatisation et d'une annotation en traits morphosyntaxiques fins, tels le cas, le genre, le nombre, etc. Cela a été fait dans le souci d'assurer des conditions optimales pour l'analyse syntaxique : en effet, le serbe est une langue à morphologie flexionnelle riche et à ordre des constituants flexible, et de nombreux travaux montrent que l'analyse syntaxique de ces langues est facilitée par l'utilisation de ces deux types d'annotation (Collins *et al.*, 1999 ; Marton *et al.*, 2013). Le corpus, qui contient 101 000 *tokens*, peut être téléchargé à partir de l'adresse suivante : `https://github.com/aleksandra-miletic/serbian-nlp-resources`.

Au-delà de la description des résultats de ce projet, nous présentons dans cet article une méthode de confection d'un corpus arboré qui vise à optimiser les ressources, par définition rares, dont on dispose dans le cas d'une langue peu dotée, qu'il s'agisse de moyens matériels (corpus et outils) ou humains requis pour le processus d'annotation. En effet, malgré l'expansion continuelle des corpus annotés depuis la fin du XX^{e} siècle, la question des standards et des bonnes pratiques quant à leur constitution reste peu abordée. Hovy et Lavid (2010) proposent une schématisation à huit étapes, articulant plusieurs phases d'annotation et d'évaluation. Pustejovsky et Stubbs (2012) proposent un schéma MATTER en sept points, soit modélisation du phénomène et création des guides, annotation, entraînement (*train*), test, évaluation et révision. À la différence de ces approches, qui soumettent l'adaptation du schéma d'annotation aux résultats des outils du TAL, Fort (2016) préconise la séparation de l'élaboration du corpus et des évaluations dans le cadre du TAL afin de garantir une validité plus générale de la ressource. Par ailleurs, Fort (2012) opère une distinction plus nette entre les périodes principales d'une campagne d'annotation, qui s'organise en travail de prépa-

1. `http://universaldependencies.org/`

ration (identification des participants, constitution du corpus, création du guide d'annotation), précampagne (création d'un corpus de référence minimal, formation des annotateurs), campagne (entraînement des annotateurs, annotation proprement dite, mise à jour du guide) et finalisation (publication du corpus). L'auteur identifie également les différents rôles à l'intérieur d'une campagne (le gestionnaire de campagne, l'expert, les annotateurs, l'évaluateur, etc.) et donne des recommandations en ce qui concerne la qualité de l'annotation. Plus précisément, Fort (2012) met en avant la méthode d'annotation agile, définie par Voormann et Gut (2008) et implémentée par Alex *et al.* (2010). Cette approche préconise une organisation cyclique du travail, présente de manière implicite dans les formalisations de Hovy et Lavid (2010) et Pustejovsky et Stubbs (2012) : le corpus est divisé en échantillons qui sont traités tour à tour ; chaque cycle d'annotation est suivi d'une étape d'évaluation, où l'accord interannotateur est calculé, permettant ainsi de contrôler la qualité de l'annotation, mais aussi de relever les problèmes dans les guides d'annotation et d'y remédier.

Dans le but de faciliter le travail des annotateurs humains et d'accélérer l'enrichissement du corpus, tout en garantissant la validité de l'annotation produite, nous adoptons la méthode de Fort (2012) et y intégrons des éléments de l'annotation agile. Plus concrètement, cette méthode est fondée sur plusieurs principes : l'adaptation de ressources existantes pour une langue proche et mieux dotée ; l'utilisation de ressources lexicales produites de façon collaborative ; la conception d'un processus d'annotation qui articule de façon optimale les phases d'annotation humaine et de préannotation automatique.

La section 2 aborde les principales caractéristiques du serbe et l'état de l'art en TAL de cette langue. Dans la section 3, nous présentons la méthode d'annotation adoptée dans sa globalité, pour détailler ensuite le travail de préparation d'outils permettant l'optimisation de l'annotation (section 4), les campagnes d'annotation manuelle (section 5) et le processus de finalisation du corpus (section 6). Enfin, la section 7 présente nos conclusions en rappelant les jalons d'une méthode optimisée pour le développement de ressources pour les langues peu dotées.

2. Le serbe : aperçu du système

Le serbe est une langue slave méridionale, parlée majoritairement en Serbie et dans les pays de l'ex-Yougoslavie par environ 8,7 millions de locuteurs (Keith, 2006). Il dispose d'un système d'écriture phonétique et a recours de façon quasi équivalente aux deux alphabets cyrillique et latin. Dans ce qui suit, nous proposons un aperçu de ses caractéristiques principales pertinentes pour le TAL et présentons ensuite l'état de l'art du TAL serbe.

2.1. *Principales propriétés linguistiques*

Le serbe exhibe toutes les propriétés phares de la famille slave : il dispose d'un système de déclinaisons relativement complexe, l'ordre des constituants est flexible, il n'y a pas d'articles, le système de l'aspect verbal est particulièrement bien développé, et la réalisation du sujet dans la phrase n'est pas obligatoire (il s'agit d'une langue *pro-drop*). Nous nous concentrerons ici sur deux d'entre elles : la morphologie flexionnelle riche et l'ordre des constituants flexible.

2.1.1. *Morphosyntaxe*

Le serbe dispose d'un système de déclinaison à 7 cas (nominatif, génitif, datif, accusatif, vocatif, instrumental et locatif), des marques du nombre (singulier ou pluriel)[2] et du genre (masculin, féminin ou neutre). Ces trois catégories caractérisent aussi bien les noms que les pronoms et les adjectifs. Par ailleurs, les adjectifs portent des marques du degré de comparaison (positif, comparatif ou superlatif). Par conséquent, on considère typiquement qu'un paradigme nominal contient 14 formes, alors qu'un paradigme adjectival en a 126. Quant à la conjugaison, le paradigme verbal prototypique compte au-delà de 120 formes.

Il existe cependant un degré de syncrétisme important (et en partie systématique), notamment dans les paradigmes de déclinaison. Par exemple, pour les noms et pour les adjectifs, le datif et le locatif sont toujours identiques (cf. la forme *detetu* – le datif ou le locatif singulier du nom *dete* 'enfant') ; pour les adjectifs, les formes du pluriel sont identiques pour les trois genres sauf au nominatif et au vocatif (cf. la forme *lepih* – le génitif pluriel du masculin, du féminin ou du neutre).

2.1.2. *Syntaxe*

Comme c'est souvent le cas, cette morphologie flexionnelle riche est accompagnée d'une flexibilité importante dans l'ordre des constituants, le système casuel prenant en charge l'encodage d'une partie des fonctions syntaxiques. Par exemple, les réalisations prototypiques du sujet, de l'objet direct et de l'objet indirect sont respectivement assurées par le nominatif, l'accusatif et le datif (exemple 1).

(1)

Filip	predstavlja	**An-u**	Alan-u
Filip.NOM.SG	présente	Ana.ACC.SG	Alan.DAT.SG

'*Filip* présente **Ana** à Alan'

L'ordre de ces constituants est très variable : même si l'ordre canonique est SVO, les 5 autres variations sont grammaticales, et l'objet indirect dispose d'un degré de

2. Le serbe exhibe également des traces de l'ancien dual (*paucal*), mais ces formes ne connaissent pas un usage systématique et ne sont typiquement pas considérées comme faisant partie du paradigme nominal canonique (Stanojčić et Popović, 2012).

flexibilité comparable. Une phrase simple comme celle donnée ci-dessus peut donc connaître de nombreuses variations (*Filip* predstavlja Alanu **Anu**, *Filip* **Anu** predstavlja Alanu, *Filip* Alanu predstavlja **Anu**, etc.).

Cette flexibilité au niveau syntaxique va jusqu'à autoriser des structures discontinues (cf. exemple 2). Ici, la seule manière de déterminer la fonction syntaxique de l'adjectif *lepu* 'beau' est de faire appel aux traits morphosyntaxiques qui participent à l'accord : comme le cas, le genre et le nombre de cette forme coïncident avec ceux du nom *knjigu* 'livre', c'est celui-ci qui est son gouverneur plutôt que le nom *Filip*.

(2)

Lep-u	je	Filip	knjig-u	kupio.
beau-ACC.SG.F	AUX	Filip.NOM.SG	livre-ACC.SG.F	acheté

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

Le serbe est donc plus riche en formes fléchies et en traits morphosyntaxiques, et plus variable au niveau syntaxique que les langues telles que l'anglais ou le français. Ces propriétés ont un effet concret sur le traitement automatique, discuté dans la section suivante.

2.2. *Traitement automatique du serbe et des langues proches*

Les langues comme le serbe, dotées d'une morphologie flexionnelle riche et d'une syntaxe flexible, posent des défis particuliers au TAL. Leur diversité aux niveaux lexical, morphosyntaxique et syntaxique se traduit par une dispersion des données : dans un corpus de taille standard, le nombre d'occurrences des phénomènes individuels reste bas, ce qui empêche les outils automatiques de les maîtriser. Ces langues sont donc souvent victimes d'un paradoxe : pour obtenir une bonne couverture des différents phénomènes qu'elles exhibent, elles doivent disposer de corpus plus larges que les langues à morphologie réduite. Or, la complexité de l'annotation exigée a un effet rédhibitoire sur la constitution de ressources et elles sont souvent relativement mal dotées en corpus annotés. Un indice en est la taille des corpus utilisés dans la campagne d'évaluation SPMRL 2013 (Seddah *et al.*, 2013). Ce fait est sans doute l'une des raisons pour lesquelles le serbe reste relativement peu doté de ressources et outils en TAL. Cependant, un autre facteur entre également en jeu : un manque de pratique du libre partage et de la diffusion des données et outils au sein de la communauté TAL serbe.

Au moment où nous avons entrepris ce projet, les seules ressources dédiées à cette langue librement diffusées comprenaient un corpus adapté à l'étiquetage morphosyntaxique (Krstev *et al.*, 2004), un corpus issu du Web doté d'annotations automatiques et par conséquent inadapté à l'entraînement des outils statistiques (Ljubešić et Klubička, 2014), un étiqueteur et lemmatiseur (Gesmundo et Samardžić, 2012) et un lexique morphosyntaxique (Krstev *et al.*, 2004). Cependant, de nombreuses autres

ressources sont citées dans les travaux existants sans être librement diffusées (Krstev et Vitas, 2005 ; Jakovljević *et al.*, 2014 ; Vitas et Krstev, 2004 ; Pavlović-Lažetić *et al.*, 2004 ; Krstev, 2008). Quant à l'analyse syntaxique, la seule tentative d'entraînement d'un analyseur syntaxique sur un corpus serbe a donné des résultats largement en dessous de l'état de l'art (LAS = 58 % et UAS = 66 %) (Jakovljević *et al.*, 2014), dus le plus probablement à la taille très limitée du corpus utilisé. À notre connaissance, ni le corpus d'entraînement ni les modèles d'analyse syntaxique développés n'ont été diffusés.

C'est dans ce cadre-là que nous avons posé les objectifs de ce travail : la création d'un corpus arboré pour le serbe, mais aussi de toute autre ressource qui pourrait faciliter le traitement automatique de cette langue, tel un lexique morphosyntaxique. Il faut néanmoins noter que cette situation s'est améliorée récemment grâce à la publication de plusieurs ressources, dont le lexique morphosyntaxique srLex (Ljubešić *et al.*, 2016) et un corpus arboré produit dans le cadre du projet UD, annoncé dans le travail de Samardžić *et al.* (2017) et publié en automne 2017.

Le croate, très proche du serbe, est mieux doté du point de vue du TAL. Avant la décomposition de l'ex-Yougoslavie, le serbo-croate était la langue officielle en Serbie, Croatie, Bosnie et au Monténégro. La création des États indépendants a mené à la proclamation des langues nationales. Leur statut est débattu depuis lors. Sans entrer dans des considérations socio-politiques complexes et sensibles, on peut résumer le rapport entre ces langues, à la suite de (Thomas, 1994), en disant que le serbe, le croate, le bosniaque et le monténégrin sont quasiment identiques aux niveaux phonologique, morphologique et syntaxique. Des différences plus importantes existent au niveau lexical, mais elles n'empêchent pas une compréhension mutuelle élevée des locuteurs sur le terrain. Parmi ces langues, c'est le croate qui est le mieux doté du point de vue du TAL (Agić *et al.*, 2013a ; Agić *et al.*, 2013b ; Agić et Ljubešić, 2014 ; Ljubešić *et al.*, 2016). Qui plus est, cette communauté pratique la libre diffusion de ressources et données. Nous nous sommes donc servis à plusieurs reprises de travaux effectués sur cette langue, ce qui sera détaillé dans la suite.

3. Constitution du corpus arboré serbe : méthode d'annotation adoptée

La création d'un corpus arboré est un processus complexe et coûteux. Avant d'entamer la constitution du corpus proprement dite, il est nécessaire de déterminer plusieurs aspects : il faut sélectionner le contenu à traiter, définir les annotations qui seront apportées au corpus, mettre en place des mécanismes pour assurer leur qualité, et optimiser le processus du point de vue du temps et de l'effort humain nécessaires. Dans cette section, nous présentons le corpus retenu (section 3.1), les principes qui ont guidé la création de notre corpus arboré (section 3.2) et la méthode qui nous a permis de les articuler (section 3.3).

3.1. *Corpus retenu*

Le contenu textuel utilisé dans ce projet provient de deux ouvrages littéraires serbes : *Bašta, pepeo* de D. Kiš et *Testament* de V. Stevanović (respectivements échantillons *basta* et *testament* dans le tableau 1)[3]. Une pratique plus courante consiste à utiliser des textes journalistiques (Marcus *et al.*, 1993 ; Abeillé *et al.*, 2003 ; Agić et Ljubešić, 2014), notamment parce que la question des droits limite souvent les possibilités de diffusion de textes littéraires. Or, nous avions déjà obtenu l'accord des ayants droit pour la diffusion non commerciale des textes concernés et avions procédé à leur étiquetage et, partiellement, à leur lemmatisation dans le cadre d'un travail antérieur (Miletic, 2013). Enfin, ce corpus littéraire apporte une diversification bienvenue aux corpus disponibles pour le croate et le serbe : SETimes.hr (Agić et Ljubešić, 2014) ainsi que les corpus UD pour le croate (Agić et Ljubešić, 2015) et le serbe (Samardžić *et al.*, 2017) sont basés sur des textes journalistiques.

La structure du corpus sélectionné et l'état de l'annotation au démarrage de la constitution du corpus arboré sont présentés dans le tableau 1.

Échantillon	*Tokens*	Étiquetage	Lemmatisation
Basta	55 783	Oui	Non
Testament	45 642	Oui	Oui
Total	101 425		

Tableau 1. *Structure et annotation préexistante de l'échantillon sélectionné*

3.2. *Principes d'annotation*

Quand il s'agit de l'annotation du corpus arboré, deux aspects principaux doivent être définis : la nature de l'annotation à apporter au corpus et les conditions dans lesquelles l'annotation manuelle se déroulera. Le premier dépend en partie de la nature de la langue traitée, mais aussi des exploitations envisagées du corpus et des contraintes temporelles du projet. L'objectif du deuxième est de garantir la qualité et l'efficacité du travail manuel.

Dans notre corpus arboré, nous avons mis en place une annotation en plusieurs couches. En effet, les exemples donnés dans la section 2.1.2 montrent que l'identification des fonctions syntaxiques en serbe repose fortement sur des traits morphosyntaxiques fins comme le cas, le nombre ou le genre. Nous avons donc effectué une annotation morphosyntaxique fine, qui inclut des traits utiles à l'analyse syntaxique. Au niveau de l'annotation syntaxique, nous avons adopté la syntaxe en dépendances (Tesnière, 1959 ; Mel'čuk, 1988). Au-delà du fait d'assurer une représentation

3. Kiš, Danilo. *Bašta, pepeo*, 2010. Podgorica : Narodna knjiga.
Stevanović, Vidosav. *Testament*, 1986. Beograd : SKZ.

aisée de structures discontinues (cf. exemple 2), ce cadre théorique devient le standard *de facto* en analyse syntaxique grâce aux campagnes d'évaluation CoNLL (Buchholz et Marsi, 2006 ; Nivre *et al.*, 2007) et au projet UD. Et comme nous nous attendions à un degré de dispersion des données important, nous avons inclus la lemmatisation pour réduire ce phénomène, suivant les travaux de Seddah *et al.* (2010) et de Le Roux *et al.* (2012). La contribution d'un lexique morphosyntaxique dans ce contexte ayant été mise en évidence (Hajič, 2000 ; Sagot, 2016), la création d'une telle ressource a également été incluse dans nos objectifs.

Quant aux conditions de l'annotation, nous avons conçu le processus de travail manuel de façon à garantir la validité de l'annotation tout en optimisant la vitesse de sa réalisation. Pour assurer la qualité et la cohérence des annotations manuelles, nous avons rédigé des guides d'annotation détaillés, qui ont été éprouvés *via* des évaluations de l'accord interannotateur. Ces techniques ont été utilisées dans la création de nombreux corpus arborés (Marcus *et al.*, 1993 ; Hajič, 2005 ; Brants, 2000).

Pour optimiser la rapidité du processus, nous avons eu recours à une préannotation automatique des données, dans la lignée de nombreux travaux qui ont démontré que cette méthode augmente la vitesse de traitement pour différents types d'annotation linguistique (Fort, 2012 ; Xue *et al.*, 2005 ; Tellier *et al.*, 2014). Dans la création des corpus arborés, cette méthode a été utilisée déjà lors de la constitution de PennTreebank (Marcus *et al.*, 1993) : ici, l'exploitation d'une préannotation automatique au niveau morphosyntaxique réduisait de moitié le temps d'annotation par rapport à l'annotation manuelle intégrale. Au niveau syntaxique, Chiou *et al.* (2001) notent que l'utilisation d'un analyseur syntaxique en constituants atteignant une précision de 82,87 % et un rappel de 81,42 % menait à une réduction du temps d'annotation de 50 %. Plus récemment, Skjærholt (2013) montre que la préannotation du norvégien avec un analyseur syntaxique d'une langue proche (le danois) permet de réduire le temps d'annotation de 50 %, et de 75 % lorsque l'analyseur est entraîné sur la même langue. Bien évidemment, l'utilisation de la préannotation entraîne le risque de biais : Fort et Sagot (2010) montrent que les erreurs d'un étiqueteur morphosyntaxique peuvent se propager dans les productions des annotateurs humains. Les auteurs concluent néanmoins que ce risque est justifié en vue des gains de temps conséquents. En témoignent également de nombreux autres projets de corpus arborés qui ont eu recours à cette approche (Hajič, 2005 ; Boguslavsky *et al.*, 2002 ; Abeillé *et al.*, 1998).

En l'absence d'outils déjà disponibles pour la préannotation du serbe, nous avons conçu une méthode itérative (section 3.3) fondée sur l'exploitation de ressources lexicales collaboratives (section 4.2) et l'adaptation de ressources disponibles pour une langue proche, le croate (section 4.3). La mise au point de ces différents volets de l'annotation s'appuie sur le principe général de l'annotation agile.

3.3. *Annotation agile basée sur un* bootstrapping *itératif multicouche*

L'organisation globale du travail est présentée dans la figure 1. Nous reprenons les quatre étapes de base identifiées par Fort (2012) : préparation (en bleu), précampagne (en jaune), campagne (en vert) et finalisation (en rouge). La phase de la campagne est plus complexe car elle est itérative et intègre des outils automatiques ; elle sera expliquée en détail dans la suite.

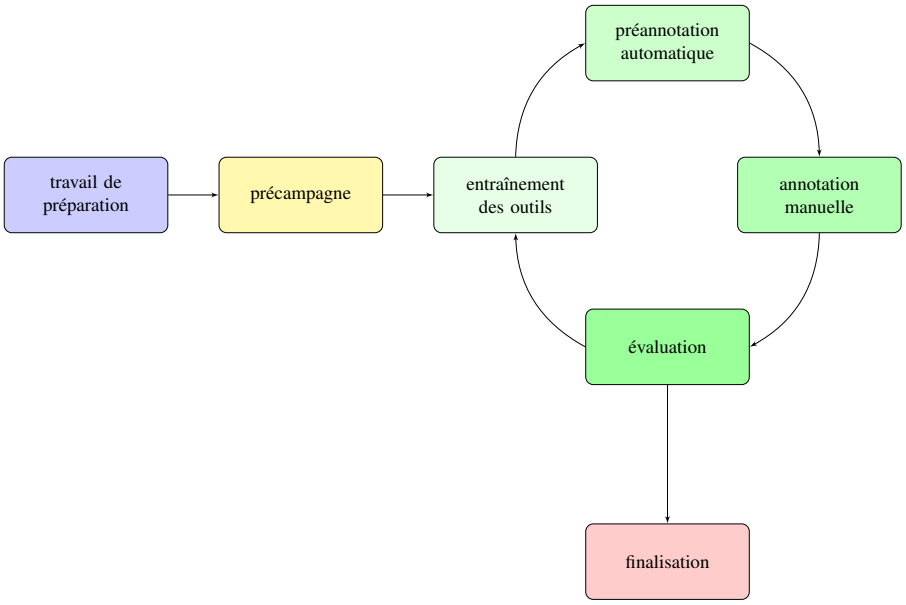

Figure 1. *Organisation du processus d'annotation*

Le travail de préparation correspond à la période de mise en place du matériel nécessaire à l'annotation du corpus. Concrètement, il s'agit de la sélection des outils automatiques à exploiter, du choix des textes qui composent le corpus, de la définition des jeux d'étiquettes, de la constitution des guides d'annotation et de leur première évaluation sur les données, ainsi que de la préparation des ressources externes (en particulier du lexique et des ressources d'entraînement initial pour les outils automatiques).

Le stade de la précampagne est dédié au recrutement et à la formation des annotateurs, qui leur permet de s'approprier les guides et les interfaces d'annotation.

L'organisation de la campagne est guidée par deux principes : l'agilité et l'utilisation d'outils automatiques. Le premier impose une organisation itérative du travail et introduit une étape d'évaluation à la fin de chaque cycle d'annotation manuelle. Le deuxième introduit deux étapes supplémentaires en début de chaque cycle : l'entraînement des outils et la préannotation automatique. Notons que ces deux étapes sont également exécutées itérativement par le recours au *bootstrapping* (v. *infra*). Lors du

premier passage par la boucle, l'entraînement des outils est effectué sur les ressources issues d'une phase minimale d'apprentissage constituées dans le stade du travail de préparation. Ces premiers modèles sont utilisés pour la préannotation du premier échantillon du corpus ; la préannotation automatique est corrigée manuellement, et l'échantillon nouvellement validé est rajouté aux ressources d'entraînement initiales. Lors du prochain passage par la boucle, les outils automatiques sont entraînés sur ces ressources augmentées, ce qui leur permet de s'améliorer à chaque itération, facilitant ainsi l'annotation manuelle. L'évaluation telle que nous la définissons diffère de ce qui est préconisé par Voormann et Gut (2008). Étant donné le temps nécessaire pour effectuer systématiquement l'annotation en double nécessaire à l'évaluation de l'accord interannotateur, nous n'intégrons pas celle-ci dans ce cycle, mais vérifions la qualité du travail des annotateurs à travers un contrôle ponctuel de la part d'un annotateur expérimenté. Cette étape contient également une séance de travail dédiée au retour d'expérience des annotateurs, qui porte notamment sur les guides d'annotation. Si les problèmes identifiés l'exigent, les guides sont modifiés en conséquence. Pour éviter les incohérences que ces modifications progressives peuvent introduire dans l'annotation, un travail d'harmonisation des annotations est réalisé dans la phase de finalisation.

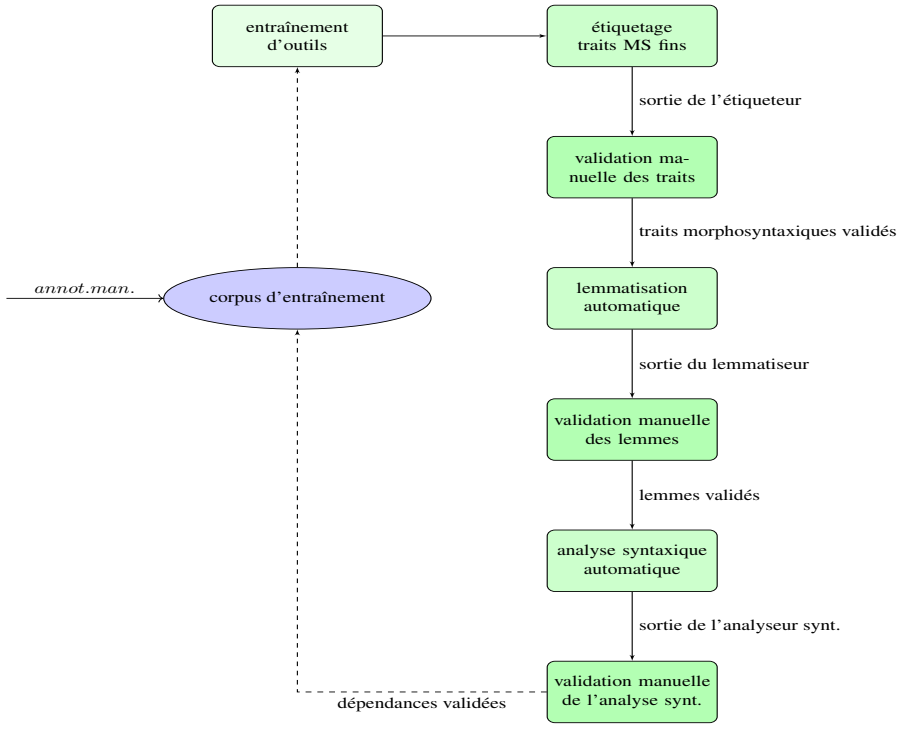

Figure 2. *Bootstrapping itératif pour une annotation multicouche*

La finalisation du corpus comprend donc le travail d'harmonisation des annotations mentionné ci-dessus. Elle porte également sur toutes les activités nécessaires à la diffusion du corpus : la conversion du corpus vers un format de diffusion standard, l'élaboration d'une documentation, la diffusion du corpus proprement dite.

Le schéma de la figure 1 correspond à l'annotation d'une seule couche d'informations. Notre corpus en exige trois. La figure 2 présente l'organisation d'une annotation multicouche, qui commence par un entraînement initial des trois outils dans l'étape de préparation. Dans le cadre de la campagne proprement dite, l'étiquetage morphosyntaxique, la lemmatisation et l'analyse syntaxique sont effectués en cascade. La sortie de chaque outil est corrigée manuellement avant d'être passée à l'outil suivant. Ainsi, chaque outil reçoit en entrée une annotation fiable, ce qui permet de maximiser ses performances et de faciliter par extension le travail des annotateurs humains. Précisons encore que l'ordre d'exécution des tâches est défini par les besoins des outils : l'analyseur syntaxique s'appuie typiquement sur les lemmes et les informations morphosyntaxiques, et le lemmatiseur exploite les catégories grammaticales, alors que l'étiqueteur n'exige pas de données supplémentaires.

4. Ressources optimisant l'annotation

Cette section est dédiée à une description du travail de préparation qui a porté sur la constitution des éléments matériels nécessaires à notre méthode.

4.1. *Création et évaluation des guides d'annotation*

Notre objectif étant ici de présenter la méthode globale, nous n'entrons pas dans les détails des schémas d'annotation adoptés pour la création du corpus arboré serbe, mais présentons plutôt les principes globaux sur lesquels ils reposent et le travail d'évaluation des guides d'annotation.

4.1.1. *Principes de constitution des jeux d'étiquettes et des schémas d'annotation*

Pour définir les jeux d'étiquettes morphosyntaxiques et syntaxiques, nous avons soumis les notions issues de la tradition grammaticale serbe à un examen détaillé théorique et empirique. Si les traitements existants ne se basaient pas sur des critères explicites, accessibles aussi bien à un annotateur humain qu'à un analyseur syntaxique, nous les avons modifiés de sorte à satisfaire cette exigence.

Au niveau morphosyntaxique, cela se traduit par l'élimination du jeu d'étiquettes d'un nombre de traits traditionnellement reconnus par les grammaires serbes et utilisés dans certains travaux de TAL (Krstev *et al.*, 2004)[4]. Il s'agit notamment de traits rele-

4. Le jeu d'étiquettes proposé dans ce travail compte 1 243 étiquettes. Une description détaillée est disponible à l'adresse suivante :`http://nl.ijs.si/ME/V4/msd/html/msd-sr.html`.

vant de distinctions qui ne sont pas utiles à l'identification des fonctions syntaxiques dans la phrase, comme l'aspect verbal, la définitude des adjectifs et l'opposition animé – non animé. Nous avons ainsi établi un jeu d'étiquettes raisonné de 1 042 étiquettes, basé sur les traits présentés dans le tableau 2.

Partie du discours	Traits encodés
Adjectif	Partie du discours, sous-catégorie, cas, nombre, genre, degré de comparaison
Nom	Partie du discours, sous-catégorie, cas, nombre, genre
Numéral	Partie du discours, sous-catégorie, cas, nombre, genre
Pronom	Partie du discours, sous-catégorie, cas, personne, nombre, genre
Verbe	Partie du discours, sous-catégorie, forme, personne, nombre, genre
Adverbe	Partie du discours, sous-catégorie, degré de comparaison
Conjonction	Partie du discours, sous-catégorie
Interjection	Partie du discours
Particule	Partie du discours
Préposition	Partie du discours

Tableau 2. *Traits morphosyntaxiques encodés dans le corpus arboré*

Au niveau syntaxique, les fonctions reconnues par les grammaires serbes (Stanojčić et Popović, 2012 ; Mrazović, 2009) ont été examinées à l'aide d'un ensemble de critères de distinction des relations syntaxiques de surface. Ces critères s'inspirent à la fois des fondements théoriques de la syntaxe en dépendances définis par Mel'čuk (1988) et du travail pratique de Burga *et al.* (2011) sur un corpus arboré espagnol. Dans ce travail, les auteurs établissent un inventaire de relations syntaxiques en espagnol en s'appuyant sur une liste de critères syntaxiques et morphologiques, comme les catégories grammaticales et les lemmes possibles du gouverneur et du dépendant, les traits de flexion du gouverneur et du dépendant, la présence de l'accord, la possibilité de pronominalisation par un clitique (dans le cas des dépendants verbaux), ainsi que les règles de linéarisation du gouverneur et du dépendant dans la phrase. Comme il s'agit de critères de surface, accessibles à un analyseur syntaxique, leur utilisation garantit non seulement une définition rigoureuse des étiquettes, mais aussi l'adaptation de ces dernières à l'analyse syntaxique. Nous avons donc adopté une démarche comparable.[5]

5. Une autre possibilité aurait consisté à adopter le jeu d'étiquettes du projet UD. Cependant, nous sommes en accord avec certaines critiques de ce schéma d'annotation proposées dans (Groß et Osborne, 2015) et nous avons opté pour une annotation basée sur le principe des têtes fonctionnelles et qui préserve un schéma particulier au serbe. Néanmoins, étant donné l'utilité du schéma UD pour un large éventail de recherches en TAL, la conversion du corpus existant vers ce format fait partie des perspectives du projet.

Cette démarche nous a permis d'identifier 48 étiquettes syntaxiques de base, complémentées par un traitement spécifique pour l'ellipse[6]. Dans le jeu, nous maintenons un noyau de fonctions de la grammaire serbe (les sujets, les objets, les prédicatifs). L'écart le plus important entre notre jeu d'étiquettes et la grammaire serbe relève de la distinction entre les arguments et les ajouts, que nous avons décidé d'ignorer, du fait des difficultés à la capter par des critères de surface. Nous avons opté pour une étiquette de dépendant sous-spécifiée. Une description détaillée du jeu d'étiquettes syntaxique est disponible dans le guide d'annotation diffusé avec le corpus[7].

4.1.2. *Mise au point des guides d'annotation*

Afin de garantir la complétude des guides d'annotation et leur adaptation à un travail sur corpus, nous les avons soumis à une évaluation de l'accord interannotateur. Nous avons calculé le taux d'accord pour l'annotation morphosyntaxique et l'annotation syntaxique, en utilisant comme mesure d'accord le *kappa* de Cohen (Cohen, 1960). Malgré des critiques quant à l'interprétation de ses différentes valeurs (Artstein et Poesio, 2008 ; Mathet *et al.*, 2012), le *kappa* de Cohen reste une mesure standard communément utilisée pour l'évaluation de l'accord interannotateur en corpus (Bhat et Sharma, 2012 ; Bond *et al.*, 2006 ; Agić et Merkler, 2013) et on considère en général que les valeurs supérieures à 0,90 représentent un accord satisfaisant.

Au niveau morphosyntaxique, l'évaluation a été effectuée par deux paires d'annotateurs. Chaque paire a eu un échantillon de 2 000 *tokens* à traiter. L'accord a été calculé aussi bien au niveau des étiquettes catégorielles (11 étiquettes instanciées) qu'au niveau des étiquettes morphosyntaxiques détaillées (205 étiquettes instanciées). Les résultats sont donnés dans le tableau 3. À titre d'illustration, le taux d'accord interannotateur au niveau morphosyntaxique dans le PennTreebank indiqué par Marcus *et al.* (1993) est de 96 % avec un jeu de 36 étiquettes, et celui dans le corpus arboré NEGRA est de 98,6 % (Brants, 2000) avec un jeu de 54 étiquettes. Notons néanmoins qu'il s'agit dans les deux cas d'un taux d'accord observé, qui n'est donc pas directement comparable à nos résultats.

Le degré de l'accord sur les étiquettes complètes est inférieur à celui qui concerne l'identification des catégories grammaticales. D'après le retour des annotateurs, ceci est en partie dû au nombre élevé de traits relatifs aux catégories grammaticales fléchies. Une analyse des matrices de confusion a montré que les distinctions les plus problématiques étaient celles entre les participes et les adjectifs déverbaux (cf. *izgu-*

6. Nous reprenons le traitement de l'ellipse mis en place dans le Prague Dependency Treebank (Hajič *et al.*, 1999, p. 204-221).

7. `https://github.com/aleksandra-miletic/serbian-nlp-resources/blob/master/ParCoTrain-Synt/`

bljen ‘perdu’) et celles entre les adverbes et les particules (cf. *jednostavno* ‘simplement’[8].). Le traitement de ces points a donc été clarifié dans le guide.

Annotation morphosyntaxique		
	Paire 1	Paire 2
kappa sur étiquettes complètes	0,90	0,91
kappa sur POS	0,96	0,97
Annotation syntaxique		
kappa sur fonctions syntaxiques	0,94	

Tableau 3. *kappa de Cohen à différents niveaux d’annotation*

Quant à l’annotation syntaxique, l’évaluation a été effectuée par une paire d’annotateurs sur un échantillon de 3 000 *tokens* (48 étiquettes instanciées) (cf. tableau 3). L’accord a été calculé en prenant en compte le rattachement étiqueté. À titre de comparaison, dans le corpus arboré des discussions Wikipédia FrWikiDisc, constitué par Urieli (2013), le *kappa* de Cohen était de 0,86 entre les deux annotateurs au niveau du rattachement étiqueté. Skjærholt (2013) indique un taux d’accord interannotateur observé de 95,3 % sur leur corpus arboré norvégien dans les mêmes conditions.

L’analyse de la matrice de confusion que nous avons effectuée a montré que c’est le traitement de différentes formes d’ellipse qui a généré le plus d’erreurs. Pour contrer cet effet, nous avons introduit des tests syntaxiques dans le guide d’annotation pour faciliter l’identification du gouverneur et du dépendant dans ces constructions.

Suite à ces modifications, les guides d’annotation ont été jugés suffisamment fiables pour être exploités dans le cadre de l’annotation manuelle.

4.2. *Création de ressources lexicales*

Comme mentionné dans la section 3.2, plusieurs travaux ont montré que l’utilisation d’un lexique externe peut faciliter l’analyse syntaxique d’une langue comme le serbe, car il permet de compléter le nombre de formes fléchies auxquelles un corpus d’apprentissage donne accès. Or, le seul lexique serbe librement disponible au début de ce projet était trop petit pour avoir un effet satisfaisant dans le cadre de l’analyse syntaxique (20 000 entrées seulement) (Krstev *et al.*, 2004). Pour assurer une couverture plus solide, nous avons constitué un lexique morphosyntaxique à partir du Wiktionnaire pour le serbo-croate, en nous inspirant des travaux de Sajous *et al.* (2013), Sagot (2014) et Sennrich et Kunz (2014), qui ont exploité la ressource libre du Wiktionnaire pour constituer des ressources électroniques dotées d’informations morphosyntaxiques.

8. Comparer l’adverbe extraprédicatif dans *Jednostavno, treba prestati* ‘Il faut simplement arrêter’ à l’adverbe intraprédicatif dans *Govori jednostavno* ‘Il parle simplement/d’une manière simple’.

Cette expérience, détaillée dans (Miletic, 2017), a abouti à la création de Wikimorph-sr, un lexique contenant 1 226 638 formes fléchies provenant de 117 445 lemmes différents, réparties en 3 066 214 triplets uniques <*forme fléchie, lemme, étiquette morphosyntaxique détaillée*>. Ce lexique est donc nettement mieux doté que le lexique existant du projet MultextEast (Krstev *et al.*, 2004) (20 000 entrées). Il est vrai que le contenu du lexique extrait n'est pas parfait : il présente un certain degré de surgénération des formes fléchies (notamment dans les paradigmes adjectivaux, qui disposent systématiquement des formes de comparatif et de superlatif, même pour les adjectifs relationnels), et ne contient pas du tout de catégories invariables, ce qui nous a amenés à effectuer des ajouts à partir de différentes ressources. Néanmoins, le lexique a été réalisé en trois semaines de travail, ce qui montre que cette approche est adaptée aux projets à durée limitée.

Ce travail a été effectué en 2015 ; en 2016, un lexique serbe de 5,3 millions d'entrées construit manuellement a été diffusé par Ljubešić *et al.* (2016). SrLex a été construit à partir des lexiques croate, serbe et bosniaque du logiciel de traduction automatique à base de règles Apertium (Forcada *et al.*, 2011), mais il a connu des extensions importantes dans le cadre d'une campagne de création manuelle d'entrées. Le lexique est librement diffusé[9], avec une licence qui autorise la redistribution. Nous avons fusionné les deux lexiques afin de maximiser leur utilité et avons ainsi obtenu une nouvelle ressource nommée ParCoLex. Ce troisième lexique contient au total 7 180 665 entrées uniques <*forme fléchie, lemme, étiquette morphosyntaxique détaillée*>, qui représentent 1 956 094 formes fléchies uniques provenant de 157 886 lemmes. Nous avons ensuite évalué la couverture des trois lexiques sur un échantillon de texte de 16 389 *tokens*, correspondant à 6 301 formes fléchies uniques. Les résultats, présentés dans le tableau 4, montrent que si SrLex a une couverture très largement supérieure à celle de Wikimorph-sr, la fusion des deux ressources offre un gain de 2,4 % sur les formes fléchies uniques, et de 4 % sur le nombre total d'occurrences des formes fléchies de l'échantillon. Ce lexique, aussi bien que Wikimorph-sr, est également diffusé avec le corpus arboré.

Lexique	Entrées	Lemmes	Couverture de l'autre lexique	Couverture échantillon	
				formes fléchies	occurrences
Wikimorph-sr	3 066 214	117 445	20,8 %	63,3 %	73,2 %
srLex	5 327 361	105 358	41,1 %	92,8 %	93,8 %
ParCoLex	7 180 665	157 886	NA	95,2 %	97,8 %

Tableau 4. *Tests de couverture avec les trois lexiques*

9. `http://nlp.ffzg.hr/resources/lexicons/srlex/`. Dernier accès : le 23 octobre 2017.

4.3. *Mise en œuvre des outils automatiques*

Notre choix d'outils de préannotation automatique a été guidé par deux critères : leurs performances sur le serbe ou sur une langue proche, et leur vitesse et ergonomie. Le premier critère s'explique par le fait qu'une préannotation de meilleure qualité facilite le travail des annotateurs humains, alors que le deuxième est dû à la nature itérative de notre méthode : comme elle prévoit plusieurs cycles d'entraînement et d'annotation, il est essentiel que les outils soient rapides et faciles à maîtriser.

À partir du travail d'Agić *et al.* (2013a), nous avons identifié l'étiqueteur HunPos (Halácsy *et al.*, 2007) et le lemmatiseur CST (Jongejan et Dalianis, 2009) comme outils offrant le meilleur compromis entre ces deux aspects. En ce qui concerne l'analyse syntaxique, les résultats signalés dans (Agić et Ljubešić, 2015) pointaient vers l'analyseur syntaxique Mate (Bohnet, 2010), basé sur un algorithme par graphes. Nous avons néanmoins préféré utiliser l'analyseur syntaxique Talismane (Urieli, 2013), basé sur un algorithme par transitions. Cet outil permet de définir avec précision l'exploitation de différents traits d'apprentissage (*tokens*, étiquettes POS, lemmes, informations morphosyntaxiques détaillées). Par ailleurs, il n'utilise pas les traits morphosyntaxiques désambiguïsés du corpus d'apprentissage, mais les puise plutôt dans un lexique externe en gardant toute l'ambiguïté rencontrée. Cette particularité est censée lui assurer une meilleure robustesse face au traitement d'un texte brut.

Une deuxième étape a consisté à assurer des ressources d'apprentissage pour les outils choisis et à les entraîner. Les travaux cités ci-dessus (Agić *et al.*, 2013a ; Agić et Ljubešić, 2015) montrent qu'il est possible de transposer un modèle entraîné sur le croate à des textes en serbe sans encourir d'importantes pertes de performances, et ceci aux trois niveaux d'annotation considérés. Comme les modèles de traitement développés par ces chercheurs sont librement disponibles, nous avons tâché de les exploiter pour la préannotation des échantillons de texte qui allaient constituer les ressources d'entraînement initiales pour notre méthode. Cette approche a été fructueuse pour l'étiquetage morphosyntaxique. Après avoir annoté notre échantillon avec le modèle de HunPos entraîné exclusivement sur le croate, nous avons constaté une exactitude moyenne du modèle de 77,95 %. Malgré la perte importante par rapport aux résultats rapportés dans (Agić *et al.*, 2013a) (où l'outil avait atteint une exactitude moyenne de 85 %), cette préannotation a facilité la correction manuelle de manière importante, permettant aux annotateurs humains de traiter 24 % de *tokens* en plus par rapport à une annotation manuelle intégrale. Cependant, le schéma d'annotation sur lequel a été entraîné le modèle croate n'est pas identique au nôtre, ce qui a entraîné des corrections supplémentaires pour adapter les traitements reproduits par l'étiqueteur à nos règles d'annotation. Ce sont ces modifications qui ont été jugées comme les plus chronophages par les annotateurs humains. Pour éviter ce type de corrections, dans l'étape suivante, le premier échantillon validé à l'aide du modèle croate a été utilisé pour ré-entraîner HunPos sur notre schéma d'annotation. Ce modèle, appris sur 20 000 *tokens*, a atteint une exactitude moyenne de 78,82 %, et la préannotation effectuée avec ce modèle a mené à une accélération du travail manuel de 60 % par rapport à l'annotation manuelle intégrale. La vitesse d'annotation moyenne dans ces conditions

était de 710 *tokens*/h. Pour comparaison, Marcus *et al.* (1993) indiquent une vitesse moyenne de 3000 *tokens*/h lors de la création du PennTreebank. Néanmoins, leur jeu d'étiquettes est beaucoup plus petit que le nôtre (36 étiquettes *vs* 1 042 étiquettes).

En revanche, cette démarche ne s'est pas avérée adaptée à la lemmatisation et à l'analyse syntaxique. En effet, le modèle croate de CST avait été entraîné sur des données étiquetées avec un jeu d'étiquettes morphosyntaxiques différent du nôtre ; par conséquent, ses performances sur nos données ont été compromises. Pour l'entraînement initial du lemmatiseur CST nous avons donc exploité un texte lemmatisé manuellement par Miletic (2013), que nous avons transformé en un lexique d'entraînement d'environ 20 000 entrées (10 000 lemmes différents). Malgré la taille restreinte de la ressource d'entraînement, ce premier modèle de CST a atteint une exactitude moyenne de 86,2 % et a permis une accélération de la lemmatisation manuelle de 41 % par rapport à la lemmatisation manuelle intégrale. Dans un deuxième temps et suite à la confection du lexique ParCoLex (cf. section 4.2), nous avons effectué un deuxième entraînement basé sur cette ressource. Le nouveau modèle a atteint une exactitude globale de 96,5 %, et la vitesse d'annotation manuelle a atteint 3 400 *tokens*/h, soit une accélération de 242 % par rapport à la lemmatisation manuelle intégrale.

Quant à l'analyse syntaxique, le modèle croate n'a pas pu être utilisé pour deux raisons : il a été entraîné sur un jeu d'étiquettes morphosyntaxiques différent et, qui plus est, il intégrait également un schéma d'annotation syntaxique très éloigné du nôtre. Nous avons donc jugé qu'une préannotation avec ce modèle n'était pas la manière la plus économe de procéder et avons favorisé une annotation manuelle. L'entraînement initial de Talismane a été effectué sur 40 000 *tokens* annotés à la main. Ce premier modèle a atteint une exactitude de 76,3 % en LAS et de 80,6 % en UAS. Ces scores ont été jugés satisfaisants, étant donné les résultats obtenus sur le serbe avec les analyseurs syntaxiques MST (LAS = 73,9 % et UAS = 80,6 %) (Agić *et al.*, 2013b) et Mate (LAS = 75,8 % et UAS = 82,4 %) (Agić et Ljubešić, 2015), notamment si l'on prend en compte le fait que les modèles en question ont été développés sur un corpus deux fois plus grand que notre échantillon (87 000 *tokens*).

5. Campagnes d'annotation manuelle

L'essentiel de l'annotation manuelle de notre corpus a été effectué dans le cadre de deux campagnes d'annotation avec des annotateurs étudiants serbophones. Les étudiants ont été majoritairement sélectionnés au département des études romanes à l'université de Belgrade à l'aide de questionnaires adaptés à la tâche à effectuer. Les campagnes d'annotation ont eu lieu à l'université de Toulouse-Jean-Jaurès. Avant d'entamer le travail d'annotation, les deux groupes d'annotateurs ont été formés à la fois sur les guides d'annotation et sur les interfaces d'annotation utilisées.

La première campagne a été dédiée à l'annotation morphosyntaxique et à la lemmatisation, effectuées en cascade et de manière itérative comme stipulé par notre méthode. En revanche, l'annotation syntaxique a été effectuée indépendamment durant

la deuxième campagne. Cette modification de notre méthode a été conditionnée par les compétences des annotateurs recrutés pour la première campagne. Les deux campagnes ont systématiquement fait appel aux outils de prétraitement présentés dans la section 4.3, entraînés selon la démarche de *bootstrapping* itératif. La tâche concrète des annotateurs consistait donc à corriger l'annotation produite par ces outils. Dans le cas de l'annotation syntaxique, nous avons exploité une fonctionnalité particulière de Talismane : cet analyseur syntaxique offre la possibilité d'accompagner chaque annotation du taux de probabilité qui lui est associé. Nous avons donc filtré la sortie de l'outil de sorte à ne retenir que les dépendances dont le taux de probabilité était supérieur à 0,85. Ainsi, les annotateurs disposaient d'une annotation partielle, mais fiable, ce qui a été jugé préférable par rapport à la possibilité de disposer d'une annotation complète, mais d'une qualité inférieure. Cette décision a été motivée par l'expérience de l'annotateur expert qui avait testé la préannotation intégrale dans l'étape de préparation. Cependant, nous n'avons pas fait d'expérience systématique sur ce point.

Les résultats des campagnes sont résumés dans le tableau 5. Comme certains annotateurs ont exprimé le souhait de poursuivre leur participation au projet, le travail d'annotation a été continué à distance (cf. les parties des deux campagnes effectuées à Belgrade). Précisons encore que l'annotation du corpus arboré a déjà été entamée, soit dans le cadre de travaux antérieurs, soit dans le cadre de l'initialisation des outils automatiques décrite dans la section 4.3. Le travail effectué par les annotateurs a donc mené à la complétion de l'annotation du corpus arboré.

Camp.	Tâche	Annotateurs	Durée	Endroit	Rendement	Pers.-heures
C1	morphosyntaxe	3 L3 + 1 M1	2 semaines	Toulouse	30 000 tok.	60 h
			6 semaines	Belgrade	30 000 tok.	50 h
	lemmatisation	3 L3 + 1 M1	2 semaines	Toulouse	35 000 tok.	25 h
C2	syntaxe	1 L3 + 1 M1	3 semaines	Toulouse	40 000 tok.	150 h
			6 semaines	Belgrade	20 000 tok.	60 h

Tableau 5. *Travail réalisé dans les campagnes d'annotation manuelle*

Nous pensons que ces résultats ont été favorisés par notre méthode globale : sans la préannotation automatique, la validation de différentes couches d'annotation aurait pris considérablement plus de temps. Par ailleurs, les annotateurs ont particulièrement apprécié le fait de pouvoir faire un retour direct de leurs expériences. Nous sommes d'avis que cette démarche a renforcé leur sentiment d'appartenance au projet, ce qui a garanti un niveau d'implication et de motivation élevé tout au long du projet.

6. Finalisation du corpus constitué

La dernière étape dans la création du corpus arboré a consisté en la finalisation et la diffusion du corpus constitué. Un annotateur expérimenté a été chargé d'harmoniser les annotations en accord avec la dernière version des guides d'annotation, établie

durant les campagnes d'annotation. Une fois ce travail finalisé, nous avons procédé à la préparation du corpus pour la diffusion.

Le corpus est diffusé dans le format CoNLL-X. Les champs contiennent les informations suivantes : l'ID du *token*, le *token*, le lemme, l'étiquette POS gros grain, l'étiquette POS plus spécifique, les traits morphosyntaxiques, l'ID du gouverneur du *token* et l'étiquette syntaxique du *token*. Le corpus a été divisé en trois sections : *train* (destinée à l'entraînement des analyseurs syntaxiques), *dev* (dédiée au paramétrage fin) et *test* (réservée à l'évaluation). Afin d'éviter tout biais relatif à la longueur des phrases dans différents segments du corpus, les phrases ont été réordonnées de manière aléatoire avant la segmentation du corpus en sections. Quelques informations sur les propriétés du corpus annoté sont données dans le tableau 6.

Les sections sont bien équilibrées au regard de plusieurs caractéristiques (longueur des phrases et profondeur de l'arbre syntaxique, représentation des parties du discours). Néanmoins, la répartition des étiquettes morphosyntaxiques détaillées est inégale. Ce fait pourrait se montrer problématique lors de l'utilisation du corpus pour l'entraînement des étiqueteurs morphosyntaxiques sur cette couche d'annotation.

Section	*Tokens*	Phrases	Formes fléchies	Lemmes	Étiq. POS	Étiq. détail.	Traits MS	Étiq. synt.	Long. phr.	Prof. ar.
all	101 029	3 861	22 739	11 251	16	679	165	67	27,16	6,98
train	80 869	3 116	19 598	10 120	16	643	159	67	26,95	6,98
test	10 162	367	4 033	2 802	15	432	134	60	28,69	7,11
dev	9 998	379	3 959	2 722	16	424	126	58	27,38	6,88

Tableau 6. *Corpus arboré constitué : statistiques de base. Long.phr. = longueur moyenne de phrase en* tokens*; Prof.ar. = valeur moyenne de la profondeur d'arbre maximale*

Le corpus est téléchargeable à partir de l'adresse suivante : `https://github.com/aleksandra-miletic/serbian-nlp-resources` sous la licence CC BY-NC-SA 3.0[10]. Ce dépôt contient également les différents modèles de traitement automatique développés durant ce projet pour l'étiquetage morphosyntaxique, la lemmatisation et l'analyse syntaxique, ainsi que la documentation des annotations apportées au corpus.

Ce corpus arboré s'est déjà montré utile dans plusieurs applications différentes. Notamment, il a permis d'entraîner un modèle d'analyse syntaxique avec Talismane qui atteint les scores de 87,48 % en LAS et de 91,22 % en UAS à partir d'un étiquetage morphosyntaxique manuel. Ceci représente un gain de 5,98 % en LAS et de 5,22 % en UAS par rapport aux meilleurs résultats préalables en analyse syntaxique du serbe, réalisés par Agić et Ljubešić (2015) avec l'analyseur syntaxique Mate. Il est vrai que certains aspects de ces expériences diffèrent de manière importante : le corpus d'entraînement utilisé par Agić et Ljubešić (2015) est journalistique alors que le nôtre est littéraire ; nos schémas d'annotation ne sont pas identiques (ils utilisent ceux du projet

10. `https://creativecommons.org/licenses/by-nc-sa/3.0/`

UD) ; différents types d'analyseurs syntaxiques ont été utilisés (Talismane est un analyseur syntaxique par transitions, alors que Mate est basé sur les graphes). Par conséquent, les résultats ne sont pas directement comparables. Néanmoins, les différences citées ne nous sont pas *a priori* favorables : les textes journalistiques sont en général considérés comme plus faciles que les textes littéraires, les analyseurs syntaxiques par graphes sont censés être plus adaptés aux langues à la morphologie flexionnelle riche, et les schémas d'annotation UD visent l'optimisation de l'analyse syntaxique indépendamment de la langue. Des évaluations plus directes sont tout de même nécessaires pour comprendre les effets de ces différents paramètres.

Notre corpus a également servi de base pour deux études en syntaxe théorique du serbe. La première version du corpus a été exploitée afin d'examiner les structures discontinues en serbe, à la fois dans une perspective monolingue et contrastive (Miletic et Urieli, 2017). La dernière version, présentée ici, a été utilisée dans une analyse approfondie de la position et de la structure du groupe adjectival gouverné par un nom (Miletic, 2018, ch. 10). Cette ressource s'est donc déjà montrée adaptée à la fois aux recherches en analyse syntaxique et en linguistique théorique.

7. Conclusions et perspectives

Dans cet article, nous avons présenté la démarche que nous avons utilisée pour constituer un nouveau corpus arboré pour le serbe, une langue peu dotée en corpus et outils du TAL. Ce corpus arboré s'accompagne également de lexiques morphosyntaxiques et de plusieurs modèles de traitement automatique. À la différence du corpus arboré serbe du projet UD, qui met en place l'annotation syntaxique indépendante de la langue préconisée par le projet, notre corpus est doté d'une annotation spécifique au serbe, qui établit une analyse plus fine, notamment au niveau des dépendants verbaux. Par ailleurs, le corpus UD est basé sur des textes journalistiques, alors que le nôtre relève du genre littéraire. Ce fait ouvre la voie aux expériences liées aux effets du genre, aussi bien en TAL qu'en linguistique serbe.

Nos résultats montrent que notre méthode, qui exploite au maximum les outils et ressources disponibles et qui accorde une attention particulière à l'organisation du travail des annotateurs, permet la réalisation aisée et relativement rapide d'une ressource complexe et polyvalente : la totalité du travail décrit ici a été effectuée entre novembre 2014 et avril 2018 dans le cadre d'un projet de thèse. Qui plus est, cette approche est basée sur des procédés généraux qui peuvent être appliqués à d'autres langues.

Tout d'abord, nous nous sommes servis de ressources existantes pour une langue proche : nous avons exploité avec succès un modèle d'étiquetage entraîné sur le croate pour faciliter la première phase d'étiquetage morphosyntaxique manuel de notre corpus. Grâce à la relation particulière entre le croate et le serbe décrite dans la section 2.2, nous avons pu effectuer cette manipulation sans faire appel à des techniques de traitement interlangue. Cependant, une approche comparable peut également être envisagée pour des langues plus éloignées. À titre d'illustration, le travail de Vergez-

Couret et Urieli (2015) montre que l'étiquetage de l'occitan bénéficie de l'ajout d'un grand corpus catalan à un corpus minimal de l'occitan lors de l'apprentissage.

Nous avons également fait appel à des ressources d'entraînement minimales : pour assurer un entraînement initial du lemmatiseur CST, nous avons exploité un lexique d'entraînement d'à peine 10 000 lemmes. Vu la richesse morphologique du serbe, ceci n'était pas prometteur. Néanmoins, malgré les performances moyennes de l'outil (86,2 % d'exactitude), la vitesse d'annotation manuelle a quasiment doublé.

Nous avons également exploité une ressource électronique créée de manière collaborative pour en extraire un premier lexique morphosyntaxique. En combinant le résultat de ce travail avec un autre lexique serbe (Ljubešić *et al.*, 2016), nous avons amélioré les résultats de la lemmatisation de manière significative. Cette approche est également transposable à toute langue dotée d'une ressource comparable au Wiktionary. À titre d'exemple, l'occitan et le picard en disposent.

Enfin, nous avons fait usage des particularités des outils utilisés pour la préannotation, et notamment de la capacité de l'analyseur syntaxique choisi à accompagner les étiquettes syntaxiques produites par les taux de probabilité associés. Ce fait nous a permis de trier sa sortie et de retenir une préannotation incomplète, mais fiable. Cette démarche est notamment importante lors d'une préannotation avec un outil relativement peu performant, autrement dit, durant les premières itérations de la démarche.

Dans l'avenir, nous poursuivrons ce travail dans deux directions principales. Premièrement, en ce qui concerne le corpus arboré créé, nous chercherons à raffiner certains points de l'annotation manuelle, et notamment le traitement des dépendants verbaux et de l'ellipse. Nous entreprendrons également la conversion du corpus vers le format UD, évoquée ci-dessus. Cette nouvelle annotation ne prendra cependant pas la place de l'annotation existante, mais sera ajoutée comme une couche d'information supplémentaire. Deuxièmement, nous chercherons à confirmer davantage la pertinence de notre méthode en la transposant sur l'occitan. Ce travail sera effectué dans le cadre du projet LINGUATEC (EFA227/16) « Développement de la coopération transfrontalière et du transfert de connaissance en technologies de la langue » du Programme de coopération territoriale Espagne-France-Andorre, POCTEFA, financé par le Fonds européen de développement régional. Cet effort démontrera le caractère transposable de notre méthode et sa capacité à favoriser la constitution des corpus arborés pour les langues peu dotées en général.

8. Bibliographie

Abeillé A., Clément L., Reyes R., « Talana annotated corpus : the first results », *Actes de The First Conference on Linguistic Resources*, Granada, 1998.

Abeillé A., Clément L., Toussenel F., « Building a treebank for French », *Treebanks*, Springer, p. 165-187, 2003.

Agić Ž., Ljubešić N., « The SETimes.HR Linguistically Annotated Corpus of Croatian », *Actes de Ninth International Conference on Language Resources and Evaluation (LREC2014)*,

European Language Resources Association (ELRA), Reykjavik, Iceland, p. 1725-1727, 2014.

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

Agić Ž., Merkler D., « Three syntactic formalisms for data-driven dependency parsing of Croatian », *Actes de 16th International Conference on Text, Speech and Dialogue*, Springer, Pilsen, Czech Republic, p. 560-567, 2013.

Agić Ž., Ljubešić N., Berović D., « Lemmatization and morphosyntactic tagging of Croatian and Serbian », *Actes de 4th Biennial International Workkshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, Sofia, Bulgaria, p. 48-57, 2013a.

Agić Ž., Merkler D., Berović D., « Parsing Croatian and Serbian by using Croatian dependency treebanks », *Actes de Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 22-33, 2013b.

Alex B., Grover C., Shen R., Kabadjov M., « Agile corpus annotation in practice : An overview of manual and automatic annotation of CVs », *Actes de Fourth Linguistic Annotation Workshop*, Association for Computational Linguistics, Uppsala, Sweden, p. 29-37, 2010.

Artstein R., Poesio M., « Inter-coder agreement for computational linguistics », *Computational Linguistics*, vol. 34, n° 4, p. 555-596, 2008.

Bhat R. A., Sharma D. M., « A dependency treebank of Urdu and its evaluation », *Actes de Sixth Linguistic Annotation Workshop*, Association for Computational Linguistics, p. 157-165, 2012.

Boguslavsky I., Chardin I., Grigorieva S., Grigoriev N., Iomdin L. L., Kreidlin L., Frid N., « Development of a Dependency Treebank for Russian and its Possible Applications in NLP », *Actes de 3rd International Conference on Language Ressources and Evaluation (LREC2002)*, LREC, Las Palmas, Canary Islands, Spain, p. 852-856, 2002.

Bohnet B., « Very high accuracy and fast dependency parsing is not a contradiction », *Actes de 23rd International Conference on Computational Linguistics (COLING2010)*, Association for Computational Linguistics, Beijing, China, p. 89-97, 2010.

Bond F., Fujita S., Tanaka T., « The Hinoki syntactic and semantic treebank of Japanese », *Language Resources and Evaluation*, vol. 40, n° 3-4, p. 253-261, 2006.

Brants T., « Inter-annotator Agreement for a German Newspaper Corpus. », *Actes de 2nd International Conference on Language Ressources and Evaluation*, LREC, 2000.

Buchholz S., Marsi E., « CoNLL-X shared task on multilingual dependency parsing », *Actes de 10th Conference on Computational Natural Language Learning*, Association for Computational Linguistics, New York City, USA, p. 149-164, 2006.

Burga A., Mille S., Wanner L., « Looking behind the scenes of syntactic dependency corpus annotation : Towards a motivated annotation schema of surface-syntax in Spanish », *Actes de DepLing 2011*, p. 104-114, 2011.

Chiou F.-D., Chiang D., Palmer M., « Facilitating treebank annotation using a statistical parser », *Actes de First international conference on Human language technology research*, Association for Computational Linguistics, p. 1-4, 2001.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, vol. 20, n° 1, p. 37-46, 1960.

Collins M., Ramshaw L., Hajič J., Tillmann C., « A statistical parser for Czech », *Actes de 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, p. 505-512, 1999.

Forcada M. L., Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J. A., Sánchez-Martínez F., Ramírez-Sánchez G., Tyers F. M., « Apertium : a free/open-source platform for rule-based machine translation », *Machine translation*, vol. 25, n° 2, p. 127-144, 2011.

Fort K., Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus, PhD thesis, Université Paris-Nord-Paris XIII, 2012.

Fort K., *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*, John Wiley & Sons, 2016.

Fort K., Sagot B., « Influence of Pre-annotation on POS-tagged Corpus Development », *Actes de 4th Linguistic Annotation Workshop*, Association for Computational Linguistics, Uppsala, Sweden, p. 56-63, 2010.

Gesmundo A., Samardžić T., « Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification », *Actes de 8th Language Resources and Evaluation Conference (LREC 2012)*, p. 2103-2106, 2012.

Groß T., Osborne T., « The Dependency Status of Function Words : Auxiliaries », *Actes de 3rd International Conference on Dependency Linguistics (DepLing2015)*, Uppsala, Sweden, p. 111-120, 2015.

Hajič J., « Morphological tagging : Data vs. dictionaries », *Actes de 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics, p. 94-101, 2000.

Hajič J., « Complex corpus annotation : The Prague dependency treebank », *Insight into Slovak and Czech Corpus Linguistics. Veda Bratislava*p. 54-73, 2005.

Hajič J., Panevová J., Buráňová E., Urešová Z., Bémová A., « Annotations at analytical level. Instructions for annotators », *UK MFF ÚFAL, Praha, Czech Republic. URL http ://ufal. mff. cuni. cz/pdt2. 0/doc/manuals/en/a-layer/pdf/a-man-en. pdf (2012-03-18)*, 1999.

Halácsy P., Kornai A., Oravecz C., « HunPos : an open source trigram tagger », *Actes de 45th annual meeting of the ACL on interactive poster and demonstration sessions*, Association for Computational Linguistics, Prague, Czech Republic, p. 209-212, 2007.

Hovy E., Lavid J., « Towards a 'science'of corpus annotation : a new methodological challenge for corpus linguistics », *International journal of translation*, vol. 22, n° 1, p. 13-36, 2010.

Jakovljević B., Kovačević A., Sečujski M., Marković M., « A Dependency Treebank for Serbian : Initial Experiments », *Actes de International Conference on Speech and Computer*, Springer, Novi Sad, Serbia, p. 42-49, 2014.

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

Keith B. (ed.), *Encyclopedia of language and linguistics*, 2006.

Krstev C., *Processing of Serbian. Automata, Texts and Electronic Dictionaries*, Faculty of Philology of the University of Belgrade, 2008.

Krstev C., Vitas D., « Corpus and Lexicon-Mutual Incompletness », *Actes de Corpus Linguistics Conference*, Birmingham, UK, p. 14-17, 2005.

Krstev C., Vitas D., Erjavec T., « MULTEXT-East resources for Serbian », *Actes de 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004*, Erjavec, Tomaž and Zganec Gros, Jerneja, 2004.

Le Roux J., Sagot B., Seddah D., « Statistical parsing of Spanish and data driven lemmatization », *Actes d'ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, p. 55-61, 2012.

Ljubešić N., Klubička F., « {bs, hr, sr} WaC–web corpora of Bosnian, Croatian and Serbian », *Actes de 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden, p. 29-35, 2014.

Ljubešić N., Klubička F., Željko Agić, Jazbec I.-P., « New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian », *in* N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Actes de Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, France, 2016.

Marcus M. P., Marcinkiewicz M. A., Santorini B., « Building a large annotated corpus of English : The Penn Treebank », *Computational linguistics*, vol. 19, n° 2, p. 313-330, 1993.

Marton Y., Habash N., Rambow O., « Dependency parsing of Modern Standard Arabic with lexical and inflectional features », *Computational Linguistics*, vol. 39, n° 1, p. 161-194, 2013.

Mathet Y., Widlöcher A., Fort K., François C., Galibert O., Grouin C., Kahn J., Rosset S., Zweigenbaum P., « Manual corpus annotation : Giving meaning to the evaluation metrics », *Actes d'International Conference on Computational Linguistics*, p. 809-818, 2012.

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

Miletic A.*, « Annotation morphosyntaxique semi-automatique d'un corpus litéraire serbe »*, Master's thesis, Université Charles de Gaulle - Lille 3, 2013.

Miletic A., « Building a morphosyntactic lexicon for Serbian using Wiktionary », *Actes des Sixièmes Journées d'études Toulousaines (JéTou2017)*, Toulouse, France, p. 30-34, 2017.

Miletic A., Un treebank pour le serbe : constitution et exploitations, PhD thesis, Université de Toulouse Jean Jaurès, 2018.

Miletic A., Urieli A., « Non-projectivity in Serbian : Analysis of Formal and Linguistic Properties », *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling2017)*, Pisa, Italy, p. 135-144, 2017.

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

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

Pavlović-Lažetić G., Vitas D., Krstev C., « Towards full lexical recognition », *Text, Speech and Dialogue*, Springer, p. 179-186, 2004.

Pustejovsky J., Stubbs A., *Natural Language Annotation for Machine Learning : A guide to corpus-building for applications*, " O'Reilly Media, Inc.", 2012.

Sagot B., « DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German », *Actes de 9th International Conference Language Resources and Evaluation (LREC2014)*, Reykjavik, Iceland, 2014.

Sagot B., « Etiquetage multilingue en parties du discours avec MElt », *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, Paris, France, 2016.

Sajous F., Hathout N., Calderone B., « Glàff, un gros lexique à tout faire du français », *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, France, p. 285-298, 2013.

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

Seddah D., Chrupała G., Çetinoğlu Ö., Van Genabith J., Candito M., « Lemmatization and lexicalized statistical parsing of morphologically rich languages : the case of French », *Actes de NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Association for Computational Linguistics, p. 85-93, 2010.

Seddah D., Tsarfaty R., Kübler S., Candito M., Choi J., Farkas R., Foster J., Goenaga I., Gojenola K., Goldberg Y. *et al.*, « Overview of the SPMRL 2013 shared task : cross-framework evaluation of parsing morphologically rich languages », *Actes de Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, Association for Computational Linguistics, 2013.

Sennrich R., Kunz B., « Zmorge : A German Morphological Lexicon Extracted from Wiktionary », *in* N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Actes de Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

Skjærholt A., « Influence of preprocessing on dependency syntax annotation : speed and agreement », *Actes de 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 28-32, 2013.

Stanojčić Ž., Popović L., *Gramatika srpskog jezika*, Zavod za udžbenike, 2012.

Tellier I., Eshkol-Taravella I., Dupont Y., Wang I., « Peut-on bien chunker avec de mauvaises étiquettes POS ? », *Actes de TALN 2014*, p. 125-136, 2014.

Tesnière L., « Eléments de syntaxe structurale », 1959.

Thomas P.-L., « Serbo-croate, serbe, croate..., bosniaque, monténégrin : une, deux..., trois, quatre langues ? », *Revue des études slaves*, vol. 66, n° 1, p. 237-259, 1994.

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

Vergez-Couret M., Urieli A., « Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan », *Actes de l'atelier TALARE 2015*, 2015.

Vitas D., Krstev C., « Intex and Slavonic morphology », *INTEX pour la linguistique et le traitement automatique des langues, Presses Universitaires de Franche-Comté*p. 19-33, 2004.

Voormann H., Gut U., « Agile corpus creation », *Corpus Linguistics and Linguistic Theory*, vol. 4, n° 2, p. 235-251, 2008.

Xue N., Xia F., Chiou F.-D., Palmer M., « The Penn Chinese TreeBank : Phrase structure annotation of a large corpus », *Natural language engineering*, vol. 11, n° 02, p. 207-238, 2005.