TALC-sef

A Manually-Revised Pos-TAgged Literary Corpus

in Serbian, English and French

***Antonio Balvet, **Dejan Stosic, *Aleksandra Miletic**
*Université Lille Nord de France
F-59000 Lille, France
UdL3, STL, F-59653 Villeneuve d'Ascq, France
CNRS, UMR 8163
Domaine Universitaire du Pont de Bois, 59653 Villeneuve d'Ascq
**Université de Toulouse II-Le Mirail
CLLE-ERSS (UMR 5263)
5, allées Antonio Machado
F-31058 Toulouse Cedex 9
Email: antonio.balvet@univ-lille3.fr, dstosic@univ-tlse2.fr, aleksandramiletic1207@gmail.com

Abstract

In this paper, we present a parallel literary corpus for Serbian, English and French, the TALC-sef corpus. The corpus includes a manually-revised pos-tagged reference Serbian corpus of over 150,000 words. The initial objective was to devise a reference parallel corpus in the three languages, both for literary and linguistic studies. The French and English sub-corpora had been pos-tagged from the onset, using the Treetagger (Schmid, 1994), but the corpus lacked, until now, a tagged version of the Serbian sub-corpus. Here, we present the original parallel literary corpus, then we address issues related to pos-tagging a large collection of Serbian text: from the conception of an appropriate tagset for Serbian, to the choice of an automatic pos-tagger adapted to the task, and then to some quantitative results. We then move on to a discussion of perspectives in the near future for further annotations of the whole parallel corpus.

Keywords: Serbian, pos-tagged corpus, parallel corpus

1. Introduction

In this paper, we present a parallel literary corpus for Serbian, English and French, the TALC-sef corpus, which includes a manually-revised pos-tagged reference Serbian corpus of over 150,000 words. This corpus was created in the framework of two projects involving researchers from Lille 3 University, Artois University and of the University of Belgrade, during the years 2007-2009 and 2010-2011. The initial objective of these projects was to devise a reference parallel corpus in the three languages, both for literary and linguistic studies. From the onset, the French and English sub-corpora had been pos-tagged using the TreeTagger (Schmid, 1994), but the corpus lacked, until now, a tagged version of the Serbian sub-corpus.

Here, we present the original parallel literary corpus, then we address issues related to pos-tagging a large collection of Serbian text: from the conception of an appropriate tagset for Serbian, to the choice of an automatic pos-tagger adapted to the task, and then to some quantitative results. We then move on to a discussion of perspectives in the near future for further annotations of the whole parallel corpus, namely a dependency annotation in the three languages.

2. Setting up a parallel literary corpus in three European languages

In 2007-2009 and 2010-2011, D. Stosic (Artois University) has led two successive projects, aiming at the constitution of a parallel corpus of literary texts in French, Serbian and English, with the participation of Lille 3 University and the University of Belgrade. The corpus contained original works in the three languages, as well as professional translations in the different languages (see Table 1 below). The main objective of this project was to set up a parallel reference corpus for literary and linguistic studies. Hence, much attention has been paid to ensure the quality of the corpus, as well as its usability: the whole corpus has been automatically aligned at the sentence level with manual corrections[1], the whole corpus will be available for research purposes via a web-based concordancing program. The table below summarizes the main quantitative elements of this corpus.

[1] Xalign, an alignment software distributed by INRIA and available at http://led.loria.fr/download/source/Xalign.zip, was used in this process. This software is based on (Church and Gale, 1993), and comes integrated into the Unitex text annotation platform (http://www-igm.univ-mlv.fr/~unitex/) which was used for preprocessing the different corpora.

	French	Serbian	English
French	300,105	332,521	353,934
Serbian	316,210	388,326	-
English	45,457	156,074	148,486
TOTAL	**661,772**	**876,921**	**502,420**

Table 1: Number of tokens of original and translated works in French, Serbian and English, in the TALC-sef corpus

As can be seen in the table above, in the present state of the TALC-sef corpus, French ↔ Serbian translations were favoured, while Serbian → English translations are not yet available. The Serbian sub-corpus alone represents 388,326 words. This is over three orders of magnitude greater than the translated version of G. Orwell's *1984* novel (104,286 words), the cesAna corpus[2] from the MULTEXT-EAST project, which, to this date, was the sole freely-available gold standard corpus for Serbian.[3] Out of these 380,000 words, we devised a manually-revised pos-tagged corpus of over 150,000 words.

The list below gives an overview of the literary works comprising the TALC-sef corpus, categorized by original language:

- Serbian: *Putnica* (B. Blagojević), *Rani jadi* (D. Kiš), *Enciklopedija mrtvih* (D. Kiš), *Grobnica za Borisa Davidoviča* (D. Kiš), *Iskupljenje* (B. Šćepanović), *Ljudi govore* (R. Petrović), *Testament* (V. Stevanović);
- French: *Les Dieux ont soif* (A. France), *Le père Goriot* (H. De Balzac), *Notre-Dame de Paris* (V. Hugo);
- English: *The Last of the Mohicans* (J. F. Cooper).

In the present version of the TALC-sef corpus, each of the Serbian novels was translated into French. French novels have all been translated into Serbian and English, while the sole English novel was translated into both French and Serbian.

3. POS-tagging Serbian: issues and solutions

Since the release of reference corpora such as the Brown corpus (Francis & Kučera, 1964) and the Penn Treebank (Marcus et al., 1993) for English, or the French Treebank (Abeillé, 2003) for French, the definition of a set of reference part-of-speech tags, or tagsets, for the annotation of large volumes of text can be considered as a settled matter for those languages. However, the same cannot be said for Serbian. This is due to two main reasons. Firstly, Serbian is under-resourced when it comes to electronic linguistic resources: as mentioned above, to this date, the sole freely accessible annotated corpus of Serbian texts was cesAna,[4] and the first tagger developed specifically for slavic languages, BTagger, was distributed in 2012 (Gesmundo and Samardžić, 2012). Reference taggers and tagsets for English and French date back to the 1990's.[5] Moreover, when pos-tagging English texts, precision scores routinely reach 97% (Shen et al., 2007), while tests conducted on Serbian remain well below the 96% standard (Gesmundo and Samardžić 2012; Popović, 2010).

Secondly, Serbian is a South Slavic language with rich inflectional morphology. It distinguishes three persons, two numbers and seven cases. Declension marks apply to nouns, adjectives, and pronouns, as well as some of the cardinal numbers. Nouns can have up to 12 different inflected forms (as opposed to 4 in French, and generally 2 in English), adjectives up to 36. Moreover, depending on tense, mood, person and gender, verbs can have more than 120 inflected forms. For this reason, Gesmundo and Samardžić (2012) use very large tagsets: over 900 different tags for Serbian, as opposed to 36 tags for English (Penn Treebank tagset) or 33 tags for French[6] (French Treebank tagset). To make matters worse, from the automatic tagging point-of-view, word order in Serbian is much less rigid than in French or English: even if the typical word order is SVO, it has numerous and frequent variations, as noted in (Stanojčić and Popović, 2011). This implies that surface ambiguity is very common, more so than in English and French, because of the vast number of inflected forms and a less constrained constituent order.

3.1. The TALC Serbian tagset: a compromise between precision and coverage

To our knowledge, two different corpora have been used in experiments on Serbian POS tagging to date. Popović (2010), and Gesmundo and Samardžić (2012) used the cesAna corpus cited above. Popović used this corpus to test 5 different taggers, among which TnT (Brants 2000) obtained the best average precision: an 85.47% score is reported. Gesmundo and Samardzic used the corpus to test their tagger, BTagger, developed for highly-inflectional languages such as Serbian. However, even this tagger performed relatively poorly on the given corpus: the obtained average precision was 86.65%.

Utvić (2011) devised a corpus of 1 million words annotated with 16 tags encoding only the main parts of speech. Using TreeTagger, precision scores of up to 96.57% were reported, which advocates in favor of a

[2] See (Krsteva *et al.*, 2004), (Erjavec, 2004) and (Ide and Véronis, 1994).

[3] Full metadata and updated versions are available at http://nl.ijs.si/ME/V4/doc/teiHeaders/ana/oana-en-teiHeader.html. This corpus is annotated at the linguistic (lemma, pos-tag) level for: Bulgarian, Czech, English, Estonian, Hungarian, Macedonian, Persian, Polish, Romanian, Serbian, Slovak, and Slovene.

[4] A 104,286 words translation of *1984* (G. Orwell), see (Krsteva et al., 2004) for more details for each language.

[5] For a reference tagset for French, see (Adda *et al.*, 1998) and (Valli and Véronis, 1999).

[6] The French Treebank distinguishes 15 base tags, which are generally expanded to 33 main category+subcategory tags (*e.g.* PRO, PRO:DEM, PRO:PER, PRO:POS, PRO:REL, for pronouns).

drastic reduction in morphosyntactic distinctions for our corpus.[7]

Given these results, we set upon the task of developing a new training corpus, tagged with a new, more balanced tagset than what was present in the cesAna corpus. This tagset was meant to be comparable in size and design to those used for the English and French sub-corpora[8]: in order to keep the three sub-corpora at the same level of granularity, it was necessary to prioritize consistency over precision in our semi-manual annotation process of the Serbian sub-corpus. Our tagset therefore contains 45 different tags encoding the main parts of speech and sub-categorization indications, as well as some morphological features for adjectives and adverbs. This was rendered necessary both for practical and theoretical reasons: one of the intended applications of the TALC-sef corpus is literary and linguistic comparison, based on a sentence-aligned as well as syntactic constituent-aligned corpus in three European languages. It was thus necessary to harmonize the tagsets for each sub-corpus.

The tagset we propose was used to manually annotate a sub-corpus REF1 of 101,000 tokens, which was used as a training corpus in the tests performed with three candidate taggers (see below).[9] After the evaluation presented below, we used the models trained on REF1 as a bootstrap for the semi-manual annotation of a larger reference corpus, REF2 of 150,000 tokens (REF1 + 50,000 new tokens). REF2 therefore integrates REF1 and 50,000 new tokens, which were manually revised as well. REF2 was not used in the evaluations summarized below.

3.2. Selecting a tagger adapted to the task

Following (Popović, 2010), (Utvić, 2011) and (Gesmundo and Samardžić, 2012), we selected the following popular pos-taggers for our own comparative study, in order to identify which tagger and possibly which tagging algorithm was best suited to the task: TreeTagger (Schmid, 1994) and TnT (Brants, 2000), as a baseline, and BTagger (Gesmundo and Samardžić, 2012). In the quantitative results presented below, we used an adapted version of the *n-fold* evaluation procedure, where *n*=4. Due to the size of our REF1 corpus (101,000 words total), were we to follow a standard *10-fold* evaluation procedure, we would have computed precision scores on very small amounts of text,[10] which could have biased the overall results. This adaptation was also rendered necessary due to BTagger's processing times for setting up a stable model[11], and for actually pos-tagging the test corpus.[12] On average, the training corpus represented 71,683 tokens, while the test corpus contained 23,896 tokens.[13] We adapted each sub-corpus in order to maintain sentential integrity, as in future experiments, we plan to test sentence-aware taggers, such as MBTagger (Daelmans et al., 1996). Sentences were nonetheless randomly shuffled in each sub-corpora so as to prevent any text structure-related bias.

3.3. Quantitative and qualitative evaluations

3.3.1. Tagging precision rates with three popular taggers

This section presents the main outcomes of our experiments on pos-tagging the Serbian sub-corpus. Precision scores are presented in Figure 1 below, for a *4-fold* cross validation using REF1 (101,000 words) as a reference corpus.

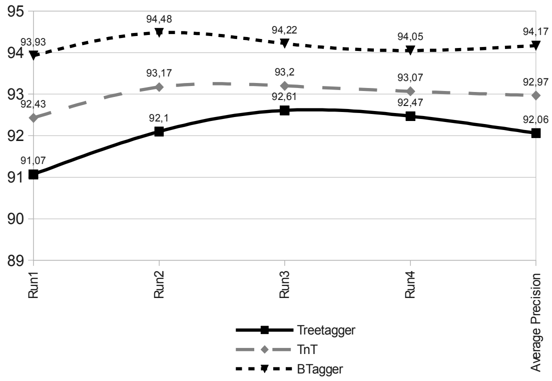

Figure 1: Average pos-tagging precision scores for Serbian

As can be seen in Figure 1, in each of the four test runs (Runs 1 to 4), BTagger performed consistently higher than the other taggers, with an average 94.17% precision. TreeTagger scored consistently below both other taggers, with an average 92.15% precision, whereas TnT's performance lies in-between those two extremes. It is worth noting that all taggers were used with "out-of-the-box" settings, which could have favoured BTagger somewhat.

The pos-tagging models for each tagger, as well as complimentary resources (tagging lexica, ngrams etc.) generated during the training phases are available for download at http://code.google.com/p/tagged-literary-corpus/.

Based on these results, BTagger was selected and used to automatically tag a new corpus of 56,093 tokens, which was checked manually as stated above. The size of this new reference corpus REF2 (REF1 + 56,093 new tokens) is therefore 157,678 tokens. A web-based searching and

[7] A more detailed account of annotation choices, as well as a quantitative and qualitative analysis of each tagger's performance is available in (Miletic, 2013).

[8] The default TreeTagger tagsets for these languages.

[9] Due to the scarcity of freely-available resources and qualified Serbian annotators, the annotation presented here relies exclusively on the work presented in (Miletic, 2013), under the supervision of D. Stosic and A. Balvet, as senior researchers.

[10] 101,000/10 = 10,100 words for each run.

[11] Over 1h30 for each run on a standard computer.

[12] Over 45 min. for each run.

[13] As the results we obtained for the corpus and tagset described here compare favorably with usual precision scores for other languages, a standard *10-fold* evaluation procedure will be set up in the near future.

concordancing interface is being setup in order to provide web access to the TALC-sef corpus, either from a monolingual or a multilingual (parallel) perspective.[14]

3.3.2. Qualitative evaluation

As mentioned above, the manually-revised portion of the Serbian sub-corpus represents 157,678 tokens, while the overall Serbian sub-corpus amounts to a total of 876,921 tokens (original + translated texts). In the version presented here, the Serbian sub-corpus is thus far from being fully manually-revised. A summary of the main tagging errors end-users will be likely to encounter in the fully-automatically tagged portion of this corpus appears thus necessary. Therefore, alongside the figures presented above, we outline in this subsection the main elements of a qualitative analysis of BTagger's main tagging errors, conducted on a sample of *Bašta, Pepeo* (D. Kiš).

Our census of BTagger's tagging errors shows that Adjectives (22.7%), Nouns (16.3%) and Verbs (13.5%) make up for the vast majority of tagging errors. Homograph words which are regularly ambiguous between two categories, like *teško*, were expected to affect precision scores: in context, this token might be considered either as a form of the Adjective *težak* (hard), or as the manner Adverb *teško* (hard).[15] Contrary to our expectations, Adjectives were mostly confused with Nouns and Verbs, while Adjective/Adverb confusions remained marginal. As for Nouns, they were mostly confused first with Adjectives, then with Verbs, probably because of BTagger's faulty detection of Adjective-like suffixes in pure non-deadjectival Nouns. For example, *čvorove* is a plural accusative form of the Noun *čvor* (knot), but the suffix *-ove* is also typical of possessive Adjectives derived from proper nouns, like *Petrove* (Petar's). Pure non-deverbal Nouns were also confused with Verbs: *saksije* is a nominative plural form of Noun *saksija* (flowerpot) and is not a homograph of any Verb, but the *-ije* suffix seems to have triggered the detection of a third person singular form of present tense Verbs like *bije* (he hits), *pije* (he drinks), *krije* (he hides) etc. As for Verbs, they were mostly confused with Adjectives, Nouns, and auxiliaries. Verb/Adjective confusions were mainly due to the homography between past participles and Adjectives, while Verb/Noun confusions were mostly due to ambiguous declensions: for example, *sinu*, a third person singular aorist of the verb *sinuti* (to shine) was confused with the dative singular form of Noun *sin* (son). Finally, most of the Verb/Auxiliary errors involve the word *jesam*, a ubiquitous element as it can either be an auxiliary verb or an attributive (copula) verb. This particular word poses a challenge to automatic pos-taggers, as the disambiguation process typically should rely on the whole sentence structure, and not just a fixed amount of context[16], as is generally the case for pos-taggers. An auxiliary verb hypothesis for *jesam* is only valid if the word is found within the scope of a main verb past participle form. As word-order in Serbian is somewhat less rigid than in languages such as English or French, the distance between an auxiliary verb and its corresponding past participle may be quite unpredictable, from a pos-tagger's point-of-view. The citation below gives an example of just such a configuration, where *sam* (present tense of *jesam* 'to be') was considered an auxiliary while it is really one of the sentence's main verbs.[17]

Zaboravljam da <u>sam</u> novorođenče i da od svih životnih senzacija, ljudskih i božanskih, najviše ako mogu da <u>osetim</u> i <u>doživim</u> scenski efekat sunca.

I forget that I <u>am</u> but a newborn baby and that, of all of life's sensations, human or divine, I can but <u>feel</u> and <u>live</u> the scenic effect of the sun.

(*Bašta, Pepeo*, D. Kiš)

Citation 1: example of a main verb/auxiliary verb confusion

Of all the tagging errors examined above, the Verb/Auxiliary confusion is maybe the most problematic, as it entails a considerable amount of manually revision of the Serbian sub-corpus. No easy correction strategy seems to be applicable, as the correct disambiguation between those two classes requires some level of syntactic analysis (*e.g.* chunk or dependency parsing at the very least), in order to determine whether all propositions come complete with a main verb, or just a faulty participle-less auxiliary verb.

4. Discussion and perspectives

In its present version, the TALC-sef corpus outlined here comprises an original 380,000 words automatically tagged Serbian sub-corpus of literary works, of which 157,678 words were manually-revised. The total (Serbian + translations into Serbian) 876,921 words Serbian sub-corpus was also automatically tagged, with a tagset aiming at a compromise between precision and coverage.

The TALC-sef corpus also contains two sub-corpora in French (661,772 words) and English (502,420 words), with original as well as translated texts, which were tagged without manual revision, with the baseline TreeTagger models for French and English.

The TALC-sef corpus is, to our knowledge, among the most extensive available sources of pos-tagged Serbian texts. This corpus is therefore of utmost interest for literary and linguistic studies, as well as for computational linguistics, offering a wide range of possible applications, from comparative corpus-based studies in syntax and morphosyntax to machine translation and Serbian NLP. It will be made accessible by way of online concordancing

[14] For reasons of copyright, no direct access to the complete original and translated texts can be provided in the version presented here.

[15] In English, a similar Adjective/Adverb homography exists in sentences like "Tagging words is hard" *vs.* "He works hard".

[16] Typically 3-grams.

[17] Underlined words are main verbs.

services in the near future.[18] In its present version, the corpus allows for parallel concordances such as the one presented in Table 2 (full-text search).

Serbian segments: sentences	*Aligned English translation*
Dok je govorio psu, gledao mu je pravo u oči i pas ga je razumeo.	This time he looked the dog straight in the eye while talking to him, and the dog understood.
Zavrteo je repom i zacvileo, nakrivivši glavu.	He wagged his tail, cocking his head and whimpering.

Table 2: Examples of aligned sentences in the TALC-sef corpus (*Rani jadi*, D. Kiš)

Serbian segments: chunks and post-tagged words	*Chunk-aligned English translations*

{Dok,dok,KON:SUB}
{je,jesam,VER}
{govorio,govoriti,VER}
{psu,pas,NOM}
	{while,while.IN}
	{talking,talk.VBG} {to,to.TO}
	{him,him.PP}

{gledao,gledati,VER}
{mu,on,PRO:PER}
{je,jesam,VER}
{pravo,pravo,ADV}
{u,u,PRP} {oči,oko,NOM}
	{This,this.DT} {time,time.NN}
	{he,he.PP} {looked,look.VBD}
	{the,the.DT}
	{dog,dog.NN}
	{straight,straight.RB} {in,in.IN}
	{the,the.DT}
	{eye,eye.NN}

{i,i,KON:COOR}
{pas,pas,NOM}
{ga,on,PRO:PER}
{je,jesam,VER}
{razumeo,razumeti,VER}
	{and,and.CC}
	{the,the.DT}
	{dog,dog.NN}
	{understood,understand.VBD}

{Zavrteo,zavrteti,VER}
{je,jesam,VER}
{repom,rep,NOM}
	{He,he.PP} {wagged,wag.VBD}
	{his,his.PP$}
	{tail,tail.NN}

{i,i,KON:COOR}
{zacvileo,zacvileti,VER}
	{and,and.CC}
	{whimpering,whimpering.NN}

{nakrivivši,nakriviti,VER}
{glavu,glava,NOM}
	{cocking,cock.VBG}
	{his,his.PP$} {head,head.NN}

Table 3: Examples of chunk-aligned segments in Serbian and English

In future versions, we plan to enhance the overall quality of the corpus primarily by adding lemmata to the Serbian sub-corpus. We also plan to retag the French and English sub-corpora using state-of-the-art pos-taggers and models for French[19] and English. Finally, as for corpora-related tasks, a major endeavor will have to be undertaken, which is rendered necessary by the philosophy of the project: adding consistent Serbian → English translations.

From the standpoint of semi-automatic corpus processing and annotation, we are in the process of providing chunk parses to the existing annotated corpora in Serbian, English and French. Adding chunk annotations on top of pos-tags and sentence delimiters will allow for more in-depth comparative studies in three European languages, representing three distinct linguistic types. Adding chunk parses to each sentence in the corpus will allow for (surface) constituent-alignments rather than mere sentence-alignments and comparisons, as is shown in Table 3. As can be seen from the table, sentence alignments and chunk alignments might exhibit some degree of variation: in the sentences above, as could be expected in human translation, constituents have been reordered between the Serbian text and its English translation. A chunk-aligned version of the corpus will allow for finer-grained queries than the sentence-aligned one.

Finally, in the near future, we plan to provide dependency parses, as well on top of chunk segments, for the three languages of the corpus, in order to offer Head/Dependent-aware search features, so as to overcome discrepancies in constituent orders and argument realizations in the different languages.

5. References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French, *in* A. Abeillé (ed.) *Treebanks*, Kluwer, Dordrecht, 2003

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

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

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system, *COMPLEX'94*, Budapest.

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)*.

Daelemans, W., Zavrel, J., Berck, P., and Gillis, S. (1996) MBT: A Memory-Based Part of Speech Tagger-Generator, *Fourth ACL Workshop on Very Large Corpora (1996)*, 14-27.

[18] A good candidate for this platform would be the IMS Corpus Workbench (Christ, 1994). Other approaches are nonetheless under evaluation, such as graph databases.

[19] The French sub-corpus would benefit from new advances in pos-tagging for French: Malttager (Hall, 2003), LGTagger (Constant and Sigogne, 2011) and (Denis and Sagot, 2009).

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.

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

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, Brown University, Providence, Rhode Island.

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

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.

Hall, J., (2003). A Probabilistic Part-of Speech Tagger with Suffix Probabilities. *MSI report 03015*. Växjö University: School of Mathematics and Systems Engineering

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

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

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*, Master's Thesis Dissertation, Université Charles de Gaulle Lille 3.

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

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

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

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

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

Valli, A., Véronis, J. (1999). Étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2), 113-133.