ParCoTrain : Training and Test Corpus for POS-tagging and Lemmatization of Serbian

1. Short description

ParCoTrain is a training and test corpus for POS tagging and lemmatization of Serbian. The corpus was developed as part of the ParCoLab project (http://parcolab.univ-tlse2.fr/), with the goal of creating NLP resources for Serbian that are freely accessible. The lemmatized part of the corpus contains 95 585 annotated tokens. The annotation was fully manual. The POS-tagged part of the corpus contains a total of 153 625 tokens, with 95 585 manually annotated tokens, and 57 977 tokens annotated automatically and then validated manually. The source texts for the corpus are contemporary Serbian novels from the second half of the 20th century. The POS-tagging used in the corpus indicates the main POS and the sub-category. Degree of comparison is also indicated for adjectives and adverbs.

2. Corpus structure

The following source texts were used to create the training and test corpus:

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

The corpus itself contains 8 files in a csv format with tab as the field separator. The character encoding used is UTF-8. The EOL character is CR-LF (\r\n). The list of the files and their content is given in Table 1.

File	Size (in tokens)	Content	Annotation level	Annotation type	Format
enciklopedija-testament.txt	95 585	“Enciklopedija mrtvih” (47 792 tokens) “Testament” (47 793 tokens)	POS lemmatization	Manual	[token][lemma][POS]
enciklopedija.txt	47 792	“Enciklopedija mrtvih”	POS lemmatization	Manual	[token][lemma][POS]
testament.txt	47 793	“Testament”	POS lemmatization	Manual	[token][lemma][POS]
enciklopedija-sample1.txt	23 908	“Enciklopedija mrtvih”	POS lemmatization	Manual	[token][lemma][POS]
enciklopedija-sample2.txt	23 885	“Enciklopedija mrtvih”	POS lemmatization	Manual	[token][lemma][POS]
testament-sample1.txt	23 908	“Testament”	POS lemmatization	Manual	[token][lemma][POS]
testament-sample2.txt	23 884	“Testament”	POS lemmatization	Manual	[token][lemma][POS]
basta.txt	57 977	“Bašta, pepeo”	POS	Automatic, manually validated	[token][POS]

Table 1: Corpus structure

enciklopedija-testament.txt is a concatenation of *enciklopedija.txt* and *testament.txt*.

enciklopedija-sample1.txt, enciklopedija-sample2.txt, testament-sample1.txt and *testament-sample2.txt* are balanced samples derived from *enciklopedija.txt* and *testament.txt*. Each of the two big files was split into 2 samples in a way that allowed to have samples as close in size as possible while maintaining sentence integrity. These files can be used for a 4-fold cross-validation, as was done in (Balvet et al., 2014). A standard cross-validation with 10 iterations may not be optimal, given the total size of the corpus.

basta.txt contains only POS tagging. This file was automatically tagged with BTagger (Gesmundo & Samardzic, 2012) trained on the 4 sample files. The output of the tagger was subsequently manually validated. The file was then used to extend the training corpus for POS-tagging and get a total of 153 625 POS-tagged tokens. This can be done as follows:

1. create a file containing only the POS annotation from *enciklopedija-testament.txt* using the following command on the command line:

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

2. concatenate this file with *basta.txt*:

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

3. POS- tagging implemented in the corpus

The tagset used for the POS-tagging of the training corpus contains 46 tags. The tags indicate the main POS and the subcategory, as well as the degree of comparison for adjectives and adverbs. All tags are given in Table 1, along with some examples of usage.

Table 2: POS tagset

Tag	Subcategory	Example
NOM:com	common noun	*vrata* 'door', *kuća* 'house', *strast* 'passion'
NOM:col	collective noun	*lišće* 'leaves', *pilad* 'chicks', *kamenje* 'stones'
NOM:nam	proper noun	*Duško, Beograd* 'Belgrade', *Afrika* 'Africa'
NOM:num	numeral noun	*dvojica* 'two men', *trojica* 'three men'
NOM:approx	approximate noun	*desetak* 'about ten', *pedesetak* 'about fifty'
VER	main verb	*jedem* 'I eat', *radio* 'worked'
VER:aux	auxiliary verb	*sam* 'I am', *ćete* 'you will'
PRO:per	personal pronoun	*ja* 'I', *mene* 'me', *ti* 'you' (sg.), *vi* 'you' (pl.)
PRO:intr	interrogative pronoun	*ko* 'who', *šta* 'what'
PRO:dem	demonstrative pronoun	*ovaj* 'this' (m.sg.), *ona* 'that' (f.sg.), *ti* 'those' (m.pl.)
PRO:ind	indefinite pronoun	*neko* 'somebody', *niko* 'nobody', *svako* 'everybody'
PRO:pos	possessive pronoun	*moj* 'mine', *naši* 'ours', *njihovi* 'theirs'
PRO:rel	relative pronoun	*koji* 'who/which/that'
PRO:ref	reflexive pronoun	*sebe, se* 'self'
PRO:num	numeral pronoun	*jedan* 'one', *drugi* 'other', 'another'
ADJ	adjective in positive	*nov* 'new', *lepa* 'beautiful'
ADJ:comp	adjective in comparative	*noviji* 'newer', *lepša* 'more beautiful'
ADJ:sup	adjective in superlative	*najnoviji* 'newest', *najlepša* 'the most beautiful'
ADJ:intr	interrogative adjective	*which* 'lequel', *kakav* 'what' (adj.), *koliki* 'of what size'
ADJ:dem	demonstrative adjective	*ovaj* 'this', *ona* 'that', *ti* 'those'
ADJ:ind	indefinite adjective	*neki* 'some', *nijedan* 'no', *svaki* 'every'
ADJ:pos	possessive adjective	*moj* 'my', *naši* 'our', *njihov*i 'their'
ADJ:rel	relative adjective	*čiji* 'whose', *kakav* 'what' (adj.), *koliki* 'of the size'
NUM:car	cardinal number	*jedan* 'one' (m.), *jedna* 'one' (f.), *dvadeset* 'twenty'
NUM:ord	ordinal number	*prvi* 'first', *druga* 'second', *dvadeseti* 'twentieth'
NUM:col	collective number	*dvoje* 'two people', *petoro* 'five people', *dvadesetoro* 'twenty people'
ADV	adverb (other than relative, interrogative or indefinite)	*pametno* 'intelligently', *nespretno* 'clumsily'
ADV:comp	adverb in comparative	*bolje* 'better', *pametnije* 'smarter'
ADV:sup	adverb in superlative	*najbolje* 'best', *najpametnije* 'smartest'
ADV:intr	interrogative adverb	*kako* 'how', *gde* 'where', *kad* 'when'
ADV:rel	relative adverb	*kako* 'how', *gde* 'where', *kad* 'when'
ADV:ind	indefinite adverb	*nekako* 'in some way', *igde* 'anywhere'
CONJ:coor	coordination conjunction	*i* 'and', *ali* 'but', *ili* 'or'
CONJ:sub	subordination conjunction	*da* 'that', *jer* 'because', *iako* 'although'
PREP	preposition	*na* 'on', *pod* 'under', *u* 'in'
PAR	particle	*da* 'yes', *ne* 'no', *čak* 'even'
INT	interjection	*ah* 'ah', *apćiha* 'achoo', *hej* 'hey'
SENT	strong punctuation	. ! ?
PONC	weak punctuation	, ; : ( )
PONC:cit	quotation marks	« » " „ “
STR	foreign word	chéri
ABR	abbreviation	dr, itd. (etc.)
LET	letter	A, p, L
NUM	litteral number	12, 252, XII
PAGE	page number	7, 10
ID	page number indicator	@@

Remarks concerning some specific cases

The choices made in the design of this tagset are presented in detail in (Miletic, 2013). The most important ones are addressed below.

3.1 Remarks on some of the annotation principles

Collective nouns are distinguished for their specific agreement properties: these nouns in Serbian follow the declension patterns of singular nouns, but their semantics is plural. This allows them to impose either plural or singular form onto the verbal phrase.

Numeral nouns typically designate a group of male humans. However, they follow the declension patterns of singular feminine nouns, so they are able to impose both singular feminine and masculine plural forms onto the adjectives attached to them.

Approximate nouns indicate an imprecise quantity and are derived from cardinal numbers.

Traditionally, Serbian grammar considers as pronouns both forms like *neko* 'somebody', *niko* 'nobody', *svako* 'everybody' and forms such as *neki* 'some', *nijedan* 'no', *svaki* 'every'. The two series do not display the same syntactic behavior: the forms of the first series function independently from nouns and are referred to as *nominal pronouns*. The forms from the second one are bound to nouns and are called *adjectival pronouns*. In other words, the first series is equivalent of the subclasses of indefinite, possessive and demonstrative pronouns, and the other one corresponds to the same subclasses of determiners. In the corpus, only the forms from the first series are treated as pronouns, whereas the forms from the second one are annotated as adjectives (since Serbian does not have determiners, the "adjective" tag was adopted). Forms that can work in both ways (cf. possessive forms such as *moj* 'my/mine', *tvoj* 'your(s)' or demonstrative forms such as *ovaj* 'this (one)', *taj* 'that (one)') are annotated depending on the context: if they appear independently of a noun, they are tagged as pronouns, and if they appear with a noun they are treated as adjectives (compare *Više mi se sviđa tvoj nego moj* 'I prefer yours to mine' to *Više mi se sviđa tvoj automobil nego moj motor* 'I prefer your car to my motorcycle').

Both *nominal* and *adjectival* pronouns in Serbian are traditionally subdivided into several subcategories such as possessive, interrogative, relative, demonstrative, indefinite, general and negative pronouns. While maintaining the first 3 subcategories, the tagset merges indefinite (*neko* 'somebody', *neki* 'some'), general (*svako* 'everybody', *svaki* 'every') and negative forms (*niko* 'nobody', *nijedan* 'no') into one subclass of indefinite for both pronouns and adjectives.

4. References

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.