SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

TALC-sef
Un corpus étiqueté de traductions littéraires en serbe, anglais et français

Balvet, Antonio, & Stosic, Dejan, & Miletic, Aleksandra

Balvet, Antonio
Miletic, Aleksandra
Université Lille Nord de France, F-59000 Lille, France
UdL3, STL, F-59653 Villeneuve d'Ascq, France
CNRS, UMR 8163
antonio.balvet@univ-lille3.fr
aleksandramiletic1207@gmail.com
Stosic, Dejan
UMR CLLE-ERSS 5263, Université Toulouse 2/CNRS
dstosic@univ-tlse2.fr

1 Introduction

Le corpus TALC-sef (pos-TAgged Literary Corpus, Serbian-English-French), initié dans le cadre de deux projets de recherche (2007-2009 et 2010-2011) impliquant les universités Lille 3, Artois et l'université de Belgrade, était à l'origine conçu comme un corpus parallèle de référence, de traductions dans le domaine littéraire, pour le serbe, l'anglais et le français. En complément d'un alignement au niveau phrastique, réalisé de façon automatique et validé manuellement, des annotations morpho-syntaxiques en parties du discours avaient été rajoutées dès les premières versions du corpus, de façon complètement automatique pour les sous-corpus français et anglais, grâce aux modèles d'étiquetage du Treetagger (Schmid, 1994). Toutefois, faute de ressources exploitables en serbe, ces annotations n'avaient pu être menées à bien. Nous détaillons dans la suite de cet article la méthodologie adoptée pour la définition d'un jeu d'étiquettes syntaxiques pour la langue serbe, les choix techniques et linguistiques que nous avons faits dans le but d'enrichir le sous-corpus serbe d'annotations syntaxiques comparables à celles des sous-corpus français et anglais, afin de permettre des recherche de cooccurrences d'étiquettes syntaxiques en parallèle dans les trois langues. Puis, nous discutons des performances d'étiquetage syntaxique enregistrées avec trois étiqueteurs librement disponibles : TnT (Brants 2000), Treetagger (Schmid, 1994) et BTagger (Gesmundo & Samardžić, 2012), avant d'aborder les perspectives d'exploitation de l'étiquetage syntaxique réalisé pour d'autres niveaux d'annotations, en soulignant l'apport de ce corpus multilingue pour la linguistique française.

2 Constitution d'un corpus littéraire parallèle en trois langues

Au cours des périodes 2007-2009, puis 2010-2011, un corpus littéraire en trois langues européennes, appartenant à des groupes typologiques distincts (base latine, base anglo-saxonne et base slave) a été constitué sous la direction de D. Stosic (Univ. d'Artois), en partenariat avec l'Université Lille 3 et l'Université de Belgrade. Pour ce corpus ont été sélectionnés des ouvrages littéraires originales dans les trois langues, chacune accompagnée de ses traductions dans les deux autres langues du corpus (cf. tableau 1). L'objectif principal du projet était la constitution d'un corpus de référence dans le domaine littéraire, dans une optique d'exploitation en linguistique, littérature et stylistique comparées. De ce fait, un soin constant a été apporté, dans la constitution du corpus, à la qualité de l'alignement et de l'annotation, d'une part, ainsi qu'à son disponibilité pour l'ensemble de la communauté des chercheurs dans ces domaines. Ainsi, chaque phrase des ouvrages originaux a été alignée automatiquement avec ses différentes traductions dans les autres langues-cibles.[1] Dans une seconde phase, les alignements proposés ont tous été corrigés et validés manuellement. Les textes originaux et leurs traductions pourront être

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801035

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

consultés en ligne par le biais d'une interface de recherche multilingue (concordancier).[2] Le tableau ci-dessous reprend les principaux éléments quantitatifs du corpus dans son ensemble.

	français	**serbe**	**anglais**
français	300 105	332 521	353 934
serbe	316 210	388 326	-
anglais	45 457	156 074	148 486
TOTAL	661 772	876 921	502 420

Tableau 1 : ventilation par langue des sous-corpus de TALC-sef

Ainsi que le montre le tableau ci-dessus, dans sa version actuelle, les traductions français ↔ serbe ont été favorisées, alors que des traductions serbe → anglais font à l'heure actuelle encore défaut.

Serbe **Auteur** **Titre**	**Français** **Auteur** **Titre**	**Anglais** **Auteur** **Titre**
►B. Blagojević *Putnica*	*Voyageuse*	-
►D. Kiš *Rani jadi*	*Chagrins précoces*	-
►D. Kiš *Enciklopedija mrtvih*	*Encyclopédie des morts*	-
►D. Kiš *Grobnica za Borisa Davidoviča*	*Un Tombeau pour Boris Davidovitch*	-
►B. Šćepanović *Iskupljenje*	*Rachat*	-
►R. Petrović *Ljudi govore*	*Conversations insulaires*	-
►V. Stevanović *Testament*	*Prélude à la guerre*	-
Bogovi su žedni	►A. France *Les Dieux ont soif*	*The Gods are Athirst*
Čiča Gorio	►H. De Balzac *Le père Goriot*	*Old Goriot*
Zvonar Bogorodične crkve u Parizu	►V. Hugo *Notre-Dame de Paris*	*The Hunchback of Notre Dame*
Poslednji Mohikanac	*Le dernier des Mohicans*	►J. F. Cooper *The Last of the Mohicans*

Tableau 2 : ouvrages alignés au niveau phrastique du corpus TALC-sef

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

Le sous-corpus serbe représente à lui seul un total de 388 326 mots, un volume de données trois fois supérieur à la traduction du roman *1984* de G. Orwell (104 286 mots), ou corpus cesAna du projet MULTEXT-EAST, qui, à ce jour, constitue le seul corpus de référence en accès libre pour le serbe.[3] À partir de ces 380 000 mots, nous avons élaboré un corpus de référence, révisé et corrigé manuellement de plus de 150 000 mots. Le Tableau 2 présente les différents ouvrages qui constituent le corpus TALC[4]. En sus des ouvrages énumérés ci-dessus, d'autres ouvrages sont intégrés au corpus, bien qu'ils n'aient pas tous fait l'objet d'un alignement manuellement corrigé. C'est le cas, notamment, du roman *Bašta, Pepeo* de D. Kiš (cf. section 2.5).

2.1 L'étiquetage du sous-corpus serbe : difficultés et propositions

Les sous-corpus français et anglais ont été étiquetés grâce aux modèles d'étiquetage du Treetagger (Schmid, 1994), appliqués de façon standard, sans relecture manuelle. En effet, la création du corpus TALC-sef visait à combler un manque en linguistique contrastive : disposer de corpus parallèles de qualité, du registre littéraire[5], pour des langues appartenant à des familles de langues distinctes, soit : base latine (français), anglo-saxonne (anglais) et slave (serbe). Outre un alignement au niveau des phrases, le projet visait dès l'origine un alignement infra-phrastique, reposant en l'occurrence sur des constituants syntaxiques. Pour ce faire, il était donc indispensable de disposer pour le serbe d'annotations syntaxiques de granularité comparable à celles du français et de l'anglais. Cependant, le serbe fait partie des langues dites « peut dotées », en termes d'outils informatiques disponibles pour leur traitement automatique, alors que le français et l'anglais sont comparativement mieux dotées, ce qui représente une source de difficultés non négligeable. En effet, pour les langues les mieux dotées, la question du choix d'un jeu d'étiquettes syntaxiques et morphologiques peut être considérée comme en partie réglée. Ainsi, cette question semble ne plus devoir faire l'objet de recherches pour l'anglais américain, depuis l'avènement des corpus de référence tels que le corpus Brown (Francis & Kučera, 1964) et le Penn Treebank (Marcus *et al.*, 1993). Pour le français, la question semble également tranchée, depuis la diffusion du corpus French Treebank (Abeillé, 2003) : en effet, la constitution même d'un corpus de type Treebank[6] présuppose la définition d'un jeu d'étiquettes syntaxiques consensuel. Toutefois, la situation est loin d'être similaire pour le serbe, qui, pour des raisons diverses (politiques, économiques...) a fait l'objet d'investissements moindres que d'autres, dans le domaine de leur traitement automatique. Dans le cas du serbe, comme indiqué plus haut, le seul corpus de référence librement disponible à ce jour consiste en une traduction du roman *1984*, pour un volume total de données linguistiques assez faible, si on le compare au million de mots d'un corpus aussi ancien que le *Brown Corpus*. Le premier étiqueteur syntaxique spécifiquement conçu pour les langues slaves à morphologie riche et à ordre des mots relativement souple, comme le serbe, n'a été diffusé publiquement qu'à partir de 2012 (Gesmundo and Samardzic, 2012), alors que des corpus de référence, révisés manuellement, existent depuis les années 1990 pour l'anglais et le français.[7] Le décalage entre les langues richement dotées et les autres se fait sentir notamment au niveau des performances d'étiquetage : les étiqueteurs pour l'anglais présentent de façon habituelle des scores de précision d'étiquetage dépassant les 97% (Shen et al., 2007), alors que des expérimentations d'étiquetage syntaxique automatique récentes du serbe restent bien en-deçà de ce niveau de performance (Gesmundo and Samardzic 2012, Popovic, 2010).

Une difficulté supplémentaire dans la constitution d'un jeu d'étiquettes syntaxiques consensuel tient à la morphologie riche du serbe. En effet, le serbe distingue trois personnes, deux nombres, trois genres et sept cas différents. Les suffixes casuels s'appliquent aux noms, aux adjectifs, aux pronoms et à certains cardinaux. Les noms peuvent avoir jusqu'à 12 formes fléchies différentes (à comparer aux 4 du français, et aux 2 de l'anglais), et les adjectifs jusqu'à 36. De plus, en fonction du temps, du mode, de la personne et du genre, les verbes serbes peuvent présenter 120 formes fléchies différentes. De ce fait, les expérimentations menées dans (Gesmundo & Samardzic, 2012) reposent sur des jeux d'étiquettes très importants : plus de 900 étiquettes morpho-syntaxiques différentes pour le serbe, à opposer aux 36 étiquettes du *Penn Treebank*, ou encore les 33 étiquettes principales du *French Treebank*. Enfin, une complexité supplémentaire s'ajoute à cette morphologie riche, et qui tient à l'ordre des mots en serbe. En effet, celui-ci est nettement moins fixe qu'en anglais ou en français : bien que l'ordre canonique soit SVO

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

(Sujet Verbe Objet), de nombreuses et fréquentes variations sont possibles (Stanojcic & Popovic, 2011). Ceci entraîne une ambiguïté potentiellement massive dans les formes de surface, bien plus importante qu'en français et en anglais.

2.2 Le jeu d'étiquettes du sous-corpus serbe : un compromis entre précision et couverture

À notre connaissance, deux corpus différents ont été mis en œuvre dans des expérimentations d'étiquetage syntaxique de la langue serbe. Le premier d'entre eux est le corpus *cesAna* cité plus haut, utilisé dans (Popovic, 2010) et (Gesmundo & Samardžić, 2012). Ce corpus comporte 906 étiquettes distinctes, codant toutes les distinctions morpho-syntaxiques, dans une optique de fidélité à une certaine tradition grammaticale serbe. Dans ses expérimentations, Popovic a fait appel à cinq étiqueteurs différents, dont l'étiqueteur TnT (Brants, 2000), qui a enregistré les meilleures performances, avec un score de précision de 85,47%. Gesmundo et Samardzic, de leur côté, ont exploité le corpus *cesAna* afin de tester BTagger, un étiqueteur de leur conception, spécifiquement pensé pour les langues slaves comme le serbe. Toutefois, cet étiqueteur, reposant sur une approche algorithmiquement bien plus complexe que TnT, est loin des scores de plus de 96% qui sont aujourd'hui la norme pour l'anglais : il plafonne ainsi à 86,65% pour le serbe, qui est d'ailleurs l'une des langues slaves pour lesquels l'étiqueteur BTagger enregistre les performances les plus basses. La seule autre alternative semble être celle de (Utvic 2011), qui a constitué un corpus de un million de mots[8], annoté grâce à un jeu d'étiquettes de seulement 16 catégories morpho-syntaxiques principales. Les performances d'étiquetage avec l'étiqueteur Treetagger sont de 96,57%, soit des scores de précision comparables à l'état de l'art pour l'anglais et le français.

Comme on peut le constater, les expérimentations d'étiquetage morpho-syntaxique du serbe semblent buter sur un écueil majeur lié à la taille du jeu d'étiquettes. Dans l'optique de l'étiquetage du sous-corpus avec un large recours à une approche manuelle dans un premier temps, il était exclu d'imposer aux annotateurs un jeu d'étiquettes aussi riches que celui du corpus *cesAna*. Par ailleurs, l'objectif étant dès le départ de permettre des expressions de recherche automatique comparables dans les trois langues, il était nécessaire d'opérer, pour le serbe, une sélection dans les étiquettes majeures, afin de rester en cohérence avec les principales distinctions faites pour le français (33 étiquettes principales) et pour l'anglais (36). Nous avons donc élaboré un jeu d'étiquettes qui représente un compromis acceptable entre précision et couverture, qui soit comparable à ceux des autres sous-corpus anglais et français, et qui ne représente pas une difficulté majeure pour les étiqueteurs syntaxiques utilisés. Ce jeu d'étiquettes comporte 45 distinctions syntaxiques principales, ainsi que les principaux traits morphologiques pour les adjectifs et les verbes.

Le jeu d'étiquettes proposé a été utilisé pour procéder à l'étiquetage manuel d'un sous-corpus dénommé REF1, comportant 101 000 mots graphiques (tokens), utilisé dans une première phase d'entraînement avec plusieurs étiqueteurs syntaxiques (voir plus bas). Afin de réduire la complexité de l'annotation manuelle, nous nous sommes concentrés sur l'étiquetage syntaxique proprement dit, sans indiquer la forme lemmatisée pour les mots étiquetés. Une fois l'évaluation comparative réalisée avec les trois étiqueteurs sélectionnés, nous avons exploité le modèle d'étiquetage du sous-corpus REF1 afin de réaliser un étiquetage automatique d'un corpus complémentaire de 50 000 mots, afin d'aboutir au sous-corpus REF2. Ce dernier corpus comporte la référence entièrement manuelle REF1 + 50 000 mots étiquetés automatiquement puis vérifiés et corrigés manuellement[9].

2.3 Quel étiqueteur pour le serbe ?

Afin de permettre des comparaisons avec les travaux mentionnés plus haut (Popovic, 2010), (Utvic, 2011) et (Gesmundo & Samardžić, 2012), nous avons sélectionné trois étiqueteurs, parmi les plus populaires : TreeTagger (Schmid, 1994) et TnT (Brants, 2000), en tant qu'étiqueteurs de référence (*baseline*), et BTagger (Gesmundo and Samardžić, 2012). Dans l'évaluation quantitative présentée plus bas, nous avons adapté le principe des procédures d'évaluation croisées (*n-fold evaluation*) : nous avons découpé le sous-

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

corpus REF1 en 4 parties égales, en gardant ¾ du sous-corpus comme corpus d'entraînement et ¼ comme corpus de test. En raison de la taille du corpus utilisé pour ces évaluations (101 000 mots), une évaluation classique reposant sur une partition en 10 sous-corpus[10] aurait potentiellement biaisé les résultats : nous aurions évalué les performances d'étiquetage sur seulement 10 100 mots pour chaque phase d'évaluation.[11] Les temps de calcul très élevés de BTagger, tant pour la phase d'apprentissage d'un modèle d'étiquetage[12] que pour celle d'étiquetage proprement dite[13], ont également pesé dans la balance, en faveur d'une évaluation croisée en seulement 4 passes, contre 10 en règle générale.

Ainsi, en moyenne, le corpus d'apprentissage représentait un volume total de 71 683 mots, et le corpus de test 23 896. Nous avons adapté chaque sous-partition du corpus REF1 de manière à préserver l'intégrité phrastique (pas de coupure au milieu d'une phrase) : en effet, à moyen terme d'autres étiqueteurs seront évalués sur notre corpus, dont MBT (Daelmans *et al.*, 1996), qui attend en entrée un corpus d'apprentissage segmenté en phrases et non découpé en mots graphiques, contrairement à la plupart des étiqueteurs syntaxiques. Les phrases soumises à l'étiqueteur ont toutefois été sélectionnées de façon aléatoire, afin de ne pas introduire de biais lié à la structure textuelle (ex. : structures syntaxiques présentatives, liées à une sur-représentation des descriptions).

2.4 Évaluation quantitative de l'étiquetage du sous-corpus serbe

La figure ci-dessous présente les principaux résultats quantitatifs de l'évaluation menée grâce au corpus REF1.

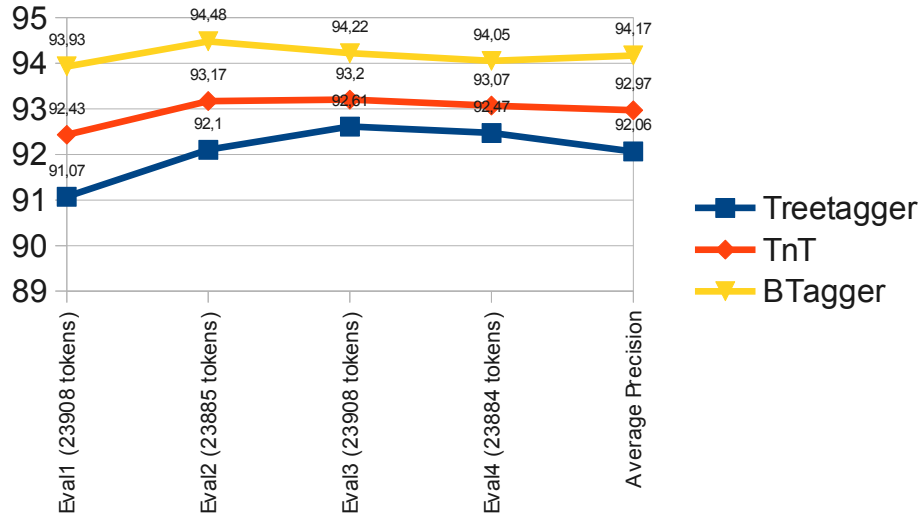

Figure 1 : scores de précision d'étiquetage du sous-corpus serbe par trois étiqueteurs

Comme on peut le voir, pour chaque phase d'évaluation (Eval1 à 4), BTagger est l'étiqueteur ayant enregistré les scores les plus élevés, avec une précision d'étiquetage moyenne de 94,17%. De façon surprenante, Treetagger présente des scores de précision d'étiquetage systématiquement en-dessous des deux autres, avec une précision moyenne de 92,15%. Enfin, TnT, reposant pourtant sur une approche algorithmique beaucoup plus fruste que Treetagger[14], enregistre des scores intermédiaires, avec une précision moyenne de 92,97%. Dans ces évaluations, les trois étiqueteurs ont été utilisés sans modifications de leurs paramétrages d'origine, ce qui a pu favoriser BTagger, un étiqueteur plus récent et spécifiquement pensé pour l'étiquetage des langues slaves.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

En nous appuyant sur ces résultats, nous avons sélectionné BTagger pour l'étiquetage automatique d'un sous-corpus de 56 093 mots, validés et corrigés manuellement dans un second temps, ce qui nous permet de proposer un corpus de référence pour le serbe REF2 de plus de 150 000 mots, soit un corpus comparable dans sa taille avec le corpus *cesAna*. Dans sa version actuelle, le corpus TALC-sef comporte donc un sous-corpus serbe de plus de 800 000 mots, étiquetés automatiquement grâce à BTagger, pour lequel une proportion d'erreurs d'étiquetage d'au plus 6% peut être attendue.

2.5 Analyse qualitative des résultats de Btagger

Outre l'évaluation quantitative synthétisée ci-dessus, dont le but était principalement d'identifier l'étiqueteur syntaxique le mieux adapté à l'étiquetage du sous-corpus serbe, une analyse qualitative de l'annotation automatique de *Bašta, Pepeo* a également été effectuée. Cette analyse avait pour but d'identifier les structures ou les mots les plus problématiques pour BTagger, ainsi qu'éventuellement de proposer des stratégies de contournement (pré- ou post-traitement des annotations). Elle est également destinée à accompagner les différentes versions du corpus TALC-sef, afin de signaler aux utilisateurs finaux quelles catégories risquent de comporter le plus d'erreurs d'étiquetage. Afin de déterminer quelles étaient les parties du discours qui causaient le plus d'erreurs d'étiquetage, un décompte par étiquette a été réalisé. Pour ce faire, nous avons identifié combien de fois un verbe principal, un adjectif ou bien un nom s'est vu attribuer une étiquette erronée. Nous avons ensuite déterminé la distribution de la confusion pour chaque étiquette (i.e., le nombre d'erreurs d'étiquetage verbe/adjectif, verbe/nom etc.).

Nos analyses ont montré que 52,5% d'erreurs concernent l'annotation du nom, de l'adjectif et du verbe principal. La distribution des erreurs sur ces trois catégories est donnée dans le tableau ci-dessous.

Partie du discours	**Nombre de tags erronés**	**Pourcentage de tags erronés sur l'ensemble des erreurs**
Adjectif	432	22,7%
Nom commun	310	16,3%
Verbe principal	258	13,5%

Tableau 3 : distribution des erreurs de l'étiqueteur BTagger

Comme on peut le voir dans le tableau, la catégorie la plus difficile à annoter pour BTagger est celle des adjectifs, alors que noms et verbes enregistrent des taux d'erreur très proches. Une partie de ces erreurs était prévisible, vu l'homonymie plus ou moins systématique entre certaines catégories grammaticales en serbe. Par exemple, certaines formes fléchies des adjectifs qualificatifs coïncident avec l'adverbe de manière correspondant : *teško* peut être une forme du genre neutre de l'adjectif *težak* 'difficile', mais aussi l'adverbe *teško* 'difficilement'. Par conséquent, nous nous attendions à ce qu'une proportion importante de fautes dans l'étiquetage des adjectifs soit due à la confusion avec les adverbes. Toutefois, contrairement à nos attentes, l'adjectif a été le plus souvent confondu avec le nom commun (56,6% du nombre total d'erreurs pour la catégorie de l'adjectif), puis avec le verbe principal (29,1%), alors que seulement 7,4% d'erreurs concernaient les adverbes. L'adjectif annoté comme nom commun avait typiquement le rôle d'épithète antéposé au nom, et faisait partie d'une suite des adjectifs : *onog jezivog dana* (lit. 'cet horrible jour'), *uzak ravan obod* (lit. 'étroit plat bord'). L'étiquette du verbe principal était le plus souvent attribuée aux formes du participe passé homonyme avec l'adjectif verbal : il n'existe apparemment pas de contraintes contextuelles suffisantes pour distinguer ces deux formes. Par exemple, les deux peuvent figurer avec le verbe attributif *jesam* ('être'), l'adjectif en tant qu'attribut, et le participe dans une forme verbale composée.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

En ce qui concerne les noms communs, les résultats de l'analyse correspondaient à notre intuition : cette catégorie était le plus souvent confondue avec celles de l'adjectif (46,1% d'erreurs) et du verbe principal (21,3%). Une partie des noms communs annotés en tant qu'adjectifs étaient dérivés par la substantivation des adjectifs et, par conséquent, étaient homonymes des formes adjectivales, ce qui explique la source des erreurs. On peut citer en tant qu'exemple la forme *mrtvih*, qui peut représenter le génitif pluriel masculin soit de l'adjectif *mrtav* ('mort'), soit du nom *mrtvi* ('les morts'). La plupart des exemples relevaient toutefois des formes nominales qui ne peuvent pas être interprétées comme adjectifs. Cependant, nous avons remarqué que les terminaisons de ces occurrences apparaissaient également dans le paradigme adjectival : par exemple, la forme *čvorove* est l'accusatif pluriel du nom *čvor* ('nœud'), mais la terminaison *-ove* est typique pour les formes fléchies des adjectifs possessifs dérivés des noms propres, tels *Petrove* ('celle(s)/ceux de Petar'). Il est donc probable que l'erreur d'étiquetage soit due à ces suffixes ambigus. Les cas de confusion entre noms communs et verbes tombent dans ces deux mêmes catégories : une partie relève des formes homonymes avec les formes verbales fléchies (*pogleda* est le génitif singulier du nom *pogled* 'regard', mais aussi la troisième personne du singulier du présent du verbe *pogledati* 'regarder'), alors que l'autre concerne les formes nominales avec les terminaisons communes au paradigme verbal. Ainsi, par exemple, *saksije* est le nominatif pluriel du nom *saksija* 'pot-de-fleurs', et n'est homonyme avec aucun verbe, mais il partage la terminaison *-ije* avec la troisième personne du singulier du présent d'une série des verbes tels *bije* 'il bat', *pije* 'il boit', *krije* 'il cache' etc.

Dans le cas du verbe, les erreurs concernaient majoritairement la confusion avec l'adjectif (40,3%), le nom commun (21,7%) et le verbe auxiliaire (26,7%). La plupart des verbes annotés comme adjectifs étaient des formes du participe passé, un cas d'homonymie assez régulier, déjà décrit ci-dessus. La confusion avec les noms communs était majoritairement due à l'homonymie des formes fléchies en question : par exemple, *sinu,* qui est la troisième personne du singulier d'aoriste du verbe *sinuti* ('briller'), est homonyme du datif singulier du nom commun *sin* ('fils'). La plupart des erreurs concernant le verbe auxiliaire relèvent des forme du verbe *jesam* ('être'), qui peut être un verbe auxiliaire ou un verbe attributif. Dans l'analyse de l'annotation, nous avons relevé nombre d'exemples dans lesquels le verbe *jesam*, le verbe principal de la phrase, portait l'étiquette du verbe auxiliaire. Autrement dit, une forme du verbe *jesam* ne devrait pas être étiquetée comme verbe auxiliaire si dans son contexte on ne trouve pas un participe. Toutefois, comme la syntaxe du serbe permet que l'auxiliaire et le participe soient séparés par plusieurs autres constituants (compléments adverbiaux, compléments d'objet), il est probable que le nombre de mots entre l'auxiliaire et le participe était supérieur à la fenêtre prise en compte par l'étiqueteur et que, par conséquent, Btagger n'ait pas pris la présence du verbe principal comme critère de désambiguïsation des formes du verbe *jesam*. Ceci peut être la raison pour laquelle dans l'exemple 1) ci-dessous la forme *sam,* qui est le présent du verbe *jesam* ('être'), a été annotée en tant que verbe auxiliaire, alors qu'il s'agit d'un verbe principal attributif et que le verbe principal le plus proche se trouve à 14 tokens de distances (*osetim*). Autrement dit, vu la distance entre *sam* et *osetim*, ainsi que les frontières de propositions entre ces deux verbes, il est impossible d'envisager que *sam* soit ici un auxiliaire.

> 1) Zaboravljam da sam novorođenče i da od svih životnih senzacija, ljudskih i božanskih, najviše ako mogu da osetim i doživim scenski efekat sunca
> J'oublie que je suis un nouveau-né et que de toutes les sensations dans la vie, humaines ou divines, je peux au plus sentir et vivre l'effet scénique du soleil.
> (Kiš, D., 1965, *Bašta, Pepeo*)

La glose de cette citation est donnée ci-dessous.

Zaboravljam	da	sam	novorođenče	i
oublier-V-prés-1p	*que-conj*	*être-V-prés-1p*	*nouveau-né-N-Nom*	*et-conj*
da	od	svih	životnih	senzacija

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

que-conj	*de-Prép*	*toutes-Adj-Gen*	*vitales-Adj-Gen*	*sensation-N-Gen*
ljudskih	i	božanskih	najviše	ako
humaines-Adj-Gen	*et-conj*	*divines-Adj-Gen*	*le plus-Adv*	*si-conj*
mogu	da	osetim	i	doživim
pouvoir-V-prés-1p	*que-conj*	*sentir-V-prés-1p*	*et-conj*	*vivre-prés-1p*
scenski	efekat	sunca		
scénique-Adj-Acc	*effet-N-Acc*	*soleil-N-Gén*		

Glose 1 : exemple d'un emploi du verbe *sam* comme attributif et non auxiliaire

L'analyse détaillée des erreurs d'étiquetage nous a permis de voir que la majorité des erreurs relevaient de deux cas de figure principaux : soit il s'agissait des formes homonymes entre deux catégories grammaticales, soit le mot lui-même n'était pas ambigu, mais sa terminaison était partagée par plusieurs paradigmes. La documentation de Btagger n'est pas suffisamment claire sur ce point (voir Gesmundo et Samardzic 2012), mais il paraît possible que, dans le processus d'étiquetage, le logiciel utilise l'analyse des suffixes effectuée par le module de la lemmatisation. Ce type d'erreurs indique aussi que le contexte immédiat des trois catégories principales, à savoir du verbe, du nom commun et de l'adjectif, n'est pas suffisamment discriminant pour permettre une désambiguïsation fiable.

De même, le contexte plus large peut imposer des contraintes importantes, telle la présence du verbe principal dans la proximité relative d'un verbe auxiliaire. Pourtant, la nature discontinue des formes verbales composées en serbe fait que la taille du contexte qui doit être analysé dépasse la taille du contexte examiné par l'étiqueteur. Ceci cause ce qui pourrait être considéré comme l'erreur d'étiquetage la plus grave que nous avons rencontrée : des phrases sans verbe principal. Pour ce dernier cas, une stratégie de contournement pourrait consister en une sorte de méta-règle destinée à signaler toute phrase sans verbe principal aux correcteurs. Malheureusement, pour les autres cas d'erreurs relevés ci-dessus, qui représentent les erreurs les plus courantes, aucune stratégie claire et économique de correction d'étiquettes ou de pré-annotation ne semble émerger des cas examinés jusqu'ici.

3 Apports d'un corpus parallèle étiqueté pour la linguistique française

Bien que la priorité, au cours des premières phases synthétisées plus haut, ait été donnée au sous-corpus serbe en raison du faible niveau d'outillage linguistique de cette langue, le corpus TALC-sef a été conçu comme un outil d'analyse contrastive pour les trois langues qui y sont représentées. Nous proposons d'illustrer ce point à travers l'exemple de l'analyse morpho-sémantique des noms déverbaux. Il est en effet bien connu que de nombreux lexèmes de ce type connaissent deux ou plusieurs lectures différentes, en particulier une lecture prédicative (ou d'action) et une lecture résultative (ex. : *investissement* = « action d'investir » vs. « sommes investies »). Toutefois, l'identification en corpus des différentes lectures d'un même nom déverbal est une tâche relativement longue et difficile, qui nécessite une analyse sémantique manuelle approfondie de chaque occurrence, reposant tant sur un jugement sémantique que sur l'application de tests syntaxico-sémantiques. À titre d'exemple, les deux déverbaux français ici retenus (*entrée* et *observation*) auront deux traductions différentes selon le type de lecture (cf. (2.b) et (3.b) pour *entrée= ulazak* vs *ulaz*, et (4.b) et (5.b) pour *observation= razgledanje* vs *primedba*).[15] Autrement dit, les exemples (2.a) et (4.a) ci-dessous, tirés du corpus TALC-sef, illustrent la lecture prédicative, les exemples (3.a) et (5.a) la lecture résultative pour des noms déverbaux français.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

(2.a) Évariste Gamelin, lui-même, bien que d'humeur sévère, en prenant sur le giron d'Élodie son couteau de six liards, récita de bonne grâce l'**entrée** de Grisbourdon aux enfers. (A. France, *Les Dieux ont soif*)
(2.b) I sam Evarist Gamlen, mada uvek ozbilxan, uzimajući iz Elodijinog krila svoj nožić od šest sua izgovori rado stihove u kojima se opeva **ulazak** Griburdona u pakao.

(3.a) Simon découvre souvent l'empreinte de leurs sandales à l'**entrée** d'un village. (D. Kiš, *Encyclopédie des morts*)
(3.b) Simon često otkriva tragove njihovih sandala na **ulazu** u neko selo.

(4.a) Il est clair pour eux que l'**observation** des timbres à la loupe n'est qu'une partie du goût réprimé de l'évasion qui se cache souvent chez les êtres calmes et stables, peu enclins aux voyages et aux aventures; (Kiš, *Encyclopédie des morts*, p. 62)
(4.b) Njima je jasno da je to **razgledanje** maraka kroz lupu samo deo one prigusxene fantazije koja se cyesto krije u mirnih i stabilnih lxudi, malo sklonih putovanjima i avanturama;

(5.a) Le visage du père Goriot, qui s'était allumé comme le soleil d'un beau jour en entendant l'étudiant, devint sombre à cette cruelle **observation** de Vautrin. (Balzac, *Le Père Goriot*)
(5.b) Lice čiča-Gorioa, koje je sinulo kao sunce na lepom danu kad je čuo studenta, zamrači se na tu svirepu **primedbu** Votrenovu.

Ainsi que le montrent les exemples de traduction des différentes occurrences de noms déverbaux français, la morphologie du serbe s'avère plus explicite que celle du français en ce qui concerne la lecture prédicative *vs.* résultative : les noms déverbaux serbes portant le suffixe *-ak* ou *-nje* semblent majoritairement relever de la lecture prédicative. Si elle s'avère régulière, comme le laisse penser une étude préliminaire, la morphologie du serbe pourrait donc compléter des approches reposant sur le jugement sémantique et le recours à des tests syntaxico-sémantique[16] afin d'identifier de façon semi-automatique les différentes lectures des noms déverbaux français. À titre de comparaison, la ressource Nomage (Balvet *et al.*, 2012) ne comporte qu'une occurrence de *observation*[17], et aucune pour *entrée*, essentiellement en raison des choix de conception de la ressource. Le corpus TALC-sef permettrait de poursuivre la dynamique de caractérisation sémantique des unités prédicatives non verbales, initiée, entre autres, dans des ressources telles que Nomage, en disposant d'indices complémentaires à l'approche manuelle. Toutefois, l'exploitation à grande échelle de ces indices présuppose d'autres niveaux d'annotation syntaxique, dont notamment une analyse en dépendances des trois sous-corpus (cf. *infra*).

4 Synthèse et perspectives

Dans cet article, nous avons présenté TALC-sef, un corpus parallèle de traductions d'ouvrages littéraires en trois langues européennes : français, anglais et serbe, alignées au niveau phrastique. Un corpus de référence pour le serbe de 150 000 mots, vérifié manuellement, a été constitué, à partir duquel des modèles d'étiquetage pour les étiqueteurs Treetagger, TnT et BTagger ont été dérivés. D'autres ressources lexicales ont également été dérivées de ce corpus de référence : bases de *n-grammes*, lexique électronique (sans lemmes pour cette version) notamment. L'ensemble des données et ressources constituées à partir des textes dans les trois langues et de leurs annotations est librement téléchargeable à l'adresse : http://code.google.com/p/tagged-literary-corpus/source/browse/. Pour des raisons de copyright, les corpus primaires ne sont pas encore librement téléchargeables, toutefois une solution technique permettant la consultation des corpus serbes alignés avec les traductions françaises et anglaises est en cours de déploiement (concordancier multilingue en ligne). Le sous-corpus serbe dans son entier a été syntaxiquement étiqueté avec un taux d'erreur estimé à 6 % maximum. Les sous-corpus français et anglais ont, eux, été étiquetés sans vérification manuelle, grâce aux modèles d'étiquetage de Treetagger, pour lesquels (Schmid, 1994) revendique des scores de précision de l'ordre de 96 % pour l'anglais et 92 % pour le français.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

La mise à disposition, dans un futur proche, de ce corpus littéraire aligné et syntaxiquement étiqueté en trois langues européennes se veut comme une première étape vers le développement d'une interface de consultation du corpus aligné tant au niveau phrastique qu'infra-phrastique. En effet, nous envisageons de poursuivre le travail d'annotation syntaxique entrepris jusqu'ici afin de proposer une version annotée non pas seulement en parties du discours, mais également en dépendances, de l'ensemble du corpus, dans les trois langues. En effet, comme le montrent les exemples ci-dessous, nous sommes convaincus qu'une telle analyse en noyaux verbaux et dépendants permet de proposer une représentation syntaxique indépendante de l'ordre des constituants, fondée sur les centres organisateurs que sont les prédicats (en l'occurrence verbaux dans les deux cas) et les fonctions de leurs dépendants respectifs. Ceci devrait permettre des comparaisons syntaxiques sur corpus, dans une optique de syntaxe comparée, comme l'illustrent les deux figures ci-dessous, dans lesquelles la phrase *Dok je govorio psu, gledao mu je pravo u oči i pas ga je razumeo*, tirée de *Rani jadi* de D. Kiš (*Chagrins Précoces*), et sa traduction française *Pendant qu'il parlait au chien, il le regardait droit dans les yeux et le chien le comprenait* sont analysées en dépendances, centrées sur les noyaux verbaux.

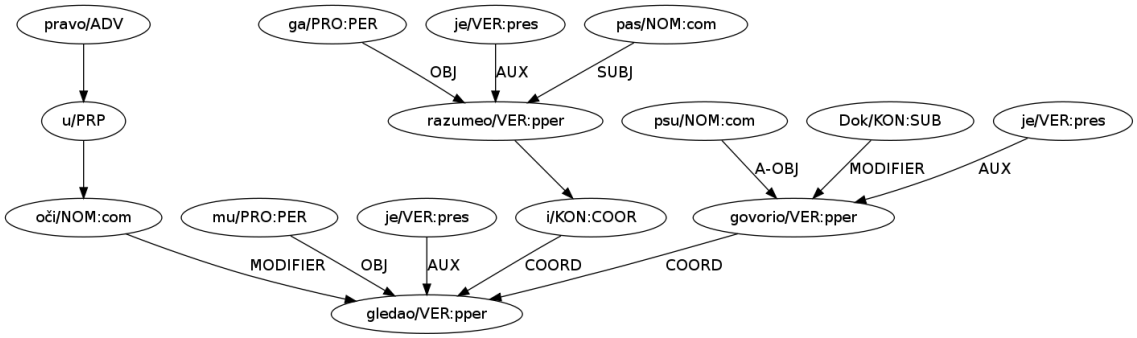

Figure 3 : analyse en dépendances d'une phrase du corpus serbe

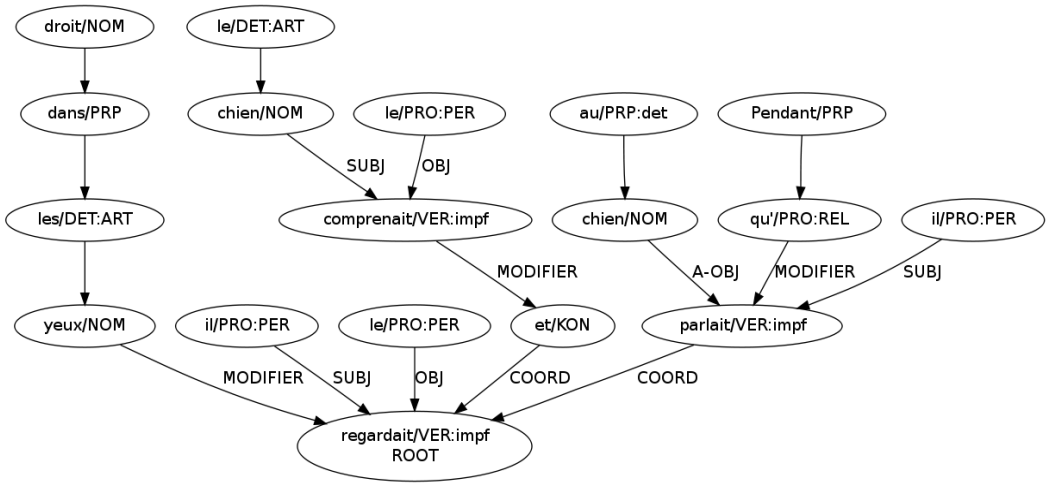

Figure 4 : analyse en dépendances de la traduction française correspondante

Dans les figures ci-dessus, les centres organisateurs que sont les trois noyaux verbaux *govorio(psu)*/*parlait(il, au chien)*, *gledao(mu)*/*regardait(il, le)* et *razumeo(ga)*/*comprenait(le chien, le)* peuvent être alignés, moyennant une analyse de leur valence respective, en tenant compte des différences morpho-syntaxiques entre le français et le serbe.[18] Ici, par exemple, *psu* (datif) et *au chien* portent la

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

même fonction A-OBJ eu égard à leur relation au verbe *govorio* / *parlait*, alors même qu'en français, cette relation se réalise par l'intermédiaire d'une préposition, et qu'en serbe elle pourrait être réalisée soit par une préposition, soit par le biais d'une marque casuelle dative comme dans la figure 3. L'alignement des autres noyaux verbaux et de leurs dépendants respectifs peut également être réalisée pour *gledao* / *regardait* et *razumeo* / *comprenait* grâce à cette analyse.

Outre l'intérêt que représentent les analyses en dépendances pour un tel projet de corpus aligné, l'analyse syntaxique à profondeur modulable de grands volumes de textes constitue une technique aujourd'hui mature dans le domaine du Traitement Automatique des Langues appliqué à la linguistique de corpus : plusieurs analyseurs automatiques robustes sont aujourd'hui disponibles, reposant majoritairement sur des approches probabilistes permettant d'induire les règles d'analyse à partir de corpus de référence annotés manuellement. D'un point de vue plus fondamental, différents algorithmes d'analyse en dépendances ont ont été publiés, comme par exemple : (Attardi, 2006), ou encore (Hall & Nivre, 2008a), (Hall & Nivre, 2008b), (Nivre, 2006), (Nivre, 2008). Dans tous les cas, les auteurs proposent des algorithmes optimisés d'induction de modèles d'analyse en dépendance dans des contextes multilingues. Dans le cas de (Hall & Nivre, 2008a et 2008b), ces algorithmes sont, par ailleurs, spécifiquement orientées vers l'analyse des langues à ordre libre telles que l'allemand ou le serbe. Enfin, des implémentations librement disponibles existent pour chacun de ces algorithmes : sous la forme de l'analyseur MaltParser décrit dans (Nivre, Hall & Nilsson, 2006) et téléchargeable sur http://www.maltparser.org/, ou encore de la plate-forme DeSR[19] pour (Attardi, 2006). Tant MaltParser que DeSR intègrent d'ores et déjà des modèles d'analyse acquis à partir de corpus de référence pour le français et l'anglais. À l'intérêt linguistique apporté par ces analyses s'ajoutent donc des garanties de faisabilité technique, raison pour laquelle nous envisageons, dans un avenir proche, d'exploiter ces plate-formes pour analyser les corpus anglais et français du corpus TALC grâce aux modèles disponibles, puis d'ajouter au sous-corpus serbe des analyses en dépendances manuelles, afin d'induire un modèle d'analyse déployable sur tout le corpus. Ceci permettra un alignement du corpus TALC au niveau infra-phrastique, en prenant comme références les noyaux verbaux et une représentation syntaxique limitée, mais suffisante, indépendante, pour une large part, de l'ordre des constituants et du mode de réalisations des différents dépendants.

Références bibliographiques

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*, 433-441, Granada: ELRA.

Agić Ž., Merkler D. & Berović D. (2013). Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, Seattle: Association for Computational Linguistics.

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, Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb.

Attardi G. (2006). Experiments with a Multilanguage Non-Projective Dependency Parser. *Proceedings of the Tenth Conference on Natural Language Learning*, New York: ACL.

Balvet, A., Barque, L., Condette, M.-H., Haas, P., Huyghe, R., Marín, R. & Merlo, A. (2012). La ressource Nomage, Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus, *TAL* 52, 129-152.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing*, 224-231, Seattle: The Association for Computational Linguistics.

Constant, M., Sigogne, A. (2011). MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources. *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11)*, Portland: The Association for Computational Linguistics.

Daelmans, W., Zavrel, J. *et al.* (1996). MBT: A memory-based part of speech tagger-generator. (E. Ejerhed, & I. Dagan, eds.), *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: City University of Hong Kong Press.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

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

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

Filipović L. (2010). The importance of being a prefix. Prefixal morphology and the lexicalization of motion events in Serbo-Croatian. In: Hasko V. & Perelmutter R. Eds. *New approaches to Slavic verb of motions*. 247–266, Amsterdam: J. Benjamins.

Francis W. N. and Kučera H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown), 1964, 1971, 1979*. Providence, Rhode Island: Brown University.

Gale, W. A., Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics* 19 (1), 75–102, Springer.

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: ELRA.

Hall J., Nivre J. (2008a). A Dependency-Driven Parser for German Dependency and Constituency Representations. *Proceedings of the ACL Workshop on Parsing German (PaGe08)*, Columbus: ACL.

Hall, J., Nivre J. (2008b). Parsing Discontinuous Phrase Structure with Grammatical Functions. *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*, Gothenburg.

Hall, J., Nilsson, J. & Nivre, J. (2010). Single Malt or Blended? A Study in Multilingual Parser Optimization. In Bunt, H., Merlo, P. and Nivre, J. (eds.) *New Trends in Parsing Technology*. Springer.

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

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

Marcus, M., Santorini B., Marcinkiewicz M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2), 313-330.

Miletic, A. (2013). *Annotation semi-automatique en parties du discours d'un corpus littéraire serbe*, Mémoire de Master 2, Université Charles de Gaulle Lille 3.

Nivre, J. (2006). *Inductive Dependency Parsing.* Springer.

Nivre, J. (2008) Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics* 34(4), 513-553.

Nivre, J., Hall J. & Nilsson J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pp. 2216-2219, Genoa: ELRA.

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

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

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

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

Stanojčić, Ž., & Popović, L. (2011). *Gramatika srpskog jezika* (ed. 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*, 523-530, Paris- Athens: ELRA.

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*, 173-180, Edmonton: ACL.

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801035

[1] L'alignement phrastique a été réalisé grâce à l'outil XAlign développé par l'INRIA (http://led.loria.fr/download/source/Xalign.zip), basé sur (Church & Gale, 1993). Cet outil est intégré à la plate-forme Unitex (http://www-igm.univ-mlv.fr/~unitex/), qui a servi à la validation des alignements proposés de manière automatique.

[2] Les différents modèles d'étiquetage et autres ressources lexicales construits à partir du corpus sont, eux, d'ores et déjà disponibles sur http://code.google.com/p/tagged-literary-corpus/source/browse/.

[3] Les méta-données complètes, ainsi que des versions actualisées de ce corpus sont disponibles sur http://nl.ijs.si/ME/V4/doc/teiHeaders/ana/oana-en-teiHeader.html. Ce corpus est annoté au niveau linguistique (lemme, étiquette syntaxique) pour plusieurs langues slaves : serbe, bulgare, tchèque, estonien, hongrois, polonais, roumain, slovaque et slovène, entre autres. La version originale anglaise est également disponible.

[4] Le signe ► indique la langue originelle de chaque œuvre.

[5] Plutôt que législatif, comme c'est le cas pour les corpus parallèles Europarl, qui constituent une norme de fait dans le domaine de par leur taille.

[6] Corpus de phrases analysées syntaxiquement, où chaque phrase est associée à un arbre syntaxique.

[7] Voir (Adda et al., 1998) et (Valli & Véronis, 1999) pour une présentation d'un jeu d'étiquettes syntaxiques de référence pour le français.

[8] Soit une taille comparable au *Brown Corpus* et au *French Treebank*.

[9] REF2 n'a pas été utilisé dans les évaluations quantitatives détaillées en section 2.4.

[10] Soit 9/10 du corpus pour l'entraînement, puis 1/10 pour l'évaluation.

[11] 10 1000/10 = 10 100 mots.

[12] Plus de 1h30 pour chaque passe.

[13] Plus de 45 min. pour chaque passe.

[14] Trigrammes contre arbres de décision.

[15] Le marquage des deux types de lecture n'est pas systématique en serbe, mais il n'en reste pas moins que dans beaucoup de cas la morphologie permet d'identifier sans ambiguïté le type de lecture.

[16] Voir par exemple la base de données du projet ANR Nomage sur http://nomage.recherche.univ-lille3.fr/nomage (Balvet *et. al.*, 2012), dans laquelle les différentes interprétations de noms déverbaux tirés du corpus French Treebank sont annotées en contexte.

[17] Caractérisée en l'occurrence comme ACTIVITÉ, donc comme ayant la lecture prédicative.

[18] Notamment phénomènes de « pro-drop », ou d'incorporation de la personne de conjugaison au verbe, en serbe.

[19] Voir notamment DeSR https://sites.google.com/site/desrparser/, un analyseur en dépendances robuste et multilingue.