Création des ressources de TAL pour une langue peu dotée : le cas du serbe

Aleksandra Miletic
UMR 5263 CLLE-ERSS
Axes CARTEL & S'caladis
aleksandra.miletic@univ-tlse2.fr

Sommaire

- En quoi le serbe est-il spécifique ?
- Application pratique du travail ?
- Annotation morphosyntaxique (master)

- 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.
- Balvet, A., Stosic, D., Miletic, A. (2014). TALC-sef, A Manually-Revised POS-Tagged Litterary Corpus in Serbian, English, and French, Proceedings of The Ninth International Conference on Language Resources and Evaluation (LREC'14), 4105-4110.
- Miletic, A. (2014). POS-Tagging a Litterary Corpus of Serbian, oral presentation at Young Linguists' International Conference. Stockholm.

- Annotation syntaxique (thèse ; en cours)

- Miletic, A., Fabre, C., Stosic, D. (2015). Construction du jeu d'étiquettes pour le parsing du serbe, Actes de TASLA à TALN 2015, 1-9.
- Miletic, A., Fabre, C., Sajous, F., Stosic, D. (soumis). Building a morphosyntactic lexicon for Serbian. LREC2016.

- Bilan et pistes

SERBE : LANGUE SLAVE PEU DOTÉE

Serbe : langue slave

- Morphologie flexionnelle riche
  - Nombre de formes fléchies élevé
  - Syncrétisme

- Syntaxe flexible
  - Ordre de constituants libre
  - Relations à longue distance
  - Constituants discontinus

Serbe : langue slave

- Flexion nominale beaucoup plus riche qu'en français :

Flexion du nom *knjiga* 'livre'		
	Singulier	Pluriel
Nominatif	knjiga	knjige
Génitif	knjige	knjiga
Datif	knjizi	knjigama
Accusatif	knjigu	knjige
Vocatif	knjigo	knjige
Instrumental	knjigom	knjigama
Locatif	knjizi	knjigama

- ~ 14 formes fléchies par lemme
- 4 classes de déclinaison

Serbe : langue slave

- Flexion adjectivale beaucoup plus riche qu'en français :

Flexion de l'adjectif *lep* 'beau' au singulier			
	Masculin	Féminin	Neutre
Nominatif	lepi	lepa	lepo
Génitif	lepog	lepe	lepog
Datif	lepom	lepoj	lepom
Accusatif	lepog	lepu	lepo
Vocatif	lepi	lepa	lepo
Instrumental	lepim	lepom	lepim
Locatif	lepom	lepoj	lepom

- 'aspect' adjectival : défini ou indéfini
- 3 degrés de comparaison
- Jusqu'à 168 formes fléchies

Serbe : langue slave

Ordre de constituants libre

- Ordre canonique : SVO

Filip čita knjigu. (’Filip lit un/le livre’)

- Ordres grammaticaux
  - SOV : *Filip knjigu čita.*
  - VSO : *Čita Filip knjigu.*
  - VOS : *Čita knjigu Filip .*
  - OVS : *Knjigu čita Filip.*
  - OSV : *Knjigu Filip čita.*

Serbe : langue slave

Filip je čitao lepu knjigu. ‘Filip lisait un beau livre.’

Serbe : langue slave

Filip je čitao lepu knjigu. ‘Filip lisait un beau livre.’

- Relations à longue distance :

Filip je lepu knjigu čitao.

Serbe : langue slave

Filip je čitao lepu knjigu. ‘Filip lisait un beau livre.’

- Relations à longue distance :

Filip je lepu knjigu čitao.

- Constituants discontinus :

Lepu je knjigu Filip čitao.

Serbe : langue slave

Filip je čitao lepu knjigu. ‘Filip lisait un beau livre.’

- Relations à longue distance :

Filip je lepu knjigu čitao.

- Constituants discontinus :

Lepu je knjigu Filip čitao.

~ C’est un beau livre que Filip lisait.

Serbe : langue peu dotée ?

Ressources lexicales	Descriptif	Taille
SrpMD (Krstev et al., 2004b)	Dictionnaire morphologique	85K lemmes 3,6M entrées
SWN (idem)	Wordnet serbe	6K synsets
AlfaNum MD (Jakovljevic et al., 2014)	Dictionnaire morphologique	100K lemmes 3,8M entrées

Corpus	Descriptif	Taille
MULTEXT-East (Krstev et al., 2004)	Étiquetage MS lemmatisation	108K tokens
SrpKor (Krstev et Vitas, 2005)	Parties du discours lemmatisation	113M tokens
AlfaNum Corpus (Jakovljevic et al., 2014)	Étiquetage MS lemmatisation	200K tokens

Serbe : langue peu dotée ?

POS-tagging / annotation morphosyntaxique / lemmatisation

Outil	Descriptif	Type	Précision
BTagger (Gesmundo et Samardzic, 2012)	Étiqueteur MS/ lemmatisateur	statistique	86%
AlfaNum Tagger (Secujski, 2009)	Étiqueteur MS	sur base de règles	93,2%

Parsing

- Pas de treebank
- Pas de modèles de parsing
- Premières expériences: Jakovljevic et al. (2014.)

PARCOLAB : CORPUS PARALLÈLE FRANÇAIS – SERBE – ANGLAIS

ParCoLab

Objectifs :

- Corpus annoté et lemmatisé interrogeable en ligne
- Corpus d'entraînement pour le serbe redistribuable
- Genres divers : littérature, textes parallèles provenant du web, sous-titres des filmes
- Applications : linguistique, didactique des langues, TAL, traductologie

ParCoLab

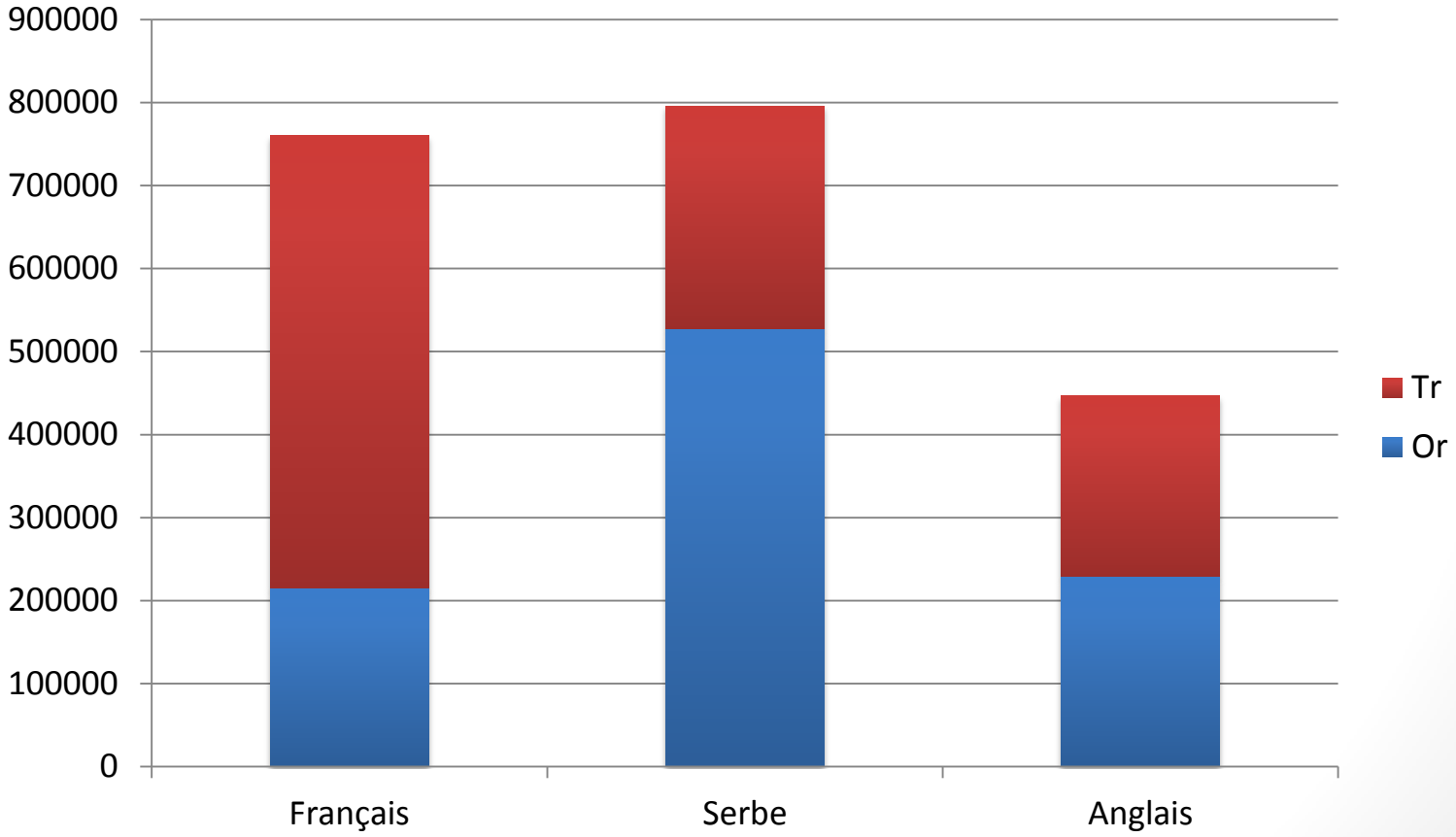
- Constitution et développement depuis 2007
- Taille : 2M tokens
- Textes majoritairement littéraires
- Alignements au niveau du paragraphe et de la phrase, vérifiés manuellement

Annotation :

- Morphosyntaxique : 157 000 tokens (mémoire de M2)
- Syntaxique : travaux en cours (thèse)

ParCoLab

- Chaque langue : langue source et langue cible



ParCoLab

- Moteur de recherche:
  - Recherche de mots
  - Recherche d'expressions complexes (avec joker(s))
  - Usage des expressions régulières
  - Usage des opérateurs booléens

- Consultable en ligne : http://parcolab.univ-tlse2.fr

ANNOTATION MORPHOSYNTAXIQUE D'UN CORPUS LITTÉRAIRE SERBE

Mémoire de Master 2 à l'Université Lille 3 (2012-2013)

Directeurs :

M. Antonio Balvet (Université Lille 3, UMR 8613 STL)

M. Dejan Stosic (Université d'Artois, Centre de recherche *Grammatica*)

Résultats

- Jeu d'étiquettes de taille modérée : 45
  - Taille vs. information
- Corpus d'entraînement pour l'annotation morphosyntaxique manuellement annoté ou corrigé
  - 157 000 tokens
- Modèles d'étiquetage disponibles en ligne
  - http://abalvet.github.io/TALC-sef/

Corpus d'entraînement existant

cesAna (MULTEXT-East ; Krstev et al., 2004)

- 108 000 tokens
- Traduction de *1984* d'Orwell
- Jeu d'étiquettes : > 900 étiquettes

Corpus d'entraînement existant

cesAna (MULTEXT-East ; Krstev et al., 2004)

- 108 000 tokens
- Traduction de *1984* d'Orwell
- Jeu d'étiquettes : > 900 étiquettes

Expériences antérieures sur *cesAna* :

- (Popovic, 2010) :
  - TnT : 85,47%
- (Gesmundo et Samardzic , 2012)
  - BTagger : 86,65%

Corpus d'entraînement existant

cesAna (MULTEXT-East ; Krstev et al., 2004)

- 108 000 tokens
- Traduction de *1984* d'Orwell
- Jeu d'étiquettes : > 900 étiquettes

Expériences antérieures sur *cesAna* :

- (Popovic, 2010) :
  - TnT : 85,47%
- (Gesmundo et Samardzic , 2012)
  - BTagger : 86,65%
- (Utvic, 2011)
  - Corpus d'entraînement : 1 000 000 tokens, 16 étiquettes
  - TreeTagger : 96,57%

Méthode

- Constitution d’un jeu d’étiquettes de taille plus adaptée
- Annotation manuelle des textes serbes originaux
- Sélection et évaluation de plusieurs étiqueteurs
- Étiquetage du corpus

Jeu d'étiquettes proposé

- 45 étiquettes
  - Partie du discours principale et la sous-catégorie
  - Degré de comparaison pour les ADJ et les ADV

- Comparable avec les jeux d'étiquettes envisagés pour les deux autres volets du corpus :
  - Anglais : PennTreebank = 36 étiquettes
  - Français : TreeTagger = 33 étiquettes

Quelques exemples

	Étiquette	Interprétation	Exemple
1.	NOM:com	Nom commun	*kuća* 'maison
2.	NOM:nam	Nom propre	*Beograd* 'Belgrade'
3.	NOM:col	Nom collectif	*pilad* 'poussins'
4.	NOM:num	Nom numéral	*dvojica* 'deux personnes'
5.	ADJ	Adjectif qualificatif au positif	*lep* 'beau'
6.	ADJ:comp	Adjectif qualificatif au comparatif	*lepši* 'plus beau'
7.	ADJ:sup	Adjectif qualificatif au superlatif	*najlepši* 'le plus beau'
8.	ADJ:dem	Adjectif démonstratif	*ovaj* 'ce'

Annotation manuelle

Corpus sélectionné :

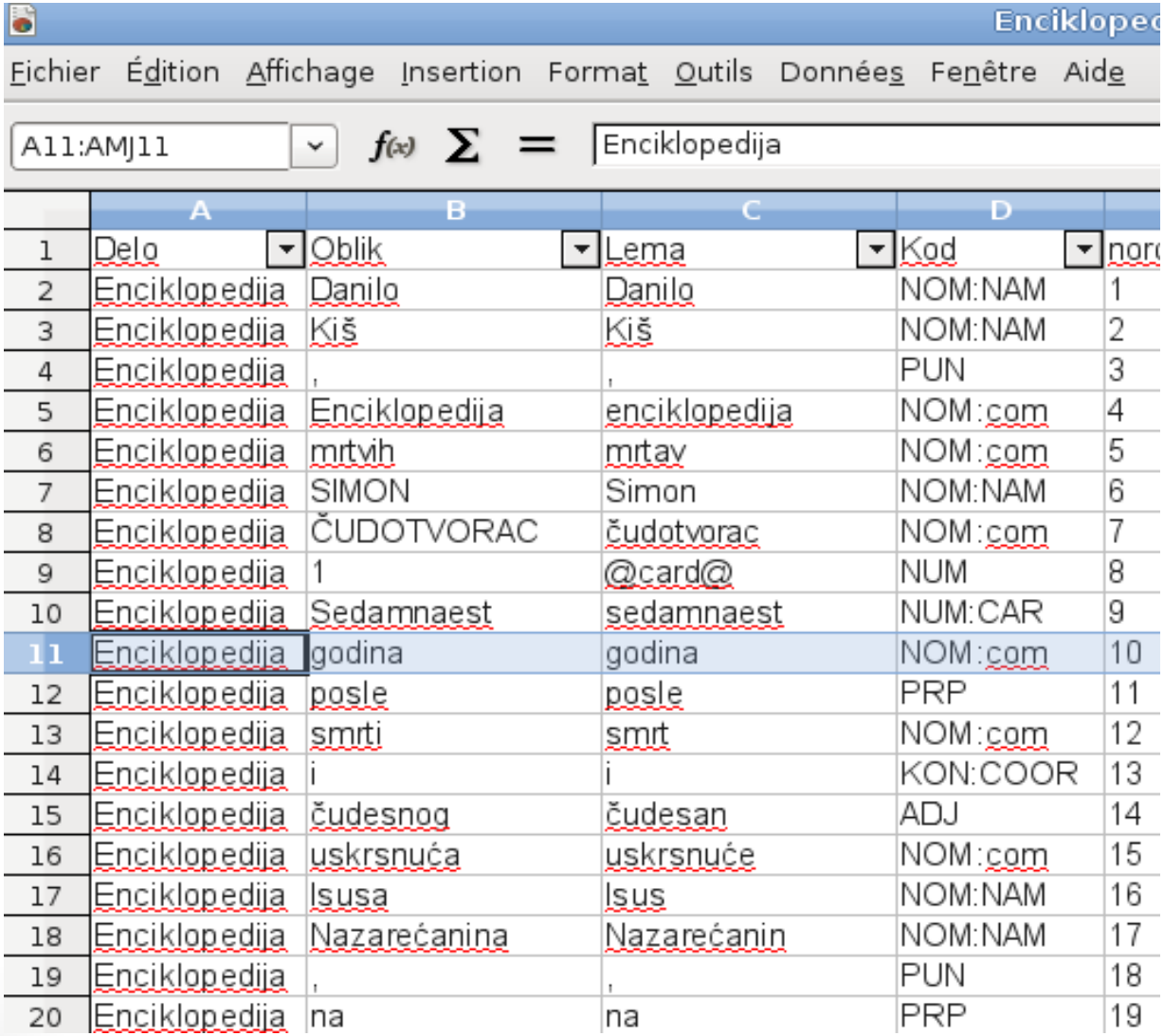
100 000 tokens

- Kiš, Danilo, *Enciklopedija mrtvih,* CD « Danilo Kiš : Sabrana dela ». Ed. française : *Encyclopédie des morts,* Gallimard, 1985.
- Stevanović, Vidosav, *Testament*, SKZ. Beograd, 1986. Ed. française : *Prélude à la guerre,* Mercure de France, 1996.

Principes d'annotation :

- Un token – une étiquette
- Désambiguïsation sur base de critères syntaxiques
- Guide d'annotation

Annotation manuelle



Enciklopec

Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

A11:AMJ11 | Enciklopedija

	A	B	C	D	
1	Delo	Oblik	Lema	Kod	nor
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
12	Enciklopedija	posle	posle	PRP	11
13	Enciklopedija	smrti	smrt	NOM:com	12
14	Enciklopedija	i	i	KON:COOR	13
15	Enciklopedija	čudesnog	čudesan	ADJ	14
16	Enciklopedija	uskrsnuća	uskrsnuće	NOM:com	15
17	Enciklopedija	Isusa	Isus	NOM:NAM	16
18	Enciklopedija	Nazarećanina	Nazarećanin	NOM:NAM	17
19	Enciklopedija	,	,	PUN	18
20	Enciklopedija	na	na	PRP	19

Étiqueteurs sélectionnés

- (Popović, 2010): TnT (Brants, 2000).
  - Précision sur le serbe : 85,47%
  - > 900 étiquettes
- (Utvić, 2011): TreeTagger (Schmidt, 1994)
  - Précision sur le serbe : 96,6%
  - 16 étiquettes
- (Gesmundo et Samardžić, 2012): BTagger (idem)
  - Précision sur le serbe : 86,65%
  - > 900 étiquettes

Évaluation des étiqueteurs

- Validation croisée avec 4 itérations :

Portion du corpus	# de tokens	Eval 1	Eval 2	Eval 3	Eval 4
Enciklopedija-1	**23 980**	Test	Entrain.	Entrain.	Entrain.
Enciklopedija-2	**23 885**	Entrain.	Test	Entrain.	Entrain.
Testament-1	**23 908**	Entrain.	Entrain.	Test	Entrain.
Testament-2	**23 884**	Entrain.	Entrain.	Entrain.	Test

Résultats 1/3

Tagger	Eval 1	Eval 2	Eval 3	Eval 4	Moyenne
TnT	*92,43%*	93,17%	*93,20%*	93,07%	**92,97%**
TreeTagger	*91,43%*	92,10%	*92,61%*	92,47%	**92,15%**
Btagger	*93,93%*	*94,48%*	94,22%	94,07%	**94,17%**

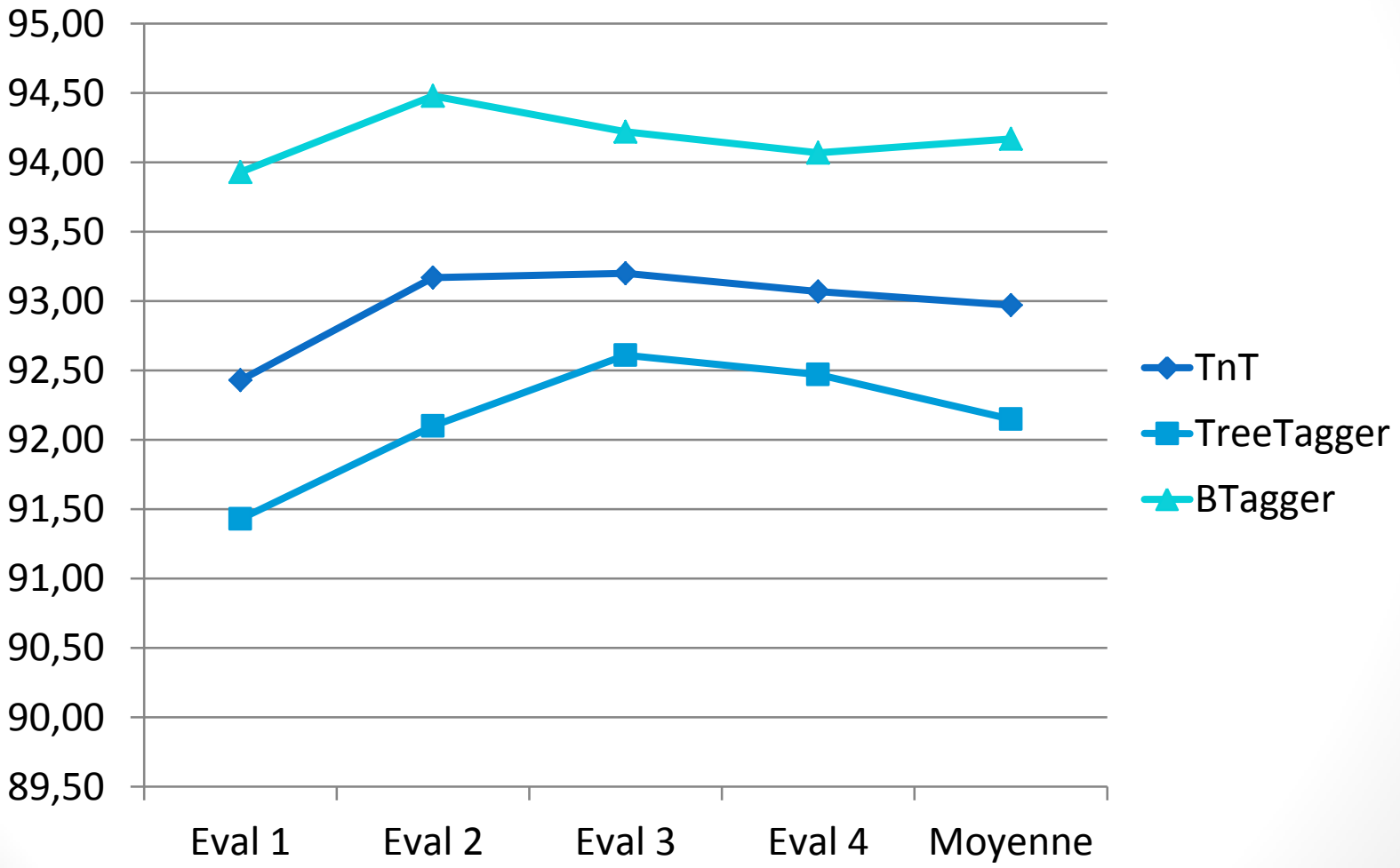
Résultats 2/3

- (Popović, 2010): TnT (Brants, 2000).
  - Précision sur le serbe : **85,47%**
  - > 900 étiquettes
- (Utvić, 2011): TreeTagger (Schmidt, 1994)
  - Précision sur le serbe : **96,6%**
  - 16 étiquettes
- (Gesmundo et Samardžić, 2012): BTagger (idem)
  - Précision sur le serbe : **86,65%**
  - > 900 étiquettes

Résultats 2/3

- (Popović, 2010): TnT (Brants, 2000).
  - Précision sur le serbe : **85,47%** **92,97%**
  - > 900 étiquettes
- (Utvić, 2011): TreeTagger (Schmidt, 1994)
  - Précision sur le serbe : **96,6%** 92,15%
  - 16 étiquettes
- (Gesmundo et Samardžić, 2012): BTagger (idem)
  - Précision sur le serbe : **86,65%** **94,17%**
  - > 900 étiquettes

Résultats 3/3



Annotation automatique de sous-corpus *Bašta*

57 000 tokens

- Source: Kiš, Danilo, *Bašta, Pepeo,* CD « Danilo Kiš : Sabrana dela ». Ed. française : *Jardin, cendre*. Gallimard 1971.

- Correction manuelle
- Précision : 96,3%

- Ajouté au corpus d'entraînement initial
- Taille finale du corpus annoté : 157 000 tokens

Analyse qualitative de l'annotation de *Bašta*

Partie du discours	% d'erreur totale	Annoté comme
Adjectif	22,7%	Nom commun
		Verbe principal
		Adverbe
Nom commun	16,3%	Adjectif
		Verbe principal
Verbe principal	13,5%	Adjectif
		Nom commun
		Verbe auxiliaire

Bilan

- Précision d'étiquetage améliorée en utilisant un jeu d'étiquettes mieux adapté à la taille du corpus
- Choix satisfaisant des étiquettes
- Comparaison des performances de 3 étiqueteurs sur le serbe : BTagger obtient des meilleurs résultats que TnT et TreeTagger

ANNOTATION SYNTAXIQUE D'UN CORPUS LITTÉRAIRE SERBE

Doctorat en Sciences du Langage (2014 à présent)

Directeurs :

Mme Cécile Fabre (UT2J, CLLE-ERSS)

M. Dejan Stosic (UT2J, CLLE-ERSS)

Objectifs et méthode

Objectif : création d'une banque d'arbres syntaxiques pour le serbe

Méthode :

- Préparation des ressources nécessaires
  - Lexique, jeu d'étiquettes, corpus d'entraînement
- Évaluation quantitative et qualitative de plusieurs parsers :
  - Talismane (Urieli, 2013)
  - MaltParser (Nivre, 2006)
  - MSTParser (McDonald, 2006)
- Annotation de la totalité du volet serbe
- Étude de cas : comparaison des réalisations d'un phénomène linguistique dans les 3 langues

Objectifs et méthode

Objectif : création d'une banque d'arbres syntaxiques pour le serbe

Méthode :

- **Préparation des ressources nécessaires**
  - Lexique, jeu d'étiquettes, corpus d'entraînement

- Évaluation quantitative et qualitative de plusieurs parsers :
  - Talismane (Urieli, 2013)
  - MaltParser (Nivre, 2006)
  - MSTParser (McDonald, 2006)

- Annotation de la totalité du volet serbe

- Étude de cas : comparaison des réalisations d'un phénomène linguistique dans les 3 langues

Ressources nécessaires

- Lexique
  - Informations morphosyntaxiques qui ont un rôle dans le fonctionnement syntaxique (cas, genre, nombre)
- Corpus d'entraînement
  - Lemmatisation, annotation MS, annotation syntaxique manuelle

Ressources existantes ?

Premières expériences : Jakovljevic et al. (2014)

- Corpus d'entraînement : 1148 phrases (7 117 tokens)
- Parser : MaltParser (Nivre, 2006)
- Précision :
  - LAS : 58%
  - UAS : 66%
- Jeu d'étiquettes : Prague Dependency Treebank (Hajic, 1988)
- Le corpus d'entraînement ni le modèle ne sont accessibles

Ressources existantes ?

Premières expériences : Jakovljevic et al. (2014)

- Corpus d'entraînement : 1148 phrases (7 117 tokens)
- Parser : MaltParser (Nivre, 2006)
- Précision :
  - LAS : 58%
  - UAS : 66%
- Jeu d'étiquettes : Prague Dependency Treebank (Hajic, 1988)
- Le corpus d'entraînement ni le modèle ne sont accessibles

⇒Constitution d'un lexique et d'un corpus d'entraînement

Lexique

Wiktionary : couverture solide, libre de droit, téléchargeable

Qualité du contenu ?

- Inspection manuelle : qualité satisfaisante
- La seule ressource disponible

Lexique

ParCoLex (titre de travail)

- 1,2M formes fléchies (126 000 lemmes)

Partie du discours	Propriétés MS dans le lexique
Verbe	temps/mode, personne, nombre, genre
Nom	genre, cas, nombre
Pronom	genre, cas, nombre
Adjectif	genre, cas, nombre, degré de comparaison, aspect
Adverbe	degré de comparaison (si applicable)

Format du lexique

trag	N_m_nom_sg_trag N_m_acc_sg_trag
traga	V_Present_3_sg_0_tragati N_m_gen_sg_trag
tragah	V_Imparfait_1_sg_0_tragati
tragahu	V_Imparfait_3_pl_0_tragati
tragaj	V_Imperatif_2_sg_0_tragati

Processus d'extraction

Deux éditions traitant des entrées en serbe :

- sh.wiktionary.org : version serbo – croate
  - 850 000 entrées
- sr.wiktionary.org : version serbe
  - 45 000 entrées

Dump utilisé : sh.wiktionary.com du 02/10/2015

Processus d'extraction

Wikicode :

- XML : macrostructure de la page
- Contenu de la page : wikicode
  - Format textuel
  - Syntaxe non formalisée
  - Variable entre éditions différentes
  - Variable à l'intérieur d'une édition

Processus d'extraction

Deux formats de page principaux :

```
==== Deklinacija ====
{{sh-imenica-
deklinacija2
|jezik|jezici
|jezika|jezika
|jeziku|jezicima
|jezik|jezike
|jeziče|jezici
|jeziku|jezicima
|jezikom|jezicima
}}
```

```
=== Flektirani oblici ===
'''guvernerskim'''

# instrumental množine ženskog roda pozitiva
određenog vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# lokativ množine ženskog roda pozitiva određenog
vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# dativ množine muškog roda pozitiva određenog
vida pridjeva
[[guvernerski#Srpskohrvatski|guvernerski]]
# instrumental množine muškog roda pozitiva
određenog vida pridjeva
```

Ressources complémentaires

Absence des mots invariables :

- Listes des prépositions créées manuellement (Stosic, 2001)
  - 107 prépositions
  - + information sur la rection
- Corpus étiqueté en parties du discours
  - 76 prépositions
  - 43 conjonctions
  - 33 interjections
  - 868 adverbes

Évaluation du lexique

- Sur le sous-corpus annoté morpho-syntaxiquement
- 157 000 tokens, 28 980 formes fléchies uniques

seuil de fréquence	No de formes fléchies uniques	Présents dans le lexique	Couverture
1	28 980	20 808	71.80%
2	10 630	8 136	76.53%
5	2 946	2 328	79.02%
10	1 241	990	79.77%

Analyse d'ambiguïté

- 2,1 étiquettes par forme fléchie
- 60% de formes fléchies sont ambigües
- > 37 000 formes fléchies ont 10 ou plus étiquettes associées
- 5 formes fléchies avec 43 étiquettes

Analyse d'ambiguïté

- POS ambigu, lemme ambigu

Krilo : nom./acc. sg. du nom *krilo* 'genous' |n.sg. du part. passé du verbe *kriti* 'cacher'

- POS ambigu, lemme univoque

Blizu : préposition *blizu* 'dans la proximité de' | adverbe *blizu* 'dans la proximité'

- POS univoque, lemme ambigu

Vrata : gén.sg. du nom *vrat* 'cou' |nom./acc.pl. du nom *vrata* 'porte'

- POS et lemme univoques, traits MS ambigus

Jastucima : dat.pl. | instr.pl. | loc.pl. du nom *jastuk* 'oreiller'

Analyse d'ambiguïté

Type d'amiguité	No de formes fléchies	% de toutes les formes fléchies ambigues
POS et lemme ambigus	15 496	2,13%
POS ambigu, lemme univoque	303	0,04%
POS univoque, lemme ambigu	19 822	2,72%
POS et lemme univoques, traits MS ambigus	691 814	95,10%

Pistes

- Couverture solide, mais amélioration souhaitable :
  - Édition serbe du wiktionary ?
  - Sur la base du sous-corpus étiqueté :
    - Liste des formes fléchies du corpus absentes du lexique
    - Tri par fréquence
    - Ajout manuel
- Soumission à LREC2016
- Mise en ligne sur REDAC

Élaboration du corpus d'entraînement

- Jeu d'étiquettes
- Guide d'annotation
- Annotation manuelle

Jeux existants pour les langues slaves

Langue	Corpus	Jeu d'étiquettes
Tchèque	Prague Dependency Treebank (PDT)	Sgall (28 étiq.)
Croate	HOBS (Croatian Dependency Treebank)	PDT (28 étiq.)
	SETimes.hr	SETimes.hr (15 étiq.)
Slovène	Slovene Dependency Treebank	PDT (28 étiq.)
	JOS Corpus	JOS Corpus (10 étiq.)
Russe	SynTagRus (Russian National Corpus)	Mel'cuk (78 étiq.)

Jeux existants pour les langues slaves

Langue	Corpus	Jeu d'étiquettes
Tchèque	Prague Dependency Treebank (PDT)	Sgall (28 étiq.)
Croate	**HOBS (Croatian Dependency Treebank)**	**PDT (28 étiq.)**
	SETimes.hr	SETimes.hr (15 étiq.)
Slovène	**Slovene Dependency Treebank**	**PDT (28 étiq.)**
	JOS Corpus/Fida+	JOS Corpus (10 étiq.)
Russe	SynTagRus (Russian National Corpus)	Mel'cuk (78 étiq.)

Jeux existants pour les langues slaves

Langue	Corpus	Jeu d'étiquettes
Tchèque	Prague Dependency Treebank (PDT)	Sgall (28 étiq.)
Croate	HOBS (Croatian Dependency Treebank)	PDT (28 étiq.)
	SETimes.hr	**SETimes.hr (15 étiq.)**
Slovène	Slovene Dependency Treebank	PDT (28 étiq.)
	JOS Corpus/Fida+	**JOS Corpus (10 étiq.)**
Russe	SynTagRus (Russian National Corpus)	Mel'cuk (78 étiq.)

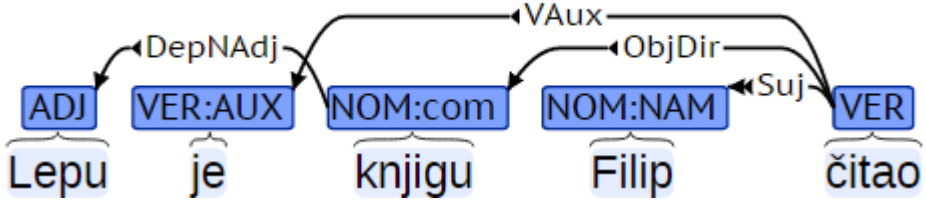
Jeux existants pour les langues slaves

Langue	Corpus	Jeu d'étiquettes
Tchèque	Prague Dependency Treebank (PDT)	Sgall (28 étiq.)
Croate	HOBS (Croatian Dependency Treebank)	PDT (28 étiq.)
	SETimes.hr	**SETimes.hr (15 étiq.)**
Slovène	Slovene Dependency Treebank	PDT (28 étiq.)
	JOS Corpus/Fida+	**JOS Corpus (10 étiq.)**
Russe	SynTagRus (Russian National Corpus)	Mel'cuk (78 étiq.)

⇒Définir un nouveau jeu pensé pour le serbe

Jeu d'étiquettes proposé

- 28 étiquettes (cf. l'article TASLA 2015)
- Critères formels
  - Type du gouverneur
  - Type du dépendant
  - Comportement combinatoire
- Etiquettes en 2 ou 3 parties
- Obj + [Dir | Indir | Prep]
- Dep + [N | V | Adj] + [Cas | Prep | Adv | Adj]



Choix des étiquettes

Deux influences dans le choix des étiquettes :

Comparabilité des résultats :

- Avec Talismane
- Avec les travaux en croate

Nature parallèle du corpus :

- Traitements uniformes
- Accessible aux deux communautés

Choix méthodologiques

Confusion terminologique entre l'attribut et *atribut* :

- Attribut du sujet et de l'objet ~ 3 fonctions de *predikativ*

Choix méthodologiques

Confusion terminologique entre l'attribut et *atribut* :

- Attribut du sujet et de l'objet ~ 3 fonctions de *predikativ*
- *Atribut* ~ épithète + complément du nom

Attributs du sujet et de l'objet direct

3 fonctions correspondantes dépendamment du verbe :

- prédicatif nominal : *être*
- prédicatif complémentaire : verbes essentiellement attributifs
- prédicatif optionnel : verbes optionnellement attributifs

- Pas de distinction formelle de surface
- Pas d'indication du lien avec le sujet ou avec l'objet

Attributs du sujet et de l'objet direct

3 fonctions correspondantes dépendamment du verbe :

- prédicatif nominal : *être*
- prédicatif complémentaire : verbes essentiellement attributifs
- prédicatif optionnel : verbes optionnellement attributifs

- Pas de distinction formelle de surface
- Pas d'indication du lien avec le sujet ou avec l'objet

⇒Introduction des étiquettes AttrSuj et AttrObj

Épithète et complément du nom

- Adjectif accordé avec le nom

lep-a	*knjiga*
beau-NOM.SG.F	livre[NOM.SG]
beau livre	

- Nom fléchi

ljubav	*majk-e*
amour[NOM.SG.]	mère-GEN.SG
amour maternel	

- Syntagme prépositionnel

kolač	*sa*	*višnj-ama*
gâteau[NOM.SG.]	avec	cerise-INSTR.PL
gâteau aux cerises		

Épithète et complément du nom

- Adjectif accordé avec le nom — DepNAdj

lep-a	*knjiga*
beau-NOM.SG.F	livre[NOM.SG]
beau livre	

- Nom fléchi — DepNCas

ljubav	*majk-e*
amour[NOM.SG.]	mère-GEN.SG
amour maternel	

- Syntagme prépositionnel — DepNPrep

kolač	*sa*	*višnj-ama*
gâteau[NOM.SG.]	avec	cerise-INSTR.PL
gâteau aux cerises		

Est-ce justifié ?

Retour de TASLA2015 :

- Abandonner la tradition slave pour se rapprocher de la tradition française ?
- Résultats de Talismane : les rapprochements faits remettent en question l'objectivité des résultats sur le serbe
- Chercheurs souhaitant faire des études contrastives doivent être en mesure de maîtriser les deux terminologies

Demande de retour des collègues serbes – en attente…

- Réflexion…

Guide d’annotation

Réflexion sur le jeu d’étiquettes …

Annotation manuelle

Piste d'accélération :

Agic et al. (2013) : modèle de MSTParser pour le croate utilisé sur le serbe et le croate

	SETimes		Wikipedia	
	Croate	Serbe	Croate	Serbe
LAS				
UAS				

Annotation manuelle

Piste d'accélération :

Agic et al. (2013) : modèle de MSTParser pour le croate utilisé sur le serbe et le croate

	SETimes		Wikipedia	
	Croate	Serbe	Croate	Serbe
LAS	76,7%	75,4%	71,9%	72,4%
UAS	81,6%	80,6%	80,0%	80,6%

Diff. max : 1,3%

Annotation manuelle

Piste d'accélération :

Agic et al. (2013) : modèle de MSTParser pour le croate utilisé sur le serbe et le croate

	SETimes		Wikipedia	
	Croate	Serbe	Croate	Serbe
LAS	76,7%	75,4%	71,9%	**72,4%**
UAS	81,6%	80,6%	80,0%	**80,6%**

Diff. max : 1,3%

Annotation manuelle

Utilisation envisagée :

- Pré-annotation avec le modèle pour le croate
- Correction manuelle

Problèmes :

- Étiquetage MS du modèle croate diffère de celui de notre corpus d'entraînement
- Jeu d'étiquettes syntaxiques différent : post-traitement (conversion vers notre jeu d'étiquettes)

Jeu d'étiquettes proposé

Comparé au jeu de SETimes.hr :

- 28 étiquettes vs. 15 étiquettes
- Augmentation de granularité : possibilité de conversion automatique

SETimes.hr	ParCoLab
Sb (sujet)	Suj
	SujLog
Obj (objet)	ObjDir
	ObjIndir
	ObjPrep
Atr (modifieur nominal)	DepNAdj
	DepNCas
	DepNPrep

La suite

- Lexique
  - Augmentation de la couverture
  - Version au format MULTEXT-East
  - Mise en ligne
- Jeu d'étiquettes
  - …
- Guide d'annotation
  - Élaboration de la grille d'annotation
  - Évaluation de l'accord inter-annotateurs
- Corpus d'entraînement
  - Tester la piste d'accélération de l'annotation manuelle
  - Entamer l'annotation

Et encore…

- Mise en ligne du corpus d'entraînement pour l'étiquetage
- Étiquetage de la totalité du volet serbe
- Lemmatisation

- Évaluation des parsers sur le serbe
- Intégration des règles dans Talismane

MERCI POUR VOTRE ATTENTION !

Références

AGIĆ, Ž., MERKLER, D., BEROVIĆ, D. (2013). Parsing Croatian and Serbian by Using Croatian Dependency Treebanks. pp. 22-33.

BRANTS, T. (2000). TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing* (pp. 224-231). Association for Computational Linguistics.

GESMUNDO, A., & SAMARDŽIĆ, T. (2012). Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification. In *LREC* (pp. 2103-2106).

HAJIČ, J. (1998). Building a syntactically annotated corpus: The prague dependency treebank. *Issues of valency and meaning*, 106-132.

JAKOVLJEVIĆ, B., KOVAČEVIĆ, A., SEČUJSKI M., & MARKOVIĆ, M. (2014). A Dependency Treebank for Serbian: Initial Experiments. In *Speech and Computer* (pp. 42-49). Springer International Publishing.

KRSTEV C., VITAS, D., & ERJAVEC, T. (2004). MULTEXT-East resources for Serbian. In *Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004*. Erjavec, Tomaž and Zganec Gros, Jerneja.

KRSTEV C., VITAS, D., STANKOVIĆ, R., OBRADOVIĆ, I., PAVLOVIĆ-LAŽETIĆ, G. (2004), Combining Heterogeneous Lexical Resources. *4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbonne, Portugal. pp.1103--1106, 4th International Conference on Language Resources and Evaluation (LREC'04)

Références

KRSTEV, C., & VITAS, D. (2005, July). Corpus and Lexicon-Mutual Incompletness. In *Proceedings of the Corpus Linguistics Conference, Birmingham* (pp. 14-17).

MCDONALD, R., LEMRAN, K., PEREIRA, F. (2006). *Multilingual Dependency Parsing with a Two-Stage Discriminative Parser*.

MILETIC, A., FABRE, C., STOSIC, D. (2015). Construction du jeu d'étiquettes pour le parsing du serbe, Actes de TASLA à TALN 2015, 1-9.

NIVRE, J., HALL, J., NILSSON, J. (2006). *MaltParser A Data-Driven Parser-Generator for Dependency Parsing.*

POPOVIĆ, Z. (2010). Taggers applied on texts in Serbian. In *Proceedings of the INFOtheca '10 Conference*.

SCHMID, H. (1994, September). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (Vol. 12, pp. 44-49).

SEČUJSKI , M. (2009). *Automatic part-of-speech tagging of texts in the Serbian language,* Novi Sad, Serbia: Faculty of Technical Sciences.

STOSIC, D. (2001). Le rôle des préfixes dans l'expression des relations spatiales. Eléments d'analyse à partir des données du serbo-croate et du français. *Cahiers de Grammaire 26*, p. 207-228.

URIELI, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit* (Doctoral dissertation, Université Toulouse le Mirail-Toulouse II).

UTVIĆ, M. (2011, December). Annotating the Corpus of contemporary Serbian. In *Proceedings of the INFOtheca '12 Conference*.