Université Lille III
Département Lettres modernes

Deuxième année de Master 2 Sciences du langage
Spécialité Lexicographie, Terminographie et Traitement Automatique des Langues et des Corpus

ANNOTATION SEMI-AUTOMATIQUE EN PARTIES DU DISCOURS D'UN CORPUS LITTÉRAIRE SERBE

Mémoire préparé sous la direction de M. Antonio Balvet (Université de Lille 3, UMR 8163 *Savoirs, Textes, Langage*) et M. Dejan Stosic (Université d'Artois à Arras, Centre de recherche *Grammatica*)

Présenté et soutenu par Aleksandra Miletic

Année universitaire 2012/2013

TABLE DES MATIÈRES

I INTRODUCTION

Comment traduit-on en serbe les prépositions locatives de l'anglais ou du français ? Existe-t-il des contraintes contextuelles, stylistiques ou d'usage portant sur les différentes stratégies de traduction, i.e. via une marque casuelle seule (cf. *marcher dans les rues* vs. *hodati ulicama*, où *hodati* est l'infinitif du verbe 'marcher', et *ulicama* l'instrumental du pluriel du nom *ulica* 'rue') vs. une préposition et le cas exigé par cette préposition (cf. *go to the cinema* vs. *ići u bioskop*, où *ići* et l'infinitif du verbe 'aller', *u* la prépostion 'dans', 'à', et *bioskop* l'accusatif du singulier du nom *bioskop* 'cinéma') ? Pour répondre à cette question, qui touche aussi bien à la linguistique contrastive (morphologie et syntaxe comparées), qu'à la traductologie, il est bien sûr possible d'avoir recours à l'intuition linguistique. Mais force est de reconnaître que, une fois les exemples prototypiques énumérés, l'intuition reste un outil limité dès qu'il s'agit d'apporter des réponses précises, quantifiées, reposant sur une démarche explicite. Pour traiter cette question simple en apparence, une autre stratégie consiste à exploiter des textes traduits dans les trois langues, alignés phrase par phrase, et annotés en parties du discours, de manière à disposer d'une réelle base de données représentative de l'usage effectif. Autrement dit, il est nécessaire de disposer d'un corpus parallèle des langues en question.

Jusqu'à récemment, une telle ressource n'existait pas. Cependant, un corpus parallèle français-serbe-anglais a été développé en 2010 dans le cadre du projet Egide *Constitution du corpus plurilingue français – serbe – anglais.* Ce projet a été réalisé par l'Université d'Artois, l'Université Lille 3 et l'Université de Belgrade, sous la direction de M. Dejan Stošić (Université d'Artois, Centre de recherche *Grammatica*). Le corpus en question a le serbe pour langue pivot du corpus, ce qui signifie qu'il est constitué des textes littéraires serbes et leurs traductions en français et en anglais. Il est annoté structurellement et parallélisé au niveau de la phrase[1]. Il contient à présent 2 041 113 tokens (mots orthographiques), dont 661 772 proviennent des textes serbes, 876 921 des textes français et 502 420 des textes anglais. Le serbe étant une langue faiblement dotée en ce qui concerne les ressources linguistiques numériques, il est clair que ce corpus constitue une ressource de premier plan, aussi bien pour les études contrastives

[1] Les paragraphes et les phrases ont été identifiés, et le lien entre les phrases correspondantes dans le trois langues du corpus a été établi.

(linguistiques ou autres) de ces trois langue, qu'en tant que base pour le développement des outils du traitement automatique du langage pour le serbe. Néanmoins, ses possibilités d'exploitation sont encore limitées, vu qu'il ne dispose pas d'annotations linguistiques.

Afin d'optimiser cette ressource et d'augmenter son utilité pour la communauté scientifique, nous avons choisi d'effectuer l'annotation morpho-syntaxique du volet serbe du corpus cité. L'utilité immédiate pour le domaine linguistique mise à part, notre travail a permis d'obtenir une ressource de première importance pour le traitement automatique du serbe, En effet, nous avons élaboré un corpus de référence, annoté au niveau morpho-syntaxique, révisé manuellement, qui peut être le point de départ pour des expérimentations dans le domaine du TAL. Avant l'élaboration de notre corpus, le seul corpus de référence du serbe était celui développé dans le cadre du projet MULTEXT-East (Krsteva *et al.* 2004). Ce corpus contient 108 000 tokens provenant de la traduction serbe de l'ouvrage 1984 par George Orwell. Le corpus de référence que nous avons élaboré compte 157 000 tokens extraits des ouvrages serbes originaux : à l'avantage de la taille se joint la certitude que les textes du corpus sont représentatifs de la langue serbe.

Ce mémoire présente le travail effectué sur l'étiquetage morpho-syntaxique du volet serbe du corpus parallèle français-serbe-anglais. La suite de cette partie définit la problématique de cette tâche et présente l'étude bibliographique de la question, ainsi que la méthode adoptée. Le Chapitre II décrit les deux étapes de l'élaboration du corpus d'entraînement, à savoir la constitution du jeu d'étiquettes et l'étiquetage manuel du corpus. Le processus de la sélection de l'étiqueteur le mieux adapté à la tâche est présenté dans le Chapitre III, alors que les analyses quantitative et qualitative des performances des étiqueteurs choisis figurent dans le Chapitre IV. Enfin, le Chapitre V contient les conclusions et les perspectives pour le travail à venir.

I.1 Positionnement linguistique et technique

L'annotation de corpus peut être définie comme une valeur ajoutée ou comme un enrichissement de données (Véronis 2000). Le type d'informations apportées peut différer selon le type de corpus et l'usage auquel il est destiné : si on reprend la classification présentée dans (*idem*, p. 114), il peut s'agir d'annotation phonétique (transcription et prosodie), grammaticale (parties de discours et syntaxe), sémantique

(annotation des mots ou du discours) ou multilingue (annotation au niveau des mots ou à celui de la phrase). Selon (McEnery 2003), les formes d'enrichissement les plus répandues aujourd'hui sont la lemmatisation[2] et l'étiquetage catégoriel[3], suivi du parsing (annotation de la structure syntaxique). Les annotations sémantique, discursive et pragmatique se développent de plus en plus ; elles représentent toutefois un niveau de complexité d'annotation tel que peu de procédures automatiques sont actuellement disponibles.

L'objectif principal de toutes ces formes d'enrichissement est d'élargir le champ des applications possibles d'un corpus et d'en faciliter l'utilisation (ou la réutilisation). La lemmatisation permet d'exécuter des requêtes en utilisant la forme canonique d'un mot, ce qui permet d'extraire du corpus toutes les occurrences, indépendamment de leur forme fléchie. Grâce à l'étiquetage morpho-syntaxique, il est possible de restreindre la requête à une seule partie de discours, ou bien seulement aux mots portant un trait morpho-syntaxique spécifique, tel le genre masculin ou l'aspect indéfini. Si une annotation syntaxique du corpus a été effectuée, il est également possible d'identifier les occurrences d'une structure syntaxique donnée, indépendamment des mots particuliers par lesquels elle est instanciée. En fonction de la puissance du système de consultation du corpus, il peut être possible de combiner plusieurs de ces paramètres dans une seule requête. Cependant, comme il a déjà été mentionné, le corpus parallèle sur lequel porte ce mémoire ne dispose pas encore de cet enrichissement. Par conséquent, les possibilités de son exploitation ne sont pas optimales.

L'état actuel du corpus permet, par exemple, d'en tirer toutes les phrases contenant les occurrences du mot *passons* et la traduction de ces phrases en serbe et en anglais. Pourtant, il est impossible d'identifier par une seule requête toutes les formes fléchies du verbe *passer* ou de faire des requêtes plus formalisées, de type *passer + préposition*. Pour que cela soit faisable, il est nécessaire que le corpus soit lemmatisé (que chaque token soit liée à son lemme) et annoté morpho-syntaxiquement (que chaque token soit accompagné de l'indication de ses propriétés morpho-syntaxiques). Comme les trois langues représentées dans le corpus montrent des différences structurelles

[2] L'ajout de sa forme canonique à chaque mot du corpus.
[3] Cette forme d'étiquetage peut simplement comprendre une annotation en parties du discours, mais elle dénote le plus souvent un encodage des propriétés morphologiques du mot traité.

importantes[4], le résultat de ces traitements rendra possible des analyses contrastives diverses apportant un éclaircissement sur les différences dans le fonctionnement morpho-syntaxique du français, de l'anglais et du serbe.

La lemmatisation et l'annotation morpho-syntaxique permettront de repérer, par exemple, toutes les occurrences de *pass + préposition* et d'avoir, par le biais des traductions, les structures équivalentes en français et en serbe. Dans ce cas particulier, la comparaison avec le serbe surtout se montrerait intéressante : comme il s'agit d'une langue à déclinaisons qui exprime certaines fonctions syntaxiques des groupes nominaux par leur seule forme fléchie, il serait possible d'établir des correspondances entre des SP identifiés en anglais et en français et des valeurs des certains cas en serbe.

Il sera aussi envisageable d'identifier dans la partie française les occurrences du verbe *faire* suivies d'un nom non précédé d'un article, telles que *faire preuve*, *faire partie* ou *faire peur* et d'examiner les moyens qu'utilisent les deux autres langues pour exprimer le même contenu.

On pourra également sortir de la partie serbe du corpus les occurrences de la suite des tokens *adjectif démonstratif + adjectif possessif +* nom, une structure fréquente en serbe et qui n'existe ni en français ni en anglais et d'identifier et comparer les structures équivalentes dans ces deux langues.

Or, comme il a déjà été mentionné, pour que cela soit faisable, il est nécessaire de lemmatiser et d'annoter morpho-syntaxiquement les trois parties du corpus. Comme ce mémoire a pour son sujet l'annotation morpho-syntaxique, nous procéderons à donner une courte description de l'état de l'art dans ce domaine, pour présenter ensuite la méthode que nous avons choisie pour notre travail.

L'étiquetage morpho-syntaxique est une forme de l'annotation grammaticale qui consiste à enrichir le corpus en ajoutant des informations sur le comportement morphologique et syntaxique des mots du corpus. La quantité d'informations apportées peut varier : on peut se limiter à déterminer seulement la partie du discours, ou bien ajouter également des précisions plus ou moins détaillées sur les traits morpho-syntaxiques des mots traités (Véronis, 2000:114). Dans tous les cas, les informations ajoutées sont portées par les étiquettes (ou tags) : il s'agit des suites des lettres encodant

[4] Le serbe est une langue slave à morphologie flexionnelle riche et à ordre des constituants libre.

les informations ajoutées. L'ensemble des étiquettes utilisées dans l'annotation d'un corpus est appelé *jeu d'étiquettes* (angl. *tagset*) (Véronis 2000:114). Le nombre d'étiquettes employées sur un corpus dépend en premier lieu de la complexité morphologique de la langue en question. Par exemple, le corpus d'anglais américain *Penn Treebank* (Marcus *et al.* 1993) dispose d'un jeu de 48 étiquettes, alors que le corpus du bulgare, une langue à morphologie flexionnelle riche, *BulTreeBank* (Simov *et al.* 2002) emploie 680 étiquettes. Cependant, dans les cas où un étiquetage moins détaillé est acceptable, typiquement quand il n'est pas un objectif en lui-même, mais une base pour un autre traitement (telle l'annotation syntaxique), ou bien quand on veut assurer un taux de précision élevé, une simplification du jeu d'étiquettes est envisageable. Ceci est le cas de (Dojchinova et Mihov 2004), qui ont travaillé sur l'étiquetage morpho-syntaxique du bulgare : ils ont réduit leur jeu d'étiquettes initial de 946 à 40 étiquettes pour obtenir un résultat plus fiable.

Les premiers logiciels pour l'étiquetage catégoriel automatique étaient basés sur l'utilisation des règles d'annotation écrites manuellement. Un des premiers exemples de cette approche est le logiciel TAGGIT (Greene et Rubin 1971), qui atteignait la précision de 77% sur le *Brown Corpus* (corpus de référence pour l'anglais américain). L'étiqueteur de Brill en 1995 a apporté une amélioration importante de la précision : ce logiciel, utilisant un algorithme d'apprentissage de règles à partir du corpus (*Transformation Based Learning*, cf. (Brill 1995) obtenait des résultats de l'ordre de 96,9% (*idem*). Deux problèmes principaux ont été rencontrés dans l'utilisation de ce logiciel : premièrement, son exécution s'est montrée beaucoup moins rapide que celle des étiqueteurs basés sur les approches statistiques développés postérieurement ; en second lieu, l'induction automatique des règles à partir d'un corpus, suivie de leur correction manuelle entraînaient des temps de traitement très importants. La rapidité d'utilisation et la possibilité d'entraînement pour plusieurs langues sont les raisons principales pour lesquelles on favorise aujourd'hui les étiqueteurs stochastiques.

Quant à cet autre type d'étiqueteurs, leur fonctionnement consiste en deux étapes : celle d'entraînement et celle d'annotation (Agić *et al*. 2009). Dans la première étape, un corpus d'entraînement, doté d'une annotation manuelle, est parcouru par le logiciel. En utilisant des analyses statistiques, l'étiqueteur détermine le modèle linguistique de la langue en question, i.e. la probabilité d'occurrence de différentes suites des tags.

Certains étiqueteurs, tel TreeTagger (Schmid 1994), utilisent également un lexique - un inventaire contenant les formes fléchies et les étiquettes valides pour chacune de ces formes. Le modèle linguistique dérivé est ensuite utilisé dans l'étiquetage : si le logiciel rencontre une forme qui lui est inconnue (qui n'existe pas dans le lexique et qu'il n'a pas rencontrée dans le corpus d'entraînement), il utilise le modèle linguistique pour déterminer l'étiquette la plus probable pour la forme en question, en prenant en compte les étiquettes des tokens dans le contexte plus ou moins immédiat du mot traité. Le nombre des étiquettes, ainsi que le contexte considéré peuvent varier : TnT (Brants 2000) utilise deux étiquettes précédentes[5] , alors que la taille par défaut du contexte considéré par TreeTagger (Schmid 1994) est également 2, mais elle peut être spécifiée par l'utilisateur, d'une part, et elle est surtout déterminée de façon automatique par le logiciel, en fonction d'un score d'ambiguïté pour chaque mot, défini à partir du corpus d'apprentissage. Les deux logiciels cités analysent le contexte gauche du mot traité et effectuent, par conséquent, un étiquetage unidirectionnel[6] Cependant, des travaux plus récents mettent en œuvre des systèmes d'apprentissage bidirectionnel, ou le contexte de droite aussi bien que le contexte de gauche est utilisé pour la création des règles statistiques. On peut citer les travaux de (Shen *et al.* 2007) et (Toutanova *et al.* 2003). Cette propriété est importante pour les langues, telles que le serbe, dans lesquelles la tête d'un constituant peut se trouver à droite des mots dépendants.

En ce qui concerne les modèles statistiques intégrés dans les logiciels, on trouve une diversité importante dans la littérature, surtout parmi les travaux centrés sur l'anglais. Par exemple, (Ratnaparkhi 1996) a développé un système fondé sur le maximum d'entropie avec une précision de 96,6% ; TnT tagger présenté dans (Brants 2000) emploie un modèle de Markov caché avec le même résultat ; (Lafferty *et al.* 2001) se servent des champs conditionnels aléatoires pour obtenir une précision de 95,7%; (Collins 2002) a développé un modèle basé sur un perceptron (*averaged perceptron discriminative sequence model*) et a dépassé le seuil de 97% avec son résultat de 97,1%. Tous les systèmes cités utilisent l'ordre d'inférence de gauche à droite ; quant aux travaux de (Toutanova *et al.* 2003) et (Shen *et al.* 2007) cités ci-dessus, ils emploient respectivement les réseaux des dépendances cycliques (*cyclic dependency network*)

[5] Autrement dit, il considère les trigrammes des tokens, d'où vient son nom Trigrams'n'Tags.
[6] L'annotation s'effectue linéairement, de gauche à droite, et le choix de l'étiquette à attribuer au mot traité est basé sur le contexte qui précède le mot.

(atteignant la précision de 97,2%) et un système d'apprentissage guidé avec classification bidirectionnelle des séquences (97,3%).

Ce survol de l'état de l'art montre que la tâche de l'annotation en parties du discours n'est pas problématique dans le cas de l'anglais, et il en est de même pour le français, dans l'étiquetage duquel (Denis *et al.* 2009) atteignent la précision de 97,7%. Ceci est conditionné par le fait que ces langues disposent déjà de ressources linguistiques informatisées importantes (lexiques et corpus d'entraînement) et plusieurs logiciels de natures différentes ont été développés pour leur étiquetage et lemmatisation. En revanche, la précision moyenne de l'étiquetage morpho-syntaxique du serbe est environ 86% (Gesmundo et Samardžić 2012, Popović 2010). Ceci s'explique par deux facteurs.

En premier lieu, le serbe est une langue faiblement dotée en ressources linguistiques. Le seul corpus annoté du serbe librement disponible est celui développé dans le cadre du projet MULTEXT-East (Krsteva *et al.* 2004). Ce corpus contient 108 000 tokens annotés provenant de la traduction serbe de l'ouvrage *1984* de G. Orwell. Si l'on compare ce corpus avec le corpus French Treebank (Abeillé *et al.* 2000), qui compte 780 000 tokens, ou avec Penn Treebank (Marcus *et al.* 1993), un corpus d'anglais qui en compte 7 000 000, on voit que le corpus *1984*constitue un échantillon relativment limité de la langue serbe, de traduction à partir de l'anglais de surcroît.. Or, nous avons vu ci-dessus que les corpus d'entraînement jouent un rôle crucial dans le paramétrage des étiqueteurs statistiques : si le corpus ne représente pas bien le comportement morpho-syntaxique d'une langue, le modèle linguistique dérivé d'un tel corpus n'aura pas la qualité nécessaire pour effectuer un étiquetage fiable. En ce qui concerne les logiciels d'étiquetage automatique, le premier étiqueteur spécifiquement développé pour le traitement du serbe n'est apparu que très récemment (Gesmundo et Samardžić 2012). Il s'agit d'un système d'apprentissage guidé bidirectionnel basé sur celui de (Shen *et al.* 2007). La précision qu'il obtient est 86,65%, ce qui est bien en-dessous des résultats de l'étiquetage du français et de l'anglais.

Deuxièmement, le serbe est une langue à morphologie riche. Elle connaît trois personnes, deux nombres, trois genres et sept cas. Les parties de discours qui se déclinent systématiquement sont le nom, l'adjectif et le pronom ; de plus, certains nombres cardinaux peuvent également se décliner. Un nom peut avoir jusqu'à 12 formes différentes, un adjectif jusqu'à 36. Le système des formes verbales est très développé :

en fonction du temps, mode, personne et genre, un verbe peut avoir plus de 120 formes différentes. Un jeu d'étiquettes prenant en compte toutes les propriétés morphologiques et syntaxiques, ainsi que quelques traits sémantiques du serbe, a été développé dans le cadre du projet MULTEXT-East (Erjavec *et al.* 2004) : il compte 906 tags. En même temps, cette prolifération des formes fléchies conditionne un degré d'ambiguïté important entre les catégories grammaticales, notamment entre les adjectifs et les adverbes, les adjectifs et les verbes et entre les noms et les verbes. Il arrive donc que la seule forme du mot traité ne suffise pas au logiciel pour déterminer l'étiquette à attribuer : pour opérer une désambiguïsation, le logiciel est obligé d'analyser le contexte.

Toutefois, même si l'ordre typique des constituants est SVO, il connaît de nombreuses variations qui sont très fréquentes (Stanojčić et Popović 2011:366-376). Ceci est conditionné par le fait que certaines fonctions syntaxiques, surtout celles des groupes nominaux, sont encodées par la forme même que le mot prend (le plus souvent par la désinence casuelle) : l'ordre des constituants strict n'est donc pas nécessaire pour l'identification des fonctions syntaxiques. Cela résulte dans un ordre des constituants largement plus libre qu'en français ou en anglais, permettant à une catégorie grammaticale d'apparaître dans un nombre élevé des contextes différents. Par conséquent, il est possible que même l'analyse du contexte n'apporte pas une désambiguïsation fiable.

Comme le montrent les travaux de (Popović 2010) et (Gesmundo et Samardžić 2012), la précision de l'étiquetage par un étiqueteur entraîné sur le corpus *1984* (Krsteva *et al.* 2004, voir *supra*) reste en-dessous de 87%. (Popović 2010) teste sur ce corpus 5 étiqueteurs différents : TnT (Brants 2000), TreeTagger (Schmid 1994), Rule-based Tagger (Brill 1995), MXPOST (Ratnaparkhi 1996), SVMTool (Gimenez et Marquez 2000). Les meilleurs résultats ont été obtenus avec TnT : il a atteint la précision de 85,47%. (Gesmundo et Samardžić 2012) ont utilisé ce corpus pour tester leur logiciel BTagger, adapté au traitement du serbe. Cependant, même les résultats de cet étiqueteur ne sont pas encourageants : la précision atteinte est 86,65%.

Comme nous l'avons vu, le serbe est une des langues dont les propriétés inhérentes rendent la tâche de l'annotation morpho-syntaxique difficile. Pourtant, les résultats présentés ci-dessus peuvent, à notre avis, être partiellement conditionnés par la nature du corpus d'entraînement. Le corpus 1984 compte 106 000 tokens annotés avec 906

étiquettes. Si la taille du corpus ne permet pas de bien représenter le jeu d'étiquettes, la dérivation des probabilités par le logiciel devient moins fiable (Schmid 1995). Par ailleurs, il ne s'agit pas d'un texte originalement écrit en serbe, mais d'une traduction. On peut donc également mettre en question la qualité du texte du corpus : une traduction risque d'être influencée par la langue source et, par conséquent, de ne pas être représentative de la langue cible (Xiao et McEnery 2002). Cette intuition est corroborée par le travail de (Utvić, 2011). A notre connaissance, c'est le seul travail sur l'étiquetage du serbe qui utilise un corpus autre que *1984*. Ici, TreeTagger a été entraîné sur un corpus des textes serbes de 1 000 000 de mots annoté avec un jeu d'étiquettes minimal comptant 16 tags qui n'encodent que la partie de discours. Le logiciel a atteint la précision de 96,57%. Ceci indique qu'un rapport plus favorable entre la taille du corpus d'entraînement et la taille du jeu d'étiquettes peut apporter une amélioration importante dans la précision de l'étiquetage. Pourtant, le jeu d'étiquettes utilisé est très peu informatif. Il est donc nécessaire de trouver un équilibre entre ces deux possibilités.

Compte tenu de ces résultats, nous avons décidé de créer un nouveau corpus d'entraînement et de construire un jeu d'étiquettes de taille modérée pour son annotation. Ce corpus, que nous avons intitulé *REF1*, est extrait de la partie serbe du corpus parallèle français-serbe-anglais et compte 101 000 tokens. Le jeu d'étiquettes que nous avons définis contient 45 tags qui encodent la catégorie et la sous-catégorie grammaticale du mot, ainsi que quelques propriétés morphologiques pour les adjectifs et les adverbes. Une fois cette étape terminée, trois étiqueteurs (TreeTagger, TnT et BTagger) ont été testés sur le corpus *REF1*. Les résultats obtenus indiquent que la méthode adoptée est efficace : TreeTagger a atteint la précision moyenne de 92,15%, TnT celle de 92,95%, et BTagger celle de 94,17%. Comme BTagger s'est montré le plus performant, il a été choisi pour l'annotation automatique d'un autre sous corpus, nommé *Bašta.* Ce corpus contient 56 000 tokens. L'annotation a ensuite été vérifiée manuellement et *Bašta* a été joint à *REF1* pour constituer le corpus d'entraînement final *REF2*, qui compte 157 000 tokens. Egalement, une analyse qualitative de l'étiquetage de *Bašta* par BTagger a été effectuée (cf. partie IV.2). Elle a permis d'identifier les points les plus problématiques dans l'étiquetage du serbe et d'envisager quelques possibilités de post-traitement dans l'objectif d'amélioration des résultats. La suite de ce mémoire présente la description détaillée de chacune des étapes citées.

II ÉLABORATION DU CORPUS D'ENTRAINEMENT ET DE TEST

II.1 Etiquetage morpho-syntaxique

Comme on a vu dans la partie I.1, l'annotation de corpus peut être définie comme une valeur ajoutée aux textes ou comme un enrichissement de données (Véronis 2000 :112). Il en existe plusieurs types selon la nature d'informations ajoutées : l'annotation phonétique, l'annotation grammaticale, l'étiquetage sémantique et l'étiquetage multilingue (*idem*, p. 114).

L'étiquetage morpho-syntaxique est une forme de l'annotation grammaticale qui consiste à ajouter des informations du niveau morphologique et syntaxique aux mots du corpus. La richesse des précisions apportées varie selon l'usage envisagé du corpus : sa forme la plus simple consiste à déterminer automatiquement la partie du discours, mais il peut également comprendre l'ajout d'information sur différents traits morphologiques des mots traités (*idem*).

Dans tous les cas, les informations ajoutées sont portées par les étiquettes ou les tags : il s'agit de suites des lettres encodant les précisions morphologiques et syntaxiques différentes. La structure des étiquettes peut largement varier. Elles peuvent être synthétiques et comprendre des informations de différents niveaux, comme c'est le cas avec les tags utilisés sur le corpus *Penn Treebank* (Marcus *et al.* 1993). Dans ce corpus, la catégorie des verbes dispose des étiquettes suivantes : VB pour la forme verbale de base (infinitif sans *to*), VBD pour le passé, VBG pour le gérondif et le participe présent, VBN pour le participe passé, VBP pour toutes les formes du présent sauf la troisième personne du singulier, et VBZ pour la troisième personne du singulier du présent. On voit que les tags encodent systématiquement le temps et le mode, mais restent incohérents en ce qui concerne la personne et le nombre. Il est visible que la définition des étiquettes a été véhiculée par les traits formels des verbes anglais, plutôt que par leurs catégories morphologiques.

Or, les étiquettes peuvent également avoir une structure plus modulaire et plus systématique, ce qui est le cas des tags employés dans les projets MULTEXT (Ide et Véronis 1994) et GRACE (Adda *et al.* 1998). Ici les étiquettes ont une structure régulière, comparable pour chaque partie du discours : la première lettre indique la partie du discours, et les autres encodent les valeurs des différents attributs pertinents pour la

catégorie grammaticale en question. Ainsi, les étiquettes GRACE pour les verbes contiennent 7 champs qui encodent la partie du discours, la distinction entre le verbe principal et le verbe auxiliaire, le mode, le temps, la personne, le nombre et le genre (Rajman *et al.* 1997). Le mot *parle* dans l'exemple *Il parle* porterait donc l'étiquette Vmip3s-, les différents éléments du tag ayant les valeurs suivantes : V - verbe, m - principal, i - indicatif, p - présent, 3 - troisième personne, s - singulier, et - pour le genre, vu que cette catégorie morphologique n'est pas pertinente pour cette forme verbale.

La position de chaque attribut dans l'étiquette est fixe, et dans le cas où l'un d'entre eux ne s'applique pas au mot en l'occurrence, cela est indiqué par un tiret. On obtient ainsi des étiquettes dont la longueur et la structure sont prévisibles, ce qui facilite largement leur traitement automatique.

L'ensemble des étiquettes utilisé dans l'annotation d'un corpus est appelé *jeu d'étiquettes* (angl. *tagset*) (Véronis 2000:114). La taille du jeu d'étiquettes dépend en premier lieu de la complexité morphologique de la langue en question. Par exemple, le corpus *Penn Treebank*, mentionné ci-dessus, dispose d'un jeu de 48 étiquettes, dont 36 pour les catégories morphologiques et 12 pour les symboles de ponctuation et de monnaie. D'un autre côté, *BulTreeBank* (corpus du bulgare présenté dans (Simov *et al.* 2002)) dispose de 680 étiquettes, alors que *Prague Dependency Treebank* (corpus du tchèque décrit dans (Hajič 1998)) en a 1400.

Le processus de l'annotation même consiste en deux étapes : l'entraînement du logiciel et l'étiquetage du corpus (Agić *et al.* 2009). L'apprentissage de l'étiqueteur s'effectue sur un corpus d'entraînement préalablement annoté, le plus souvent manuellement. A partir des couples forme-étiquette rencontrés, le logiciel dérive un modèle linguistique. Il se sert ensuite de ce modèle pour déterminer quelle étiquette doit être attribuée à chaque token du corpus en train de développement.

II.2 Cas des langues à morphologie flexionnelle riche

La précision de l'étiquetage morpho-syntaxique atteint aujourd'hui des valeurs très élevés : dans (Véronis 2000), on cite le seuil de 95% comme une valeur standard de la précision des étiqueteurs, alors que les approches les plus récentes montent jusqu'à 97% (Shen *et al.* 2007). Or, la plupart des expérimentations sur la précision de l'étiquetage morpho-syntaxique est centrée sur l'anglais ; les résultats dans le traitement des langues à morphologie flexionnelle plus complexe sont aujourd'hui encore

largement inférieurs à ceux cités ci-dessus (Agić *et al.* 2009, Popović 2010). On cite le plus souvent deux raisons principales de ce phénomène.

Tout d'abord, ayant des distinctions morpho-syntaxiques plus nombreuses, ces langues demandent un jeu d'étiquettes beaucoup plus important que les langues à morphologie flexionnelle réduite : comme nous venons de le voir, l'anglais peut être étiqueté à l'aide de 48 étiquettes alors que les jeux d'étiquettes pour le bulgare et le tchèque utilisent 680 et 1400 tags, respectivement. Cette prolifération des formes conditionne des cas d'ambiguïtés fréquents et qui ne peuvent pas être résolus sans l'analyse du contexte. Or, l'ordre des mots dans les langues à flexion riche connaît beaucoup moins de contraintes qu'en anglais ou en français : comme certaines fonctions syntaxiques, surtout celles des groupes nominaux, sont encodées par la forme même que le mot prend (le plus souvent par la désinence casuelle), l'ordre des constituants strict n'est pas nécessaire pour l'identification des fonctions syntaxiques. Par conséquent, il est possible que même l'analyse du contexte n'apporte pas une désambiguïsation fiable.

A ces causes inhérentes au système de la langue se joint le fait que les langues en question sont souvent sous- ou pauvrement dotées et ne disposent pas de grands corpus d'entraînement annotés manuellement (Agić *et al.* 2009, Резникова 2008). Il est fort probable que les corpus petits ou de taille moyenne, accompagnés de jeux d'étiquettes larges et de modèles linguistiques complexes, ne permettent pas un apprentissage optimal pour les étiqueteurs.

Pour atteindre la précision de 96%, il est le plus souvent nécessaire de recourir à différentes stratégies de compensation : (Hajič *et al.* 2001) ont combiné le modèle de Markov caché avec les règles grammaticales pour arriver à une précision de 95,2% en utilisant un jeu qui compte plus de 1400 étiquettes sur le corpus *Prague Dependency Treebank*. Dans l'étiquetage de l'islandais, (Dredze et Wallenberg 2008) ont eu une précision de 92,1% avec un jeu de 639 étiquettes en dissociant l'étiquetage en deux étapes, dont la première a consisté à déterminer les parties du discours de base, et la seconde à ajouter des spécifications morpho-syntaxiques plus détaillées. Dans le cas de bulgare, on recourt à la réduction du jeu d'étiquettes : (Dojchinova et Mihov 2004) ont réduit leur jeu d'étiquettes initial de 946 à 40, ce qui leur a permis d'avoir une précision de 95,5% avec le Brill tagger et 98,4% avec les règles grammaticales écrites manuellement.

Vu la nature plurilingue du corpus sur lequel nous avons travaillé, notre tâche avait deux impératifs : non seulement atteindre une qualité satisfaisante de l'étiquetage, mais aussi, rendre le jeu d'étiquettes pour le serbe comparable à ceux employés pour le français et l'anglais. Pour répondre à ces deux exigences, une analyse contrastive des morphologies de l'anglais, du français et du serbe a été effectuée dans le but d'identifier les différences, mais aussi les points communs de ces trois systèmes morpho-syntaxiques Nous avons ensuite étudié les jeux d'étiquettes disponibles pour le serbe, d'une part, et de l'autre nous avons analysé comment les morpho-syntaxes française et anglaise sont reflétés par les jeux d'étiquettes choisis pour l'étiquetage de ces deux langues[7]. Ces analyses, présentées dans les sections suivantes de ce mémoire, ont permis de définir les principes de base de la constitution du jeu d'étiquettes pour le serbe

II.3 Définition du jeu d'étiquettes

II.3.1 Courte présentation de la morpho-syntaxe de l'anglais, du français et du serbe

L'anglais étant une langue germanique, le français une langue romane, et le serbe une langue slave, le fait que leurs fonctionnements morphologiques et syntaxiques diffèrent n'étonne pas. Pourtant, en tant que langues indo-européennes, les trois langues partagent également des ressemblances importantes. L'objectif de cette partie est de rendre compte des propriétés partagées par ces langues, ainsi que de leurs points de divergence, en prenant le serbe pour langue pivot.

Comme les trois traditions grammaticales distinguent les mêmes parties du discours principales (nom, verbe, adjectif, adverbe), on procédera par classe grammaticale, en donnant tout d'abord une présentation introductive du serbe.

II.3.1.1 Description générale de la langue serbe

La différence principale entre le serbe d'une part et le français et l'anglais de l'autre repose dans la richesse de la morphologie flexionnelle. Alors que l'anglais est proche d'une langue isolante avec un petit nombre d'affixes flexionnels, le serbe est une langue slave avec une flexion très développée. Le système verbal du serbe peut être comparé à celui du français, mais, à la différence de cette langue et autres langues romanes, le domaine nominal dispose également d'une multitude de formes fléchies.

[7] Il s'agit des jeux d'étiquettes de TreeTagger par défaut pour ces deux langues (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ . Dernier accès : 01 août 2013).

Les mots du groupe nominal sont marqués pour le nombre (singulier ou pluriel), le genre (masculin, féminin ou neutre) et le cas (nominatif, génitif, datif, accusatif, vocatif, instrumental et locatif), avec l'adjectif ayant aussi la catégorie de l'« aspect » (défini ou indéfini). En fonction des paramètres énumérés, un nom peut avoir jusqu'à 14 formes différentes, un adjectif jusqu'à 36. Le nom, l'adjectif et le pronom se déclinent systématiquement, alors que certains nombres cardinaux (de un à quatre) peuvent se décliner aussi.

Le système des formes verbales est très développé : en fonction du temps, mode, personne et genre, un verbe peut avoir plus de 120 formes différentes.

L'ordre typique des constituants est SVO, mais il connaît de nombreuses variations qui sont très fréquentes. Deux autres spécificités par rapport à l'anglais et le français résident dans le fait qu'elle n'a pas d'article et qu'il s'agit d'une langue *pro-drop*. Cette dernière propriété signifie que le sujet de la phrase peut ne pas avoir une réalisation syntaxique. Ainsi, la phrase

On	čita	novine.
pron.per. 3p. sg. m.	3p. sg. prés.	nom fém.
il	lit	journal

'Il lit le journal'

Exemple glosé 1: Sujet réalisé

peut également avoir la forme suivante :

Čita	novine.
3p. sg. prés.	nom fém.
lit	journal

'Il lit le journal'

Exemple glosé 2: Sujet non réalisé

Une description plus détaillée des traits principaux du serbe sera donnée dans la suite.

II.3.1.2 Nom

La morphologie flexionnelle des noms en français, et surtout en anglais, est réduite. Le nom anglais connaît le nombre, mais pas le genre ni le cas (Blevins 2006:507). La grande majorité des noms construisent la forme du pluriel selon le schéma régulier (suffixation en *-s*), les exceptions étant les noms au pluriel irrégulier (i.e. *mouse - mice*, *child - children*), et les noms empruntés (i.e. *criterion - criteria*, *index - indices*) (*idem*, p. 512).

Quant au nom français, il est marqué pour le nombre (singulier ou pluriel), mais aussi pour le genre (masculin ou féminin) (Riegel *et al.* 2009:320). Le pluriel est le plus souvent marqué par le suffixe *-s* (*mur - murs*, *garçon - garçons*), remplacé par *-x* dans certains contextes phonétiques (*tuyau - tuyaux*, *manteau - manteaux*, *hibou - hiboux*). Il en existe également qui ont une marque de pluriel synthétique : *cheval - chevaux, travail - travaux*, *œil - yeux* (*idem*, p. 333). Les noms désignant des entités non - animées ont un genre arbitraire : *la voiture*, *le lit* etc. En revanche, les noms qui dénotent des référents animés exhibent une distinction de genre qui correspond en général à la distinction des sexes. Dans la plupart des cas, le genre féminin se construit par un ajout ou une modification d'un suffixe : *rival - rivale*, *vendeur - vendeuse*, mais il peut également être marqué par une différence lexicale avec le genre masculin : *homme - femme*, *cheval - jument* (*idem*, p. 330).

Les propriétés morphologiques du nom serbe sont largement plus développées : on distingue les mêmes deux nombres qu'en français et en anglais (*kuća* 'maison', sg. - *kuće* 'maisons', pl. ; *noć* 'nuit', sg. - *noći* 'nuits', pl.), alors qu'aux genres masculin et féminin se joint le genre neutre (*muškarac* 'homme', m. - *žena* 'femme', f. - *dete* 'enfant' n.). La différence la plus saillante reste la déclinaison : le nom serbe connaît sept cas : nominatif, génitif, datif, accusatif, vocatif, instrumental et locatif. Ces formes sont porteuses de sémantismes et de fonctionnements syntaxiques différents : le nominatif marque typiquement la fonction du sujet et de l'attribut de sujet, l'accusatif celle de l'objet direct, le datif et le locatif celle de l'objet indirect (Stanojčić et Popović 2011:78-79). La déclinaison des noms féminins qui se terminent par *-a* est donnée dans le Tableau 1.

Genre/Cas	Singulier	Pluriel
Nominatif	žen**a**	žen**e**
Génitif	žen**e**	žen**a**
Datif/Locatif	žen**i**	žen**ama**
Accusatif	žen**u**	žen**e**
Vocatif	žen**o**	žen**e**
Instrumental	žen**om**	žen**ama**

Tableau 1: Déclinaison des noms féminins en -a

Ce Tableau illustre un de 4 modèles principaux de déclinaison en serbe. Les catégories du genre, du nombre et de cas y sont fortement liées : l'appartenance d'un nom à une déclinaison dépend de la terminaison qu'il exhibe au nominatif, mais aussi de son genre (i.e. les noms masculins qui se terminent par une consonne appartiennent à la déclinaison I, alors que les noms qui se terminent par une consonne mais qui sont de genre féminin suivent le modèle de déclinaison IV (Stanojčić et Popović 2011:82,91)), et la forme du pluriel dépend du modèle de déclinaison (les noms de la déclinaison III forment le nominatif du pluriel en *-e :* (*žena* 'femme' - *žene* 'femmes', ceux de la déclinaison IV en *-i* : *ljubav* 'amour'- *ljubavi* 'amours', etc.)

Ce système flexionnel permet l'existence des groupes nominaux faits des noms juxtaposés, tel

ljubav	majke
n.f. nom. sg.	n.f. gén.sg.
amour	mère

'amour d'une / de la mère'

ou bien :

masaža	kamenjem
n.f. nom. sg.	n.n. gén.sg.
massager	pierre

'massage à pierre'

Le deuxième nom de la suite se trouve dans une forme fléchie, définissant un rapport avec le premier nom. Dans le premier cas, il s'agit d'une relation de provenance, marquée par le génitif du nom *majka*, alors que dans le deuxième exemple l'instrumental du nom *kamenje* indique l'instrument employé dans la réalisation du processus désigné par le premier nom.

Cette propriété du serbe rend également possible un ordre des constituants largement plus libre qu'en français ou en anglais. Considérons la phrase *Ivan kupuje knjigu*, avec l'ordre des constituants habituel SVO :

Ivan	kupuje	knjigu.
N nom.sg.	V prés. 3e p. sg.	N acc.sg.
Ivan	achète	livre

'Ivan achète un/le livre.'

Exemple glosé 3: Ordre des constituants SVO

Cet ordre peut être facilement transformé en OVS sans que le sens change[8] :

Knjigu	kupuje	Ivan
N acc.sg.	V prés. 3e p. sg.	N nom.sg.
Livre	achète	Ivan

'Ivan achète un/le livre.'

Exemple glosé 4: Ordre des constituants OVS

Le sens de la phrase reste inchangé grâce au fait que les fonctions syntaxiques sont indiqués par les cas (le nominatif étant typiquement le cas du sujet, l'accusatif celui de l'objet direct). Les ordres VOS (*Kupuje knjigu Ivan*) et VSO (*Kupjue Ivan knjigu*) sont tout aussi possibles et correspondent chacun à une thématisation différente de la phrase.

II.3.1.3 Verbe

Un verbe anglais a typiquement 4 formes simples : la base verbale (i.e. *talk*), la troisième personne du singulier du présent (*talks*), le participe présent/le gérondif (*talking*), et le participe passé/prétérite (*talked*). Ce nombre peut monter jusqu'à 5 pour les verbes dits 'forts' dont les formes du participe passé et du prétérite ne coïncident pas (i.e. pour le verbe *eat*, la forme du participe est *eaten*, alors que celle du prétérite est *ate*) (Blevins 2006:516). La distinction du genre dans le paradigme verbal n'existe pas, alors que le marquage du nombre et de la personne est réduit : il n'existe qu'au présent, avec la forme en *-s* (*idem*, p. 530). Les autres catégories morphologiques, tel l'aspect progressif ou perfectif, le temps futur, ou le mode passif, sont exprimés par les formes composées, i.e. par la combinaison de différentes formes des verbes auxiliaires et des participes et de l'infinitif (*idem*, p. 519).

Le verbe français exprime plus systématiquement la personne et le nombre (Riegel *et al.* 2009:441), et même le genre peut être marqué dans des formes verbales composées (*idem*, p. 501). Les marques du temps et du mode peuvent être exprimées par des

[8] Une modification pareille de l'ordre des constituants sert à obtenir une autre thématisation de la phrase (Stanojčić et Popović 2011:367).

suffixes dans les formes simples (*idem*, p. 440), ou bien par des formes verbales composées (*idem*, p. 450). Selon le résumé du marquage de la personne et du nombre donné dans (Riegel *et al.* 2009:441), la première et la deuxième personne du pluriel sont toujours marquées par *-ons* et *-ez*, respectivement. La désinence de la troisième personne du pluriel peut être *-t* (*fon-t*), *-nt* (*chante-nt*), *-ent* (*chantai-ent*), ou *-rent* (*fini-rent*). La deuxième personne du singulier est marquée par *-s* (*chante-s*,, *parlai-s*), ou par *-x* (*veu-x*,). La première personne peut avoir la désinence *-s* (*fini-s, parlai-s*), *-x* (*veu-x*), *-ai* (*parler-ai*), ou bien ne pas être marquée (*parle-*). Ces désinences se combinent avec une marque de temps : *-er* pour le futur et le conditionnel présent, *-ai* pour l'imparfait ; ou bien se soudent simplement à la base verbale, comme dans le cas du présent ou du passé simple (*idem*, p. 440).

Le système verbal serbe est assez proche du système français : les verbes portent les marques de la personne et du nombre, ainsi que du temps et du mode (Stanojčić et Popović 2011:109). On distingue 9 formes verbales simples (présent, imparfait, aoriste, impératif, participe présent, participe passé, participe passif, participe actif et infinitif) et 4 formes composées (parfait, plus-que-parfait, futur antérieur, conditionnel). Le futur a deux paradigmes et peut être aussi bien une forme composée qu'une forme simple (Stanojčić et Popović 2011:120-128). Les verbes serbes ont deux radicaux : radical infinitival et radical du présent. Selon la terminaison de ces deux formes, les verbes sont classifiés dans un de sept modèles de conjugaison (*idem*, p. 114). A titre d'illustration, nous donnons ici la conjugaison du verbe *raditi* 'faire' au présent, parfait et futur. Pour sa conjugaison complète, voir l'Annexe 1..

présent	personne	singulier	pluriel
	1°	radim	radimo
	2°	radiš	radite
	3°	radi	rade
futur	personne	singulier	pluriel
	1°	ću raditi / radiću	ćemo raditi / radićemo
	2°	ćeš raditi / radićeš	ćete raditi / radićete

	3°	će raditi / radiće		će raditi / radiće
parfait	personne	genre	singulier	pluriel
	1°	masculin	sam radio	smo radili
		féminin	sam radila	smo radile
		neutre	sam radilo	smo radila
	2°	masculin	si radio	ste radili
		féminin	si radila	ste radile
		neutre	si radilo	ste radila
	3°	masculin	je radio	su radili
		féminin	je radila	su radile
		neutre	je radilo	su radila

Tableau 2: Conjugaison du verbe *raditi* 'travailler'

Ces trois temps verbaux, avec le conditionnel et l'impératif, sont les seuls à être en usage actif. Cela est possible grâce au système très développé de l'aspect verbal : les verbes en serbe peuvent être de l'aspect perfectif, imperfectif, ou bien ils sont bi-aspectuels (Stanojčić et Popović 2011:109-110). Les verbes imperfectifs ont leurs correspondants perfectifs, dérivés le plus souvent par préfixation : *čitati* 'lire'- *pročitati* 'avoir lu', *učiti* 'apprendre' - *naučiti* 'avoir appris'. Cette distinction minimalise le besoin d'avoir plusieurs temps du passé pour marquer la différence aspectuelle. Des analyses plus extensives du système aspectuel serbe peuvent être trouvées dans les travaux de P.-L. Thomas, notamment dans (Thomas 1993) et (Thomas 1998).

De point de vue de la syntaxe, une différence importante existe entre les formes verbales composées en français et en anglais d'une part et en serbe de l'autre : dans la phrase serbe, le participe et le verbe auxiliaire peuvent se trouver dans l'ordre inverse (participe suivi de l'auxiliaire). Si le sujet d'une forme composée est omis, la forme du participe occupe la position initiale dans la phrase. Cette permutation est conditionnée par le fait que le verbe auxiliaire, étant une forme clitique, ne peut pas se trouver au début d'un groupe prosodique. Ainsi, la phrase

On	je	kupio	knjigu.
PRO 3p.nom.sg.m.	VA prés.3p.sg.	V PP sg.m.	N acc.sg.
Il	est	acheté	livre

Il a acheté le livre.

Exemple glosé 5

dans sa variante sans sujet exprimé devient

Kupio	Je	knjigu.
PP sg.m.	VA prés.3p.sg.	N acc.sg.
acheté	est	livre

Il a acheté le livre.

Exemple glosé 6

la phrase **Je kupio knjigu* étant agrammaticale.

II.3.1.4 Adjectif

Les adjectifs anglais ne connaissent ni le genre, ni le nombre. Leur seule catégorie flexionnelle est celle de comparaison, pour les adjectifs qualificatifs monosyllabiques et certains adjectifs dissyllabiques : *smart - smarter - smartest*, *pretty - prettier - prettiest*. (Blevins 2006:523). En revanche, la plupart des adjectifs qualificatifs français ont une comparaison analytique avec *plus* ou *moins* (hormis *bon* et *mauvais*), mais ils distinguent le genre et le nombre (Riegel *et al.* 2009:597)[9]. Le genre féminin est en général signalé par l'ajout du suffixe *-e* : *vrai - vraie*, *dur - dure*, qui peut être accompagné par des modifications phonétiques : *sec - sèche*, *heureux* - heureuse (*idem*, pp. 606-607). Le marquage du nombre pour le genre masculin est comparable à celui des noms : le suffixe est le plus souvent *-s* (*bon - bons*, *vert - verts*), avec quelques cas de figure qui utilisent *-x* (*beau - beaux*, *hébreu - hébreux*), et certains adjectifs en *-al* qui exhibent une modification de la terminaison (*général - généraux*, *brutal - brutaux*). Le genre féminin a systématiquement le pluriel en *-s*. (*idem*, pp. 608-609).

A la différence de l'adjectif français, l'adjectif serbe distingue trois genres - masculin, féminin et neutre, et deux nombres - singulier et pluriel. Il connaît les mêmes sept cas que le nom et il est également marqué par l'aspect adjectival, qui peut être défini ou indéfini (Stanojčić et Popović 2011: 94). Le sémantisme de cette distinction est proche

[9] Cette remarque ne s'applique pas aux adjectifs relationnels qui ne connaissent pas la comparaison (Riegel *et al.* 2009:598).

de l'opposition entre l'article défini et l'article indéfini en anglais ou en français. Considérons l'exemple suivant :

Ivan	kupuje	plav	džemper.
N nom.sg.	V prés. 3p.sg.	Adj. acc. m.sg.ind.	N acc.sg.
Ivan	achète	bleu	pull

'Ivan achète un pull bleu.'

Exemple glosé 7: Aspect adjectival indéfini

L'utilisation de l'aspect indéfini de l'adjectif *plav* 'bleu' conditionne une interprétation non-déterminée du nom *džemper* 'pull', ce qui est également visible dans la traduction ('un pull bleu'). Si, en revanche, on avait employé l'aspect défini *plavi*, la phrase *Ivan kupuje plavi džemper* aurait le sens *Ivan achète le pull bleu*.

L'aspect adjectival est morphologiquement marqué seulement au singulier des genres masculin et neutre, alors que le singulier du féminin et le pluriel des trois genres marquent cette opposition par des moyens prosodiques. Par exemple, la forme de l'accusatif du genre féminin de l'adjectif *lep* 'beau' est *lepa* pour les deux aspects adjectivaux, mais l'accent porté par la voyelle *e* est long ascendant dans l'aspect indéfini, et long descendant dans l'aspect défini. Il faut encore souligner que le paradigme de l'aspect indéfini des genres masculin et neutre disparaît de l'usage actif : seule la forme du nominatif de l'indéfini est préservée ; pour les autres cas, on utilise les formes de l'aspect défini.

Le Tableau 2 illustre la réalisation des catégories du genre, nombre, cas et aspect de l'adjectif *lep* 'beau'.

Singulier			
Indéfini	**Masculin**	**Neutre**	**Féminin**
	'un beau chemin'	'un beau champ'	'une belle rue'
Nom.	lep (put)	lepo (polje)	lepa (ulica)
Gén.	lepa (puta)	lepa (polja)	lepe (ulice)
Dat./Loc.	lepu (putu)	lepu (polju)	lepoj (ulici)
Acc.	lep (put)	lepo (polje)	lepu (ulicu)

Voc.	lepi (putu)	lepo (polje)	lepa (ulico)
Instr.	lepim (putem)	lepim (poljem)	lepom (ulicom)
Défini	**Masculin**	**Féminin**	**Neutre**
	‘le beau chemin’	‘le beau champ	‘la belle rue’
Nom.	lepi (put)	lepo (polje)	lepa (ulica)
Gén.	lepog (puta)	lepog (polja)	lepe (ulice)
Dat./Loc.	lepom (putu)	lepom (polju)	lepoj (ulici)
Acc.	lepi (put)	lepo (polje)	lepu (ulicu)
Voc.	lepi (putu)	lepo (polje)	lepa (ulico)
Instr.	lepim (putem)	lepim (poljem)	lepom (ulicom)
Pluriel			
	Masculin	**Neutre**	**Féminin**
	‘de/les beaux chemins’	‘de/les beaux champs’	‘de/les belles rues’
Nom.	lepi (putevi)	lepa (polja)	lepe (ulice)
Gén.	lepih (puteva)	lepih (polja)	lepih (ulica)
Dat./Loc.	lepim (putevima)	lepim (poljima)	lepim (ulicama)
Acc.	lepe (puteve)	lepa (polja)	lepe (ulice)
Voc.	lepi (putevi)	lepa (polja)	lepe (ulice)
Instr.	lepim (putevima)	lepim (poljima)	lepim (ulicama)

Tableau 3: Déclinaison de l'adjectif serbe *lep* ‘beau’

Tout comme en anglais ou en français, l’adjectif en serbe a deux positions typiques dans la phrase : soit il a le rôle de l’épithète et se trouve alors antéposé au nom qu’il

détermine (*lepa žena* 'belle femme', où *lepa* est l'adjectif *belle*, et *žena* le nom *femme*), soit il exerce la fonction de l'attribut du sujet et figure derrière le verbe attributif (*Žena je lepa* 'La femme est belle', *žena* étant le nom *femme*, *je* le verbe attributif *est*, et *lepa* l'adjectif *belle*). Cependant, les deux fonctions permettent des variations syntaxiques. Un attribut épithète peut être séparé de son nom par d'autres constituants :

Lepu	sam	kuću	kupila.
ADJ acc.sg.f.	VA prés.1p.sg.	N acc.sg.	V ppas. act.
belle	suis	maison	acheté

'J'ai acheté une belle maison.'

Exemple glosé 8

De même, la construction attributive permet des variations importantes. Ainsi, l'adjectif et le nom peuvent échanger leur positions de sorte à avoir *Lepa je kuća* (lit. 'Belle est maison') au lieu de *Kuća je lepa* (lit. 'Maison est belle'). La présence d'un constituant adverbial permet même une distribution où l'adjectif et le nom se trouvent tous les deux après le verbe attributif :

Stvarno	je	lepa	kuća.
adv.	ver.	adj.	nom com.
vraiment	est	belle	maison

'La maison est vraiment belle.'

Exemple glosé 9

II.3.1.5 Pronom

La classe des pronoms est la seule partie du discours en anglais qui exhibe des traces d'une flexion casuelle. Il s'agit plus précisément des pronoms personnels et de certains pronoms relatifs. On y retrouve deux séries de formes : *I* vs. *me*, *we* vs. *us*, *she* vs. *her*, *who* vs. *whom*. Bien que la classification traditionnelle de la première série en tant que formes du nominatif, de la deuxième en tant que formes de l'accusatif peut être contestée (Blevins 2006:514), le fait qu'il s'agit d'une forme de flexion n'est pas remis en question.

On constate un phénomène comparable dans la morphologie du français : les pronoms personnels, hormis *nous* et *vous*, peuvent avoir des formes différentes en fonction de leur rôle syntaxique. Ainsi, les formes *je*, *tu*, *il*, *elle*, *ils* et *elles* peuvent occuper seulement la position du sujet. Pour la première et la deuxième personne du singulier, la fonction du COD et du COI est exercée par les mêmes formes *me* et *te*, alors que la troisième

personne fait la distinction entre ces deux rôles syntaxiques : au singulier, la forme du COD est *le* et *la* pour le masculin et le féminin respectivement, et le COI a la forme *lui* pour les deux genres. Au pluriel, il n'y a pas de distinction de genre : le COD a la forme *les*, et le COI la forme *leur*. On rencontre également une série des formes dites 'disjointes' (Riegel *et al.* 2009:367), capables de faire partie d'un sujet composé, d'être employés indépendamment, ainsi que de figurer dans un groupe pronominal. Il s'agit des formes *moi*, *toi*, *lui*, *elle*, *eux* et *elles*.

La classe des pronoms est l'une des plus complexes en serbe. On y distingue le plus souvent deux grandes sous-catégories : les pronoms nominaux et les pronoms adjectivaux (Stanojčić et Popović 2011:100). Le premier ensemble de formes a un véritable fonctionnement pronominal : ces formes remplacent les noms et les groupes nominaux. Le second groupe a un comportement adjectival : antéposés au nom, les pronoms adjectivaux forment des groupes nominaux avec lui. Les sous-catégories sémantico-syntaxiques de ces deux ensembles, telles que présentées dans (*idem*, pp. 100-101), sont données dans le Tableau 4.

Pronoms nominaux		Pronoms adjectivaux	
personnels	*ja* 'je', *ti* 'tu', *oni* 'ils'	**possessifs**	*moj* 'mon', *tvoj* 'ton', *njihov* 'leur'
interro-relatifs	*ko* 'qui', *šta* 'quoi'	**interro-relatifs**	*koji* 'quel/qui/lequel', *čiji* 'de qui'
indéfinis	*neko* 'quelqu'un, *nešto* 'quelque chose'	**indéfinis**	*neki* 'certain', *nečiji* 'de quelqu'un'
négatifs	*niko* 'personne', *ništa* 'rien'	**négatifs**	*nikoji* 'aucun', *ničiji* 'de personne'
généraux	*svako* 'tout le monde', *svašta* 'tout'	**généraux**	*svaki* 'chaque', *svačiji* 'de tout le monde'
pronom personnel pour toutes les personnes	*sebe, se* 'soi-même'	**démonstratifs**	*ovaj* 'ce' (proximal), *onaj* 'ce' (distal)

Tableau 4: Sous-catégorisation des pronoms

Quant aux catégories grammaticales exhibées par les pronoms en serbe, parmi les pronoms nominaux seuls les pronoms personnels distinguent la personne et le nombre, ainsi que le genre à la troisième personne. Il y en a 10 au total : *ja* 'je', *ti* 'tu', '*on* 'il', *ona*

'elle', *ono* - genre neutre de la troisième personne du singulier, *mi* 'nous', *vi* 'vous', *oni* 'ils', *one* 'elles', et *ona* - le neutre de la troisième personne du pluriel. Les autres pronoms nominaux ne connaissent ni la personne, ni le nombre, ni le genre. En ce qui concerne la déclinaison, celle des pronoms personnels est spécifique pour chaque pronom, alors que celle des autres groupes suit le même modèle. A titre d'exemple, on trouve la déclinaison du pronom de la première personne du singulier *ja* 'je' et du pronom indéfini *neko* 'quelqu'un' dans le Tableau 5.

ja 'je'	**Forme accentuée**	**Forme clitique**
Nominatif	ja	
Génitif	mene	me
Datif/Locatif	meni	mi
Accusatif	mene	mi
Vocatif	-	
Instrumental	mnome	mnom
neko 'quelqu'un'	**Forme**	
Nominatif	neko	
Génitif	nekoga	
Datif/Locatif	nekome	
Accusatif	nekoga	
Vocatif	neko	
Instrumental	nekime	

Tableau 5: Déclinaison des pronoms *ja* 'je' et *neko* 'quelqu'un'

Les pronoms adjectivaux possessifs sont marqués pour la personne et distinguent les deux nombres et les trois genres aux trois personnes : le nominatif de la forme pour la première personne du singulier *moj* 'mon' est donné dans le Tableau 6.

moj 'mon'	**Singulier**	**Pluriel**
Masculin	moj	moji
Féminin	moja	moje
Neutre	moje	moja

Tableau 6: Formes du nominatif du pronom possessif *moj* 'mon'

Les autres pronoms adjectivaux distinguent le nombre et le genre. La déclinaison des pronoms adjectivaux est illustrée par l'exemple de *neki* 'certain' dans le Tableau 7.

Singulier			Pluriel	
	Masculin/Neutre	Féminin	Masculin/Neutre	Féminin
Nominatif	neki	neka	neki/neka	neke
Génitif	nekog	neke	nekih	
Datif/Locatif	nekom	neku	nekim	
Accusatif	nekog/neko	neku	neke/neka	neke
Vocatif	neki	neka	neki/neka	neke
Instrumental	nekim	nekom	nekim	

Tableau 7: Déclinaison du pronom *neki* 'certain'

II.3.1.6 Déterminant

On ne cherchera pas ici à décrire la catégorie complexe qu'est le déterminant en français et en anglais, présentée en détails dans (Riegel *et al.* 2009:276-319) et (Aarts et Haegemann 2006:119-122). On remarquera seulement que les pronoms adjectivaux serbes décrits ci-dessus (cf. partie II.3.1.5) contiennent des sous-classes sémantiques identifiées comme déterminants en anglais et en français : les possessifs (*mon*, *ton*, *notre* en français ; *my*, *your*, *our* en anglais), les démonstratifs (*ce*, *cette*, *ces* en français ; *this*, *that*, *these*, *those* en anglais), les indéfinis (*certain* en français, *some* en anglais) etc. De même, leur comportement est comparable : ils sont antéposés aux noms qu'ils définissent et forment avec eux des groupes nominaux. Il existe pourtant une différence essentielle entre ces formes en serbe et la catégorie des déterminants : en serbe, ces formes ne sont pas obligatoires pour constituer un groupe nominal valide. En effet, les noms serbes peuvent fonctionner (et fonctionnent souvent) sans aucune détermination. C'est principalement pour cette raison que ces deux classes, celle des pronoms adjectivaux en serbe et celle des déterminants en anglais et en français, ne peuvent pas être considérés comme la même partie du discours.

II.3.1.7 Homonymie entre différentes parties du discours

La prolifération des formes causée par le système flexionnel riche du serbe conditionne l'existence de nombreux cas d'homonymie entre différentes parties du discours. Dans une certaine mesure, le français et l'anglais exhibent le même

phénomène : *être* est un verbe et un nom, ce qui est également le cas avec *decrease* en anglais ; le mot *mort* peut être un adjectif ou un nom, et il en est de même pour le mot anglais *dead*. Cependant, ces occurrences sont plus sporadiques qu'en serbe, qui connaît quelques cas de figure de recoupement systématique des paradigmes des différentes parties du discours. L'homonymie peut être due aux processus de dérivation (surtout la conversion d'une partie du discours vers une autre), ou bien au syncrétisme des formes.

La catégorie de l'adjectif en serbe exhibe un degré important d'homonymie avec les catégories d'adverbe et de verbe. Dans le premier cas, ce sont les adjectifs qualificatifs et les adverbes de manière qui sont touchés : la forme de cette sous-catégorie d'adverbes coïncide avec celle du genre neutre (ou, dans certains cas, du genre masculin) de l'adjectif qualificatif correspondant. Par exemple, la forme *dobro* est en même temps le genre neutre de l'adjectif *bon* (*dobro dete* 'bon enfant') et l'adverbe *bien* (*plivati dobro* 'nager bien'), alors que le mot *stoički* correspond au genre masculin de l'adjectif *stoïque* (*stoički otpor* 'resistance stoïque') et à l'adverbe *stoïquement* (*stoički trpeti* 'endurer stoïquement').

Le deuxième cas de figure concerne les adjectifs qualificatifs dérivé des formes du participe passé actif et passif par adjectivation. Ce type d'homonyme est également connu en français et en anglais : les adjectifs tels *intéressant* et *intéressé* en français, et *boring* et *bored* en anglais dérivent des participes passé et présent. Ainsi en serbe la forme *načinjen* peut être le participe passé passif du verbe *faire, causer*, et, dans ce cas-là, elle fait partie d'un temps composé : *Šteta je već bila načinjena* 'Les dégâts avaient déjà été faits'. Elle peut également être un adjectif et exercer la fonction d'épithète : *načinjena šteta* 'les dégâts faits'[10]. Il faut souligner que le participe, étant un nom verbal, ne se décline pas, mais il connaît les trois genres et les deux nombres et a, par conséquent six formes. Le Tableau 8 présente les formes du participe *polomljen* 'brisé'.

	Singulier	Pluriel
Masculin	polomljen	polomljeni
Féminin	polomljena	polomljene

[10] Noter que l'interprétation verbale de cet exemple n'est pas possible en serbe : alors qu'on français on peut avoir la construction *les dégâts faits par les cambrioleurs*, en serbe le syntagme équivalent n'est pas grammatical : **načinjena šteta od strane provalnika*. Pour que le complément d'agent (et, par conséquent, l'interprétation verbale) soit admissible, l'attribut doit être postposé au nom : *šteta načinjena od strane provalnika*.

Neutre	polomljeno	polomljena

Tableau 8: Formes du participe *polomljen* 'brisé'

Ces formes correspondent aux formes du nominatif de l'aspect indéfini de l'adjectif dérivé. L'adjectif lui-même se décline et, par conséquent, chacune des six formes citées est accompagnée de tout le paradigme selon le modèle de flexion présenté dans la partie II.3.1.4). On peut conclure que, si le recoupement du paradigme du participe avec celui de l'adjectif est total, l'inverse n'est pas vrai : ce n'est que le nominatif (et les cas dont la forme est homonyme du nominatif) de l'adjectif qui soient homonymes avec le participe correspondant.

Pour la catégorie des noms communs, les cas d'homonymie avec d'autres parties du discours les plus répandus concernent les adjectifs et les verbes. L'homonymie avec les adjectifs est conditionnée par un mécanisme de dérivation assez répandu : un adjectif et un nom commun partagent le même paradigme si le nom en question a été dérivé de l'adjectif qualificatif par substantivation. Ainsi, dans l'exemple *mrtvo lišće* 'feuilles mortes', le mot *mrtvo* représente une forme fléchie de l'adjectif *mrtav* 'mort', alors que dans le syntagme *enciklopedija* mrtvih 'encyclopédie des morts' il s'agit d'une forme du nom commun *mrtvi* 'les morts'.

Le recoupement des formes avec le paradigme verbal est plus sporadique : typiquement, le nominatif, l'accusatif ou le génitif (ou une autre forme fléchie) de certains noms communs coïncide avec la première ou la deuxième personne du singulier du présent du verbe ayant la même base dérivationnelle. Ainsi la forme *kazni* correspond en même temps au datif/locatif du singulier et au génitif du pluriel du nom *kazna* 'punition' (i.e. *Pričali su o kazni* 'Ils ont parlé de la punition') et à la troisième personne du singulier du présent et à la deuxième personne du singulier de l'impératif du verbe *kazniti* 'punir' (i.e. *Kazni ga* 'Punis-le').

L'homonymie est un paramètre important dans la tâche de l'étiquetage automatique car elle augmente le degré d'ambiguïté rencontré dans le corpus d'entraînement : dans le cas où une forme fléchie peut appartenir à plusieurs catégories différentes (cf. il existe plus qu'une étiquette valide pour cette forme), le logiciel doit avoir recours à un mécanisme supplémentaire pour décider quelle étiquette attribuer. L'analyse d'homonymie permet ainsi de prévoir quelles parties du discours seront les plus problématiques dans l'étiquetage.

II.3.2 Jeux d'étiquettes disponibles pour le serbe

Les spécifications morpho-syntaxiques pour la description de la langue serbe ont été définies dans le cadre du projet MULTEXT-East (Krstev *et al.* 2004). Ce jeu d'étiquettes est conforme aux principes de structuration des descriptions morpho-syntaxiques utilisés dans la totalité du projet (Erjavec *et al.* 2004). Il s'agit des étiquettes positionnelles, où la première lettre encode la partie du discours, alors que les positions suivantes sont occupées par les valeurs de différents attributs associés à la partie du discours en question. Ainsi, l'étiquette d'adjectif a 8 position, encodant la catégorie, la sous-catégorie (ou le type), le degré de comparaison, le genre, le nombre, le cas, la définitude et l'animéité, dans cet ordre. Les valeurs possibles de ces attributs sont données dans le Tableau 9.

Attribut	Valeurs possibles	Code
Catégorie	Adjectif	A
Type	qualificatif	f
	possessif	s
	ordinal	o
Degrée de comparaison	positif	p
	comparatif	c
	superlatif	s
	élatif	e
Genre	masculin	m
	féminin	f
	neutre	n
Nombre	singulier	s
	pluriel	p
Cas	nominatif	n
	génitif	g
	datif	d
	accusatif	a

	vocatif	v
	locatif	l
	instrumental	i
Définitude	non (no)	n
	oui (yes)	y
Animéité	non (no)	n
	oui (yes)	y

Tableau 9: Valeurs des attributs dans l'étiquette d'adjectif dans MULTEXT-East[11]

Par conséquent, un adjectif qualificatif au positif, qui est au nominatif du singulier du masculin, et qui est indéfini et inanimé, telle la forme *žut* ('jaune'), porte l'étiquette suivante : Afpmsnnn.

Un attribut occupe toujours la même position dans l'étiquette : s'il n'est pas applicable au mot en question, on l'indique en utilisant un tiret. Par exemple, le degré et la définitude ne sont pas applicables aux adjectifs possessifs. Par conséquent, la forme du nominatif du singulier du genre masculin d'un adjectif possessif inanimé, tel *Pavlov* 'appartenant à Pavle', porte l'étiquette As-msn-n.

Ces étiquettes visent à encoder la totalité des propriétés morpho-syntaxiques du serbe, ainsi que des traits sémantiques, ce qui résulte dans un jeu de 906 étiquettes. Ce jeu a été utilisé à annoter le corpus serbe de MULTEXT-East (traduction serbe de « 1984 » de Orwell, présenté dans (Krsteva *et al.* 2004)).

Etant donné que *1984* est le seul corpus du serbe annoté et disponible en ligne, la plupart des expérimentations dans l'étiquetage automatique du serbe ont été effectuées sur ce corpus, en utilisant le jeu d'étiquettes associé. Les résultats obtenus restent bien en-dessous du seuil de 96%. On peut citer notamment (Gesmundo et Samardžić 2012), qui ont utilisé *1984* pour tester leur logiciel BTagger et ont obtenu une précision d'étiquetage de 86,65%, aussi bien que (Popović 2010), qui a effectué plusieurs expérimentation avec différents étiqueteurs et a atteint la précision de 85,47%.

À notre connaissance, le seul travail dans lequel un autre jeu d'étiquettes a été employé est celui de (Utvić 2011) : un jeu minimal de 16 étiquettes, qui n'encodent que

[11] Ce tableau a été repris du site officiel du projet MULTEXT-East (http://nl.ijs.si/ME/V4/msd/html/msd.A-sr.html). Dernier accès : 30 juillet 2013.

la partie du discours principale, a été utilisé pour atteindre la précision de 96,57%. Ces résultats peuvent également avoir été influencés par la taille du corpus d'entraînement : à la différence de *1984*, qui contient environ 108 000 tokens, le corpus utilisé dans (Utvić 2011) en compte 1 000 000. Cependant, des expérimentations réalisées par (Agić *et al.* 2009) dans l'étiquetage du croate confirment l'importance de la taille du jeu d'étiquettes. Les tests ont été effectués sur le corpus du croate *Croatia Weekly*, comptant environ 100 000 tokens (Tadić 2000). L'annotation a été réalisée en utilisant le jeu d'étiquettes développé dans la version 3 de MUTLEXT-East (Erjavec *et al.* 2004). Par conséquent, les tags ont la structure décrite ci-dessus : la première position encode la catégorie grammaticale, suivie des valeurs de différentes propriétés morphologiques. Environ 890 étiquettes sont représentées dans le corpus. Les auteurs ont effectués 6 évaluations de précision d'étiquetage en réduisant à chaque fois le jeu d'étiquettes utilisé. Chaque réduction du nombre d'étiquettes a résulté dans une réduction d'erreur, pour obtenir finalement une précision de 96,23% avec 13 étiquettes, comparé à 84,80% avec le jeu d'étiquettes initial.

Prenant en compte les résultats des travaux présentés ci-dessus, nous avons décidé de construire un jeu d'étiquettes plus simple que celui du projet MULTEXT-East. Pourtant, nous avons souhaité préserver plus d'information que ce que proposent les jeux d'étiquettes minimalistes, tel celui de (Utvić 2011) qui encode seulement la partie du discours principale. Aussi, il a été décidé de construire un ensemble de tags qui indiquent la catégorie et la sous-catégorie grammaticale, ainsi que quelques propriétés morphologiques pour les adjectifs et les verbes. Deux version de ce jeu d'étiquettes ont été développées, dont la plus large contient 71 étiquettes, alors que la plus petite en a 45. La description détaillée des deux versions et de leur utilisation sera donnée dans la partie II.3.4.

II.3.3 Analyse des jeux d'étiquettes choisis pour l'anglais et le français

Le jeu d'étiquettes utilisé par TreeTagger pour l'étiquetage de l'anglais est celui du corpus *PennTreebank*. Il compte 36 tags, encodant différentes parties de discours et précisions morpho-syntaxiques. Par exemple, pour les catégories de l'adjectif (JJ) et de l'adverbe (RB), on note le degré de comparaison (JJR : adjectif au comparatif, JJS : adjectif au superlatif ; RBR : adverbe au comparatif, RBS : adverbe au superlatif). Pour les noms, on distingue les noms propres (NP) et les noms communs (NN), et pour les deux sous-

catégories on marque le pluriel (NPS et NNS, respectivement). La partie du discours avec les étiquettes les plus élaborées est celle des verbes. Il existe 24 étiquettes verbales au total. On marque le temps et le mode, sans indiquer la personne ni le nombre sauf dans le cas de la troisième personne du singulier du présent simple : VV - base de l'infinitif, VVD - le passé simple, VVG - le gérondif/le participe présent, VVZ - le présent simple, la troisième personne de singulier, VVP - le présent simple, toutes les autres formes.

Quant aux autres parties du discours, leur encodage ne semble pas très systématique. Dans la catégorie des pronoms, on distingue les pronoms personnels (PP), possessifs (PP$), les pronoms commençant par *wh-* (i.e. *who*, *what* ; WP), et le pronom possessif de la même série (i.e. *whose* ; WP$). On marque les conjonctions de coordination (CC), mais les catégories des conjonctions de subordination et des prépositions sont fusionnées sous une seule étiquette (IN).

Le jeu d'étiquettes français par défaut de TreeTagger est plus restreint : il compte 33 étiquettes, dont 4 sont utilisées pour les tokens non lexicaux (ponctuations et symboles différents). A la différence du jeu d'étiquettes pour l'anglais, ici on ne fait aucune des distinctions morphologiques pour les adjectifs, adverbes et noms. Comme on ne marque pas le degré de comparaison pour les adverbes et les adjectifs, chacune de ces parties du discours dispose d'une seule étiquette (ADV et ADJ, respectivement). Également, le nombre dans la catégorie des noms n'est jamais encodé, tout en gardant la distinction entre les noms communs (NOM) et les noms propres (NAM). Il est cependant visible que le lien formel entre ces deux étiquettes n'existe pas. La seule exception se trouve dans la catégorie des verbes : le jeu d'étiquettes français encode également le temps et le mode pour les formes verbales simples : conditionnel (VER:cond), futur simple (VER:futu), impératif (VER:impe), imparfait (VER:impf), participe passé (VER:pper), participe présent (VER:ppre), présent (VER:pres), passé simple (VER:simp), subjonctif d'imparfait (VER:subi) et subjonctif de présent (VER:subp). En revanche, on ne marque jamais la personne ni le nombre dans les étiquettes verbales.

Quant aux autres parties du discours, les différences les plus importantes concernent les déterminants et les pronoms : tandis que pour l'anglais il n'existe qu'une étiquette DT consacrée aux articles, dans le jeu français on trouve DET:ART pour les articles et DET:POS pour les déterminants possessifs. Dans le jeu d'étiquettes anglais, les déterminants et les pronoms possessifs sont les deux étiquetés comme pronoms possessifs. Dans la

catégorie des pronoms, on distingue les sous-catégories sémantiques suivantes : pronoms démonstratifs (PRO:DEM), indéfinis (PRO:IND), personnels (PRO:PER), possessifs (PRO:POS) et relatifs (PRO:REL).

Les étiquettes du jeu français étant mieux structurées et plus intuitives, on a décidé d'adopter ce format pour le jeu d'étiquettes serbe. On a décidé de retenir les distinctions morphologiques pour les adverbes et les adjectifs trouvées dans le jeu d'étiquettes pour l'anglais, alors que pour les catégories des verbes et des pronoms on a suivi la logique du jeu d'étiquettes français. La description détaillée du jeu d'étiquettes proposé pour le serbe se trouve dans la partie suivante. Pour la liste intégrale des jeux d'étiquettes pour l'anglais et le français, consulter le site officiel de TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/).

II.3.4 Proposition d'un nouveau jeu d'étiquettes pour le serbe

Comme il a déjà été souligné, le corpus serbe sur lequel ce travail a été effectué fait partie d'un corpus parallèle trilingue français-serbe-anglais. L'usage principal des corpus parallélisés étant l'étude contrastive (linguistique ou autre), il a été nécessaire d'assurer un certain degré de comparabilité entre les trois jeux d'étiquettes utilisés pour les trois langues, malgré les différences structurelles dont elles font preuve. Pour ce faire, certaines distinctions catégorielles traditionnellement acceptées dans la morphosyntaxe serbe ont dû être modifiées ou négligées. Chaque occurrence de ce procédé est expliquée et justifiée ci-dessus.

Comme on a décidé d'adopter la structure des tags utilisée dans le jeu d'étiquettes pour le français, les tags pour le serbe ont la forme suivante : la catégorie principale est indiquée par les trois premières lettres, toujours en majuscules, alors que la sous-catégorie sémantique et/ou les informations morpho-syntaxiques sont données ensuite, en utilisant les deux points comme séparateur.

La description détaillée des étiquettes par partie du discours est donnée dans la suite.

II.3.4.1 Nom

4 étiquettes ont été utilisées pour l'étiquetage de la catégorie des noms : NOM:com, NOM:col, NOM:NAM, et NOM:NUM. Leur distribution est donnée dans le Tableau 10.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	NOM:com	nom commun	vrata ‘porte’, kuća ‘maison’, strast ‘passion’
2.	NOM:col	nom collectif	lišće ‘feuillage’, pilad ‘poussins’, kamenje ‘rochers’
3.	NOM:NAM	nom propre	Duško, Beograd ‘Belgrade’, Afrika ‘Afrique’
4.	NOM:NUM	nom numérique	dvojica ‘les deux’, trojica ‘les trois’

Tableau 10: Etiquettes pour la catégorie du nom

Selon (Stanojčić et Popović 2011), la sémantique serbe distingue 6 sous-catégories des noms : noms propres, communs, massifs, abstraits et déverbaux (équivalentes des mêmes sous-catégories en français et en anglais) et les noms collectifs, qui en serbe, à la différence du français et de l’anglais, comprennent les noms ayant la forme et le comportement d’un nom au singulier, mais le sémantisme du pluriel.

Les noms massifs (*voda* ‘eau’, *brašno* ‘farine’), abstraits (*ljubav* ‘amour’, *mržnja* ‘haine’) et déverbaux (*crtanje* ‘dessin’, *pevanje* ‘chant’) ont le même comportement morphologique et syntaxique que les noms communs : ils suivent les mêmes modèles de déclinaison et ont la même distribution dans la phrase. Par conséquent, il a été décidé d’appliquer la même étiquette NOM:com aux 4 sous-catégories. La distinction avec les noms propres a été gardée : dans la suite de l’enrichissement du corpus, cette propriété peut être exploitée dans la reconnaissance des entités nommées. Les sous-catégories des noms collectifs et des noms numériques portent des tags distincts pour leur comportement spécifique dans l’accord. En effet, les noms collectifs, qui désignent les ensembles des entités, ont en serbe la forme du singulier et la sémantique du pluriel. Par exemple, le nom *pilad* se décline comme le singulier des noms de la déclinaison IV, alors qu’il désigne un ensemble des poussins (*pile* signifie *poussin*). Par conséquent, ce nom peut imposer la forme du singulier (accord grammatical), aussi bien que celle du pluriel (accord sémantique) au groupe verbal (Stanojčić et Popović 2011:306) La phrase ‘Tous les poussins sont jaunes’ peut donc avoir deux formes :

Sva	pilad	je	žuta.
Adj. nom.sg.f.	N.f. nom.sg.	V. 3p.sg.prés.	Adj. nom.sg.f.
tout	poussins	est	jaunes

'Tous les poussins sont jaunes.'

Sva	pilad	su	žuta.
Adj. nom.sg.f.	N.f. nom.sg.	V. 3p.pl.prés.	Adj. nom.sg.f.
tout	poussins	est	jaunes

'Tous les poussins sont jaunes.'

sans que le sens de la phrase change[12].

On remarque que les adjectifs épithète et attribut de sujet (respectivement les formes *sva* 'tous' et *žuta* 'jaunes') connaissent seulement l'accord grammatical (ils ont la forme du singulier du féminin) et que c'est seulement le verbe qui admet les deux types d'accord.

Quant aux noms numériques tels *dvojica* 'deux', *trojica* 'trois', ils ont le comportement morphologique du singulier des noms féminins qui se terminent par *-a*, alors que sémantiquement ils désignent le pluriel masculin Par conséquent, un adjectif attribut de ces noms peut avoir soit la forme du pluriel du féminin, soit celle du pluriel du masculin. Ce phénomène est illustré dans les exemples ci-dessous.[13]

Dvojica	su	pametna
N. nom.sg.f.	V. 3p.pl. prés.	Adj. nom.sg.f.
deux	sont	intelligente

'Deux en sont intelligents.'

Exemple glosé 10: Noms numériques : accord grammatical

Dvojica	su	pametni
N. nom.sg.f.	V. 3p.pl. prés.	Adj. nom.pl.m.
deux	sont	intelligents

'Deux en sont intelligents.'

Exemple glosé 11: Noms numériques : accord sémantique

[12] Il faut souligner que le nom pile 'poussin' dispose également d'une forme de pluriel *pilići*, qui se décline comme le pluriel des noms masculins qui se terminent par une consonne.

[13] La catégorie des noms numériques ne figure pas dans la classification des noms dans (Stanojčić et Popović 2011:78-80). Pourtant, les formes *dvojica*, *trojica* etc. sont définies comme noms numériques dans (Stanojčić et Popović 2011:105).

II.3.4.2 Verbe

Deux ensembles d'étiquettes pour l'annotation des verbes ont été développées. La version plus élaborée compte 26 étiquettes qui encodent la distinction entre le verbe principal et le verbe auxiliaire, ainsi que le temps et l'aspect, alors que l'autre version garde seulement deux tags : VER et VER:AUX, pour les verbes principaux et les verbes auxiliaires, respectivement. Les étiquettes détaillées ont été construites lors de la première étape du projet Egide « Constitution du corpus parallèle français-serbe-anglais » en 2010. L'objectif était de répondre aux besoins des chercheurs travaillant sur les études contrastives de la temporalité en français et en serbe. Ce système des tags a été utilisé dans la phase initiale de l'annotation manuelle du corpus d'entraînement. Il a ensuite été abandonné, vu que son utilisation était trop coûteuse du point de vue de temps, alors que la distinction entre les verbes principaux et auxiliaires a été jugée suffisante pour un étiquetage initial. La partie du corpus annotée avec les étiquettes élaborées a tout de même été sauvegardée, pour être exploitée plus tard dans la suite de l'enrichissement du corpus. Les étiquettes retenues sont présentées dans le Tableau 11. Pour la présentation des 26 tags détaillés, voir Annexe 2.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	VER	verbe principal	jedem 'je mange', hodajući 'en marchant', radio 'travaillé'
2.	VER:AUX	verbe auxiliaire	sam 'je suis', ćete 'vous voulez'

Tableau 11: Etiquettes utilisées pour la catégorie des verbes

II.3.4.3 Pronom

Les pronoms en serbe comprennent deux sous-catégories : les pronoms nominaux et les pronoms adjectivaux (Stanojčić et Popović 2011:98-99). Vu que les pronoms adjectivaux ont le comportement adjectival, et non pas pronominal, il a été décidé de grouper ces formes avec les adjectifs et les étiquettes utilisées pour leur annotation sont présentées dans la partie consacrée aux adjectifs. Il a déjà été montré dans la partie II.3.1.6 que ce groupe des formes est très proche de la catégorie des déterminants en anglais et en français. Cependant, nous avons décidé de ne pas les étiqueter en tant que

déterminants : le corpus élaboré est censé être utilisé par les linguistes serbes, et la tradition grammaticale serbe ne reconnaît pas l'existence des déterminants dans la langue serbe. Un jeu d'étiquettes intégrant cette catégorie risquerait donc d'être contre-intuitif.

Ainsi notre jeu d'étiquettes ne considère comme pronoms que les pronoms nominaux (et quelques usages spécifiques de certains pronoms adjectivaux). A l'intérieur de cette sous-classe, on fait les distinctions suivantes : les pronoms personnels, interrogatifs, négatifs, généraux et indéfinis, ainsi que le pronom réfléchi (en serbe il n'en existe qu'un seul ; cf. partie II.3.1.5). Afin de rendre les descriptions morpho-syntaxiques du serbe plus proches de celles du français et de l'anglais, les pronoms généraux, négatifs et indéfinis ont été regroupés sous le tag des pronoms indéfinis. Les pronoms personnels et interrogatifs, ainsi que le pronom réfléchi, ont été retenus, alors qu'on a ajouté la catégorie des pronoms numériques pour les emplois pronominaux des mots *jedan* 'un' et *drugi* 'deuxième', 'autre', tel *Jedan peva, a drugi igra* 'L'un chante, et l'autre danse'.

Les formes des possessifs et des démonstratifs sont considérées des pronoms adjectivaux (Stanojčić et Popović 2011:99), ce qui peut être justifié, vu que leur comportement est typiquement adjectival : *Hoću ovu knjigu* 'Je veux ce livre', *Poznajem njegovu majku* 'Je connais sa mère'. Pourtant, ils peuvent également avoir un fonctionnement purement pronominal : *Hoću ovu, neću njegovu* 'Je veux celle-ci, je ne veux pas la sienne'. C'est pour annoter ces occurrences-là qu'on a introduit les étiquettes des pronoms démonstratifs et des pronoms possessifs.

De même, la classification des pronoms serbes généralement admise qualifie les relatifs comme pronoms adjectivaux. Même si le serbe dispose de plusieurs adjectifs relatif (*čiji* 'dont', 'duquel', *koliki* 'de quelle taille', *kakav* 'comment' (adj.)), le relatif *koji* est équivalent de *qui* en français ou de *who* en anglais et a la nature d'un pronom (cf. *čovek koji je pričao*, *l'homme qui parlait*, *the man who was talking*). L'étiquette du pronom relatif est consacrée aux formes de ce relatif.

Les étiquettes et les exemples de leur usage envisagé sont donnés dans le Tableau 12.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	PRO:PER	pronom personnel	ja ‘je’, mene ‘me’, ti ‘tu’, vi ‘vous’
2.	PRO:INTR	pronom interrogatif	ko ‘qui’, šta ‘quoi’
3.	PRO:DEM	pronom démonstratif	ovaj ‘celui-ci’, ona ‘celle-là’, ti ‘ceux-là’
4.	PRO:IND	pronom indéfini	neko ‘quelqu’un’, niko ‘personne’, svako ‘tout le monde’
5.	PRO:POS	pronom possessif	moj ‘le mien’, naši ‘les nôtres’, njihovi ‘les leurs’
6.	PRO:REL	pronom relatif	koji ‘qui’
7.	PRO:REF	pronom réfléchi	sebe, se ‘soi-même’
8.	PRO:NUM	pronom numérique	jedan ‘un’, drugi ‘deuxième’,’ autre’

Tableau 12: Etiquettes utilisées pour la catégorie des pronoms

II.3.4.4 Adjectif

L’adjectif serbe connaît plusieurs sous-catégories. On distingue les adjectifs qualificatifs (*lep* ‘beau’, *hrabar* ‘courageux’), possessifs (*Markov* ‘qui appartient à Marko’, *školski* ‘qui appartient à l’école’), massifs (*zlatan* ‘en or’, *drven* ‘en bois’), temporels (*današnji* ‘d’aujourd’hui’, *godišnji* ‘annuel’) et spatiaux (*desni* ‘de droite’, *gornji* ‘supérieur’, ‘d’en haut’) (Stanojčić et Popović 2011:91). Or, étant donné que cette classification repose sur des critères sémantiques, et que tous les types cités des adjectifs énumérés partagent le même comportement morpho-syntaxique, il a été décidé de ne pas introduire ces distinctions sémantiques dans notre jeu d’étiquettes. On a gardé une étiquette pour la catégorie traditionnelle des adjectifs, en ajoutant deux tags pour marquer le comparatif et le superlatif.

Comme il a été mentionné dans la partie ci-dessus, nous avons décidé de considérer comme adjectifs les pronoms adjectivaux. Cette décision a été motivée par le fait que les formes citées ont le comportement adjectival typique : elles sont antéposées au nom et

ont le rôle d'épithète, sauf dans le cas des interrogatifs et des relatifs, qui ont des fonctions spécifiques (cf. partie II.3.1.5). Quant aux démonstratifs, possessifs et indéfinis, ces formes correspondent par leur sémantisme et par leur fonctionnement à la catégorie des déterminants en français et en anglais. Il a déjà été mentionné que ces groupes sont traditionnellement classifiés comme pronoms. Or, une telle organisation mène inévitablement à la confusion entre les véritables valeurs pronominales de ces formes (cf. partie II.3.4.3) et leurs emplois en tant que déterminants. Leur extraction de la catégorie des pronoms facilite la comparaison avec leurs équivalents dans les deux autres langues du corpus.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	ADJ	adjectif au positif	nov 'neuf', lepa 'belle'
2.	ADJ:KOM	adjectif au comparatif	noviji 'plus neuf', lepša 'plus belle'
3.	ADJ:SUP	adjectif au superlatif	najnoviji 'le plus neuf', najlepša 'la plus belle'
4.	ADJ:INTR	adjectif interrogatif	koji 'lequel', kakav 'comment' (adj.), koliki 'de quelle taille'
5.	ADJ:DEM	adjectif démonstratif	ovaj 'ce', ona 'celle', ti 'ces'
6.	ADJ:IND	adjectif indéfini	neki 'certain', nijedan 'aucun', svaki 'tout'
7.	ADJ:POS	adjectif possessif	moj 'mon', naši 'notre', njihovi 'leurs'
8.	ADJ:REL	adjectif relatif	čiji 'de qui', 'dont', kakav 'comment' (adj.), koliki 'de quelle taille'

Tableau 13: Etiquettes utilisées pour la catégorie des adjectifs

II.3.4.5 Nombre

En serbe, les nombres constituent une partie du discours à part (Stanojčić et Popović 2011:105). Ils englobent les nombres cardinaux, ordinaux et collectifs. Les deux premiers groupes correspondent aux mêmes sous-catégories en français et en anglais, avec la différence que les cardinaux de un à quatre se déclinent, et que les ordinaux distinguent le genre. Les nombres collectifs sont, quant à eux, étrangers à l'anglais et au français : il s'agit des formes spéciales qui dénotent le nombre exact des êtres dénotés par un nom désignant les petits, ou le nombre exact des êtres humains de sexe différent (Stanojčić et Popović 2011:105). Comme les nombres collectifs influencent les règles d'accord, et que les cardinaux et les ordinaux ont des comportements distincts l'un de

l'autre, il a été décidé de reprendre la même classification dans notre jeu d'étiquettes. Les tags et les exemples d'usage sont présentés dans le Tableau 14.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	NUM:CAR	nombre cardinal	jedan 'un', jedna 'une', dvadeset 'vingt'
2.	NUM:ORD	nombre ordinal	prvi 'premier', druga 'deuxième', dvadeseti 'le vingtième'
3.	NUM:COL	nombre collectif	dvoje 'deux', petoro 'cinq', dvadesetoro 'vingt'

Tableau 14: Etiquettes utilisées pour la catégorie des nombres

II.3.4.6 Adverbe

La classification des adverbes en serbe repose sur des critères sémantiques : on différencie les adverbes spatiaux, temporels, causaux, les adverbes de manière et de quantité (Stanojčić et Popović 2011:130). Cependant, cette catégorisation ne reflète aucune spécificité fonctionnelle. Pour cette raison, nous avons choisi d'adopter les distinctions présentes en français et de marquer les adverbes indéfinis, interrogatifs et relatifs avec des tags spéciaux, alors que tous les autres porteraient l'étiquette générale d'adverbe, avec la possibilité d'annoter les degrés de comparaison des adverbes de manière. La liste des étiquettes et les exemples d'usage se trouvent dans le Tableau 15.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	ADV	adverbe (autre que relatif, interrogatif ou indéfini)	pametno 'intelligemment', nespretno 'maladroitement'
2.	ADV:KOM	adverbe au comparatif	bolje 'mieux', pametnije 'plus intelligemment'
3.	ADV:SUP	adverbe au superlatif	najbolje 'le mieux', najpametnije 'le plus intelligemment'
4.	ADV:INTR	adverbe interrogatif	kako 'comment', gde 'où', kad 'quand'
5.	ADV:REL	adverbe relatif	kako 'comme', gde 'où', kad 'quand'
6.	ADV:IND	adverbe indéfini	nekako 'n'importe comment', igde 'n'importe où'

Tableau 15: Etiquettes utilisées pour la catégorie des adverbes

II.3.4.7 Mots non lexicaux

Il existe 3 classes de mots grammaticaux en serbe : les conjonctions, les prépositions, et les particules (Stanojčić et Popović 2011:208). Les deux premières catégories sont comparables aux classes portant le même nom en français ou en anglais, alors que les particules sont des mots qui n'ont pas de fonction syntaxique dans la phrase, mais marquent le rapport de l'énonciateur envers l'énoncé. S'y joint encore une partie du discours non lexicale : les interjections. Dans la catégorie des conjonctions, on a marqué la différence entre les conjonctions de coordination et de subordination. Quant aux autres, une seule étiquette a été consacrée à chacune d'entre elles (voir Tableau 16).

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	KON:COOR	conjonction de coordination	i 'et', ali 'mais', ili 'ou'
2.	KON:SUB	conjonction de subordination	da 'que', jer 'parce que', iako 'bien que'
3.	PRP	préposition	na 'sur', pod 'sous', u 'dans'
4.	PAR	particule	da 'oui', ne 'non', čak 'même'
5.	INT	interjection	ah 'ah', apćiha 'atchoum', hej 'hé'

Tableau 16: Etiquettes utilisées pour les mots non lexicaux

II.3.4.8 Ponctuations et autres étiquettes spéciales

Étant donné qu'il est important de préserver la notion de phrase pour le processus de la constitution des échantillons, les ponctuations fortes (celles qui marque la fin de la phrase) ont été annotée avec le tag SENT, alors que toutes les autres portent le tag PUN (hormis les doubles guillemets et les guillemets français, qui sont différenciés par l'étiquette PUN:cit).

Les mots provenant des langues autres que le serbe et qui ne sont pas des emprunts (à savoir, qui ne sont pas adaptés au système morphologique serbe) portent l'étiquette STR (abréviation du serbe *strano* 'étranger').

Les abréviations, les lettres individuelles, les nombres écrits en chiffres et les numéros de page disposent chacun d'une étiquette spécialisée.

Le Tableau 17 contient la description et les exemples pour les étiquettes présentées dans cette partie.

	Etiquette	Sous-catégorie grammaticale	Exemple
1.	SENT	ponctuation forte	. ! ?
2.	PUN	ponctuation faible	, ; : ( )
3.	PUN:cit	ponctuation de citation	« » " „ “
4.	STR	mot étranger	chéri
5.	ABR	abréviation	dr, itd. (etc.)
6.	LET	lettre	A, p, L
7.	NUM	nombre écrit en chiffres	12, 252, XII
8.	PAGE	numéro de page	7, 10

Tableau 17: Autres étiquettes

II.4 Étiquetage du corpus d'entraînement

II.4.1 Descriptif du corpus

Le corpus d'entraînement a été dérivé de la totalité de la partie serbe du corpus trilingue. Il contient trois ouvrages : *Enciklopedija mrtvih*[14] et *Bašta, pepeo*[15] de Danilo Kiš et *Testament*[16] de Vidosav Stevanović. Chacun de ces ouvrages représente un sous-corpus dans la structure générale de la partie serbe du corpus trilingue. Pour faciliter leur identification, on leur a attribué la dénomination officieuse suivante : *Enciklopedija*, *Bašta* et *Testament*.

Le corpus d'entraînement compte au total 157 687 tokens, dont 53 661 d'*Enciklopedija*, 47 924 de *Testament* et 56 093 de *Bašta*. Il faut souligner que le roman *Testamen*t contient en effet 79 596 tokens, dont seulement la première partie a été incluse dans le corpus d'entraînement. Les tokens des sous-corpus *Enciklopedija* et

[14] Kiš, Danilo, *Enciklopedija mrtvih*, CD « Danilo Kiš : Sabrana dela ». Ed. française : *Encyclopédie des morts*, Gallimard, 1985.

[15] Kiš, Danilo, *Bašta, Pepeo*, CD « Danilo Kiš : Sabrana dela. Ed. française : *Jardin, cendre*. Gallimard, 1971. »

[16] Stevanović, Vidosav, *Testament*, SKZ. Beograd, 1986. Ed. française : Prélude à la guerre, Mercure de France, 1996.

Testament forment les 101 585 tokens du premier corpus de référence, *REF1*, qui a été manuellement annoté dans sa totalité. C'est sur ce corpus-là que les premiers tests quantitatifs d'étiquetage ont été effectués. Afin d'accélérer le processus de constitution du corpus d'entraînement, le sous-corpus *Bašta* a été annoté automatiquement par le logiciel BTagger entraîné et testé sur *REF1*. L'étiquetage a ensuite été vérifié manuellement et le sous-corpus joint à *REF1* pour constituer le corpus d'entraînement *REF2* qui sera utilisé pour l'apprentissage final de l'étiqueteur choisi.

Le tableau donné ci-dessous présente la structure du corpus *REF2*, la distribution des tokens par sous-corpus et la nature de l'étiquetage effectué.

Nom de (sous-)corpus			Nombre total de tokens	Nombre de tokens annotés	Type d'annotation
REF2	REF1	Enciklopedija	53 661	53 661	manuelle
		Testament	79 596	47 924	manuelle
	Bašta		56 093	56 093	automatique, vérifiée manuellement
Total :			191 352	157 678	

Tableau 18 : Structure du corpus REF2

II.4.2 Principes d'étiquetage

L'un des problèmes principaux dans l'élaboration des corpus annotés manuellement est la question de la cohérence (McEnery 2003:460). Il est fort possible que l'étiqueteur humain ne fasse pas le même choix pour toutes les occurrences d'un cas de figure, surtout s'il s'agit d'une ambiguïté ou d'un usage qui ne correspond pas aux acceptations généralement admises d'un token. Pour assurer la cohérence dans l'attribution des annotations dans un corpus, il est nécessaire d'élaborer un guide d'annotation, y compris les principes de base, ainsi que les règles définissant la résolution des cas de figure spécifiques. La partie suivante de ce mémoire présentera les principes suivis dans l'étiquetage et la vérification effectués.

II.4.2.1 Correspondance tag-token 1:1

Le principe de base qui a été utilisé dans l'annotation du corpus d'entraînement a été le suivant : un token ne peut porter qu'une seule étiquette, et une étiquette ne peut être attachée qu'à une seule forme à la fois. Ce principe permet la réalisation d'un étiquetage

du premier niveau (chaque forme orthographique se voit attribuer un tag), qui, du point de vue linguistique, n'est pas toujours correct : on dissocie ainsi les unités polylexicales et on leur impose une interprétation analytique. Pourtant, un étiquetage qui prendrait en compte les unités polylexicales serait très problématique, premièrement pour la question de la définition des unités polylexicales, mais aussi à cause de l'existence des unités discontinues. L'approche adoptée permet d'éviter ces problèmes et assure ainsi une plus grande facilité dans la maintenance de la cohérence.

II.4.2.2 Définition contextuelle des cas ambigus

Comme il a déjà été démontré dans la partie II.3.1.7, il existe un degré important d'ambiguïté entre différentes parties du discours en serbe. Pour résoudre ces cas de figure, on a employé le critère syntaxique : si un token, qui appartient à une partie du discours, prend dans le contexte le comportement d'une autre catégorie grammaticale, il est annoté selon la partie du discours dont il a pris le rôle. L'exemple typique est celui des adjectifs qui se trouvent nominalisés dans le contexte : le mot *mrtav* 'mort' est un adjectif ; cependant, dans l'exemple *razmišljati o mrtvima* 'réfléchir aux morts', il est employé indépendamment d'un groupe nominal et fonctionne lui-même comme un nom. Par conséquent, il sera annoté comme un nom commun. Ce principe peut être critiqué car il augmente le nombre d'étiquettes valables pour une forme, mais il permet en même temps une meilleure distinction entre les contextes valides pour différentes parties du discours, ce qui devrait faciliter l'étiquetage des formes inconnues.

II.4.2.3 Références

Pour vérifier les décisions prises, on a utilisé comme ouvrage de référence le dictionnaire électronique *Srpski elektronski rečnik* de M. Simić (Simić 2005). La consultation de cette ressource nous a par exemple aidé à déterminer le traitement du mot *kao* 'comme' : le dictionnaire identifie cette forme comme une conjonction dans les structures de comparaison *(trči kao sumanut* 'il court comme un fou'), et comme un adverbe dans le cas de figure suivant : *došao sam kao prijatelj* - 'je suis venu en tant qu'ami'.

II.4.2.4 Quelques cas spéciaux

II.4.2.4.1 Pronoms indéfinis discontinus

Il existe une série des pronoms indéfinis en serbe dérivés des pronoms interrogatifs *ko* 'qui' et *šta* 'quoi' par préfixation en *ni-* et en *i-* (cf. *niko* 'personne', *ništa* 'rien', *iko* 'qui que ce soit', *išta* 'quoi que ce soit'). Comme tous les autres pronoms en serbe, ces formes se déclinent : *nikoga* est l'accusatif de *niko*, *ničemu* le datif/locatif de *ništa*, *ikome* le datif/locatif de *iko*, *ičim* l'instrumental de *išta*. Cependant, si ces formes se trouvent dans un syntagme prépositionnel, elles deviennent discontinues : le préfixe se détache de la base et la préposition vient s'insérer entre les deux : *ni za koga* 'pour personne', *ni o čemu* 'de rien', *i prema kome* 'envers qui que ce soit', *i sa čim* 'avec quoi que ce soit', la préposition étant soulignée à chaque fois[17]. Il devient donc impossible d'identifier la totalité de la forme fléchie en tant que pronom indéfini.

Pour ces cas de figure, la solution suivante a été adoptée : le préfixe est annoté comme particule, la base du pronom porte l'étiquette du pronom indéfini, alors que la préposition est étiquetée de manière habituelle. La suite des tokens *i prema kome* correspond donc à la série d'étiquettes suivante : PAR PRP PRO:IND.

II.4.2.4.2 Participes actif et passif

Les formes dites du participe passif, telles que *otvoren* 'ouvert', *donet* 'apporté', et celles du participe actif comme *zalutao* 'égaré', *procvetao* 'fleuri', ont en serbe deux fonctionnements distincts : elles peuvent faire partie des formes verbales composées et être réellement des participes, ou avoir le rôle d'épithète ou d'attribut et être en effet des adjectifs. Le critère de distinction qui a été adopté est le suivant : on considère comme participe seulement les formes accompagnées d'un verbe auxiliaire (et qui, par conséquent, font clairement partie d'une forme verbale composée), comme *Hotel je otvoren juče* 'L'hôtel a été ouvert hier' ou *Pas je zalutao* 'Le chien s'est égaré', ainsi que les occurrences où il s'agit de l'ellipse de l'auxiliaire, comme *Juče otvoren novi hotel* 'Nouvel hôtel ouvert hier'. Toutes les autres occurrences sont considérées comme adjectifs.

[17] La variation formelle des pronoms (*ko : koga : kome*, *šta : čemu : čim*) est due à la déclinaison, la forme casuelle étant régie par la préposition.

II.4.2.4.3 Verbes auxiliaires dans les formes surcomposées

Le serbe dispose de deux formes verbales surcomposées : le potentiel passé (équivalent dans un certain degré du conditionnel passé français) et le plus-que-parfait. Comme c'est seulement le plus-que-parfait qui connaît une utilisation active, on expliquera le principe d'annotation des temps surcomposés sur son exemple. La forme du plus-que-parfait se compose de la forme du parfait du verbe auxiliaire et du participe passé du verbe principal[18]. Comme le parfait est lui-même un temps composé, le plus-que-parfait contient trois formes, dont deux sont celles du verbe auxiliaire. Dans l'exemple *On je bio došao* 'Il était venu', le plus-que-parfait *je bio došao* est constitué de *je bio*, parfait du verbe *jesam* 'être', et de *došao*, participe passé actif du verbe *doći* 'venir'. Le parfait du verbe *jesam* lui-même consiste en *je*, présent du verbe *jesam*, ce qui est la forme de l'auxiliaire, et en *bio*, participe passé du verbe *jesam*, le verbe principal. Si on suit le principe d'annotation utilisé dans l'étiquetage des temps composés, le parfait du verbe *jesam* devrait être annoté en tant que verbe auxiliaire faisant partie du plus-que-parfait. Or, l'un des principes d'étiquetage fondamentaux qui ont été adoptés est d'effectuer une annotation en établissant la correspondance 1:1 entre les tokens et les étiquettes sans chercher à marquer les unités polylexicales dans cet étiquetage initial. Il a donc été décidé d'attribuer l'étiquette du verbe auxiliaire à chacune des formes du verbe *jesam*, alors que la forme du participe passé porte celle du verbe principal, comme c'est le cas avec les temps composés. En reprenant l'exemple du plus-que-parfait *je bio došao*, cette suite de tokens porterait les étiquettes suivantes : VER:AUX VER:AUX VER.. Dans la suite de l'enrichissement du corpus, cette formule pourra être exploitée à fin d'annoter cette forme verbale : il sera possible d'apprendre au logiciel que la suite d'étiquettes citée peut être un plus-que-parfait

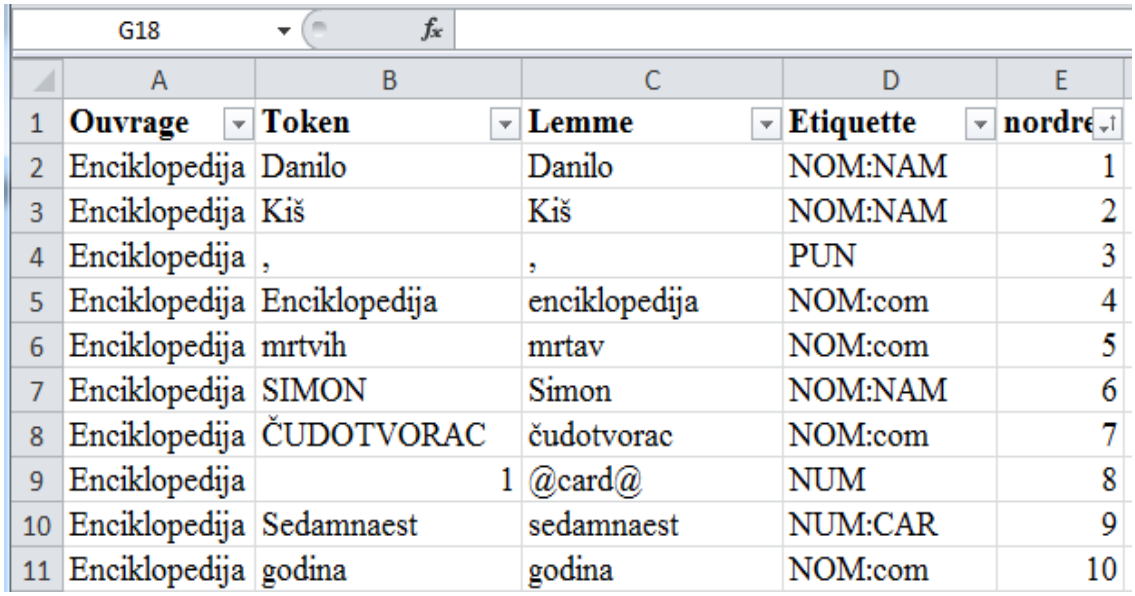
II.4.3 Étiquetage du corpus

Pour faciliter la tâche d'annotation, le corpus d'apprentissage non-étiqueté a été tokénisé (la segmentation en mots a été faite) et présenté sous forme d'un tableau Excel ayant la forme suivante : les fichiers textuels ont été verticalisés et intégrés dans une table Excel, de sorte que chaque ligne contient les informations sur un mot. Les deux premières colonnes contiennent l'indication de l'ouvrage d'où provient le token et le

[18] Le plus-que-parfait serbe peut également être constitué de l'imparfait du verbe auxiliaire et du participe passé du verbe principal (Stanojčić et Popović 2011:125). Pourtant, cette forme est désuète et très rarement utilisée, même dans la langue littéraire.

token lui-même, les deux suivantes sont destinées à contenir le lemme et l'étiquette morpho-syntaxique, alors que la dernière colonne contient la numérotation des tokens suivant leur ordre d'apparition dans le corpus.

On peut voir les 10 premières lignes du corpus d'entraînement dans la Figure 1.



	A	B	C	D	E
1	**Ouvrage**	**Token**	**Lemme**	**Etiquette**	**nordre**
2	Enciklopedija	Danilo	Danilo	NOM:NAM	1
3	Enciklopedija	Kiš	Kiš	NOM:NAM	2
4	Enciklopedija	,	,	PUN	3
5	Enciklopedija	Enciklopedija	enciklopedija	NOM:com	4
6	Enciklopedija	mrtvih	mrtav	NOM:com	5
7	Enciklopedija	SIMON	Simon	NOM:NAM	6
8	Enciklopedija	ČUDOTVORAC	čudotvorac	NOM:com	7
9	Enciklopedija	1	@card@	NUM	8
10	Enciklopedija	Sedamnaest	sedamnaest	NUM:CAR	9
11	Enciklopedija	godina	godina	NOM:com	10

Figure 1: Corpus d'entraînement

II.4.3.1 Étiquetage initial

Comme la numérotation donnée dans la colonne E du tableau Excel montré ci-dessus est unique pour chaque token (elle commence avec le premier token du premier ouvrage et continue ininterrompue jusqu'au dernier token du dernier ouvrage), elle a permis de manipuler l'ordre d'affichage des tokens et d'accélérer ainsi une partie d'annotation.

En effet, les tokens ont été triés selon l'ordre alphabétique. Cela a permis d'avoir en bloc toutes les occurrences d'une forme, d'en annoter la première et ensuite copier le tag attribué dans le reste de la suite. Une fois le corpus parcouru de cette manière, l'ordre initial des tokens a été rétabli en effectuant un triage selon la numérotation dans la dernière colonne. L'étiquetage a ensuite été vérifié et complété token par token.

Cette démarche s'est montrée pratique dans le traitement des formes non ambiguës, permettant d'annoter des dizaines des occurrences d'une forme en quelques secondes au lieu de retaper la même information à chaque fois que le mot en question serait rencontré dans le contexte. Elle devient problématique dans le cas des formes ambiguës : comme les tokens sont triés alphabétiquement, on n'a plus accès à leur

contexte immédiat qui permettrait de désambiguïser. Dans cette situation, on a opté pour l'interprétation plus fréquente du token en question, en notant bien ces choix potentiellement problématiques. Une attention particulière a ensuite été portée à ces formes au cours de la vérification une fois l'ordre des tokens initial rétabli. On peut citer l'exemple de la forme *bude*, qui correspond à la fois au présent du verbe auxiliaire *biti* 'être' et au présent du verbe *buditi* 'réveiller'. Vu que le temps composé dont la forme *bude* fait partie est assez répandue (il s'agit du futur antérieur), on a choisi d'étiqueter toutes les occurrences du token en question comme verbe auxiliaire. Cette annotation a ensuite été vérifiée dans la deuxième étape de l'étiquetage.

II.4.3.2 Étiquetage en contexte

Une fois l'étiquetage par bloc d'occurrences terminé, les tokens ont été remis en ordre selon la numérotation figurant dans la dernière colonne du fichier Excel. Cela a donné comme résultat un étiquetage incomplet, avec des 'trous' : un certain nombre de mots dans chaque phrase avait déjà été annoté, alors que les autres restaient non traités. Par ailleurs, l'étiquetage effectué devait être vérifié : même si une forme semblait univoque lors de l'étiquetage initial, le contexte peut montrer qu'une autre interprétation est mieux adaptée. Il a donc été nécessaire de parcourir la totalité du corpus *REF1* en vérifiant l'annotation attribuée dans la première étape et en ajoutant les étiquettes aux mots qui n'avaient pas été traités. Le corpus *REF1* a également été lemmatisé manuellement ; cependant, les tests et les analyses ont été centrés sur l'étiquetage seul, vu les contraintes de temps.

Pour que la possibilité de commetre une erreur de frappe en entrant les étiquettes soit réduite au minimum, les valeurs valides pour la colonne censée contenir les tags morpho-syntaxiques ont été définies dans une liste fermée. Il était donc possible de visualiser tous les tags existants sous forme d'un menu déroulant et d'en sélectionner l'étiquette correcte en la cliquant. Dans Excel 2003, cette liste peut être définie en utilisant le dispositif Data Validation dans l'onglet Data. Cela est fait en choisissant List comme valeur du paramètre Allow et en définissant la liste des valeurs valides dans la boîte Source. La façon dont ce dispositif se présente dans le tableau lui-même est donnée dans la Figure 2.

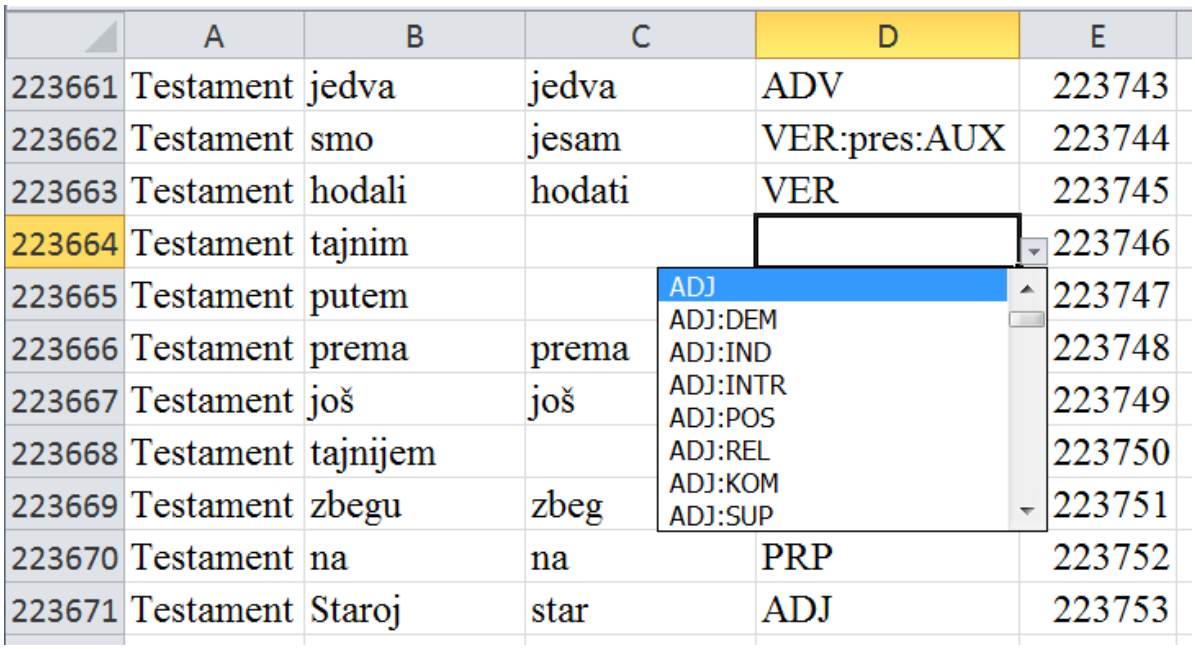


Figure 2: Menu déroulant avec les étiquettes valides

Il était pourtant plus rapide d'entrer les étiquettes en frappant : comme la fréquence de certaines étiquettes était grande, il suffisait d'entrer quelques premières lettres du tag pour que le logiciel propose l'étiquette correspondante (Figure 3).

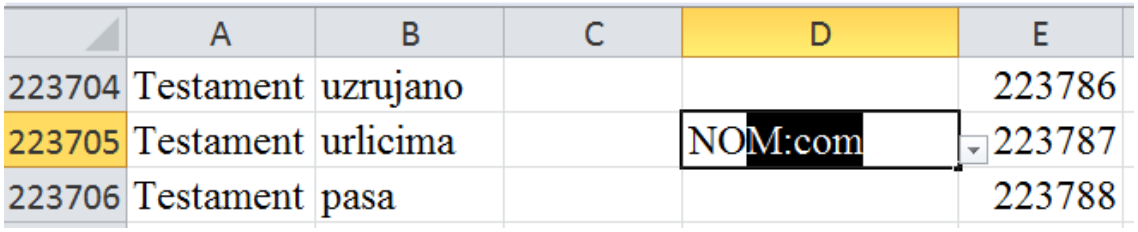


Figure 3: Etiquette proposée par le logiciel

Quant à la lemmatisation, le seul moyen d'entrer les informations était de taper le lemme manuellement. Cependant, les lemmes les plus fréquents étaient proposés par le logiciel selon le mécanisme décrit dans le paragraphe ci-dessus.

Malgré les dispositifs qu'on vient de décrire et qui ont apporté une accélération importante au processus d'étiquetage et de lemmatisation, ce travail s'est avéré long et chronophage. La vitesse moyenne de traitement étant environ 600 tokens par heure, l'annotation et la lemmatisation du corpus *REF1* a pris environ 170 heures de travail. Le volume de ce corpus a été jugé suffisant pour effectuer les évaluations préliminaires des étiqueteurs. Comme ces résultats ont été satisfaisants (cf. partie IV.1.2), la dernière partie du corpus *REF2*, le sous-corpus *Bašta*, a été étiqueté automatiquement et

l'annotation vérifiée ensuite manuellement. La démarche utilisée est décrite dans la partie suivante.

II.4.3.3 Étiquetage automatique et vérification manuelle du sous-corpus Bašta

Le sous-corpus en question contient environ 57 000 tokens. Il a été annoté par BTagger et le résultat de ce processus a été stocké dans le fichier Excel Bašta_BTagger-tagged_accord_BTagger-Manuel.xls.

Pour faciliter la correction manuelle de l'étiquetage, le contenu du fichier Excel a été organisé de manière suivante : le texte du corpus étant verticalisé, chaque ligne du fichier fait figurer le traitement d'un token, contenant le token lui-même, l'étiquette attribuée par BTagger et l'étiquette attribuée lors de l'étiquetage manuel initial, ainsi qu'un indicateur dans le cas où l'étiquette attribuée par BTagger et celle provenant de l'étiquetage manuel initial diffèrent. Cet indicateur, sous forme d'un dièse (#), a été ajouté automatiquement. L'objectif de cette démarche était de faciliter le repérage des erreurs potentielles, mais vu que l'étiquetage manuel initial contenait un nombre non négligeable des erreurs conditionnées par l'homonymie, elle ne s'est pas montrée fructueuse. L'organisation du tableau Excel est illustrée dans la Figure 4.

	A	B	C	D	E
1	Token	BTagger	Manuel	Accord BT/M	Tags Validés
17580	ponovo	ADV	ADV		ADV
17581	zazvoniše	VER	VER		VER
17582	i	KON:COOR	KON:COOR		KON:COOR
17583	mi	PRO:PER	PRO:PER		PRO:PER
17584	ugledasmo	VER		#	VER
17585	očevu	ADJ		#	ADJ
17586	zvezdu	NOM:com	NOM:com		NOM:com

Figure 4: Organisation du fichier de vérification d'étiquetage

La correction manuelle a consisté à ajouter une dernière colonne, destinée à contenir l'annotation finale du fichier. Les étiquettes proposées par BTagger ont été vérifiées une par une. Si l'étiquetage était correct, la même étiquette a été insérée dans la dernière colonne ; dans le cas contraire, le tag correct a été entré dans la cellule correspondante.

Le sous-corpus a ensuite été joint au corpus de référence *REF1* et forme avec lui le deuxième corpus d'entraînement *REF2*. Grâce à ce procédé, le corpus d'entraînement

pour l'étiquetage final de la totalité de la partie serbe du corpus contient environ 160 000 tokens.

III CHOIX DE L'ETIQUETEUR

III.1 Différentes approches dans l'étiquetage automatique

Les premiers logiciels pour l'étiquetage catégoriel automatique étaient basés sur des règles d'annotation écrites manuellement. Un des premiers exemples de cette approche est le logiciel TAGGIT, développé par Greene et Rubin en 1971. Ce logiciel atteignait la précision de 77% sur le *Brown Corpus*, qui est le corpus de référence pour l'anglais américain Une amélioration importante dans la précision a été atteinte avec l'étiqueteur de Brill en 1995. Il s'agit d'un logiciel qui utilise un algorithme d'apprentissage de règles à partir de corpus.Cet étiqueteur obtenait des résultats de l'ordre de 96,9% (Brill 1995). Cependant, cet étiqueteur s'est montré beaucoup moins rapide que les étiqueteurs basés sur les approches statistiques, développées ensuite. Un autre grand défaut réside dans le fait que l'induction automatique de règles à partir d'un corpus, suivie de leur correction manuelle, entraînait des temps de traitement importants. La rapidité d'utilisation et la possibilité d'entraînement pour plusieurs langues sont les raisons principales pour lesquelles on favorise aujourd'hui les étiqueteurs stochastiques.

Quant à ce dernier type d'étiqueteur, une grande partie des expérimentations est centrée sur l'anglais ; différentes méthodes statistiques permettent aujourd'hui d'atteindre une précision entre 95% et 96%, parfois même plus élevée. On peut citer les travaux suivants : (Ratnaparkhi 1996) a développé un système fondé sur le maximum d'entropie avec une précision de 96,6% ; TnT tagger présenté dans (Brants 2000) emploie un modèle de Markov caché avec le même résultat ; (Lafferty *et al.* 2001) se servent des champs conditionnels aléatoires pour obtenir une précision de 95,7%; (Collins 2002) a développé un modèle basé sur un perceptron (*averaged perceptron discriminative sequence model*) et a dépassé le seuil de 97% avec son résultat de 97,1%. Tous les systèmes cités utilisent l'ordre d'inférence de gauche à droite (l'annotation s'effectue linéairement, de gauche à droite, et le choix de l'étiquette à attribuer au mot traité est basé sur le contexte qui précède le mot). On rencontre plus récemment des modèles d'apprentissage bidirectionnel, tels que (Toutanova *et al.* 2003) avec les réseaux des dépendances cycliques (*cyclic dependency network*) atteignant le résultat de 97,2%, ou (Shen *et al.* 2007) qui met en œuvre un système d'apprentissage guidé avec classification bidirectionnelle des séquences pour obtenir une précision de 97,3%.

Cependant, les outils développés pour l'anglais donnent le plus souvent des résultats inférieurs une fois appliqués sur des langues à morphologie riche. Nous avons vu que l'une des spécificités de l'annotation de ces langues et la taille importante du jeu d'étiquettes (cf. partie II.2).Souvent, des stratégies spécifiques sont nécessaires pour atteindre la précision de 95-96%. (Hajič *et al.* 2001) ont combiné le modèle de Markov caché et les règles grammaticales pour arriver à une précision de 95,2% en utilisant un jeu qui compte plus de 1400 étiquettes sur le corpus *Prague Dependency Treebank*. (Habash et Rambow 2005) ont utilisé SVM (*support vector machines*) sur le corpus *ArabicTreebank* avec 139 étiquettes et ont atteint une précision de 97,6%. (Dredze et Wallenberg 2008) ont eu une précision de 92,1% en utilisant un jeu de 639 étiquettes. Ils ont utilisé la classification bidirectionnelle des séquences de (Shen *et al.* 2007) et divisé la tâche d'étiquetage en deux étapes, dont la première a consisté à déterminer les parties du discours principales, et la seconde à ajouter des informations morpho-syntaxiques plus détaillées. Le même mécanisme statistique a été utilisé dans la construction du premier logiciel adapté au traitement du serbe. Cet étiqueteur a été présenté dans (Gesmundo et Samardžić 2012) et il sera décrit en détail dans la partie suivante.

III.2 Expérimentations dans l'étiquetage automatique du serbe

Comme il a déjà été mentionné, le premier étiqueteur consacré au serbe a été distribué en 2012. Les travaux antérieurs à ce moment sont peu nombreux et ils sont centrés sur l'évaluation des performances de différents étiqueteurs généralistes. Les expérimentations les plus extensives ont été effectuées par (Popović 2010). 5 logiciels ont été choisis : TnT Tagger (Brants 2000), Tree Tagger (Schmid 1994), Brill tagger (Brill 1995), MXPOST (Ratnaparkhi 1996) et SVMTool (Giménez et Màrquez 2004). Ils ont été testés sur 3 corpus, qui comptent respectivement 7 500, 75 000 et 105 000 tokens et utilisent des jeux d'étiquettes de taille différente (79, 129 et 908 tags, respectivement). Le corpus de 105 000 tokens et 908 tags est en effet le corpus serbe *1984* du projet MULTEXT-East (Krsteva *et al.* 2004). Les résultats obtenus varient en fonction du corpus : la meilleure précision sur les corpus 1 et 3 est atteinte par TnT Tagger (86,18% et 85,47% respectivement), alors que c'est TreeTagger qui est le plus performant sur le corpus 2 (94,39%).

TreeTagger a également été utilisé dans un autre travail : (Utvić 2011) s'en est servi pour annoter un corpus du serbe contemporain. Le logiciel a été entraîné sur un corpus de 1 000 000 tokens annotés manuellement avec un jeu de 16 étiquettes. Dans ces conditions, TreeTagger améliore de manière significative sa performance et atteint la précision de 96,6%. A notre connaissance, c'est la seule expérimentation dans l'étiquetage automatique du serbe où le seuil de 96% de précision a été dépassé.

Le premier logiciel adapté au traitement du serbe est BTagger (Gesmundo et Samardžić 2012). Il a été entraîné et testé sur le corpus serbe *1984* du projet MULTEXT-East, qui a également été utilisé par (Popović 2010) en tant que corpus 3. Il a montré une précision de 86,65%, ce qui représente une réduction d'erreur de 8,12% par rapport au meilleur résultat obtenu sur le même corpus dans (Popović 2010).

Dans les deux premiers travaux cités, on trouve également des appréciations qualitatives sur les étiqueteurs testés : (Popović 2010) souligne la rapidité d'apprentissage et d'étiquetage de TnT, ainsi que sa simplicité d'usage, alors que dans (Utvić 2011) TreeTagger a été choisi au lieu de TnT pour sa gestion des mots inconnus (meilleure précision que TnT) et pour sa capacité de lemmatisation, dont TnT n'est pas doté. Basé sur les résultats qualitatifs décrits ci-dessus et ces évaluations qualitatives, ces deux logiciels ont été retenus pour être testé sur notre corpus d'entraînement. Étant le premier étiqueteur spécialisé pour le traitement du serbe, et vu les résultats qu'il a obtenu dans les tests, BTagger a également été inclus dans notre sélection des étiqueteurs.

III.3 Présentation des étiqueteurs sélectionnés

III.3.1 TnT (Trigrams'n'Tags)

Ce logiciel a été présenté dans (Brants 2000). L'étiqueteur utilise le modèle de Markov caché de deuxième ordre. Il détermine quelle étiquette doit être attribuée à un token en calculant quel est le tag le plus probable dans le trigramme en question (le trigramme étant défini comme une suite de trois tokens, dont les deux premiers sont les deux tokens avant le token traité). Son mécanisme pour l'annotation des mots inconnus repose sur une analyse des suffixes : la partie du discours d'un mot est déterminée selon la terminaison qu'il exhibe. Ce logiciel a atteint la précision de 96,7% sur l'anglais et l'allemand. Pour plus de détails, voir (Brants 2000).

III.3.2 TreeTagger

La première version de ce logiciel est décrite dans (Schmid 1994). TreeTagger est un étiqueteur basé sur un modèle de Markov qui utilise un arbre de décision pour avoir des estimations plus fiables sur des paramètres contextuels. Pour traiter les mots inconnus, il utilise un lexique des suffixes généré automatiquement sur la base du corpus d'entraînement. La version du logiciel de 1995 emploie également un lexique des préfixes, important pour les langues comme l'allemand, qui dispose des préfixes flexionnels. Ce logiciel a obtenu la précision de 97,53% dans l'étiquetage de l'allemand et 96,34% dans le traitement de l'anglais. La description détaillée de cet outil peut être trouvée dans (Schmid 1994) et (Schmid 1995).

III.3.3 BTagger

BTagger est le plus récent des logiciels choisis : il a été présenté dans (Gesmundo et Samardžić 2012). A la différence de TnT et de TreeTagger, l'algorithme d'entraînement de BTagger n'utilise pas le modèle de Markov, mais le maximum d'entropie (plus précisément, il s'agit de *averaged perceptron algorithm* utilisé dans (Collins 2002)). Une autre différence importante par rapport aux étiqueteurs présentés ci-dessus repose dans le fait que BTagger ne dispose pas d'un ordre d'inférence prédéterminé, alors que pour TnT et TreeTagger on utilise exclusivement le contexte gauche dans l'étiquetage. BTagger est basé sur Bidirectional Tagger de (Shen *et al.* 2007), qui permet d'effectuer un étiquetage itératif : dans chaque passe, seulement les étiquettes les moins ambiguës sont attribuées ; dans la passe suivante, elles sont utilisées pour déterminer l'étiquette des mots non traités. L'étiqueteur prend en compte les deux contextes (celui de gauche aussi bien que celui de droite). Plus de détails sur le fonctionnement de ce logiciel peuvent être trouvés dans (Gesmundo et Samardžić 2012).

IV TESTS ET RESULTATS

IV.1 Tests et évaluations effectués sur le corpus de référence REF1 du 25-05-2013

Pour effectuer l'évaluation quantitative des étiqueteurs choisis, une adaptation de la méthode dite de la validation croisée (*k-fold cross-validation*) a été employée. La validation croisée consiste à diviser le corpus de référence en k parties et d'en utiliser k-1 pour l'entraînement de l'étiqueteur, et la partie restante pour l'évaluation. On répète le procédé autant de fois que nécessaire pour que chacune des k parties soit utilisée comme corpus d'évaluation exactement une fois. Vu que la valeur la plus fréquente de k est 10, la méthode d'évaluation quantitative généralement acceptée comprend qu'on effectue 10 fois l'apprentissage et l'évaluation, en prenant la partie 1 comme corpus d'évaluation pour la première évaluation, la partie 2 pour la deuxième et ainsi de suite. Pour chaque évaluation les mesures des performances sont calculées. Finalement, la valeur moyenne de ces résultats sert d'indicateur de la performance des étiqueteurs.

Cependant, le nombre des sous-corpus peut également être conditionné par la taille du corpus de référence. Si la division du corpus en dix parties génère des sous-corpus dont le nombre des tokens n'est pas suffisamment élevé, cela risque de compromettre le processus d'apprentissage des logiciels. Ainsi (Gesmundo et Samardžić 2012) divisent le corpus *1984* (Krstev *et al.* 2004) en 5 parties et en utilisent 4 pour l'entraînement, et une pour l'évaluation. Vu que la taille de notre corpus de référence est relativement proche du corpus cité (101 349 vs. 108 805 de *1984*), nous avons décidé de diviser le corpus de référence en 4 sous-corpus et d'effectuer, par conséquent, 4 évaluations pour chacun des taggers. Ce procédé représente un compromis entre les impératifs de l'évaluation quantitative et la taille du corpus de référence disponible.

IV.1.1 Constitution des échantillons d'entraînement et de test

Notre corpus de référence compte 101 398 tokens au total, dont 53 561 proviennent du sous-corpus *Enciklopedija* et 47 837 du sous-corpus *Testament*. Le corpus a été étiqueté manuellement dans sa totalité. Il a le format d'un fichier Excel dans lequel chaque ligne correspond au traitement d'un token : elle contient cinq champs indiquant les informations suivantes : l'ouvrage source, la forme rencontrée dans le texte, le

lemme, l'étiquette morpho-syntaxique et le numéro identificateur du token, qui a été utilisé dans l'étiquetage manuel initial décrit dans la partie II.4.3.1 (voir ci-dessus).

Avant de procéder à la constitution des échantillons d'entraînement et de test, il a été nécessaire de corriger une erreur de tokénisation relevée lors de l'annotation manuelle : les occurrences du symbole de ponctuation « … » avaient été systématiquement segmentées en trois tokens, dont chacun portait l'étiquette de ponctuation de fin de phrase. Comme ce sont des échantillons des phrases, et non simplement des tokens, qui allaient être utilisés pour la constitution des sous-corpus, il a été indispensable d'éliminer cette erreur : dans le cas contraire, le tirage aléatoire des phrases aurait pu sélectionner celles ne contenant aucun autre token que le point final. Cela aurait risqué de perturber les algorithmes d'apprentissage des étiqueteurs. La tokénisation et l'étiquetage erronés ont été corrigés grâce à un script perl.

Comme il a été mentionné dans le paragraphe précédent, l'unité de base dans la constitution des échantillons destinés aux expérimentations devait être la phrase, et non le token. On a décidé de préserver l'unité phrastique pour deux raisons principales : premièrement, on envisage à terme des tests avec d'autres étiqueteurs que ceux employés ici, dont certains, tel Memory Based Tagger (Daelemans *et al.* 1996), prennent comme entrée des corpus d'entraînement segmentés en phrases (contrairement aux étiqueteurs utilisés pour les tests décrits ci-dessous, qui exigent des corpus segmentés en tokens). Deuxièmement, il a été jugé important de préserver l'unité phrastique dans le processus d'apprentissage. Vu le volume total des échantillons constitués, il peut être avancé que les séquences de n-grammes et de n-classes suffisent pour un apprentissage réussi. Cependant, si on avait décidé de constituer le corpus d'entraînement en sélectionnant des n-grammes, il est fort probable que l'échantillon ainsi obtenu aurait comporté des suites de tokens syntaxiquement discontinues, ce qui aurait perturbé les performances des étiqueteurs. En revanche, le procédé utilisé, à savoir la constitution des sous-corpus d'apprentissage et d'évaluation par un tirage aléatoire des phrases des deux ouvrages annotés, garantit que les suites des tokens utilisées pour l'entraînement sont grammaticales, ce qui assure la cohérence des règles apprises.

Le fichier Excel de départ, contenant la totalité du corpus de référence REF1, a été transformé en deux fichiers csv : Enciklopedija_v1.3.csv et Testament_ref_V1.0.csv. Le format csv (*comma separated values*) est un format texte avec les données structurées de

telle façon que chaque ligne de texte correspond à une ligne d'un tableau, avec les valeurs des différents champs séparés par un délimiteur, typiquement un point-virgule. Ce format facilite le traitement automatique des fichiers par des scripts perl ou autres.

Ces fichiers ont ensuite été traités pour obtenir Enciklopedija_v1.4_TreeTagger-format.csv et Testament_ref_V1.1_TreeTagger-format.csv. Ces deux fichiers respectent le format exigé par TreeTagger : sur une ligne, on trouve le token, le lemme et l'étiquette, séparés par des tabulations.

Ces fichiers sont verticalisés (i.e., il y figure un mot par ligne) et ne disposent pas de la notion de phrase. Il a donc été nécessaire de les déverticaliser, de manière à obtenir une phrase par ligne. En le faisant, les fichiers suivants ont été générés : Enciklopedija_v1.4_phrases.csv et Testament_ref_V1.1_phrases.csv. Ici, une ligne contient tous les tokens qui constituent une phrase, accompagnés du lemme et de l'étiquette correspondants. Les éléments constituant le traitement d'un token sont toujours séparés par tabulation, alors que la fin du traitement d'un token est marquée par un dièse (#). Ce délimiteur a été exploité lors de la réverticalisation des fichiers, pour restituer le format de données qui contient le traitement d'un token par ligne. Une illustration du format utilisé est donnée dans la Figure 5.

Les fichiers Enciklopedija_v1.4_phrases.csv et Testament_ref_V1.1_phrases.csv contiennent 2 110 et 2 216 phrases respectivement. Il est intéressant de noter que bien que le sous-corpus *Enciklopedija* comporte plus de tokens que celui de *Testament* (53 557 vs. 47 792), ce dernier compte plus de phrases que le premier (2 216 vs. 2 110).

Ili ili KON:COOR#, , PUN#tačnije tačno ADV:KOM#rečeno reći VER#, , PUN#nije jesam VER:AUX#bio biti VER#prevodiv prevodiv ADJ#u u PRP#reči reč NOM:com#i i KON:COOR#rečenice rečenica NOM:com#, , PUN#svako svaki ADJ:IND#tumačenje tumačenje NOM:com#nizova niz NOM:com#slika slika NOM:com#ponovo ponovo ADV#se se PRO:REF#pretvaralo pretvarati VER#u u PRP#sliku slika NOM:com#, , PUN#još još ADV#jednu jedan PRO:NUM#u u PRP#tom taj ADJ:DEM#nizu niz NOM:com#. . SENT

Figure 5: Format des fichiers déverticalisés

Afin d'éviter de biaiser l'apprentissage en sur-représentant les tokens provenant de l'un ou de l'autre ouvrage, il a été décidé de constituer deux échantillons aussi proches

que possible par le nombre des tokens contenus. Comme le sous-corpus *Testament* contient moins de tokens que celui de *Enciklopedija*, il a été gardé en totalité. En revanche, un sous-échantillon d'*Enciklopedija* a dû être dérivé. Pour ce faire, on a effectué un tirage aléatoire de phrases grâce à la commande linux shuf. Ce dispositif permet de réarranger de façon aléatoire les lignes d'un fichier et d'en extraire une quantité déterminée. Grâce à cette commande, 1 900 phrases ont été extraites au hasard du sous-corpus *Enciklopedija*, puis ramenées à un nombre de phrases qui permette d'obtenir un nombre de mots étiquetés aussi proche que possible de celui du sous-corpus *Testament*.

En l'occurrence, les deux sous-corpus représentent :

- 47 793 tokens pour Enciklopedija_v1.4_ref-sample.csv
- 47 792 tokens pour Testament_ref_v1.1_ref-sample.csv

soit un total de 95 585 tokens étiquetés manuellement.

Pour assurer qu'aucune partie du sous-corpus *Testament* (les *n* premières phrases, ou les *n* dernières) ne soit sur-représentée dans les échantillons finaux, les phrases de cet ouvrage ont été triées de façon aléatoire.

Comme il a déjà été souligné, cette stratégie a pour objectif d'éviter de biaiser l'apprentissage en sur-représentant les tokens provenant de l'un ou de l'autre ouvrage, ou bien d'une partie de l'un des ouvrages. La contrepartie de ce procédé est qu'on dispose de moins de tokens étiquetés que le volume total possible. Toutefois, le volume sélectionné reste proche de la taille du corpus *1984* élaboré dans le cadre du projet MULTEXT-East (Krsteva *et al.* 2004).

Chacun des deux sous-corpus cités ci-dessus a ensuite été partitionné en deux, afin de disposer de 4 sous-corpus de référence. On obtient ainsi 4 sous-échantillons, dont les volumes respectifs de tokens sont aussi proches que possibles les uns des autres. Chaque échantillon a été segmenté à la phrase la plus proche pour préserver l'unité phrastique.

Sous-échantillon	Nombre de tokens
Enciklopedija_v1.4_ref-subsample1.csv	23 908
Enciklopedija_v1.4_ref-subsample2.csv	23 885

Testament_ref_V1.1_ref-subsample1.csv	23 908
Testament_ref_V1.1_ref-subsample2.csv	23 884

Tableau 19 : Distribution des tokens par sous-échantillon

IV.1.2 Tests et résultats

Pour chaque expérimentation, 3 fichiers de référence ont été utilisés pour l'entraînement du logiciel, alors que le quatrième fichier a été employé pour mesurer la différence entre l'étiquetage manuel et l'étiquetage automatique effectué par l'étiqueteur en question. Cette opération a été réitérée de sorte que chaque fichier a été utilisé exactement une fois comme fichier de test. Par conséquent, 4 évaluations par étiqueteur ont été réalisées. Les scénarios suivants ont été utilisés (la lettre *A* désigne les corpus utilisés pour l'apprentissage, la lettre *T* celui de test) :

Evaluation 1		
T	23 980	Enciklopedija_v1.4_ref-subsample1.csv
A	23 885	Enciklopedija_v1.4_ref-subsample2.csv
A	23 908	Testament_ref_V1.1_ref-subsample1.csv
A	23 884	Testament_ref_V1.1_ref-subsample2.csv

Tableau 20 : Evaluation 1

Evaluation 2		
A	23 980	Enciklopedija_v1.4_ref-subsample1.csv
T	23 885	Enciklopedija_v1.4_ref-subsample2.csv
A	23 908	Testament_ref_V1.1_ref-subsample1.csv
A	23 884	Testament_ref_V1.1_ref-subsample2.csv

Tableau 21 : Evaluation 2

Evaluation 3		
A	23 980	Enciklopedija_v1.4_ref-subsample1.csv
A	23 885	Enciklopedija_v1.4_ref-subsample2.csv
T	23 908	Testament_ref_V1.1_ref-subsample1.csv
A	23 884	Testament_ref_V1.1_ref-subsample2.csv

Tableau 22 : Evaluation 3

Evaluation 4		
A	23 980	Enciklopedija_v1.4_ref-subsample1.csv
A	23 885	Enciklopedija_v1.4_ref-subsample2.csv
A	23 908	Testament_ref_V1.1_ref-subsample1.csv
T	23 884	Testament_ref_V1.1_ref-subsample2.csv

Tableau 23 : Evaluation 4

Comme les trois étiqueteurs choisis pour être testés n'exigent pas le même format des données (TreeTagger nécessite que le fichier d'entrée contienne le token suivi de l'étiquette suivi du lemme, alors que TnT et BTagger exigent que le token soit suivi du lemme suivi de l'étiquette), des adaptations nécessaires des fichiers d'entrée ont été faites pour chaque logiciel. De manière générale, dans ces expérimentations centrées sur l'étiquetage seul, les lemmes n'ont pas été pris en compte. Des expérimentations ultérieures pourront intégrer ce paramètre et mesurer son influence sur les performances de l'étiquetage.

IV.1.2.1 TreeTagger

TreeTagger exige un lexique qui recense toutes les étiquettes possibles pour un token donné, ainsi qu'un fichier contenant les tags possibles pour un token inconnu (*open class file*). Ces fichiers ont été générés grâce aux utilitaires fournis avec cet outil. Dans le cas particulier du fichier avec les tags possibles pour les tokens inconnus, les mêmes tags ont été fournis quelle que soit l'expérimentation : NOM:NAM (nom propre), NOM:com (nom commun), VER (verbe principal), VER:AUX (verbe auxiliaire), ADJ (adjectif qualificatif au positif) et ADV (adverbe). Ces tags correspondent aux classes ouvertes en serbe.

Une fois les étapes préparatoires faites, l'exécution de l'apprentissage et de l'étiquetage est rapide, voire instantanée.

Les résultats de l'évaluation quantitative des performances de TreeTagger sont présentés dans le Tableau 24. On peut remarquer que la précision varie jusqu'à 1,18 points : la valeur la plus élevée est obtenue dans le scénario d'évaluation 3 et compte 92,61%, alors que le pourcentage des tags corrects est le plus bas dans Evaluation 1 et a la valeur de 91,43%. La précision moyenne calculée à partir des résultats des quatre évaluations est 92,15%. Même si cette valeur reste en-dessous du seuil de 96%, elle est nettement plus élevée que la précision présentée dans (Gesmundo et Samardžić 2012).

Egalement, même si cette précision ne dépasse pas le meilleur résultat obtenu par (Popović 2010) pour cet étiqueteur (94,39% sur un corpus de 75 000 mots avec un jeu de 127 étiquettes), elle est supérieure à celle des deux autres expérimentations présentées dans le même travail (85,44% sur un corpus de 7 500 tokens avec un jeu de 79 étiquettes et 79,65% pour un corpus 106 000 tokens dérivé du corpus *1984* et annoté avec un jeu de 906 étiquettes).

TreeTagger				
Scénario d'évaluation	No de tokens du fichier de test	No des étiquettes correctes	Précision	
Evaluation 1	23 908	21 773	**91,43%**	
Evaluation 2	23 885	21 998	92,10%	
Evaluation 3	23 908	22 141	**92,61%**	Précision moyenne :
Evaluation 4	23 884	22 085	92,47%	**92,15%**

Tableau 24: Résultats par évaluation : TreeTagger

IV.1.2.2 TnT tagger

TnT Tagger ne nécessite pas de fichiers spécifiques pour effectuer l'apprentissage. La rapidité du logiciel est comparable à celle de TreeTagger : l'exécution ne prend que quelques secondes.

Le Tableau 25 présente les résultats de l'évaluation quantitative des performances de TnT tagger. Cet étiqueteur a atteint la précision moyenne de 92,97%, avec la valeur maximale de 93,20% obtenue dans l'Evaluation 3, et la valeur la plus basse de 92,43% dans l'Evaluation 1, avec un écart de 0,77% entre ces deux valeurs. Aussi bien que TreeTagger, TnT n'a pas atteint le seuil de 96% de précision, mais les résultats obtenus sont meilleurs que ceux présentés dans (Gesmundo et Samardžić 2012), aussi bien que ceux présentés dans (Popović 2010) pour les corpus annotés avec 79 et 809 étiquettes (respectivement 86,18% et 85,47%). En revanche, ce résultat est inférieur à celui obtenu pour le corpus annoté avec 129 étiquettes (94,11%, *idem*).

TnT				
Scénario d'évaluation	No de tokens du fichier de test	No des étiquettes correctes	Précision	
Evaluation 1	23908	22097	**92,43%**	
Evaluation 2	23885	22253	93,17%	
Evaluation 3	23908	22282	**93,20%**	Précision moyenne :

Evaluation 4	23884	22229	93,07%	**92,97%**

Tableau 25: Résultats par évaluation : TnT Tagger

IV.1.2.3 BTagger

A la différence de TreeTagger et de TnT Tagger, l'apprentissage de BTagger est incontestablement plus lent : chaque processus d'apprentissage a duré plus de 1h20, et chaque étiquetage plus de 40 minutes. Cela est dû au fait que BTagger est paramétré pour effectuer 10 itérations de l'apprentissage pour compléter le processus. Il serait probablement possible de diminuer le temps d'exécution en réduisant le nombre de cycles d'apprentissage, mais il est également probable que cette action affecterait les résultats.

La précision moyenne que ce logiciel a atteinte est 94,17%. Les résultats dans les évaluations individuelles varient entre 93,93% (Evaluation 1) et 94,48% (Evaluation 2). Cela donne un écart de 0,55% entre les précisions la plus et la moins élevée. Les résultats par évaluation sont présentés dans le Tableau 26.

Les seuls résultats disponibles pour BTagger jusqu'à présent sont ceux donnés dans (Gesmundo et Samardžić 2012). Les auteurs y citent une précision générale moyenne de 86,65%. Nos résultats (94,17%) présentent une amélioration nette de la précision d'étiquetage, avec une réduction d'erreur de 56,33%. Tout de même, il faut tenir compte du fait que (Gesmundo et Samardžić 2012) utilise un jeu de 906 étiquettes.

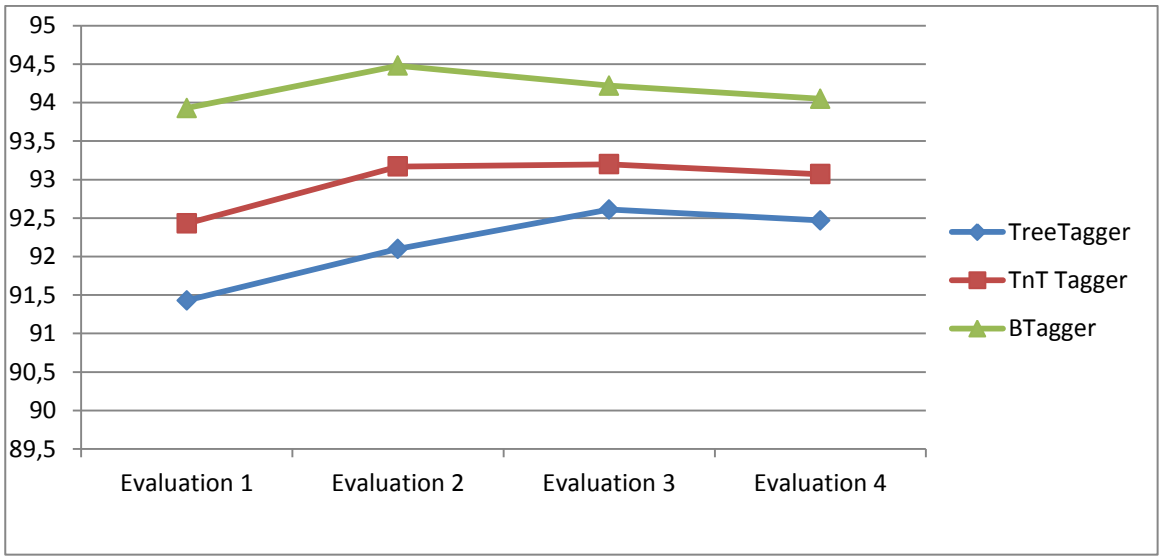
BTagger				
Scénario d'évaluation	No de tokens du fichier de test	No des étiquettes correctes	Précision	
Evaluation 1	23908	22457	**93,93%**	
Evaluation 2	23885	22566	**94,48%**	
Evaluation 3	23908	22525	94,22%	Précision moyenne :
Evaluation 4	23884	22464	94,05%	**94,17%**

Tableau 26: Résultats par évaluation : BTagger

IV.1.3 Analyse des résultats

Parmi les étiqueteurs testés, BTagger a atteint le meilleur résultat. La précision moyenne qu'il a obtenue est 94,17%, ce qui est de 1,20% supérieur au résultat moyen de TnT tagger, et de 2,02% supérieur à celui de TreeTagger. On remarque également que la

précision la moins élevée de BTagger (93,93%) est meilleure que les résultats les plus élevés de TnT Tagger et TreeTagger (93,20% et 92,61% respectivement). Ces résultats sont illustrés dans le Graphe 1.



Graphe 1 : Précision des étiqueteurs par évaluation

Il est intéressant de noter que TnT Tagger et TreeTagger tous les deux ont des performances les moins bonnes dans l'Evaluation 1, et le mieux dans l'Evaluation 3. Vu que la taille du corpus d'entraînement et de test est la même dans les deux scénarios (71 677 et 23 908 tokens respectivement), on peut supposer que c'est le contenu du corpus de test (Enciklopedija_v1.4_ref-subsample1.csv) qui s'est trouvé problématique pour les étiqueteurs. BTagger a également atteint la précision la moins élevée sur ce corpus de test.

Comme BTagger est relativement neuf (il n'est paru qu'en 2012), les seuls résultats disponibles jusqu'à présent pour cet étiqueteur sont ceux présentés dans (Gesmundo et Samardžić 2012). Pour les objectifs de ce travail, le logiciel a été paramétré sur le corpus *1984* (Krstev *et al.* 2004) qui contient environ 108 000 tokens et qui est annoté avec 906 étiquettes. Trois quart du corpus ont été utilisés pour l'entraînement, et un pour le test. La précision moyenne de 86,65% a été obtenue. Le résultat obtenu par BTagger dans notre mémoire est nettement supérieur : 94,18% de précision, ce qui représente une réduction d'erreur de 55,58%. S'agissant dans les deux cas des corpus littéraires, la seule différence qui pourrait expliquer cette amélioration significative de la précision est la réduction de la taille du jeu d'étiquettes : tandis que le corpus *1984* contient 906 tags différents, notre corpus n'en utilise que 45. Ces résultats corroborent notre hypothèse

de départ que la réduction du nombre d'étiquettes amènera une hausse de précision. Cela justifie les choix faits dans la construction du jeu d'étiquettes.

Quant à TnT Tagger et TreeTagger, les résultats que nous avons obtenus avec eux (parties III.3.1 et III.3.2 ci-dessus) se trouvent parmi les meilleurs signalés jusqu'à présent dans le traitement du serbe. En effet, de trois tests effectués dans (Popović 2010), seul le deuxième atteint une précision supérieure à celle signalée ici, à savoir 94,39% pour TreeTagger et 94,11% pour TnT Tagger. Le corpus utilisé contient 75 000 tokens annotés avec 129 étiquettes. Il faut souligner qu'il ne s'agit pas d'un corpus littéraire, mais d'un corpus des textes législatifs et administratifs. Il est possible que la différence dans les résultats obtenus par rapport à notre travail soit due à la différence dans la nature des textes. Les deux autres tests effectués dans le même travail donnent les résultats suivants : 85,44% pour TnT Tagger et 86,18% pour TreeTagger dans le premier test, et 79,65% pour TnT Tagger et 84 ,57% pour TreeTagger dans le troisième. Le corpus de la première expérimentation compte 7 500 tokens annotés avec 79 étiquettes, alors que celui utilisé dans la troisième est basé sur le corpus *1984* du projet MULTEXT-East : il contient 105 000 tokens et emploie 908 étiquettes différentes.

TreeTagger a également été utilisé dans (Utvić 2011). Dans ce travail, l'étiqueteur a été entraîné sur un corpus de référence de 1 100 000 tokens, annoté manuellement avec un jeu d'étiquettes minimal comptant 16 étiquettes et n'encodant que les parties de discours principales. Dans ces conditions, TreeTagger a atteint ce qui est, à notre connaissance, la précision la plus élevée qui a été signalée jusqu'à présent dans l'annotation morpho-syntaxique du serbe : 96,57%.

Les résultats obtenus par TreeTagger et TnT Tagger confirment encore une fois ce qui a été démontré dans les travaux précédents sur l'étiquetage morpho-syntaxique du serbe : les logiciels généraux, développés pour des langues à morphologie flexionnelle réduite et/ou à un ordre de constituants relativement fixe, ne sont pas bien adaptés pour une langue tel le serbe, exhibant une richesse des formes fléchies et peu de contraintes dans l'ordre des constituants : pour dépasser la précision de 96%, ils nécessitent des larges quantités de données pour l'apprentissage et un jeu d'étiquettes réduit autant que possible (cf. Utvić 2011). Dans tous les autres scénarios, le seuil de 96% leur reste difficilement atteignable.

Cependant, on peut conclure que la démarche décrite ici, à savoir l'étiquetage avec BTagger en utilisant le jeu d'étiquettes que nous avons proposé, présente un compromis raisonnable : elle a permis d'obtenir un taux de précision suffisamment élevé pour avoir un étiquetage fiable et en même temps bien exploitable, grâce à la structure du jeu d'étiquettes utilisé.

IV.2 Analyse des résultats de l'étiquetage automatique du sous-corpus *Bašta*

Prenant en compte les considérations présentées dans la partie précédente, il a été décidé de retenir BTagger pour l'étiquetage automatique de la totalité du corpus : malgré l'inconvénient de la lenteur d'exécution, cet outil atteint des résultats nettement supérieurs à ceux des deux autres étiqueteurs testés.

Afin d'assurer les conditions optimales pour l'étiquetage de la partie serbe du corpus, il a été décidé d'élargir le corpus de référence de base. Vu que l'annotation strictement manuelle était jugée trop coûteuse du point de vue du temps, il a été décidé d'effectuer une annotation semi-automatique d'un sous-corpus qui n'avait pas été traité. Le sous-corpus *Bašta* a été choisi ; il a été annoté automatiquement en utilisant BTagger et l'étiquetage résultant a été vérifié et corrigé manuellement. Ce sous-corpus a ensuite été joint au corpus de référence *REF1* présenté dans la description des tests quantitatifs. Grâce à ce procédé, le corpus d'entraînement pour l'étiquetage final de la totalité de la partie serbe du corpus contient environ 160 000 tokens.

Le sous-corpus *Bašta*, présenté dans la partie II.4.1, contient environ 57 000tokens. Il a été annoté par BTagger et le résultat de ce processus a été stocké dans le fichier Excel Bašta_BTagger-tagged_accord_BTagger-Manuel.xls. Le processus de la correction manuelle a été décrit dans la partie II.4.3.3. Une fois cette démarche complétée, une analyse des performances de BTagger sur ce texte inconnu a été conduite.

Pour évaluer la performance de BTagger dans cette tâche, le nombre des étiquettes erronées données par le logiciel a été déterminé à l'aide d'une macro VisualBasic. Comme l'on a déjà montré dans la partie II.4.3.3, le fichier Excel contenant l'annotation corrigée a le format suivant :

	A	B	C	D	E
1	Token	BTagger	Manuel	Accord BT/M	Tags Validés
17580	ponovo	ADV	ADV		ADV
17581	zazvoniše	VER	VER		VER
17582	i	KON:COOR	KON:COOR		KON:COOR
17583	mi	PRO:PER	PRO:PER		PRO:PER
17584	ugledasmo	VER		#	VER
17585	očevu	ADJ		#	ADJ
17586	zvezdu	NOM:com	NOM:com		NOM:com

Figure 6: Sous-corpus 'Basta'

Le script VisualBasic compare la colonne contenant l'étiquetage produit par BTagger (colonne B) avec celle qui fait figurer l'étiquetage corrigé manuellement (colonne E) et compte les occurrences où les valeurs indiquées ne coïncident pas. Le résultat du script indique 1 903 tags incorrects, ce qui correspond à un taux d'erreur de 3,3%. Ce résultat est meilleur de celui obtenu sur le corpus de référence, où le taux d'erreur moyen était 5,83%.

Le fait que le logiciel atteigne un meilleur résultat sur un texte inconnu que sur le corpus de référence est inhabituel. Pourtant, il est possible que cette hausse de performances soit due au fait que l'auteur du sous-corpus en question est le même que l'auteur de sous-corpus *Enciklopedija*, qui fait la moitié du corpus de référence utilisé pour le paramétrage de BTagger. On peut donc supposer un degré important de ressemblance entre le corpus d'entraînement et le sous-corpus *Bašta* au niveau du lexique et des structures syntaxiques.

IV.2.1 Distribution d'erreurs par partie du discours

Afin de déterminer des stratégies éventuelles de prétraitement ou de posttraitement qui pourraient diminuer davantage le nombre d'erreurs d'étiquetage, nous avons procédé à une analyse qualitative de l'annotation automatique de *Bašta*. Un second objectif de ce procédé était d'identifier les cas de figure critiques dans l'étiquetage morpho-syntaxique du serbe, i.e. les structures ou les mots qui se sont avérés problématiques pour l'étiqueteur.

Pour déterminer quelles étaient les parties de discours qui causaient le plus grand nombre de fautes, un comptage d'erreurs par étiquette a été effectué (i.e., combien de fois un nom commun, un verbe ou un adjectif s'est vu attribuer une étiquette erronée). Nous avons ensuite déterminé la distribution de la confusion pour chaque étiquette (i.e.,

combien de fois un nom commun a été annoté comme adjectif, verbe ou une autre partie de discours). Ces analyses ont également été faites à l'aide d'une macro Visual Basic.

Comme on a vu dans la partie II.3.1, la richesse des formes fléchies en serbe est associée à un degré d'homonymie non négligeable, qui peut être intra-catégorielle (entre les formes d'un paradigme, tels le nominatif, l'accusatif et le vocatif des noms du genre neutre) ou bien inter-catégorielle (entre les formes de deux paradigmes, par exemple entre les noms communs et les verbes). Etant donné que l'homonymie intra-catégorielle n'est pas pertinente pour notre jeu d'étiquettes, qui n'encode pas les propriétés flexionnelles, notre attention a été accordée à l'homonymie inter-catégorielle seule. L'analyse effectuée dans la partie II.3.1.7 a montré que les parties de discours les plus affectées par ce phénomène sont celles des noms, des verbes, des adjectifs et des adverbes. Les résultats du comptage des erreurs par étiquette confirment cette observation : les (sous-)classes de mots les plus difficiles à annoter sont celles d'adjectif, de nom commun et de verbe principal. En effet, les erreurs commises dans l'étiquetage de ces trois catégories comprennent 52,5% des 1 903 étiquettes erronées relevées. Le pourcentage d'erreurs trouvées dans ces catégories est présenté dans le Tableau 27.

Pour la distribution d'erreurs sur la totalité du jeu d'étiquettes, voir l'Annexe 3.

Partie de discours	Nombre de tags erronés	Pourcentage de tags erronés sur 1903 étiquettes erronées
Adjectif	432	22,7%
Nom commun	310	16,3%
Verbe principal	258	13,5%

Tableau 27 : Distribution d'erreurs par catégorie principale

Afin d'identifier les cas de figure générant le plus de confusion, pour chacune des catégories problématiques identifiées, la distribution d'erreurs par partie du discours a été analysée. Cela signifie qu'on a déterminé combien de fois un nom commun a été confondu avec le nom propre, le verbe, l'adjectif, etc. Les résultats sont présentés ci-dessous.

IV.2.1.1 Adjectif

Comme il a déjà été mentionné dans la partie II.3.1.7, la catégorie d'adjectif en serbe exhibe un degré important d'homonymie avec les catégories d'adverbe et de verbe. Les adverbes de manière ont la même forme que le genre neutre (ou, dans certains cas, le genre masculin) de l'adjectif qualificatif correspondant. On rappelle l'exemple de la forme *dobro*, qui est en même temps le genre neutre de l'adjectif *bon* (*dobro dete* 'bon enfant') et l'adverbe *bien* (*dobro plivati* 'nager bien').

Quant au recoupement avec les verbes, un nombre important d'adjectifs qualificatifs est dérivé des formes du participe passé actif et passif par adjectivation. La forme *načinjen* peut être le participe passé passif du verbe *načiniti* 'faire, causer', et, dans ce cas-là, elle fait nécessairement partie d'un temps composé : *Šteta je već bila načinjena* 'Les dégâts avaient déjà été faits'. Elle peut également avoir le comportement adjectival : *načinjena šteta* 'les dégâts faits'[19].

Par conséquent, notre hypothèse était que la confusion entre l'adjectif et l'adverbe serait à l'origine de la majorité des erreurs. Cependant, les résultats des comptages réalisés montrent que l'adjectif est le plus souvent confondu avec le nom commun (56,6%) et le verbe (29,1%), la confusion avec l'adverbe ne faisant que 7,4% du nombre total des erreurs commises dans l'étiquetage de l'adjectif.

Catégorie attribuée	Nombre d'occurrences de confusion	% du nombre total d'erreurs pour ADJ
NOM:com	224	56,6%
VER	126	29,1%
ADV	32	7,4%

Tableau 28 : Distribution de confusion pour ADJ

IV.2.1.2 Nom commun

Quant aux noms communs, l'homonymie avec d'autres parties du discours est la plus répandue avec les adjectifs et les verbes. Un adjectif et un nom commun peuvent

[19] Noter que l'interprétation verbale de cet exemple n'est pas possible en serbe : alors qu'en français on peut avoir la construction *les dégâts faits par les cambrioleurs*, en serbe le syntagme équivalent n'est pas grammatical : **načinjena šteta od strane provalnika*. Pour que le complément d'agent (et, par conséquent, l'interprétation verbale) soit admissible, l'attribut doit être postposé au nom : *šteta načinjena od strane provalnika*.

partager le même paradigme si le nom en question a été dérivé d'un adjectif qualificatif par substantivation. Ainsi, dans l'exemple *mrtvo lišće* 'feuilles mortes', le mot *mrtvo* représente une forme fléchie de l'adjectif *mrtav* 'mort', alors que dans le syntagme *enciklopedija mrtvih* 'encyclopédie des morts' il s'agit d'une forme du nom commun *mrtvi* 'les morts'.

Le recoupement des formes avec le paradigme verbal est plus sporadique : typiquement, le nominatif, l'accusatif ou le génitif (ou une autre forme fléchie) de certains noms communs peut coïncider avec la première ou la deuxième personne du singulier du présent du verbe ayant la même base dérivationnelle. Ainsi la forme *kazni* correspond en même temps au datif/locatif du singulier du nom *kazna* 'punition' et à la troisième personne du singulier du présent et à la deuxième personne du singulier de l'impératif du verbe *kazniti* 'punir'. Ainsi, l'exemple *pričati o kazni* signifie 'parler de la punition', avec la forme *pričati* étant l'infinitif du verbe 'parler', *o* étant la préposition 'de', et *kazni* le locatif du singulier du nom commun *kazna* 'punition'. Notre hypothèse était donc que la majorité des erreurs dans l'annotation des noms communs serait causée par la confusion avec les catégories du verbe et de l'adjectif. Les résultats des analyses quantitatives effectuées la confirment : le nom commun a été annoté comme adjectif 143 fois, et comme verbe principal 99 fois, ce qui fait au total 67,4% des erreurs dans l'annotation des noms communs. Ces résultats sont donnés dans le Tableau 29.

Classe attribuée	Nombre d'occurrences de confusion	% du nombre total d'erreurs pour NOM :com
ADJ	143	46,1%
VER	99	21,3%

Tableau 29 : Distribution de confusion pour NOM:com

IV.2.1.3 Verbe principal

L'homonymie de la catégorie du verbe avec celles de l'adjectif et du nom commun a déjà été décrite ci-dessus. S'ajoute à ces deux cas de figure un troisième, plus spécifique : le jeu d'étiquettes que nous avons défini fait la distinction entre les verbes principaux et les verbes auxiliaires. Comme c'est le cas en français avec le verbe *être*, le verbe *jesam* 'être' a deux fonctionnement : il peut être un verbe auxiliaire et participer dans une forme verbale composée, mais il peut également fonctionner comme un verbe attributif et avoir, par conséquent, le statut du verbe principal. La première possibilité peut être

illustrée par l'exemple suivant : *Marko je došao* 'Marko est venu', alors que la deuxième correspond à *Marko je lekar* 'Marko est médecin' ou bien *Marko je pošten* 'Marko est honnête', le mot souligné étant à chaque fois la forme du verbe *jesam*.

La désambiguïsation de ces formes repose donc sur la présence ou l'absence dans l'entourage de *jesam* d'un autre verbe conjugué. Dans l'exemple cité ci-dessus, où la forme de l'auxiliaire est directement suivie de celle du participe, cela ne pose pas de problème, pourvu que le participe soit reconnu en tant que verbe. Cependant, comme la syntaxe du serbe permet que l'auxiliaire et le participe soient séparés par plusieurs autres constituants, les exemples comme celui qui suit ne sont pas rares :

Marko	je	svoj	ručak	u	rancu	zaboravio.
n. propre	v.aux.	adj. pos.	n.com.	prép.	n.com.	v.princ.
'Marko	est	son	déjeuner	dans	sac à dos	oublié.'

'Marko a oublié son déjeuner dans son sac à dos'.

Exemple glosé 12 : Discontinuité des formes verbales composées

Dans le cas où le sujet de la phrase ne serait pas exprimé, il est même possible que le participe et l'auxiliaire se trouvent dans un ordre inversé :

Zaboravio	je	svoj	ručak	u	rancu
v. princ.	v.aux.	pron.réfléchi	n.com.	prép.	n.com.
oublié	est	son	déjeuner	dans	sac à dos

'Il a oublié son déjeuner dans son sac - à - dos.'

Exemple glosé 13 : Ordre inversé du verbe principal et verbe auxiliaire

Il était donc fort probable que cette flexibilité de la structure phrastique pose des difficultés dans l'étiquetage. Ainsi nous attendions-nous à un certain degré de confusion entre les catégories du verbe principal et du verbe auxiliaire.

Les résultats obtenus sont conformes aux intuitions citées ci-dessus : le verbe principal est effectivement le plus souvent confondu avec l'adjectif, le nom commun et le verbe auxiliaire. En effet, la confusion avec ces catégories constitue 94,1% de la totalité des erreurs faites dans l'annotation du verbe principal. La distribution des erreurs par catégorie est présentée dans le Tableau 30.

Classe attribuée	Nombre d'occurrences de confusion	% du nombre total d'erreurs pour VER
ADJ	104	40,3%
NOM:com	70	27,1%
VER:AUX	69	26,7%

Tableau 30 : Distribution de confusion pour VER

IV.2.2 Analyse des exemples par catégorie

Pour avoir une meilleure compréhension des formes et des structures posant le plus de problèmes dans l'étiquetage du serbe, on procédera à une analyse des erreurs prototypiques, en essayant à chaque fois de déterminer quels facteurs ont pu causer la mauvaise décision de l'étiqueteur. Nous consacrerons le plus d'attention aux catégories qui se sont avérées critiques dans les analyses quantitatives, avec un ajout possible d'autres exemples jugés pertinents et illustratifs.

IV.2.2.1 Adjectif

IV.2.2.1.1 Confusion avec nom commun

Comme on a vu dans la partie II.3.1.4, l'adjectif serbe a deux positions canoniques : antéposé au nom qu'il détermine en tant qu'épithète (*lepa kuća - belle maison*) ou dans la position de l'attribut du sujet, derrière le verbe attributif (*Kuća je lepa,* où *kuća* est le nom 'maison', *je* le verbe 'être', et *lepa* l'adjectf 'beau' au féminin). Pourtant, les positions en question ne sont pas strictement adjectivales : dans la majorité des cas, la position occupée par un adjectif admettrait également un nom commun. On peut ainsi avoir deux noms communs juxtaposés comme dans *procena štete* (*estimation des dégâts*), une construction rare, en français contemporain, restreinte aux emplois de type *tarte maison* ou *service client*. En serbe, cette construction est très productive, et elle est possible grâce à la flexion : le deuxième nom est toujours décliné, et il se trouve ici au génitif du singulier. Quant à la construction attributive, la situation est comparable au français : les phrases *Marko je lekar* 'Marko est médecin' et *Marko je lep* 'Marko est beau' sont les deux non seulement grammaticales, mais très fréquentes.

Effectivement, on trouve dans le corpus des occurrences qui correspondent à ces cas de figure. Exemple glosé 14 illustre les occurrences où un adjectif qui a la fonction de l'attribut de sujet a été annoté comme nom commun, ce qui est une interprétation

acceptable de point de vue syntaxique. En l'occurrence, c'est le token *bogat* qui a été mal annoté :

kako	sam	u	svojim	snovima	bio	*bogat*
conj. sub.	ver. aux.	prép.	dét. pos.	nom com.	ver.	adj.
comme	suis	dans	mes	rêves	été	riche

'que j'étais riche dans mes rêves'

Exemple glosé 14

Néanmoins, la plupart des exemples analysés concernent les adjectifs antéposés aux noms communs et qui se trouvent donc dans la position d'épithète. Les occurrences suivantes ont été rencontrées (la forme adjectivale mal annotée est soulignée dans tous les exemples) : *uzak ravan obod* (litt. 'étroit plat bord'), *onog jezivog dana* (litt. 'cet horrible jour'), *te mutne jesenje večeri* (litt. 'ces embrumées automnales soirées'). L'interprétation de BTagger est acceptable en ce qui concerne la distribution des parties du discours dans le contexte immédiat, vu que le système casuel serbe permet la juxtaposition des noms et/ou des groupes nominaux. On peut conclure donc qu'il n'existe pas de distinction nette entre les contextes nominal et adjectival dans la structure de la phrase serbe.

IV.2.2.1.2 Confusion avec verbe principal

Une partie de ces erreurs relève des adjectifs homonymes avec les formes du participe passé.

Comme il a déjà été mentionné, selon les principes d'étiquetage que nous avons adoptés dans la constitution du corpus d'entraînement, la différence entre le participe et l'adjectif est purement syntaxique : si la forme en question participe à un temps composé, il s'agit d'un participe, alors qu'elle est interprétée comme adjectif si employée indépendamment d'un verbe auxiliaire.

Le recoupement entre les paradigmes verbal et adjectival a été présenté dans la partie II.3.1.4. Nous avons vu qu'il est impossible d'utiliser les indices morphologiques pour distinguer un participe d'un adjectif, étant donné que les formes du participe passé sont identiques aux formes du nominatif de l'adjectif correspondant. Or, dans le sens inverse, le recoupement des paradigmes n'est pas total : ce sont seulement les formes du nominatif d'un adjectif qui coïncident avec les formes du participe correspondant. Par conséquent, pour que la confusion d'un adjectif avec un participe soit justifiée, l'adjectif

doit être à l'une des six formes du nominatif qu'il connaît (trois genres et deux nombres) .Par exemple, la forme *obučenog,* qui est le génitif du singulier du genre masculin de l'adjectif 'vêtu' ne peut pas être interprétée comme le participe passé *obučen* parce que le participe connaît seulement les formes suivantes : *obučen* (nom.sg.m.), *obučena* (nom.sg.f. et nom.pl.n.), *obučeno* (nom.sg.n.), *obučeni* (nom.pl.m.), *obučene* (nom.pl.f.); Les occurrences des erreurs repérées dans le corpus indiquent que BTagger a identifié cette règle : les adjectifs auxquels le logiciel a attribué l'étiquette du verbe sont en effet au nominatif. Ce cas de figure est illustré dans Exemple glosé 15, avec l'adjectif mal annoté souligné et mis en italiques.

njegova	nema	opomena
ADJ pos. nom.sg.f.	ADJ nom.sg.f..	N nom.sg..
son	muet	avertissement

'son avertissement muet'

Exemple glosé 15

Cet exemple confirme encore une fois que la distinction entre ces deux classes ne peut pas reposer sur les indices morphologiques. Quant à l'analyse du contexte, on peut mentionner le fait qu'un participe passé doit être accompagné par un verbe auxiliaire avec lequel il constitue une forme verbale composée. Pourtant, la seule présence d'un verbe auxiliaire n'est pas un paramètre suffisant, vu qu'il existe des exemples comme le suivant :

Zaboravljeno	klatno	je	stajalo	u	uglu	sobe.
ADJ nom.sg.n.	N sg.n.	VA 3p.sg.prés.	V pp. sg.n.	PREP	N dat.sg.	N gén.sg.
oublié	pendule	est	trouvé	dans	coin	chambre

'La pendule oubliée se trouvait dans le coin de la chambre.'

Exemple glosé 16

On voit qu'ici, la forme *zaboravljeno* (adjectif déverbal *zaboravljen* 'oublié') a la fonction d'épithète du nom *klatno* 'pendule' ; cependant, le prédicat de la phrase est une forme verbale composée *je stajalo* (parfait du verbe *stajati* 'se tenir') : le verbe auxiliaire *je* est séparé de l'adjectif déverbal par un seul token (*klatno*). La présence du verbe auxiliaire dans le contexte de l'adjectif déverbal ne peut donc pas être exploitée comme un paramètre fiable.

IV.2.2.1.3 Confusion avec adverbe

Nous avons déjà vu que la catégorie des adverbes exhibe une homonymie systématique avec les adjectifs qualificatifs : la forme des adverbes de manière est équivalente à celle du nominatif du genre neutre,ou dans certains cas du genre masculin,de l'adjectif qualificatif correspondant. Ainsi, la forme dobro peut correspondre au genre neutre de l'adjectif *dobar* 'bon' (cf. *dobro rešenje* 'bonne solution'), mais aussi à l'adverbe *dobro* 'bien' (cf. *dobro plivati* 'nager bien'). Dans le cas de stoički, il peut s'agir aussi bien du genre neutre de l'adjectif *stoički* 'stoïque' (*stoički otpor* 'resistence stoïque') et de l'adverbe *stoički* 'stoïquement' (*stoički trpeti* 'endurer stoïquement'). Les erreurs analysées montrent que la confusion entre les adjectifs et les adverbes est presque exclusivement causée par ce phénomène. Une majorité écrasante des exemples relèvent précisément des cas de figure cités. Dans l'exemple suivant, la forme a été identifiée comme adverbe, alors qu'il s'agit d'un adjectif :

Otkrivam	samo	*duboko*	ćutanje.
V 1p.sg.prés.	ADV	ADJ acc.sg.n.	N acc.sg.
découvre	seulement	profond	silence

'Je ne découvre qu'un silence profond.'

Exemple glosé 17

Néanmoins, cette forme peut également être un adverbe, comme dans l'exemple *kopati duboko*, lit. 'creuser profondément'.

Les indicateurs morphologiques ayant une telle influence, on se demande si l'analyse du contexte peut permettre la distinction entre ces deux parties de discours. En effet, les deux catégories disposent d'une mobilité importante : les adverbes peuvent être antéposés ou postposés au verbe, et même en être détachés. Cela est illustré dans les Exemples glosés 18-20.

Marko	lepo	peva
N nom.sg.	ADV	V 3p.sg.prés.
Marko	bien	chante

'Marko chante bien.'

Exemple glosé 18

Marko	peva	lepo
N nom.sg.	V 3p.sg.prés.	ADV
Marko	chante	bien

‘Marko chante bien.’

Exemple glosé 19

Lepo	Marko	peva
ADV.	N nom.sg	V 3p.sg.prés.
bien	Marko	chante

‘Marko chante bien.’

Exemple glosé 20

Une flexibilité syntactique existe également entre les adjectifs et les noms qu’ils déterminent : un adjectif épithète est normalement antéposé au nom (cf. *lep čovek* ‘bel homme’, où *lep* est l’adjectif beau, et *čovek* le nom ‘homme’), mais il peut également en être détaché :

Lepog	sam	čoveka	videla.
ADJ acc.sg.m.	VA 1p.sg.prés.	N acc.sg.	V pp. sg.f.
beau	est	homme	vu

‘J’ai vu un bel homme.’

Exemple glosé 21

Pour un humain, il semble difficile d’opposer les contextes adjectival et adverbial sur la base des critères distributionnels fiables qui faciliteraient leur discrimination et il est clair, vu les erreurs rencontrées, que le logiciel ne parvient pas à établir un mécanisme de distinction parfait. Néanmoins, il ne faut pas oublier que le nombre de ces erreurs reste bas : sur la totalité de 57 000 tokens du sous-corpus annoté, seulement 32 occurrences de confusion d’un adjectif avec un adverbe ont été détectées. Cela indique que le logiciel gère bien l’ambiguïté entre ces deux parties de discours.

IV.2.2.2 Nom commun

IV.2.2.2.1 Confusion avec les adjectifs

Une partie de ces erreurs représente le cas de figure prévu dans l'analyse de la morphologie du serbe et évoqué dans la partie précédente : les noms dérivés des adjectifs par substantivation. Les exemples incluent les formes suivantes : *žuto* (genre neutre de l'adjectif *žut* 'jaune' et le nom neutre *le jaune*), *tajna* (genre féminin de l'adjectif *tajan* 'secret' et nom féminin *le secret*), *nevini* (pluriel masculin de l'adjectif *innocent* et le nom *un innocent*).

Comme dans ces cas la forme du nom et celle de l'adjectif coïncident, le logiciel est obligé de s'appuyer sur le contexte pour déterminer quelle catégorie accorder au mot en question. Cependant, le contexte lui-même n'est pas nécessairement suffisamment contraignant. Dans l'exemple suivant :

čitava	vojska	<u>zlih</u>	i	<u>dobrih</u>,	<u>grešnih</u>	i	<u>nevinih</u>
adj.	nom.com.	nom.com	conj. coor.	nom.com.	nom.com.	conj. coor.	nom.com.
toute	armée	méchants	et	bons	pécheurs	et	innocents

'toute une armée de méchants et de bons, de pécheurs et d'innocents'

Exemple glosé 22 : Noms dérivés des adjectifs

les termes soulignés sont des noms dérivés des adjectifs par conversion, ayant la place et le fonctionnement des noms. Pourtant, ce contexte admettrait sans difficultés des adjectifs à leur place :

čitava	vojska	<u>opremljena</u>	i	<u>naoružana,</u>	<u>organizovana</u>	i	<u>disciplinovana</u>
adj.	nom. com.	adj.	conj. coor.	adj.	adj.	conj. coor.	adj.
toute	armée	équipée	et	armée	organisée	et	disciplinée

'toute l'armée équipée et armée, organisée et disciplinée'

Exemple glosé 23 : Contexte non contraignant

L'absence des indices morphologiques est donc accompagnée d'un contexte ambigu, ce qui rend la bonne détermination de la catégorie grammaticale extrêmement difficile.

Notons cependant que les noms obtenus par conversion des adjectifs sont sources d'erreurs moins souvent qu'on ne l'avait initialement prévu : seulement 22 occurrences

en ont été repérées, ce qui fait 7% de toutes les erreurs faites dans l'étiquetage des noms communs. Les autres erreurs proviennent en grande majorité de véritables noms, tels *đerdan* 'collier', *čvorove* 'nœuds' ou *sapun* 'savon'.

Vu que les trois formes citées ne sont pas polysémiques et peuvent être interprétées seulement comme noms, la question sur la cause de ces erreurs se pose. Une première explication possible se trouve au niveau morphologique : même s'il s'agit des noms, les terminaisons de ces trois formes apparaissent également dans le paradigme adjectival. La désinence *-an* qu'on trouve dans *đerdan* existe également dans les adjectif *mračan* 'sombre', *tužan* 'triste', *bitan* 'important' etc. Une partie des adjectifs possessifs dérivés des noms utilise le suffixe *-ov* et quelques-unes de leurs formes fléchies se terminent par *-ove* : *Petrove* 'qui appartient à Petar', *Pavlove* 'qui appartient à Pavle', *lavove* 'qui appartient au lion' etc, toutes les formes citées étant au génitif du singulier du féminin/accusatif du pluriel du masculin ou du féminin, ces trois formes étant homonymes. Ces formes se terminent donc de la même façon que la forme *čvorove,* l'accusatif du pluriel du nom *čvor*. Pour le dernier exemple cité, celui du nom *sapun*, sa terminaison correspond à la forme de l'adjectif *pun* 'plein'. Comme la préfixation en *sa-* est une forme fréquente de dérivation en serbe, on peut supposer que le token *sapun* a été reconnu comme un dérivé de l'adjectif *pun*.

Si l'on analyse plus en détails l'exemple du nom *đerdan* 'collier', on remarque également une influence possible du contexte. En effet, l'entourage immédiat de ce token est comme suit :

čitav	đerdan	*malih*	*limenih*	okaca
adj.	nom. com.	adj.	adj.	nom com.
tout	collier	petits	métalliques	ronds

'Tout un collier de petits ronds métalliques'

Exemple glosé 24 : Suite d'adjectifs antéposés au nom

Les termes soulignés dans l'Exemple glosé 24 sont des adjectifs dont le premier détermine le nom *đerdan* 'collier', tête de ce groupe nominal, alors que les deux autres sont épithètes du nom *okce* 'ronds'. Pourtant, comme les suites des adjectifs épithètes antéposés au nom ne sont pas rares en serbe, la position occupée par le nom *đerdan* permet également un quatrième adjectif. On remarque donc qu'ici aussi l'interprétation

adjectivale est suggérée aussi bien par l'indice morphologique que par le contexte du token.

Cependant, dans le cas du token *sapun*, le contexte ne justifie pas l'erreur faite par l'étiqueteur. Le mot en question figure dans le contexte suivant :

prostorija	koja	miriše	na	petrolej,	sapun,	cikoriju	i	čaj	od	kamilice
nom com.	pron. rel.	ver.	prép.	nom com.	nom com.	nom com.	conj. coor.	nom com.	prép.	nom com.
pièce	qui	sent	sur	pétrole lampant	savon	chicorée	et	tisane	de	camomille

'La pièce qui sent le pétrole lampant, le savon, la chicorée et la tisane de camomille'

Exemple glosé 25

Comme le nom *sapun* se trouve ici coordonné à deux autres noms *petrolej* 'pétrole lampant' et *cikorija* 'chicorée'(dans la phrase dans la forme de l'accusatif du singulier *cikoriju*) et à un groupe nominal *čaj od kamilice* 'tisane de camomille', ce contexte admettrait difficilement un adjectif à sa place. Il existe donc deux explications possibles : soit le logiciel n'a pas détecté la contrainte contextuelle qui interdit l'adjectif à cette position, soit l'indice morphologique a prévalu sur le résultat de l'analyse distributionnelle.

Quant au troisième exemple présenté ci-dessus (*čvorove*), il se trouve dans le contexte suivant :

vezujući	u	duple	kočijaške	čvorove	debelu	nit	svog	astralnog	sna
ver.	prép.	adj.	adj.	nom com.	adj.	nom com.	adj. pos.	adj.	nom com.
nouant	en	double	de cocher	nœuds	gros	fil	son	astral	rêve

'nouant le gros fil de son rêve astral en doubles nœuds de cocher'

Exemple glosé 26

Ici, la cause de l'erreur est plus difficile à identifier, vu que l'étiquetage erroné de la forme *čvorove* a été accompagné de plusieurs autres erreurs : BTagger a traité la forme *kočijaške* comme nom commun au lieu de l'analyser comme adjectif, *čvorove* comme adjectif au lieu de nom commun, *debelu* comme nom commun au lieu d'adjectif et *nit* comme conjonction de coordination au lieu de nom commun. La comparaison entre la

suite des étiquettes correctes et celle qui a été attribuée aux tokens en question se trouve dans le Tableau 31, avec les interprétations erronées de BTagger mises en gras.

Tokens	u	duple	kočijaške	čvorove	debelu	nit	svog	astralnog	sna
Etiquetage erroné	PRP	ADJ	**NOM:com**	**ADJ**	**NOM:com**	**KONJ:COOR**	ADJ:POS	ADJ	NOM:com
Etiquetage correct	PRP	ADJ	ADJ	NOM:com	ADJ	NOM:com	ADJ:POS	ADJ	NOM:com

Tableau 31 : Etiquetage erroné et étiquetage correcte du contexte du token 'čvorove'

Quant aux contraintes contextuelles, la syntaxe serbe admet la suite des étiquettes erronée. Pourtant, si la dernière des erreurs énumérées est due à l'homonymie entre la conjonction de coordination *nit(i)* 'ni*',* 'non plus' et le nom commun *nit* 'fil*'*, et que la forme *debelu* puisse également correspondre à un nom, l'interprétation nominale est peu probable pour le token *kočijaške*, et même dans le cas de *debelu* l'acceptation adjectivale est plus probable, vu qu'il s'agit d'un adjectif. Vu le nombre élevé des erreurs dans cet exemple, il est difficile de déterminer quel forme ou quelle partie du contexte a déclenché cette série de mauvaises décisions de BTagger.

IV.2.2.2.2 Confusion avec les verbes

Comme il a été annoncé dans l'analyse de l'homonymie entre les noms communs et les verbes, une partie de la confusion entre ces deux catégories est liée au cas où une forme du paradigme verbal coïncide avec une forme fléchie d'un nom commun. Les exemples suivants ont été relevés : *osvete* (nom *vengeance* et verbe *venger*), *pogleda* (nom *regard* et verbe *regarder*), *odnosi* (nom *rapport* et verbe *emporter*), *pravila* (nom *règle* et verbe *créer, faire*).

Il existe un autre cas de figure qui semble être lié à la morphologie. Il s'agit des formes nominales qui ne peuvent pas être interprétées comme verbes, mais dont la terminaison appartient également au paradigme verbal. Par exemple, le token *vanile* désigne une forme fléchie du nom *vanila* 'vanille'. Sa terminaison *-ile* coïncide avec celle du pluriel du féminin du participe passé actif de certains verbes (cf. *radile, pile, krile* - le pluriel du féminin du participe passé actif des verbes *raditi* 'travailler', *piti* 'boire' et *kriti* 'cacher', respectivement). Le token *povrće* correspond à la forme canonique du nom commun *povrće* '*légumes*', mais la terminaison *-će* est typique du futur (cf. *radiće* 'il travaillera', *pričaće* 'il parlera', *imaće* 'il aura'). De même, le token *saksije* est une forme fléchie du

nom *saksija* ‘pot-à-fleurs’, mais il existe de nombreux verbes dont la troisième personne du singulier du présent se termine par *-ije* : *pije* ‘il boit’, *bije* ‘il bat’, *krije* ‘il cache’.

Pour les deux cas de figure mentionné, la même question se pose : les contextes verbal et nominal ne sont-ils pas suffisamment distincts pour prévenir ce type de confusion ? En effet, il semble que la réponse est négative.

Dans le cas du token *osvete*, le contexte à considérer est le suivant :

kao	žrtva	osvete	bogova	cveća
con. sub.	nom com.	nom com.	nom com.	nom com.
comme	victime	vengeance	dieux	fleurs

‘comme une victime de la vengeance des dieux des fleurs’

Exemple glosé 27

Pourtant, la syntaxe serbe admettrait également la suite des étiquettes où la troisième position serait occupée par un verbe. Vu que même les propriétés formelles du token *osvete* justifient cette interprétation, la décision de l’étiqueteur n’est pas étonnante.

Cependant, le cas du mot *pogleda* semble moins justifié. Ce token figure dans le contexte suivant :

suma	nove	religije	i	novog	pogleda	na	svet
nom com.	adj.	nom com.	conj. coor.	adj.	nom com.	prép.	nom com.
somme	nouvelle	religion	et	nouveau	regard	sur	monde

‘la somme de la nouvelle religion et du nouveau regard sur le monde’

Exemple glosé 28

Ici, la position occupée par la forme nominale *pogleda* admettrait difficilement un verbe. Pourtant, cette contrainte semble ne pas être suffisamment forte, vu qu’elle n’a pas été assimilée par le logiciel.

L’exemple du token *saksije* se trouve entre les deux premiers cas de figure. Le contexte dans lequel figure ce mot est donné dans l’Exemple glosé 29.

ormari,	posuđe,	saksije	s	fikusima,	saksije	s	oleandrima
nom com.	nom com.	nom com.	prép.	nom com.	nom com.	prép.	nom com.
armoires	vaisselle	pots-à-fleurs	avec	ficus	pots-à-fleurs	avec	lauriers roses

‘des armoires, de la vaisselle, des pots-à-fleurs avec des ficus, des pots-à-fleurs avec des lauriers roses’

Exemple glosé 29

Dans ce cas, le contexte immédiat (cf. le groupe prépositionnel qui suit) admettrait également un verbe à la place de la forme fléchie du nom *saksija*. Cependant, ces syntagmes verbaux s’intégreraient difficilement dans le contexte plus large, qui ne contient qu’une juxtaposition des groupes nominaux.

Un troisième cas de figure, comprenant un petit nombre d’occurrences, concerne les erreurs où la forme annotée en tant que verbe est complètement étrangère au paradigme verbal. Les occurrences relevées incluent les exemples suivants : *potop* (nom commun *déluge*), *dodir* (nom commun *toucher*), *jesen* (nom commun *automne*).

Comme ces formes ne peuvent appartenir qu’à la catégorie des noms, et que leurs propriétés formelles ne permettent pas de les interpréter comme des verbes, la seule explication possible des erreurs faites se trouve dans le contexte. Le token *potop* figure dans l’environnement immédiat suivant :

potop	je	još	samo	daleka	uspomena
nom com.	ver.	adv.	adv.	adj.	nom com.
déluge	est	encore	seulement	lointain	souvenir

‘Le déluge n’est plus qu’un souvenir lointain.’

Exemple glosé 30

En effet, la position initiale dans la phrase contenant une forme verbale composée est souvent réservée au participe passé du verbe principal : *Došao je juče* ‘Il est venu hier’, *Pričala sam sa Marijom* ‘J’ai parlé avec Marija’, où les formes *došao* et *pričala* sont celles du participe passé actif des verbes *doći* ‘venir’ et *pričati* ‘parler’ respectivement. Par conséquent, il est probable que le logiciel a interprété la forme en question comme le verbe principal. Cette hypothèse est confirmée par le fait que le token *je* (le présent du verbe *jesam* ‘être’) a été annoté comme verbe auxiliaire, et non pas comme verbe principal.

La situation est semblable dans les cas de figure représentés par l'exemple du token *dodir*. Son contexte immédiat est donné dans l'Exemple glosé 31..

Pričalo	se	da	dodir	njegova	štapa	ima	čarobnu	moć
ver.	pron.	conj.	nom	adj.	nom	ver.	adj.	nom
	réf.	sub.	com.	pos.	com.			com.
Disait	se	que	toucher	sa	canne	a	magique	pouvoir

'On disait que le toucher de sa canne avait un pouvoir magique.'

Exemple glosé 31

On peut voir que le nom *dodir* qui a été mal annoté comme verbe suit directement la conjonction de subordination *da*, introductrice des propositions complétives. Comme le serbe est une langue *pro-drop*, ces propositions commencent souvent par la forme verbale : *Rekao je da učestvuje (Il a dit qu'il participait)*, *Mislim da dolazi (Je pense qu'il vient)* etc., où les formes *učestvuje* et *dolazi* sont respectivement les formes du présent des verbes *učestvovati* 'participer' et *dolaziti* 'venir'. Il est possible que la mauvaise interprétation de la forme *dodir* par BTagger soit liée au fait qu'elle se trouve à une position fréquemment occupée par des verbes. Pourtant, cette position n'est pas exclusivement réservée aux formes verbales, et il est difficile de déterminer si elle favorise réellement l'une de ces deux catégories. Notre interprétation reste donc une simple tentative d'analyse et d'explication.

Quant au troisième exemple évoqué, celui du token *jesen*, il se trouve dans le contexte suivant :

Jesen	te	godine,	po	odlasku	mog	oca,	došla	je	u	znaku	tišine
nom	adj.	nom	prép.	nom	adj.	nom	ver.	ver.	prép.	nom	nom
com.	dem.	com.		com.	pos.	com.		aux.		com.	com.
Automne	cette	année	après	départ	mon	père	arrivée	est	dans	signe	silence

'L'automne de cette année, après le départ de mon père, est arrivée sous l'enseigne de silence.'

Exemple glosé 32

On voit que le contenu de la proposition principale dans l'Exemple glosé 32est discontinu : le syntagme prépositionnel inséré (*po odlasku mog oca*) sépare le sujet (*jesen te godine*) du groupe verbale (*došla je u znaku tišine*). Or, vu que le serbe est une langue *pro-drop*, le groupe verbal mentionné peut être interprété comme une proposition indépendante (*elle est venue sous l'enseigne de silence*) ; dans ce cas-là, la

première partie de la phrase devrait avoir une forme verbale à elle, et la seule position disponible serait celle du token *jesen*.

IV.2.2.2.3 Confusion avec les adverbes

Même si les occurrences de la confusion avec cette catégorie ne sont pas nombreuses (on n'en a identifié que 14), elles sont illustratives de l'importance des indices morphologiques dans la prise de décision de BTagger. Ces exemples sont autant plus important vu le fait que la position des adverbes dans la phrase serbe est extrêmement variable et que, par conséquent, l'analyse du contexte ne permet pas de désambiguïser ces formes.

Quelques erreurs causées par l'homonymie ont été identifiées: la forme *čas* correspond aux noms communs *heure* et *cours*, mais participe également à la construction *čas*... *čas*... 'tantôt... tantôt...', où elle est un adverbe. Cependant, 10/14 occurrences de confusion d'un nom commun avec un adverbe relèvent des cas où le nom en question ne peut pas appartenir à la classe des adverbes mais se termine par *-o*, ce qui est la terminaison typique des adverbes de manière. Ce morphème appartient également à la flexion nominale et adjectivale. Les cas relevés incluent les exemples suivants : *železo* 'fer', *koleno* 'genou', *sudbino* 'destin'.

IV.2.2.3 Verbe

IV.2.2.3.1 Confusion avec l'adjectif

De tous les cas de confusion étudiés dans ce travail, la confusion des verbes avec les adjectifs montre le degré le plus important de régularité. Il s'agit majoritairement des occurrences du participe passé actif ou passif, dont la forme existe également en tant qu'adjectif. Comme il a déjà été mentionné (cf. partie IV.2.2.1.2), il n'est pas possible de distinguer ces deux formes en utilisant des indices morphologiques, vu que le participe passé correspond parfaitement à certaines formes de l'adjectif qui en est dérivé. Pourtant, une désambiguïsation basée sur le contexte devrait être possible. En effet, le participe passé fait partie d'une forme verbale composée et est le plus souvent accompagné d'un verbe auxiliaire dans son contexte plus ou moins immédiat[20]. Ce

[20] On exclue de cette règle les formes composées des verbes pronominaux, qui perdent la forme du verbe auxiliaire à la troisième personne du singulier : dans l'exemple *Ona se rodila* 'Elle est née', le pronom personnel *ona* 'elle' est suivi du pronom réfléchi *se* 'se' et du participe passé *rodila*, forme du verbe *roditi* 'naître' (qui est pronominal en serbe).La variante avec le verbe auxiliaire **Ona je se rodila* n'est pas grammaticale. On exclue

facteur devrait être exploitable pour déterminer si la forme ambiguë est un adjectif ou un participe. Or, cela n'est pas le cas. Dans la plupart des cas où le participe a été annoté comme adjectif le token ambigu occupe une position dans la phrase qui permet les deux interprétations. Dans l'Exemple glosé 33, la forme mal annotée est *zabranjen*, participe passé passif du verbe *zabraniti* 'interdire'.

gde	je	nama	bio	*zabranjen*	svaki	pristup
adv. rel.	ver. aux.	pron. per.	ver. aux.	ver.	adj. ind.	nom com.
où	avait	nous	été	interdit	tout	accès

'où tout accès nous avait été interdit'

Exemple glosé 33

La forme en question se trouve dans une position ambiguë : elle peut appartenir à la forme verbale composée (a), mais elle peut également être occupée par un premier adjectif épithète du groupe nominal dont la tête est le nom *pristup* (b). Les deux possibilités sont illustrées dans le Tableau 32.

a) 'zabranjen' participe passé	gde **je** nama **bio zabranjen** svaki pristup
b) 'zabranjen' adjectif épithète	gde je nama bio **zabranjen svaki pristup**

Tableau 32

Il faut souligner que dans la phrase en question la deuxième interprétation n'est pas acceptable, vu la sémantique du groupe nominal. Elle est possible cependant dans un exemple comme 'gde su bila polupana neka kola', qui peut avoir deux interprétations :

gde	su	bila	polupana	neka	kola
adv. rel.	ver. aux.	ver. aux..	ver. part.pass.	adj. ind.	nom.com.
où	sont	été	abîmée	certaine	voiture

'où une voiture avait été abîmée'

Exemple glosé 34: Interprétation A

gde	su	bila	polupana	neka	kola
adv. rel.	ver. aux.	ver. part.pass.	adj.	adj. ind.	nom.com.
où	sont	été	abîmée	certaine	voiture

'où il y avait une voiture abîmée'

Exemple glosé 35: Interprétation B

également les formes du parfait dit 'journalistique' : *Skupština usvojila novi zakon* 'Le Parlement a adopté la nouvelle loi', où la forme pleine du parfait serait *je usvojila*.

Pourtant, comme il a déjà été expliqué, le logiciel n'a pas accès aux éléments du sens ; les seuls facteurs qu'il utilise pour prendre une décision sont les propriétés morphologiques et les suites des étiquettes rencontrées dans l'entraînement. Vu qu'ici la forme du mot ne permet pas de désambiguïser, le logiciel doit s'appuyer sur la seule analyse distributionnelle et, comme on l'a vu, dans le cas de figure en question, le contexte lui-même n'est pas univoque.

Quelques autres types d'homonymie ont également été détectés. Il s'agit d'adjectifs dont une forme fléchie coïncide avec la forme verbale rencontrée dans le texte. On peut citer les exemples de *lepe* - adjectif *lep* 'beau' et verbe *lepiti* 'coller', *tamni* - adjectif *taman* 'sombre' et verbe *tamneti* 'sombrir', et *oštri* - adjectif *oštar* 'aigu' et verbe *oštriti* 'aiguiser'. La majorité de ces cas suivent le même modèle, illustré par l'Exemple glosé 36, la forme soulignée étant le token ambigu.

kako	mi	se	<u>*lepe*</u>	kapci	od	sna	i	umora
conj.	pron.	pron.	ver.	nom.	prép.	nom	conj.	nom
coor.	per.	réf.		com.		com.	coor.	com.
que	me	se	collent	paupières	de	sommeil	et	fatigue

'que mes paupières collaient du sommeil et de la fatigue'

Exemple glosé 36

Si on analyse l'étiquetage attribué à cette suite des tokens par BTagger, on voit que la forme *lepe*, qui a été annotée comme adjectif, figure directement devant un nom (*kapci* 'paupières'). Ceci étant l'une des positions les plus typiques de l'adjectif en serbe, le contexte immédiat n'exclue donc pas cette interprétation. Pourtant, elle laisse la proposition sans forme verbale et la rend par conséquent agrammaticale. Il est possible que ce problème soit résolu dans un contexte plus large : il se peut que la proposition ait été interprétée en tant qu'un groupe nominal et attaché, dans l'interprétation du logiciel, à une forme verbale plus éloignée. Une autre explication possible serait que BTagger ne soit pas capable de détecter les frontières entre les propositions et d'utiliser la contrainte d'avoir une forme verbale par proposition dans la prise de décision.

IV.2.2.3.2 Confusion avec le nom commun

Un cas problématique prévu concerne les verbes et les noms dont certaines formes fléchies sont homonymes. En effet, un certain nombre des occurrences de ce type a été détecté : *hita* -gén.sg. du nom *hit* ('hit') et 3p. sg. prés. duverbe *hitati* ('se dépêcher'), *sinu* - dat.sg. du nom *sin* ('fils') et 3p. sg .aoriste du verbe *sinuti* ('apparaître'*otpada* - gén. sg.

du nom *otpad* (‘déchets’) et 3.p.sg. prés. du verbe *otpadati* (‘se détacher’). La structure de ces occurrences est illustrée par l’Exemple glosé 37.

pozlata	otpada	u	tankim	finim	ljuspicama
nom com.	ver.	prép.	adj.	adj.	nom com.
dorure	s’écaille	en	minces	fines	paillettes

‘La dorure s’écaille en minces paillettes fines’

Exemple glosé 37

Comme c’était le cas concernant la confusion des verbes avec les noms communs, ici aussi le contexte immédiat admet des mots appartenant à la catégorie attribuée par BTagger. Pourtant, cette interprétation a pour conséquence que la proposition en question reste sans forme verbale. Les mêmes explications qui ont été données pour les noms communs sont applicables dans ce cas : soit le contexte plus large permet la combinatoire des étiquettes proposées par BTagger et ce choix est justifié, soit l’étiqueteur ne réussit pas à identifier la présence de la forme verbale dans la proposition comme contraignante et il s’agit alors d’un véritable problème d’apprentissage.

Un cas particulier dans ce groupe d’erreurs est représenté par les occurrences du participe passé passif qui ont été étiquetées comme noms communs, telle la forme *naslikana*, singulier du féminin du participe passé passif du verbe *naslikati* ‘peindre’ dans l’exemple suivant :

na	kojoj	je	bila	naslikana	jedna	gospođica
PREP	PRO:REL	VA	VA pp. sg.f.	V. pp. sg.f.	ADJ:IND	N nom.sg.
	loc.sg.f.	3p.sg.prés.			nom.sg.f.	
sur	laquell	est	été	peinte	une	mademoiselle

‘sur laquelle une mademoiselle était peinte’

Cette forme pourrait appartenir à la catégorie des noms communs par une double conversion : premièrement, le participe passé devrait être adjectivé, pour que l’adjectif ainsi obtenu soit ensuite substantivé. Pourtant, même si des exemples existent, ce procédé n’est pas fréquent dans le corpus, et la présence du verbe auxiliaire devrait permettre au logiciel de prendre la bonne décision. Dans une partie des participes mal annotés, l’étiquetage du contexte immédiat, qui contient souvent une autre erreur, présente une interprétation correcte du point de vue syntaxique. Pour la suite des

tokens *koja je bila bolesno vezana za telo* ‘qui était maladivement attachée au corps’, l’étiquetage erroné, accordé par BTagger, et l’étiquetage corrigé manuellement sont donnés dans le Tableau 33.

Tokens	koja	je	bila	bolesno	vezana	za	telo
BTagger	pron. rel.	ver. aux.	**ver.**	**adj.**	**nom com.**	prép.	nom com.
Manuel	pron. rel.	ver. aux.	ver. aux.	adv.	ver.	prép.	nom com.

Tableau 33

La proposition en question fait figurer une forme verbale surcomposée *je bila vezana* ; elle contient par conséquent deux verbes auxiliaires : *je* et *bila*, le premier étant la forme du présent du verbe auxiliaire *jesam* et l’autre son participe passé (voir les principes d’étiquetage dans la partie II.4.2). Or, BTagger a interprété la forme *bila* comme le verbe principal, en étiquetant en même temps l’adverbe *bolesno* comme adjectif (ce qui est justifié de point de vue morphologique - l’adverbe et l’adjectif sont homonymes) et le véritable verbe principal *vezana* comme nom commun. On obtient donc une phrase où le verbe principal est le verbe *jesam* ‘être’ au parfait, avec un groupe nominal dans la position de l’attribut de sujet. Même si du point de vue sémantique cette interprétation n’est pas acceptable, elle l’est en ce qui concerne la syntaxe.

IV.2.2.3.3 Confusion avec le verbe auxiliaire

L’hypothèse que la confusion entre le verbe principal et le verbe auxiliaire allait surtout porter sur les formes du présent du verbe *jesam* s’est avérée vraie : une grande majorité des occurrences concerne ce cas de figure, illustré par l’Exemple glosé 38.

da	*je*	on	taj	koji	se	žrtvuje
conj. sub.	ver.	pron. per.	pron. dem.	pron. rel.	pron. réf.	ver.
que	est	il	celui	qui	se	sacrifie

‘que c’est lui celui qui se sacrifie’

Exemple glosé 38

Dans l’exemple cité, la forme *je* est le verbe principal de la première proposition, *žrtvuje* fonctionne de la même manière dans la proposition relative subordonnée *koji se žrtvuje*. Pourtant, BTagger a identifié le verbe *je* comme étant un verbe auxiliaire ; comme la syntaxe serbe permet de séparer l’auxiliaire du verbe principal par un nombre important de mots, il est possible que le logiciel ait interprété les formes *je* et *žrtvuje* comme constituant une forme verbale composée. Cela indiquerait encore une fois que le logiciel n’est pas capable de distinguer les limites des propositions et qu’il n’a pas

intégré la règle de nécessité d'avoir un verbe principal par proposition. Cela est accentué davantage par d'autres exemples où la distance entre la forme du verbe *jesam* erronément annotée comme verbe auxiliaire se trouve à une distance de 14 ou 16 tokens de la forme verbale principale la plus proche. Dans la phrase suivante, la forme mal annotée et la forme du verbe principal la plus proche sont soulignées : *Zaboravljam da sam novorođenče i da od svih životnih senzacija, ljudskih i božanskih, najviše ako mogu da osetim i doživim scenski efekat sunca.*14 autres tokens séparent ces deux formes.

Les erreurs où la forme d'un verbe principal a été annotée comme le verbe auxiliaire incluent également un type d'exemple que nous n'avions pas prévu : il s'agit du participe du verbe *jesam* dans les temps surcomposés, tel le plus-que-parfait.

Le verbe auxiliaire y figure lui-même à un temps composé, ce qui veut dire que le temps surcomposé comporte un auxiliaire complexe (sa forme personnelle et son participe) et d'un verbe principal. Comme nous l'avons déjà expliqué dans la partie II.4.2 les principes d'annotation adoptés pour l'étiquetage du corpus d'entraînement ne prévoient pas de système de marquage des unités polylexicales. Par conséquent, il était impossible d'annoter l'auxiliaire *jesam* des temps surcomposés comme une unité. Il a été décidé d'étiqueter les deux formes du verbe *jesam* comme verbe auxiliaire : dans l'exemple *On je bio došao* 'Il était venu', les tokens composant la forme du plus-que-parfait *je bio došao* porteraient les étiquettes suivantes : VER:AUX VER:AUX VER.

Il s'est avéré que cette stratégie d'annotation mène à une confusion avec les exemples où le verbe *jesam* est un verbe attributif (le verbe principal de la phrase) et se trouve conjugué au parfait. Ce cas de figure est illustré par l'Exemple glosé 39..

teškoće	su	bile	ogromne
nom com.	ver. aux.	ver.	adj.
difficultés	sont	été	énormes

'Les difficultés étaient énormes.'

Exemple glosé 39

Ici, le verbe *jesam* est un verbe attributif. Il est conjugué au parfait (*su bile*), ce qui signifie que la forme *su* a la fonction du verbe auxiliaire, alors que *bile*, étant le participe passé, est le verbe principal. Or, la forme *bile* a été annotée par BTagger comme verbe auxiliaire. Pourtant, le logiciel a également indiqué que la forme *ogromne*, qui est un adjectif qualificatif ayant la fonction de l'attribut de sujet, était le verbe principal. Cette

interprétation de la phrase est incorrecte, mais la suite des tokens générée par le logiciel est syntaxiquement acceptable.

On pourrait donc supposer que l'annotation incorrecte des formes du verbe *jesam* est liée aux erreurs dans l'identification du verbe principal : la reconnaissance d'un token comme forme verbale principale induirait le logiciel à reconnaître la forme du parfait du verbe *jesam* comme deux auxiliaires faisant partie d'un temps surcomposé. Pourtant, tous les exemples de la mauvaise annotation du participe du verbe *jesam* ne correspondent pas à ce cas de figure.

En analysant l'annotation de l'Exemple glosé 40 proposée par BTagger, on peut voir que les formes *bio* et *je* portent l'étiquette du verbe auxiliaire, alors que la phrase ne dispose pas d'autres formes verbales. Ce fait montre les limites du logiciel : il n'est pas capable d'identifier la présence d'un verbe principal dans la phrase comme une nécessité.

Marksov	Kapital	*bio*	je	jedan	od	temelja	ove	nove	kosmogonije
adj.	nom com.	ver.	ver aux.	pron. ind.	prép.	nom com.	adj. ind.	adj.	nom com.
de Marx	Capital	été	est	un	de	fondation	cette	nouvelle	cosmogonie

'Le Capital de Marx était l'une des bases de cette nouvelle cosmogonie'

Exemple glosé 40

IV.2.2.3.4 Confusion avec d'autres parties du discours

Des occurrences de confusion du verbe principal avec d'autres parties de discours tels le pronom et la conjonction ont été repérées. La raison de ces erreurs n'est pas claire : si dans le cas de la forme *da*, qui est une conjonction de subordination, mais aussi une forme du présent du verbe *dati* (*donner*), il s'agit évidemment de l'homonymie, et si pour la forme verbale *ticalo* on peut dire qu'il existe une ressemblance formelle avec un certain nombre des formes fléchies du pronom démonstratif *taj* 'celui' (nominatif et vocatif du pluriel du masculin *ti*, génitif du pluriel des trois genres *tih*, datif, instrumental et locatif des trois genres *tim*) la cause des autres occurrences est beaucoup moins transparente1e personne du singulier du présent du verbe *kmečati* 'gémir' *kmečim* a été identifiée comme un pronom indéfini, alors que la 3e personne du singulier du présent du verbe *mukati* 'meugler' *muče* a été annotée comme un pronom personnel. Les indices contextuels ne donnent pas davantage d'informations que le fait que ces fautes sont le

plus souvent liées à d'autres erreurs, typiquement dans la reconnaissance de la forme du verbe principal.

Les deux occurrences d'homonymie entre la forme du présent du verbe *dati* 'donner', 'permettre' et de la conjonction de subordination *da* apparaissent dans le même contexte : les deux formes se suivent directement, et les deux sont indiquées comme conjonction de subordination. Ce contexte ne permet pas de décider laquelle des deux formes est le verbe et laquelle est la conjonction : *on ne da da pričamo* 'il ne permet pas qu'on parle', où la conjonction suit la forme verbale, est aussi possible que *on neće da da odobrenje* 'il ne veut pas donner l'autorisation', où la conjonction précède le verbe. Pourtant, si ces deux formes se suivent directement, il n'est pas possible qu'il s'agisse de deux conjonctions de subordination. Le logiciel devrait être obligé à étiqueter une d'entre elles comme verbe.

Quant à la confusion avec les pronoms personnels et indéfinis, ce type d'erreurs pourrait être corrigé par un post-traitement. La démarche consisterait à définir les membres de différentes sous-catégories des pronoms et de vérifier si les tokens portant ces tags appartiennent bien aux formes listées dans la définition de l'étiquette correspondante.

Même si l'analyse de ces erreurs offre des informations supplémentaires sur le fonctionnement du logiciel utilisé, il faut souligner que ces fautes ne font qu'une partie minime des fautes d'annotation : 7 erreurs sur la totalité du sous-corpus *Bašta*.

IV.2.3 Conclusion provisoire

L'analyse qualitative de l'étiquetage présentée ci-dessus a éclairci deux propriétés inhérentes à la langue serbe qui ont une influence importante sur les résultats de l'étiquetage automatique de cette langue : l'ambiguïté morphologique et l'ordre des constituants syntaxiques flexible.

On a remarqué que la plupart des erreurs commises par BTagger relevaient des cas d'homonymie entre deux parties du discours (i.e. la forme en question peut appartenir à plusieurs catégories) et des occurrences où la désinence du token en question est partagée par plusieurs paradigmes. La documentation sur BTagger n'est pas suffisamment détaillée sur l'exécution même de la tâche d'annotation, mais les auteurs indiquent que le logiciel dispose d'un module de lemmatisation, et les tâches de

l'étiquetage et de lemmatisation ne sont pas strictement délimitées, mais complémentaires (voir Gesmundo et Samardžić 2012). Il paraît donc possible que le module d'annotation de BTagger utilise l'analyse des suffixes effectuée par le module de lemmatisation dans le processus d'étiquetage.

Ce phénomène a mis en évidence le fait que les paradigmes de différentes parties de discours en serbe partagent un nombre important de terminaisons, ce qui rend difficile une correction d'étiquetage fondée sur les indicateurs formels. En effet, le seul type d'erreurs rencontré dans notre analyse qualitative qui permettrait un tel posttraitement concerne les classes fermées des pronoms. Il faut souligner que ces erreurs ne sont pas nombreuses : 74 occurrences en ont été détectées sur la totalité du sous-corpus *Bašta*. Pourtant, une vérification de l'annotation de ces catégories permettrait d'obtenir une réduction d'erreur de 3,85%.

Pour les autres types de confusion rencontrés, nous avons vu que la grande richesse des distributions possibles pour chacune des parties du discours conditionne que le contexte immédiat d'un token soit peu contraignant. On rappelle l'exemple des noms communs, qui sont typiquement précédés d'un verbe ou d'un adjectif, mais qui peuvent également se trouver derrière une préposition, un adverbe ou un autre nom. L'entourage immédiat d'un token est donc difficilement exploitable dans la vérification de l'annotation ou dans l'étiquetage lui-même.

On a également pu voir que le contexte plus large peut imposer des contraintes importantes (cf. la présence du verbe principal dans la phrase, la présence du verbe auxiliaire dans le cas d'ambiguïté entre les adjectifs et les participes passés). Cependant, la discontinuité des formes verbales composées fait que la largeur du contexte qui doit être considéré semble excéder souvent les limites du logiciel. Cela cause ce qui pourrait être considéré comme l'erreur d'étiquetage la plus grave : des phrases sans verbe principal. Or, si ce type de faute n'est pas corrigeable automatiquement, un mécanisme de détection des phrases restées sans l'étiquette du verbe principal est envisageable. Cette démarche permettrait une correction manuelle de l'étiquetage au niveau de la phrase.

L'importance évidente de l'homonymie comme cause des erreurs d'étiquetage souligne un autre facteur crucial qu'est le corpus d'entraînement. Si toutes les parties de discours valables pour une forme ne figurent pas dans le corpus d'apprentissage, le

logiciel ne disposera pas de la totalité des étiquettes possibles pour le mot en question. Si le logiciel rencontre le token *obučen* (‘vêtu’, adjectif et participe passé) seulement en tant qu’adjectif, et jamais comme verbe principal, cette forme sera pour lui univoque et chaque occurrence rencontrée dans les nouveaux textes sera annotée comme adjectif. Comme on ne peut pas s’attendre à une couverture parfaite des formes homonymes dans le corpus d’entraînement, un certain nombre d’erreurs de ce type est inévitable.

V CONCLUSION

Nous avons vu dans ce mémoire le processus de l'étiquetage morpho-syntaxique d'un corpus littéraire du serbe, une langue pauvrement dotée en ressources linguistiques. La démarche choisie ne visait pas à atteindre une précision exceptionnelle : cet objectif s'inscrit dans une logique des chercheurs informaticiens du TAL. Même si les tentatives de réaliser l'étiqueteur le plus fiable, une différence de 1% de précision n'est pas nécessairement significative pour les utilisateurs moyens des corpus. Dans ce sens ; notre mémoire s'inscrit plutôt dans une perspective de la linguistique de corpus : notre objectif était d'aboutir à un étiquetage suffisamment fiable pour rendre le corpus serbe exploitable dans le domaine de la recherche, en ouvrant en même temps la voie pour continuer le processus d'enrichissement. Les résultats obtenus, à savoir la précision moyenne d'étiquetage de 94,17% montre que cet objectif est atteint, ce qui confirme le fait que la méthode choisie était appropriée à la tâche.

Le travail effectué peut être divisé en trois tâches distinctes : la définition d'un nouveau jeu d'étiquettes pour l'annotation du serbe, l'élaboration d'un corpus d'entraînement et l'identification du logiciel le mieux adapté à l'étiquetage du serbe.

Dans la construction du jeu d'étiquettes, quelques principes fondamentaux ont été suivis. Tout d'abord, comme il s'agit d'un corpus parallèle, nous avons trouvé nécessaire de rendre comparables les jeux d'étiquettes utilisés dans l'annotation des trois langues du corpus. Il a donc été nécessaire d'effectuer une analyse contrastive des systèmes morpho-syntaxiques serbe, français et anglais, afin d'identifier les divergences, mais aussi les points communs. Pour rapprocher les tags utilisés pour le serbe de ceux employés dans l'annotation du français et de l'anglais, nous avons dû modifier la définition traditionnelle de certaines (sous-)catégories grammaticales, notamment celle des pronoms. Il a en même temps été nécessaire de veiller à ne pas trop heurter la tradition grammaticale serbe : un jeu d'étiquettes contre-intuitif rendrait le corpus non-utilisable pour les chercheurs serbes. Nous avons également établi que le jeu d'étiquettes serbe le plus utilisé n'avait pas donné des résultats satisfaisants dans les expérimentations antérieures et qu'il était difficile à appliquer dans un étiquetage manuel à cause de sa taille (906 étiquettes). En revanche, un deuxième jeu d'étiquettes existant n'a pas été jugé suffisamment informatif : il n'encodait que la partie du discours principale ; nous avons donc cherché un équilibre entre la quantité des informations

encodées par les étiquettes et le nombre des tags. Notre tentative de respecter les exigences citées a résulté dans un jeu de 45 étiquettes, encodant la catégorie et la sous-catégorie des mots, ainsi que quelques propriétés morphologiques pour les adjectifs et les adverbes.

Cet ensemble d'étiquettes a été appliqué dans l'étiquetage du corpus d'entraînement *REF2*, comptant 157 000 tokens. L'élaboration de ce corpus a été effectuée en deux étapes : premièrement, le sous-corpus *REF1* (101 000) a été annoté manuellement ; ensuite, le sous-corpus *Bašta* (56 000) a été annoté automatiquement, et le résultat de ce processus a été vérifié manuellement. Les deux principes fondamentaux adoptés dans ce processus ont été les suivants : la correspondance token-tag doit être 1:1. Cela signifie qu'un token peut porter une seule étiquette, et qu'une étiquette peut être appliquée à un seul token à la fois. Ce procédé peut sembler incorrect dans le cas des unités polylexicales et des formes fléchies composées, mais il rend plus cohérent le traitement et assure une annotation systématique du premier niveau. Le deuxième principe employé était que les cas ambigus sont interprétés selon leurs propriétés syntaxiques : si une forme typiquement adjectivale fonctionne comme nom dans un contexte donné, elle sera annotée comme nom. Cette démarche permet de faire une meilleure distinction entre les contextes valides pour différentes parties du discours. Des règles d'annotation plus précises, définissant le traitement des cas de figure spécifiques, ont été définies au fur et à mesure de l'annotation et appliquées rigoureusement pour assurer la cohérence de l'étiquetage.

Pour identifier l'étiqueteur le plus performant dans l'annotation morpho-syntaxique du serbe, une sélection préliminaire des étiqueteurs disponibles a été faite basée sur une analyse bibliographique. TreeTagger, TnT et BTagger ont été choisis pour être testés : afin d'identifier le logiciel le mieux adapté à la tâche en question, nous avons effectué une série de 4 expérimentations sur le corpus *REF1* annoté manuellement. Le corpus a été divisé en 4 parties. Dans chacun des tests, les logiciels ont été entraînés sur 3 parties et testés sur la partie restante, de sorte que chacune des parties a été utilisée comme corpus de test exactement une fois. A partir des résultats obtenus, la précision moyenne a été calculée pour chacun des étiqueteurs. Ces valeurs confirment que notre démarche a été fructueuse : avec TreeTagger et TnT nous avons obtenus des résultats proches des meilleurs signalés jusqu'à présent dans l'étiquetage du serbe (92,15% et 92,97%

respectivement, alors que les seuls résultats antérieurs qui dépasse le seuil de 90% sont obtenus sur un corpus des textes non-littéraires ou avec des jeux d'étiquettes extrêmement réduits). Avec BTagger une amélioration nette de la précision a été obtenue : 94,17% par rapport à 86% des travaux antérieurs.

Une analyse qualitative de l'étiquetage du sous-corpus *Bašta* a permis d'identifier les propriétés du serbe posant le plus de problème dans l'étiquetage automatique. L'analyse des erreurs les plus fréquentes a montré que les paradigmes des différentes catégories grammaticales partagent non seulement un nombre important des formes, mais aussi des désinences. En même temps, nous avons pu voir que les contextes valides pour une partie de discours ne sont souvent pas suffisamment discriminants pour être exploités dans la désambiguïsation. Ces deux faits conditionnent un degré de confusion élevé ce qui conditionne un degré de confusion élevé entre les adjectifs et les adverbes, les noms et les verbes, les adjectifs et les noms. Nous avons vu que les erreurs les plus importantes concernent l'identification des formes verbales principales et auxiliaires car elles risquent de générer une annotation qui laisse la phrase sans verbe principal. Nous avons également identifié quelques possibilités de post-traitement dans le but de l'amélioration des résultats, surtout dans l'annotation des pronoms et, précisément, dans l'identification des formes verbales.

La seule expérimentation dans l'étiquetage du serbe qui atteigne un meilleur résultat sur un corpus littéraire est effectuée avec un jeu d'étiquettes minimaliste, encodant seulement la partie principale du discours. Nous jugeons donc que notre démarche offre le meilleur équilibre entre la précision atteinte et la quantité d'informations encodées dans le jeu d'étiquettes. Le travail effectué ouvre également de nombreuses pistes à poursuivre : premièrement, différentes stratégies de post-traitement peuvent être mises en place pour améliorer la qualité de l'annotation. Vu la nature des erreurs, il s'agirait de méthodes semi-automatiques. La confusion d'autres parties de discours avec les pronoms personnels et démonstratifs peut être éliminée en utilisant une liste fermée des formes fléchies de ces deux sous-catégories : si un token qui porte le tag de pronom personnel ou indéfini ne figure pas sur la liste, la phrase serait signalée pour être vérifiée manuellement. De même, il est possible d'identifier les phrases qui n'ont pas de forme verbale principale et les corriger ensuite manuellement. En dernier lieu, pour améliorer la distinction entre le participe passé actif du verbe *biti* 'être' en tant que

verbe principal (dans les temps composés) ou verbe auxiliaire (dans les temps surcomposés), on peut envisager de vérifier si la phrase sans forme verbale fait figurer deux tokens annotés comme verbe auxiliaire et si l'une de ces formes est le participe en question. Dans ce cas, l'annotation du token doit être changée en verbe principal.

D'autres suites sont également envisageables : la lemmatisation manuelle effectuée sur la totalité du corpus *REF2* peut être exploitée dans l'immédiat pour assurer cette forme d'enrichissement au corpus serbe. Les résultats de BTagger signalés pour cette tâche (97,72% de précision) qualifient ce logiciel pour cette tâche. On peut envisager d'approfondir l'annotation existante des formes verbales en exploitant la partie du corpus *REF1* annotée avec les étiquettes détaillées, qui encodent le temps et l'aspect des formes verbales. Une autre suite qui permettrait une meilleure exploitation du corpus dans la perspective contrastive serait une analyse en dépendances, une forme d'analyse syntaxique qui consiste à identifier les termes régissants d'une phrase et les constituants qui en dépendent. Cette démarche permettrait d'identifier les éléments correspondants au niveau infra-phrastique dans les trois langues du corpus.

BIBLIOGRAPHIE

Aarts, B., & Haegeman, L. (2006). English word classes and phrases. Dans B. Aarts, & A. McMahon (Éds.), *The Handbook of English linguistics* (pp. 117-145). Blackwell Publishing.

Abeillé, A., Clément, L. *et al*. (2000). Building a treebank for French. *Proceedings of the 2nd Conference on Linguistic Resources.* Athens.

Adda, G., Mariani, J. *et al*. (1998). The GRACE French part-of-speech tagging evaluation task. *Proceedings of the First International Conference on Language Resources and Evaluation*, (pp. 433-441).

Agić, Ž., Tadić, M. *et al*. (2009). Tagset reductions in morphosyntactic tagging of Croatian texts. *The Future of Information Sciences: Digital Resources and Knowledge Sharing*, 289-298.

Blevins, J. (2006). English inflection and derivation. Dans B. Aarts, & A. McMahon (Éds.), *The Handbook of Engish Linguistics* (pp. 507-536). Blackwell Publishing.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing*, (pp. 224-231). Seattle.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist., 21*, pp. 543-565.

Collins, M. (2002). Discriminative trainin methods for hidden Markov models: Theory and experiments with perceptron algorithms. *Proceedings of The 2002 Conference on Empirical Methods on Natural Language Processing*, (pp. 1-8). Philadelphia.

Daelmans, W., Zavrel, J. *et al*. (1996). MBT: A memory-based part of speech tagger-generator. (E. Ejerhed, & I. Dagan, Éds.) *Fourth Workshop on Very Large Corpora*, 14-27.

Denis, P., & Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. *Proceedings of the Pacific Asia Conference on Language, Information and Computation.* Hong Kong.

Dojchinova, V., & Mihov, S. (2004). High performance part-of-speech tagging of Bulgarian. *Lecture notes in computer science, 3192*, pp. 246-255.

Dredze, M., & Wallenberg, J. (2008). Icelandic data driven part of speech tagging. *Proceedings of the 44th Annual Meeting of the Association of Computational Linguistics*, (pp. 33-36). Columbus.

Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. *Fourth International Conference on Language Resources and Evaluation*, *4*, pp. 1535-1538.

Gesmundo, A., & Samardžić, T. (2012). Lemmatising Serbian as a category tagging task with bidirectional sequence classification. *Proceedings of the Eighth International Conference on Language Resources and Evaluation.* Istanbul.

Giménez, J., & Màrquez, L. (2004). SVMtool: A general POS tagger generator based on support vector machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, (pp. 43-46). Lisbon.

Greene, B., & Rubin, G. (1971). *Automatic grammatical tagging of English.* Providence: Department of Linguistics, Brown University.

Habash, N., & Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morpholgical disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, (pp. 573-580). Ann Arbor.

Hajič, J. (1998). Building a syntactically annotated corpus: The Prage dependency treebank. (E. Hajičova, Éd.) *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevova*, 12-19.

Hajič, J., Krbec, P. *et al*. (2001). Serial combination of rules and statistics: A case study in Czech tagging. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, (pp. 268-275). Toulouse.

Ide, N., & Véronis, J. (1994). MULTEXT (Multilingual text tools and corpora). *Proceedings of the 15th International Conference on Computational Lignuistics*, (pp. 588-592).

Krsteva, C., Vitas, D. *et al*. (2004). MULTEXT-East resources for Serbian. *Proceedings of 8th Informational Society - Language Technologies Conference*, (pp. 108-114). Ljubljana.

Lafferty, J., McCallum, A. *et al*. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning*, (pp. 282-289). San Francisco.

Marcus, M., Marcinkiewicz, M. *et al*. (1993). Building a large annotated corpus of English: the Penn Treebank. *Comput. Linguist., 19*, pp. 313-330.

McEnery, T. (2003). Corpus Linguistics. Dans R. Mitkov (Éd.), *The Oxford handbook of computational linguistics* (pp. 448-463). Oxford University Press.

Popović, Z. (2010). Taggers applied on texts in Serbian. *INFOtheca, 2*(XI), pp. 21-38.

Rajman, M., Lecomte, J. *et al*. (1997). *Format de description lexicale pour le français. Partie 2: Description morpho-syntaxique.* Révision du GTR-3-2.1 suite à la journée atelier d'avril 1997 à Avignon.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. (E. Ejerhed, & I. Dagan, Éds.) *Fourth Workshop on Very Large Corpora*, 133-142.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing*, (pp. 44-49). Manchester.

Shen, L., Satta, G. *et al*. (2007). Guided learning for bidirectional sequence classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (pp. 760-767). Prague.

Simić, M. (2005). Srpski elektronski rečnik. Récupéré sur http://www.rasprog.com

Simov, K., Osenova, P. *et al*. (2002). Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. *Proceedings of LREC 2002.* Canary Islands.

Stanojčić, Ž., & Popović, L. (2011). *Gramatika srpskog jezika* (éd. 14). Beograd: Zavod za udžbenike.

Tadić, M. (2000). Building the Croatian-English parallel corpus. *Proceedings of the Second International Conference on Language Resources and Evaluation*, (pp. 523-530). Paris-Athens.

Thomas, P.-L. (1993). Bilan des recherches sur l'aspect en serbo-croate. *Revue des Etudes Slaves, 65*(3), pp. 537-550.

Thomas, P.-L. (1998). Remarques sur l'aspect en serbo-croate. Dans A. Borillo, C. Vetters, & M. Vuillaume (Éds.), *Regards sur l'aspect* (pp. 231-243). Amsterdam: Rodopi.

Toutanova, K., & Klein, D. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 173-180). Edmonton.

Utvić, M. (2011). Annotating the Corpus of contemporary Serbian. *INFOtheca, 12*(II), pp. 36-47.

Véronis, J. (2000). Annotation automatique de corpus : panorama et état de la technique. Dans J.-M. Pierrel (Éd.), *Ingénierie des langues* (pp. 111-120). Paris: Hermès Sciences Publications.

Xiao, Z., & McEnery, A. (2012). A corpus based approach to tense and aspect in English-Chinese translation. *International Symposim on contrastive and translation studies between Chinese and English.* Shanghai.

Резникова, Т. И. (2008). Корпуса славянских языков в интернете: Обзор ресурсов. *Die Welt der Slaven*(LIII).

ANNEXE 1

Les tableaux ci-dessous présentent la conjugaison complète du verbe serbe *raditi* ‘travailler’.

	personne	singulier	pluriel
présent	1°	radim	radimo
	2°	radiš	radite
	3°	radi	rade

	personne	genre	singulier	pluriel
parfait	1°	masculin	sam radio	smo radili
		féminin	sam radila	smo radile
		neutre	sam radilo	smo radila
	2°	masculin	si radio	ste radili
		féminin	si radila	ste radile
		neutre	si radilo	ste radila
	3°	masculin	je radio	su radili
		féminin	je radila	su radile
		neutre	je radilo	su radila

	personne	singulier	pluriel
imparfait	1°	rađah	rađasmo
	2°	rađaše	rađaste
	3°	rađaše	rađahu

	personne	genre	singulier	pluriel
plus-que-parfait	1°	masculin	sam bio radio	smo bili radili
		féminin	sam bila radila	smo bile radile
		neutre	sam bilo radilo	smo bila radila

	2°	masculin	si bio radio	ste bili radili
		féminin	si bila radila	ste bile radile
		neutre	si bilo radilo	ste bila radila
	3°	masculin	je bio radio	su bili radili
		féminin	je bila radila	su bile radile
		neutre	je bilo radilo	su bila radila

	personne	singulier	pluriel
futur simple	1°	ću raditi / radiću	ćemo raditi / radićemo
	2°	ćeš raditi / radićeš	ćete raditi / radićete
	3°	će raditi / radiće	će raditi / radiće

	personne	genre	singulier	pluriel
futur antérieur	1°	masculin	budem radio	budemo radili
		féminin	budem radila	budemo radile
		neutre	budem radilo	budemo radila
	2°	masculin	budeš radio	budete radili
		féminin	budeš radila	budete radile
		neutre	budeš radilo	budete radila
	3°	masculin	bude radio	budu radili
		féminin	bude radila	budu radile
		neutre	bude radilo	budu radila

	personne	singulier	pluriel
impératif	1°	-	radimo
	2°	radi	radite
	3°	-	-

conditionnel	personne	genre	singulier	pluriel

	1°	masculin	bih radio	bismo radili
		féminin	bih radila	bismo radile
		neutre	bih radilo	bismo radila
	2°	masculin	bi radio	biste radili
		féminin	bi radila	biste radile
		neutre	bi radilo	biste radila
	3°	masculin	bi radio	bi radili
		féminin	bi radila	bi radile
		neutre	bi radilo	bi radila

infinitif	raditi

adverbe déverbal présent	radeći

adverbe déverbal passé	radivši

participe passé actif	genre	singulier	pluriel
	masculin	radio	radili
	féminin	radila	radile
	neutre	radilo	radila

participe passé passif	genre	singulier	pluriel
	masculin	rađen	rađeni
	féminin	rađena	rađene
	neutre	rađeno	rađena

ANNEXE 2

Le tableau ci-dessous présente les étiquettes verbales détaillées, élaborées dans le cadre du projet Egide Constitution du corpus parallèle français-serbe-anglais en 2010. Le format des étiquettes est comme suit : VER:(temps/mode):aspect.

Temps/Mode	Aspect	Etiquette	Exemple
Infinitif	imperfectif	VER:infi:imp	čitati 'lire'
	perfectif	VER:infi:per	pročitati 'avoir lu'
Présent	imperfectif	VER:pres:imp	čitam 'je lis'
	perfectif	VER:pres:per	pročitam 'je lis' (interpr. perfective)
Aorist	imperfectif	VER:aor:imp	on pisa 'il écrivit', oni pisaše 'ils écrivirent' (interpr. imperfective)
	perfectif	VER:aor:per	on napisa 'il écrivit', oni napisaše 'ils écrivirent'
Imparfait	imperfectif	VER:impf:imp	on pisaše 'il écrivait', oni pisahu 'ils écrivaient'
Impératif	imperfectif	VER:impe:imp	čitaj 'lis'
	perfectif	VER:impe:per	pročitaj 'ais lu'
Participe passé actif	imperfectif	VER:pper:imp	čitao 'lu' (intepr. imperfective et active)
	perfectif	VER:pper:per	pročitao 'lu' (interpr. perfective et active)
Participe passé actif	imperfectif	VER:pperpv:imp	čitan 'lu', brisan 'essuyé' (interpr. imperfective)
	perfectif	VER:pperpv:per	pročitan 'lu'
Futur simple	imperfectif	VER:futu:imp	čitaću 'je lirai'
	perfectif	VER:futu:per	pročitaću 'j'aurai lu'
Adverbe déverbal présent	imperfectif	VER:ppre:imp	čitajući 'lisant'
Adverbe déverbal passé	perfectif	VER:pprep:per	pročitavši 'ayant lu'
Verbe auxiliaire	toutes formes	VER:AUX	sam 'suis', bi 'serait', ću 'veux'

ANNEXE 3

Le tableau ci-dessous présente la distribution de 1903 erreurs relevées dans l'étiquetage du sous-corpus Bašta par BTagger. La distribution est donnée sur la totalité du jeu d'étiquettes : les lignes indiquent les parties du discours, alors que dans les colonnes on retrouve les étiquettes erronées. Ainsi, la première ligne indique que le nom commun (NOM:com) a été identifié comme nom propre (NOM:NAM) 10 fois, comme verbe principal (VER) 99 fois, comme verbe auxiliaire (VER:AUX) 5 fois, et ainsi de suite.

	NOM:com	NOM:NAM	VER	VER:AUX	ADJ	ADJ:POS	ADJ:DEM	ADJ:IND	ADJ:INTR	ADJ:POS	ADJ:KOM	ADJ:SUP	ADV	ADV:KOM	ADV:SUP	ADV:IND	ADV:INTR	ADV:REL	NUM:CAR	NUM:ORD	NUM:ZBR	KON:COOR	KON:SUB	NOM:col	PRO:DEM	PRO:IND	PRO:INTR	PRO:PER	PRO:POS	PRO:REF	PRO:REL	PRO:NUM	PRP	LET	SENT	PUN	PUN:cit	NUM	STR	INT	PAR
NOM:com	/	10	99	5	143	0	5	1	0	0	0	4	14	0	0	0	0	0	2	0	0	2	0	0	1	1	0	3	0	0	1	0	13	0	0	2	0	0	3	0	1
NOM:NAM	30	/	2	0	4	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	2	0	0
VER	70	1	/	69	104	2	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	1	2	0	1	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0
VER:AUX	0	0	145	/	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ	244	13	126	3	/	0	1	3	0	0	0	0	32	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	0	0	0	4	0	0	0	0	0	0	0	0
ADJ:POS	0	0	0	0	0	/	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ:DEM	2	0	0	0	0	0	/	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	5	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ:IND	0	0	0	0	3	3	0	/	0	0	0	0	3	0	0	0	0	0	13	0	0	0	0	0	0	13	0	0	0	0	2	3	0	0	0	0	0	0	0	0	1
ADJ:INTR	1	0	0	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ:POS	0	0	0	0	0	0	0	0	0	/	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ:KOM	17	0	2	0	15	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADJ:SUP	7	0	2	1	0	0	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV	26	0	24	0	68	0	0	8	0	0	0	0	/	1	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	2	0	2	0	0	0	0	0	0	0	1
ADV:KOM	5	0	3	0	0	0	0	0	0	0	1	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV:SUP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV:IND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ADV:INTR	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	/	4	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ADV:REL	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	/	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUM:CAR	3	0	1	0	1	0	0	30	0	0	0	0	4	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUM:ORD	1	0	2	0	4	0	0	9	0	0	0	0	2	0	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUM:ZBR	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KON:COOR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KON:SUB	0	0	1	0	0	0	2	0	0	0	0	0	1	0	0	0	4	0	0	0	0	0	/	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
NOM:col	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

PRO:DEM	0	0	0	0	0	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO:IND	**1**	0	**1**	0	0	0	0	**4**	0	0	0	0	**3**	0	0	0	0	0	0	0	0	0	0	0	0	**/**	**1**	0	0	0	0	**1**	0	0	0	0	0	0	0	0	**1**
PRO:INTR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRO:PER	0	**2**	**3**	**30**	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	**2**	0	0	**/**	0	0	0	0	**1**	0	0	0	0	0	**1**	0	0
PRO:POS	**1**	0	**1**	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0	0	0	0	0
PRO:REF	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0	0	0	0
PRO:REL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**8**	0	0	0	**2**	0	0	0	**/**	0	0	0	0	0	0	0	0	0	0
PRO:NUM	0	0	0	0	0	0	0	**27**	0	0	0	0	0	0	0	0	0	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0	0
PRP	**5**	0	**3**	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0	0
LET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0	0
SENT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0	0	0
PUN	0	0	**1**	**21**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**14**	**/**	0	0	**2**	0	0
PUN:cit	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0	0
NUM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**	0	0	0
STR	**64**	**42**	**3**	0	**9**	0	**1**	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	**14**	0	0	0	0	**/**	**1**	0
INT	**22**	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**2**	0	0	**5**	0	0	0	0	0	0	0	0	0	0	0	**/**	
PAR	**15**	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**/**