ParCoTrain : corpus d'entraînement et de test pour l'étiquetage et la lemmatisation du serbe

1. Description brève

ParCoTrain est un corpus d'entraînement et de test pour l'étiquetage en parties du discours et la lemmatisation du serbe. Le corpus a été développé dans le cadre du projet ParCoLab (http://parcolab.univ-tlse2.fr/) ayant pour objectif de produire des ressources librement accessibles pour le traitement automatique du serbe et pour sa comparaison avec le français et l'anglais. La partie lemmatisée du corpus contient 95 585 tokens annotés manuellement. La partie du corpus dotée d'un étiquetage en parties du discours compte 153 625 tokens au total, dont 95 585 ont été annotés manuellement, alors que les 57 977 tokens restants ont été annotés automatiquement pour être ensuite validés manuellement. Les textes source pour le corpus sont des romans serbes contemporains datant de la deuxième moitié du 20^{e} siècle. L'étiquetage en parties du discours utilisé dans le corpus indique la catégorie principale et la sous-catégorie, ainsi que le degré de comparaison pour les adjectifs et les adverbes.

2. Structure du corpus

Les textes source suivants ont été utilisés pour créer le corpus :

- Kiš, Danilo. "Enciklopedija mrtvih", 2000. Beograd: BIGZ
- Stevanović, Vidosav. "Testament", 1986. Beograd: SKZ.
- Kiš, Danilo. "Bašta, pepeo", 2010. Podgorica: Narodna knjiga.

Le corpus lui-même contient 8 fichiers au format csv avec la tabulation comme séparateur de champ. L'encodage des caractères utilisé est UTF-8, et le caractère de fin de ligne CR-LF (\r\n). La liste des fichiers et de leur contenu est donnée dans la suite.

Fichier	Taille (tokens)	Contenu	Niveaux d'annotation	Type d'annotation	Format
enciklopedija-testament.txt	95 585	"Enciklopedija mrtvih" (47 792 tok.) "Testament" (47 793 tok.)	POS lemmatisation	Manuelle	[token][lemme][POS]
enciklopedija.txt	47 792	"Enciklopedija mrtvih"	POS lemmatisation	Manuelle	[token][lemme][POS]
testament.txt	47 793	"Testament"	POS lemmatization	Manuelle	[token][lemme][POS]
enciklopedija-sample1.txt	23 908	"Enciklopedija mrtvih"	POS lemmatisation	Manuelle	[token][lemme][POS]
enciklopedija-sample2.txt	23 885	"Enciklopedija mrtvih"	POS lemmatisation	Manuelle	[token][lemme][POS]
testament-sample1.txt	23 908	"Testament"	POS lemmatisation	Manuelle	[token][lemme][POS]
testament-sample2.txt	23 884	"Testament"	POS lemmatisation	Manuelle	[token][lemme][POS]
basta.txt	57 977	"Bašta, pepeo"	POS	Automatique, validée manuellement	[token][POS]

enciklopedija-testament.txt est le résultat de la concaténation des fichiers *enciklopedija.txt* et *testament.txt*.

enciklopedija-sample1.txt, enciklopedija-sample2.txt, testament-sample1.txt et *testament-sample2.txt* sont des échantillons équilibrés dérivés à partir des fichiers *enciklopedija.txt* et *testament.txt*. Chacun des deux grands fichiers a été divisé en deux de sorte à avoir des échantillons aussi équilibrés que possible tout en préservant l'intégrité phrastique. Ces fichiers peuvent être exploités dans le cadre d'une validation croisée à 4 itérations.

basta.txt contient seulement l'annotation en parties du discours. Ce fichier a été annoté de manière automatique avec BTagger (Gesmundo & Samardzic, 2012) entraîné sur les 4 échantillons équilibrés (fichiers *sample*). La sortie de l'étiqueteur a ensuite été manuellement validée. Ce fichier a été utilisé par la suite afin d'augmenter le corpus d'entraînement pour l'étiquetage POS et disposer de 153 625 tokens étiquetés au total. Si l'on souhaite établir un fichier avec la totalité des données annotées en POS, ceci peut être fait de manière suivante :

1. créer un fichier qui retient seulement l'étiquetage en parties du discours du fichier enciklopedija-testament.txt en utilisant la commande suivante sur la ligne de commande :

$ cut –f 1,3 enciklopedija-testament.txt > enciklopedija-testament-POS.txt

2. concaténer ce fichier avec basta.txt :

$ cat enciklopedija-testament-POS.txt basta.txt > enciklopedija-testament-basta-POS.txt

3. Étiquetage en parties du discours implémenté dans le corpus

Le jeu d'étiquettes utilisé pour l'étiquetage en parties du discours contient 46 étiquettes. Les étiquettes indiquent la catégorie principale et la sous-catégorie, ainsi que le degré de comparaison pour les adjectifs et les adverbes. Toutes les étiquettes sont présentées dans la table 1, accompagnées des exemples d'usage.

Table 1 : Jeu d'étiquettes des parties du discours

Etiquette	Sous-catégorie grammaticale	Exemple
NOM:com	nom commun	*vrata* 'porte', *kuća* 'maison', *strast* 'passion'
NOM:col	nom collectif	*lišće* 'feuillage', *pilad* 'poussins', *kamenje* 'rochers'
NOM:nam	nom propre	*Duško*, *Beograd* 'Belgrade', *Afrika* 'Afrique'
NOM:num	nom numéral	*dvojica* 'les deux', *trojica* 'les trois'
NOM:approx	nom approximatif	desetak 'dizaine', pedesetak 'cinquantaine'
VER	verbe	*jedem* 'je mange', *radio* 'travaillé'
VER:aux	verbe auxiliaire	*sam* 'je suis', *ćete* 'vous voulez'
PRO:per	pronom personnel	*ja* 'je', *mene* 'me', *ti* 'tu', *vi* 'vous'
PRO:intr	pronom interrogatif	*ko* 'qui', *šta* 'quoi'
PRO:dem	pronom démonstratif	*ovaj* 'celui-ci', *ona* 'celle-là', *ti* 'ceux-là'
PRO:ind	pronom indéfini	*neko* 'quelqu'un', *niko* 'personne', *svako* 'tout le monde'
PRO:pos	pronom possessif	*moj* 'le mien', *naši* 'les nôtres', *njihovi* 'les leurs'
PRO:rel	pronom relatif	*koji* 'qui'
PRO:ref	pronom réfléchi	*sebe*, *se* 'soi-même'
PRO:num	pronom numéral	*jedan* 'un', *drugi* 'deuxième',' autre'
ADJ	adjectif au positif	*nov* 'neuf', *lepa* 'belle'
ADJ:comp	adjectif au comparatif	*noviji* 'plus neuf', *lepša* 'plus belle'
ADJ:sup	adjectif au superlatif	*najnoviji* 'le plus neuf', *najlepša* 'la plus belle'
ADJ:intr	adjectif interrogatif	*koji* 'lequel', *kakav* 'comment' (adj.), *koliki* 'de quelle taille'
ADJ:dem	adjectif démonstratif	*ovaj* 'ce', *ona* 'celle', *ti* 'ces'
ADJ:ind	adjectif indéfini	*neki* 'certain', *nijedan* 'aucun', *svaki* 'tout'
ADJ:pos	adjectif possessif	*moj* 'mon', *naši* 'notre', *njihovi* 'leurs'
ADJ:rel	adjectif relatif	*čiji* 'de qui', 'dont', *kakav* 'comment' (adj.), *koliki* 'de quelle taille'
NUM:car	numéral cardinal	*jedan* 'un', *jedna* 'une', *dvadeset* 'vingt'
NUM:ord	numéral ordinal	*prvi* 'premier', *druga* 'deuxième', *dvadeseti* 'le vingtième'
NUM:col	numéral collectif	*dvoje* 'deux', *petoro* 'cinq', *dvadesetoro* 'vingt'
ADV	adverbe (autre que relatif, interrogatif ou indéfini)	*pametno* 'intelligemment', *nespretno* 'maladroitement'
ADV:comp	adverbe au comparatif	*bolje* 'mieux', *pametnije* 'plus intelligemment'
ADV:sup	adverbe au superlatif	*najbolje* 'le mieux', *najpametnije* 'le plus intelligemment'
ADV:intr	adverbe interrogatif	*kako* 'comment', *gde* 'où', *kad* 'quand'
ADV:rel	adverbe relatif	*kako* 'comme', *gde* 'où', *kad* 'quand'
ADV:ind	adverbe indéfini	*nekako* 'n'importe comment', *igde* 'n'importe où'
CONJ:coor	conjonction de coordination	*i* 'et', *ali* 'mais', *ili* 'ou'
CONJ:sub	conjonction de subordination	*da* 'que', *jer* 'parce que', *iako* 'bien que'
PREP	préposition	*na* 'sur', *pod* 'sous', *u* 'dans'
PAR	particule	*da* 'oui', *ne* 'non', *čak* 'même'
INT	interjection	*ah* 'ah', *apćiha* 'atchoum', *hej* 'hé'
SENT	ponctuation forte	. ! ?
PONC	ponctuation faible	, ; : ( )
PONC:cit	ponctuation de citation	« » " „ “
STR	mot étranger	*chéri*
ABR	abréviation	*dr*, *itd*. (etc.)
LET	lettre	A, p, L
NUM	nombre écrit en chiffres	12, 252, XII
PAGE	numéro de page	7, 10
ID	indicateur du numéro de page	@@

3.1 Quelques remarques sur des cas particuliers

Les choix opérés dans la constitution de ce jeu d'étiquettes ont été présentés en détail dans (Miletic, 2013). Quelques-uns d'entre eux sont illustrés dans la suite.

Les noms collectifs sont distingués pour leur comportement morpho-syntaxique spécifique : ils se déclinent comme des noms au singulier, alors que sémantiquement ils désignent un référent collectif et/ou pluriel. Pour la plupart dérivés, ils ont des particularités morphologiques qui permettent de les identifier par les suffixes qui permettent de les construire.

Les noms numéraux désignent typiquement un ensemble d'individus. Or, ils se déclinent comme les noms féminins au singulier. Ils peuvent donc imposer le singulier aussi bien que le pluriel aux adjectifs qui les modifient.

Les noms approximatifs indiquent une quantité imprécise et sont dérivés des adjectifs numéraux cardinaux tels *dvadesetak* 'une vingtaine'.

Traditionnellement, la grammaire serbe considère comme pronoms les formes telles que *neko* 'quelqu'un', *niko* 'personne' *svako* 'chacun' aussi bien que les formes comme *neki* '(un) certain', *nijedan* 'aucun', *svaki* 'chaque'. Le comportement syntaxique des deux séries n'est pas le même : les formes de la première, appelées *pronoms nominaux*, sont indépendantes des noms, alors que celles de la deuxième, nommées *pronoms adjectivaux,* accompagnent toujours un nom. La première série est donc équivalente de la classe des pronoms indéfinis, démonstratifs et possessifs en anglais et français, alors que la deuxième correspond à différentes sous-classes des déterminants en français et en anglais. Dans le corpus, seules les formes de la première série sont traitées comme pronoms dans le corpus, alors que les formes de la deuxième sont assimilées à la catégorie des adjectifs (le serbe n'ayant pas de déterminant, l'étiquette « adjectif » a été privilégiée). Les formes qui peuvent avoir les deux fonctionnements (cf. formes possessives telles *moj* 'mon/le mien', *tvoj* 'ton/le tien' ou les formes démonstratives comme *ovaj* 'ce/celui' (proximal), *onaj* 'ce/celui' (distal) sont annotées en fonction du contexte : si elles apparaissent de manière autonome, elles sont traitées comme pronoms, et si elles accompagnent un nom, elles sont considérées comme des adjectifs (comparer *Više mi se sviđa tvoj nego moj* 'Je préfère le tien au mien' avec *Više mi se sviđa tvoj automobil nego moj motor* 'Je préfère ta voiture à ma moto').

La grammaire serbe distingue traditionnellement les sous-classes suivantes des pronoms : possessifs, interrogatifs, relatifs, démonstratifs, indéfinis, généraux et négatifs. Tout en maintenant les 3 premières sous-catégories, le jeu d'étiquettes rassemble les formes indéfinies (*neko* 'quelqu'un', *neki* '(un) certain'), générales (*svako* 'chacun', svaki 'chaque') et négatives (*niko* 'personne', *nijedan* 'aucun') en une sous-classe des indéfinis. La même sous-catégorisation est opérée pour les adjectifs.

4. Références

Balvet, A., Stosic, D., & Miletic, A. (2014). TALC-sef, Un corpus étiqueté de traductions littéraires en serbe, anglais et français. In *SHS Web of Conferences* (Vol. 8, pp. 2551-2563). EDP Sciences.

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.

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