Mise au point d'une méthode d'annotation morphosyntaxique fine du serbe

Aleksandra Miletic[1], Cécile Fabre[1], Dejan Stosic[1]
(1) CLLE, Université de Toulouse, CNRS, 5, allées A. Machado, 31058 Toulouse
{prénom.nom}@univ-tlse2.fr

RESUME

Cet article présente une expérience d'annotation morphosyntaxique fine du volet serbe du corpus parallèle ParCoLab (corpus serbe-français-anglais). Elle a consisté à enrichir une annotation existante en parties du discours avec des traits morphosyntaxiques fins, afin de préparer une étape ultérieure de *parsing*. Nous avons comparé trois approches : 1) annotation manuelle ; 2) pré-annotation avec un étiqueteur entraîné sur le croate suivie d'une correction manuelle ; 3) ré-entraînement de l'outil sur un petit échantillon validé du corpus, suivi de l'annotation automatique et de la correction manuelle. Le modèle croate maintient une stabilité globale en passant au serbe, mais les différences entre les deux jeux d'étiquettes exigent des interventions manuelles importantes. Le modèle ré-entraîné sur un échantillon de taille limité (20K tokens) atteint la même exactitude que le modèle existant et le gain de temps observé montre que cette méthode optimise la phase de correction.

ABSTRACT

Developping a method for detailed morphosyntactic tagging of Serbian
This paper presents an experience in detailed morphosyntactic tagging of the Serbian subcorpus of the parallel Serbian-French-English ParCoLab corpus. We enriched an existing POS annotation with finer-grained morphosyntactic properties in order to prepare the corpus for subsequent parsing stages. We compared three approaches: 1) manual annotation; 2) pre-annotation with a tagger trained on Croatian, followed by manual correction; 3) retraining the model on a small validated sample of the corpus (20K tokens), followed by automatic annotation and manual correction. The Croatian model maintains its global stability when applied to Serbian texts, but due to the differences between the two tagsets, important manual interventions were still required. A new model was trained on a validated sample of the corpus: it has the same accuracy as the existing model, but the observed acceleration of the manual correction confirms that it is better suited to the task than the first one.

MOTS-CLES : Annotation morphosyntaxique, corpus d'entraînement, serbe.
KEYWORDS: Morphosyntactic tagging, training corpus, Serbian

1 Introduction

Cette étude s'inscrit dans un projet de constitution d'un corpus parallèle serbe – français – anglais, doté d'annotations morphosyntaxiques et syntaxiques de qualité, ParCoLab (Stosic, 2015),

susceptible d'être exploité en tant que corpus de référence pour des expériences en étiquetage morphosyntaxique et en *parsing* ainsi que pour des études linguistiques. Le travail que nous décrivons se focalise sur le volet serbe de ce corpus, dont un échantillon de 150K tokens dispose d'une annotation en parties du discours et en sous-catégories, détaillée dans (Miletic, 2013). Nous utilisons cet échantillon comme point de départ pour le développement d'un corpus d'entraînement pour le *parsing*. Cette deuxième phase nécessite d'enrichir l'annotation existante. En effet, le serbe, comme toutes les langues slaves, manifeste un ordre de constituants très flexible et les fonctions syntaxiques sont en grande partie encodées dans les indices morphosyntaxiques. Par conséquent, la partie du discours et sa sous-catégorie constituent des informations insuffisantes pour permettre à un *parser* de dériver des règles d'analyse syntaxique fiables. Notre objectif est donc d'enrichir le corpus de traits morphosyntaxiques plus fins avant d'aborder l'analyse syntaxique.

Malgré des développements récents, le serbe reste une langue relativement peu dotée en ressources et outils pour le traitement automatique. Si quelques corpus annotés ont été créés (Krstev et al., 2004, Utvić, 2011, Agić et Klubička, 2014, Jakovljević et al., 2014), ils sont encore difficiles d'accès ou peu adaptés à l'entraînement d'outils. La situation est comparable en ce qui concerne les outils. Ainsi, (Jakovljević et al., 2014) mentionnent AlfaNum POS tagger de Sečujski (2009), mais il ne semble pas librement accessible. À notre connaissance, le seul étiqueteur statistique consacré au serbe est BTagger (Gesmundo et Samardžić, 2012), mais les tests que nous avons réalisés ont montré que sa vitesse d'exécution était insuffisante. Nous nous sommes donc intéressés au modèle de HunPos (Halácsy et al., 2007) pour le croate, développé par Agić et al. (2013). Les auteurs signalent que ce modèle, entraîné seulement sur des données croates, annote des textes serbes avec une exactitude très proche (87% pour le croate *vs* 85% pour le serbe), ce qui s'explique par la proximité prononcée de ces deux langues[1].

Notre objectif est de dégager une méthode optimale pour l'enrichissement de ce corpus serbe en traits morphosyntaxiques, le critère retenu étant le temps nécessaire à l'obtention d'un corpus étiqueté validé. La première approche, l'annotation manuelle, sert de point de comparaison pour étudier le gain apporté par le recours à une pré-annotation automatique avec HunPos. Nous avons testé cet outil selon deux modalités : tout d'abord, nous avons expérimenté l'utilisation du modèle d'Agić et al. (2013), entraîné sur le croate, suivie d'une phase de correction manuelle ; dans un deuxième temps, nous avons ré-entraîné le même outil sur le corpus corrigé et appliqué ce nouveau modèle sur de nouveaux textes, avant de soumettre l'annotation produite à la correction manuelle. Nous décrivons dans les trois sections suivantes les différentes expériences menées et les résultats obtenus, avant de dégager les conclusions de ce travail et les pistes de prolongement.

2 Etablissement du jeu d'étiquettes et annotation manuelle

Pour effectuer l'annotation morphosyntaxique, nous avons établi un jeu de 910 étiquettes. Ce jeu a été construit sur une double base : le traitement des parties du discours a été repris de l'annotation existante, décrite dans (Miletic, 2013), avec un ajout des traits morphosyntaxiques utilisés dans le jeu d'étiquettes pour le serbe du projet MULTEXT-East (Krstev et al., 2004). Le jeu d'étiquettes

[1] Cette proximité est étudiée par exemple par Thomas (2002).

final ne reprend cependant pas toutes les propriétés présentes dans le jeu de MULTEXT-East : la priorité a été accordée aux traits les plus à même d'avoir un rôle au niveau syntaxique (cf. table 1). Certains traits comme l'aspect adjectival (qui marque la définitude en serbe) et l'aspect verbal n'ont donc pas été inclus. Il faut souligner encore que les parties du discours étant traitées avec le jeu d'étiquettes de (Miletic, 2013), le traitement de certaines catégories (notamment les pronoms et les adjectifs) se différencie de celui de MULTEXT-East.

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

TABLE 1: Traits morphosyntaxiques encodés dans le jeu d'étiquettes retenu

Ce jeu d'étiquettes et le schéma d'annotation correspondant ont été utilisés dans une première tentative d'annotation manuelle. La taille importante du jeu d'étiquettes, ainsi que le nombre élevé de traits à encoder pour certaines parties du discours (jusqu'à 7 pour les verbes, cf. table 1) ont rendu cette démarche très coûteuse en temps : la vitesse moyenne atteinte par un annotateur expérimenté était de 500 tokens/h, alors qu'un annotateur nouvellement initié ne dépassait pas 80 tokens/h. Même en considérant la vitesse d'annotation de l'annotateur expérimenté, 300 heures de travail seraient nécessaires pour annoter la totalité de l'échantillon de 150K tokens. Afin de diminuer ce temps, nous avons décidé de tester la possibilité d'une pré-annotation automatique qui serait ensuite corrigée manuellement, reprenant ici une démarche qui a fait ses preuves pour la constitution de corpus à différents niveaux d'annotation (Marcus et al., 1993, Abeillé et al., 2003, Xue et al. 2005, Fort et Sagot, 2010, Tellier et al., 2014).

3 Annotation avec le modèle croate pour HunPos

Pour effectuer une pré-annotation automatique de l'échantillon, nous avons sélectionné HunPos de Halácsy et al. (2007). À notre connaissance, cet étiqueteur n'a pas été entraîné sur le serbe, mais un modèle pour le croate a été développé par Agić et al. (2013). Ce modèle a été entraîné sur un corpus journalistique de 87K tokens (SETimes.hr, ibid). Il a été évalué sur le croate et le serbe, sur des textes journalistiques, mais aussi sur des textes issus de Wikipédia. Les taux d'exactitude pour les deux langues et les deux types de textes indiqués dans (Agić et al., 2013) sont donnés dans la table

2. Le modèle atteint 87,72% d'exactitude sur le croate, et 85,56% sur le serbe. Un fait remarquable se dégage en comparant les résultats obtenus sur les deux langues : le modèle semble plus affecté par un changement de domaine (journalistique *vs* encyclopédique) que par le changement de langue (croate *vs* serbe), ce qui peut s'expliquer par la forte proximité de ces deux langues (cf. Thomas, 2002).

croate		serbe	
presse	Wikipedia	presse	Wikipedia
87,72%	81,52%	85,56%	82,79%

TABLE 2 : Résultats originaux de HunPos sur le serbe et le croate

Nous avons utilisé ce modèle pour annoter automatiquement un échantillon de 20K tokens avec HunPos, avant de procéder à une correction manuelle. Les évaluations décrites dans ce qui suit ont été effectuées au niveau de granularité le plus fin de l'annotation, intégrant tous les traits afférents. Une étiquette est considérée comme incorrecte dès que l'un de ses traits est erroné.

Le score de base du modèle croate sur le texte de ParCoLab était de 77,95%[2], ce qui représente une baisse de 8% par rapport aux résultats signalés dans (Agić et al., 2013). Nous pouvons faire l'hypothèse, qui va dans le sens des observations précédentes, que cette baisse a été provoquée par le changement de domaine (presse *vs* littérature) plutôt que par le changement de langue, d'autant que l'écart entre les domaines est dans notre cas probablement plus marqué. Par ailleurs, le jeu d'étiquettes intégré au modèle de (Agić et al., 2013) est le jeu croate du projet MULTEXT-East (Erjavec, 2012)[3], fondé sur les mêmes principes de base que le jeu d'étiquettes serbe du même projet. Par conséquent, le jeu d'étiquettes et le schéma d'annotation du modèle croate ne coïncident pas parfaitement avec celui que nous avons adopté pour cette expérience. Même si une conversion automatique des étiquettes produites par HunPos vers notre jeu d'étiquettes a pu être effectuée, les divergences entre les deux schémas d'annotation n'ont pas pu être réduites par une correction automatique, le traitement correct étant fortement dépendant du contexte. La correction manuelle incluait donc non seulement la correction des erreurs d'étiquetage proprement dites, mais aussi les interventions nécessaires pour éliminer les différences entre les deux schémas d'annotation, ce qui correspondait au taux total de 26,2% d'étiquettes à corriger.

Malgré ce fait, cette méthode a été effectivement plus rapide que l'annotation manuelle intégrale : l'annotateur expérimenté traite en moyenne 620 tokens/h avec la pré-annotation, soit environ 24% de tokens par heure de plus. L'accélération est cependant moins importante qu'on ne le souhaiterait et ceci peut être imputé à la combinaison des deux facteurs évoqués ci-dessus : la baisse d'exactitude de HunPos et les différences entre le schéma d'annotation du modèle croate et le nôtre.

---

2 Cette évaluation a été effectuée en prenant en compte le schéma d'annotation sur lequel le modèle a été entraîné, et non pas celui de ParCoLab, pour s'assurer de mesurer seulement les performances du modèle.

3 Le jeu d'étiquettes croate est présenté en détail à l'adresse suivante : http://nl.ijs.si/ME/V4/msd/html/msd-hr.html.

4 Annotation avec le modèle ré-entraîné de HunPos

Nous avons cherché à augmenter le gain de vitesse de correction de l'annotation produite avec le modèle d'Agić et al. (2013) en testant une nouvelle piste : se servir des 20K tokens validés pour ré-entraîner HunPos et utiliser le modèle obtenu pour pré-annoter un nouvel échantillon de texte. L'apprentissage a été effectué dans les mêmes conditions que pour le modèle croate : aucune ressource externe n'a été utilisée. Notre inquiétude principale était qu'un échantillon de cette taille ne serait pas suffisant pour entraîner un modèle aussi performant que le modèle existant, à fortiori sur un jeu d'étiquettes étendu. Nous avons donc évalué le modèle ré-entraîné aussi bien du point de vue de ses performances (évaluation quantitative), que de ses effets sur les temps de correction.

4.1 Evaluation quantitative du modèle ré-entraîné

Pour évaluer les performances de base du modèle ré-entraîné, nous avons effectué une validation croisée à 10 itérations en utilisant comme corpus les 20K tokens issus de l'annotation avec HunPos après correction manuelle. On constate que la variation est assez importante, entre 70% et 83% (cf. table 3), ce qui était prévisible étant donné la taille très limitée du corpus d'entraînement. Néanmoins, malgré la différence de taille entre les corpus d'entraînement pour le modèle croate (87K tokens) et le modèle ré-entraîné (20K tokens), les performances de base des deux modèles sur les textes de ParCoLab sont très proches (respectivement 77,95% et 78,82% d'exactitude).

	Test1	Test2	Test3	Test4	Test5	Test6	Test7	Test8	Test9	Test10	Moyenne
Exactitude	76,23	82,97	83,52	79,83	80,68	70,91	83,28	78,47	74,52	77,73	**78,82**
Taille moyenne du corpus d'entraînement							18 370 tokens				
Taille moyenne du corpus d'évaluation							2 040 tokens				

TABLE 3 : Résultats de la validation croisée du modèle ré-entraîné

Ultérieurement, nous avons recalculé l'exactitude du modèle sur un échantillon de 2000 tokens inconnus (non inclus dans le corpus d'entraînement) : elle était de 77%. Autrement dit, les performances de l'outil n'ont pas baissé de manière importante en passant à un texte inconnu. HunPos s'est également montré très satisfaisant en ce qui concerne la vitesse d'exécution.

4.2 Evaluation de la rapidité de la correction manuelle

Dans le cadre de notre tâche, l'avantage principal du modèle ré-entraîné résidait dans le fait qu'il éliminait le besoin de combiner, dans la phase de correction, l'ajustement du jeu d'étiquettes et la correction des erreurs d'étiquetage. Pour confirmer que cet avantage est significatif pour la correction manuelle, nous avons mesuré la vitesse de correction de la sortie de ce nouveau modèle.

Nous avons utilisé le modèle ré-entraîné de HunPos pour annoter un nouvel échantillon de 20K tokens, non compris dans le corpus d'entraînement. Cette fois-ci, la rapidité d'annotation est augmentée de manière importante par rapport à l'annotation manuelle sans pré-annotation : l'annotateur expert traite en moyenne 800 tokens/h, et l'annotateur novice atteint la vitesse de 325 tokens/h. Il s'agit donc d'une augmentation du nombre de tokens traités de respectivement 60% et plus de 300% par rapport à l'annotation manuelle intégrale. D'après les témoignages des annotateurs, cette accélération est en grande partie due au fait qu'il n'était plus nécessaire d'adapter le traitement d'une sous-classe des pronoms au schéma d'annotation de ParCoLab, ce qui était ressenti comme le type de correction le plus chronophage dans l'étape précédente, par rapport à laquelle on note une accélération de correction de 29%.

Ayant jugé ces résultats concluants, nous avons retenu le modèle ré-entraîné comme outil de travail. Il sera utilisé pour entamer un processus de *bootstrapping* : un nouvel échantillon de 20K tokens sera annoté avec ce modèle, puis corrigé manuellement et joint ensuite au corpus d'entraînement initial pour un nouveau cycle de ré-entraînement. Ce procédé sera réitéré jusqu'à l'annotation de la totalité de l'échantillon de 150K tokens.

5 Conclusions et pistes

Nous avons présenté une démarche pour l'annotation morphosyntaxique riche d'un corpus serbe, fondée sur 3 principes : 1) la mise au point d'un nouveau jeu d'étiquettes enrichi, conçu pour faciliter la phase ultérieure de *parsing* ; 2) l'exploitation d'un modèle croate, langue typologiquement très proche du serbe, pour une pré-annotation automatique initiale afin d'accélérer l'annotation manuelle, et 3) le ré-entraînement d'un modèle sur un échantillon validé du corpus afin de minimiser le besoin d'intervention manuelle dans la phase de correction.

Les résultats présentés confirment la stabilité globale du modèle de HunPos pour le croate de Agić et al. (2013) lors du passage au serbe, même si une baisse d'exactitude a été constatée. La pré-annotation avec ce modèle a apporté une accélération du traitement de tokens de 24% par rapport à l'annotation manuelle intégrale. Après validation, un échantillon de cette annotation a été utilisé pour ré-entraîner HunPos. Les résultats obtenus (exactitude moyenne de 78,82%) montrent que, malgré la richesse du jeu d'étiquettes utilisé, il est possible d'entraîner un modèle statistique de manière satisfaisante à partir d'un corpus annoté de taille réduite, notamment pour la tâche en question : la pré-annotation automatique avec le modèle ré-entraîné apporte un gain très important en temps par rapport à l'annotation manuelle intégrale (60% pour un annotateur expérimenté, et plus de 300% pour un annotateur novice).

Dans l'immédiat, ce travail nous a permis d'entamer le cycle de *bootstrapping* : nous avons annoté un nouvel échantillon de 20K tokens, dont la correction est en cours. Ces tokens seront ensuite joints à l'échantillon initial pour un nouveau cycle d'entraînement et d'évaluation. Cette augmentation du corpus d'entraînement devrait permettre d'améliorer les performances de l'outil et de faciliter la correction manuelle. Grâce à cette méthode, nous pourrons atteindre plus rapidement l'objectif global, à savoir l'annotation de la totalité du volet serbe en traits morphosyntaxiques indispensables au démarrage du *parsing*.

Références

ABEILLÉ, A, CLÉMENT, L., TOUSSENEL, F. (2003). Building a Treebank for French. In A. Abeillé (éd.), *Treebanks: Building and using parsed corpora* (pp. 165-184). Dodrecht : Kluwer.

AGIĆ Ž., LJUBEŠIĆ N., MERKLER D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. Actes de *4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, 48-57. Sofia, Bulgaria.

ERJAVEC T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, 46(1), 131-142.

FORT, K., SAGOT, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. Actes de *4th ACL Linguistic Annotation Workshop*, 56-63. Uppsala, Sweden.

GESMUNDO A., SAMARDŽIĆ T. (2012). Lemmatising Serbian as a category tagging with bidirectional sequence classification. Actes de *8th Language Ressources and Evaluation Conference (LREC2012)*, 2103-2106. Istanbul, Turkey.

HALÁCSY P., KORNAI A., ORAVECZ C. (2007). HunPos: an open source trigram tagger. Actes de *45th annual meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration sessions (ACL'07)*, 209-212. Prague, République Tchèque.

JAKOVLJEVIĆ B., KOVAČEVIĆ A., SEČUJSKI M., MARKOVIĆ M. (2014). A Dependency Treebank for Serbian: Initial Experiments. *Speech and Computer Lecture Notes in Computer Science* 8773, 42-49.

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

LJUBEŠIĆ N., KLUBIČKA F. (2014) {bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian. Actes de *9th Web as Corpus Workshop (WaC-9)*, 29-35**.** Göteborg, Suède.

MARCUS, M. P., MARCINKIEWICZ, M. A., SANTORINI, B. (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, Université Charles de Gaulle Lille 3*.

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

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

TELLIER, I., ESHKOL-TARAVELLA, I, DUPONT, Y., WANG, I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS ?. Actes de *Traitement automatique de langage naturel (TALN2014)*. Marseille, France.

THOMAS, P.-L. (2002) Le serbo-croate (bosniaque, croate, monténégrin, serbe) : de l'étude d'une langue à l'identité des langues. *Revue des études slaves, 74(2-3)*, 311-325.

UTVIĆ, M. (2011). Annotating the Corpus of contemporary Serbian. *INFOtheca 2(12)*, 36-47.

XUE, N., XIA, F., CHIOU, F. D., PALMER, M. (2005). The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering, 11(02)*, 207-238.