Initial Steps in Building Serbian Treebank: Morphological Annotation

Bojana Đorđević

Azbukum – Centre for Serbian Language and Culture
bojana@lingvistika.org

Abstract. Serbian has a well developed resource for morphosyntactic text analysis, namely the Morphological Electronic Dictionary of Serbian (MEDS), as well as the Contemporary Serbian Language Corpus (around 122 million words), but it is completely lacking resources for syntax analysis. We have decided that one of the important steps in that direction is building the Serbian treebank, which would enable us to eventually induce a rich formal grammar of Serbian to be used for parsing of Serbian texts. Currently, our focus is on the basic level of sentence annotation – morphological annotation. On this level, each word is described by two pieces of information – the lemma of the word, together with its inherent morphological, stylistic and semantic values, and the realized values of the lemma used in the text. In this paper, we will be comparing the morphological annotation used in MEDS with the morphological annotation of the so far largest Slavic treebank – Prague Dependency Treebank (PDT). Each of the systems has their own annotation scheme and a varied set of tags which do not always match, especially when it comes to a variety of semantic tags. After the comparison, preliminary decisions will be made on the tags and the annotation system to be used in the future Serbian treebank.

Keywords: Serbian Treebank, Morphological Annotation, Morphosyntactic Description, Prague Dependency Treebank, Morphological Electronic Dictionary of Serbian

1 Introduction

Treebanks are usually defined as corpora coded with syntactic information. Apart from the essential syntactic layer, each of them is always coded on the word level. Word-level coding implies tagging or annotation of part of speech information (POS), enriched with morphosyntactic, and sometimes derivational and semantic information.

In this paper, we will be focusing on the morphosyntactic level, more precisely, on the morphosyntactic description (MSD) of Serbian, and validate it against the MSD of the Prague Dependency Treebank (PDT). The reason for choosing PDT as the standard owes to it being the oldest Slavic treebank (version 1.0 was released in late nineties), the fact that it is the largest Slavic treebank

(around 90,000 sentences), constantly growing and being improved, and also that it served as a benchmark for some other treebanks in the Slavic world (Croatian and Russian). Moreover, it is an MSD created specifically for a Slavic language and for the purpose of a treebank.

Nowadays, when deciding to make a treebank, there are several choices when it comes to choosing an MSD. Traditional MSDs are usually not appropriate for usage in NLP, so languages that do not already have a formal MSD can either make a new one, or accept and adapt one of the existing standards. Even though there is considerable effort to unify MSDs across languages ([15]), every now and then a choice is made to create or pursue a unique MSD that suits the purpose of a specific language, and that is already in use by applications in the given language. Such is the case of Serbian and the description it offers within the Morhological Electronic Dictionary of Serbian (MEDS).

Broadly speaking, MSDs can be of either positional or attributive type. Even though there is no data that prove that one or the other are more appropriate for use in treebanks, treebank creators often opt for the positional one. Positional tags are characterized by having the same predefined length (either across POS or within POS) and positions for each of the categories. Their categories and values are represented by single letters or numbers which can vary in meaning depending on their position. On the other hand, attributive tags do not have a predefined tag length, strict positions of categories and are more descriptive. They are also typically more user friendly, in that they can be understandable without having to consult guidelines, unlike positional tags.

Of Slavic MSDs, the Czech ([9]), Croatian ([18]), Russian ([16]), Bulgarian ([17]), Slovene ([19]) and Slovak ([7]) ones have positional tags, while the Polish ([14]) one is attributive. The Czech annotation system was adapted for Croatian and Russian, the MULTEXT-East standard for Bulgarian and Slovene, while the Slovak one represents a unique positional MSD.

2 Serbian MSDs

Serbian is a South Slavic language with a rich system of inflection. Its noun categories include case (Nominative, Genitive, Dative, Accusative, Vocative, Instrumental, Locative), gender (masculine, feminine or neutral) and number (singular, plural and remainders of paucal). Apart from case, gender and number, adjectives are additionally characterized by three grades of comparison (positive, comparative and superlative), which they share with some adverbs, and adjectival aspect (definite and indefinite). Pronouns are described by case, number, gender and person (first, second and third). Verb categories include person, number, gender in certain forms, tense (present, past, future I and II, aorist, imperfect, pluperfect), mood (imperative, potential, future II), verbal aspect (perfective and imperfective) and voice (active and passive). Serbian has a series of irregular phenomena, such as discrepancies in natural and grammatical gender and number of nouns.

MSD for Serbian has so far been covered by two major projects – one at the Faculty of Mathematics at the University of Belgrade, Serbia, within which a comprehensive electronic morphological dictionary of Serbian (MEDS) was made ([12]). That dictionary covers both simple words (130,000) and multi-word units (11,000).

The other one is the MULTEXT-East project (MTE), whose result was a common tagset for a number of East European languages (currently 16 – Bulgarian, Croatian, Czech, Estonian, English, Hungarian, Romanian, Serbian, Slovene, Resian dialect of Slovene, Macedonian, Persian, Polish, Russian, Slovak, and Ukrainian), which included Serbian from version three ([4]). MEDS can be transformed into MTE format with some loss of information ([8]) and is already used to tag Orwell's 1984.

However, it was argued by [5] that representation of MTE is not the most suitable for Serbian due to a lack of certain values and categories, interdependence of values and categories that is hard to grasp using a positional system and inconsistencies in types of categories. Similar comments come from ([14]). We will not be looking into features of MTE in more detail here.

3 MEDS

MEDS contains a thorough and unambiguous representation of Serbian morphosyntax, originally developed within the corpus processing system Intex and currently used within Unitex ([13]). This dictionary is in LADL/DELA format ([2]), based on the methodology of finite state automata. Such dictionaries, originally created for French, exist for several Slavic languages (Russian, Bulgarian, Polish and Macedonian), and several European languages (English, German, Spanish, Greek and Portuguese).

MEDS contains two sets of dictionaries: dictionaries of simple words DELAS/DELAF and dictionaries of multi-word units DELAC/DELACF. DELAS and DELAC dictionary entries present lemmas together with finite state automata that describe their inflection paradigm. Additional morphological, syntactic, semantic and derivational information related to the lemma is also added. The second set of dictionaries, DELAF and DELACF contains word forms and word-form properties, together with DELAS/DELAC information. These entries are the ones that could potentially be used to annotate the Serbian treebank so they will be presented in more detail in the following sections.

3.1 Representation of Morphosyntactic Information in DELAF

Each DELAF entry has the following structure:

form,lemma.K[+SynSem][:categories]**

It contains two sets of information – lemma and its description, and word form and its MSD. In this formula, *form* stands for word form, and *lemma* is where

the lemma of the word form is stated. If the word form matches the lemma, only the word form is present in the entry. *K* stands for POS. There are 10 POS in DELAF – N (noun), V (verb), A (adjective), ADV (adverb), PRO (pronoun), NUM (numeral), PREP (preposition), PAR (particle), CONJ (conjunction), INT (interjection).

SynSem stands for syntactic and semantic properties of the lemma in the broadest sense of the word. SynSem labels are descriptive and their meaning is most of the time intuitive. There is no defined order between them. This part specifies:

- **Word subcategories** (collective and proper for nouns, possessive for adjectives, spatial for adverbs, auxiliary for verbs, personal for pronouns etc.).
- **Semantic tags** are numerous and their number is constantly growing. Some of them are Top (toponym), Hum (human), Hyd (hydro), Bot (botanical) etc..
- **Derivational information** is sometimes also given – as Der (derived) alone, or specifying the type of derivation – DerArisatiIrati for example, which states that a verb can take both *–arisati* and *–irati* suffixes.
- **Syntactic information** exists for nouns, verbs and prepositions. For nouns, labels are added when there is a difference between the natural and grammatical gender (MG/FG/NG – natural male/female/neutral gender). Pluralia and singularia tantum are also labeled. For verbs, aspect information is given (perfective and imperfective), potential reflexivness (reflexive and irreflexive), transitivity information (transitive and intransitive), which all state the potential of the lemma, and not the actual realization in the text. Information that the verb is an auxiliary one is also added. The required case is added for prepositions.
- **Dialect variant** – Tag Ek (ekavian) is added when a word can be different in the iekavian variant. Cr (Croatism)is added for words preferred in Croatian.
- **Original transcription** of proper names of foreign origin is given in the following form:
  Dejli,. N+NProp+Hum+Last+EN+Val=Dailey+Val=
  Daily+Val=Daley+Val=Daly
- **Negation** – tag Neg is added to negative forms, whether verbal or nominal.

Categories stand for morphosyntactic categories of the word form. Each category is represented by a single letter or number with a specific meaning regardless of its position within the entry. Although some are more or less intuitive (s and p for numbers, m, f and n for genders), there are some that cannot be interpreted without guidelines.

The MSD part contains the following categories and features:

- **Number** (singular, plural, paucal)
- **Gender** (male, female, neutral)
- **Case** (nominative, genitive, dative, accusative, vocative, instrumental and locative)

- **Animacy** (animate, inanimate, no consequence)
- **Grade of comparison** (positive, comparative, superlative)
- **Person** (first, second, third)
- **Tense** (present, aorist, imperfect, future)
- **Verb form** (infinitive, imperative, active past participle, passive past participle, present gerund, past gerund)
- **Clitic** (positive or negative)
- **Definiteness** (definite, indefinite)

Some examples of DELAF entries are given below:

braće,braća.N+Hum+MG+Pl:fs2v : noun, human, natural male gender, plural; feminine, singular, genitive, animate
bavio,baviti.V+Imperf+It+Ref:Gsm : verb, imperfect, intransitive, reflexive; active past participle, singular, masculine
planinski,.A+PosQ:adms1g : adjective, possessive; positive, definite, masculine, singular, nominative, no consequence
ga,on.PRO+PrsJG:msz2i : pronoun, personal, 3rd person singular; masculine, singular, 3rd person, genitive, positive
jedan,.NUM+v1:ms1g : numeral, number 1; masculine, singular, nominative, no consequence
koliko,.ADV+Amm+Quant+Pro+Wh : adverb, amount, quantity, pronoun, question word
kod,.PREP+p2 : preposition, takes genitive
se,.PAR : particle
mada,.CONJ : conjunction
ih,.INT : interjection

Two things can be noted here. Even though paucal forms match genitive singular in modern Serbian, they prove to be distinct in some pronouns ([20]) so their inclusion in MEDS is justified.

Tenses and moods such as past tense, future II and potential are composite tenses in Serbian so there are no tags bearing their names. They are extracted and assigned a composite tag using local grammars.

3.2 MWUs in MEDS

MWUs are collected into two LADL/DELA dictionaries – DELAC, which gives base forms of MWUs, and DELACF, which is a dictionary of MWU forms, created automatically from DELAC. MWUs in DELACF are defined as words with separators (blanks and other non-alphabetical characters), which can each inflect following their own rules, do not allow inserts and represent a concept different from the one of its constituents ([12]).

MEDS covers named entities: proper names (persons, organizations, locations), time expressions and numerical expressions. It also covers fixed expressions and combinations of prepositions and nouns that often collocate in texts.

Their representation in DELACF is described with a formula such as the one in DELAF, with an addition of tag +Comp or +C in the SynSem part, which stands for 'compound':

form,lemma.K+C|Comp[+SynSem][:categories]**

Its POS is mostly defined by the head of the MWU, but sometimes also by its use in the text, which is the case with some nominal and prepositional constructions. The *Categories* part of the tag lists the features of the head in the MWU. The remaining words in the MWU are either assumed to agree in those categories or do not have inflection at all (as with prepositional MWUs). Some examples of MWUs are shown in Table 1.

Table 1: Some MWUs in DELACF

celu noć,.ADV+C+AN+Temp (all night, A+N –> ADV)
godišnjem dobu,godišnje doba.N+Comp:ns3q (season, A+N –> N)
u blizini,.ADV+C (in vicinity, PREP+N –> ADV)
u blizini,.PREP+C+p2 (in vicinity, PREP+N –> PREP)

4 PDT Morphological Annotation

The Prague Dependency Treebank (PDT) morphological description was created by researchers at Charles University in Prague and Masaryk University in Brno, as part of the three-layer annotation designed for PDT ([3]). The current version of PDT is 3.0 and even though it introduced novelties regarding MWUs starting from version 2.5 ([1]), there are no significant differences in annotation on the morphological level (m-level).

The PDT tag set contains about 4,000 tags. 2 million tokens were annotated on the morphological level in version 3.0.

In PDT, every word is unambiguously described by two pieces of information – the lemma and the tag, which are given as separate XML tags. The lemma part has more of a descriptive structure, often intended for the human reader, and carries information about terms, style and homonymy, derivational information, comments on derivation, reference to other lemmas, category and additional comments. The tag itself is positional and has 15 fixed positions. Of those, 2 are reserved for the future use and the remaining 13 positions are:

- **Part of speech** – Czech, as well as Serbian, recognizes 10 POS: N (noun), V (verb), A (adjective), D (adverb), C (numeral), P (pronoun), R (preposition), J (conjunction), I (interjection) and T (particle). Two more categories are added here – X for undetermined, unknown or unclassifiable forms and Z for punctuation.

- **Detailed Part of Speech** very broadly corresponds to a combination of the SynSem section in the Serbian MEDS and its MSD. Some of the examples are – types of nouns (general), pronouns (possessive, demonstrative, clitical form), adjectives (possessive, nominal etc.), verbs (imperative, infinitive, active past participle etc.), numerals (Roman grapheme, generic, cardinal etc.)
- **Gender** – even though Czech distinguishes three genders – male, female and neutral, there are 10 different values present in this category. Special tags are assigned to male animate and male inanimate nouns, while there is another male gender tag for either animate and inanimate nouns, used for past participles. Other tags stand for either a pair of categories (feminine or neutral, feminine or masculine), or for a negation of a category (not feminine).
- **Number** – apart from singular and plural, in some cases Czech clearly distinguishes dual. Apart from those three, there is another cumulative category here (singular for feminine, plural for neutral).
- **Case** – nominative, genitive, dative, accusative, vocative, locative and instrumental.
- **Possessor's Gender** – a category of importance for possessive pronouns and adjectives. Values are feminine, masculine animate and not feminine.
- **Possessor's Number** – plural, singular, any.
- **Person** – first, second, third or any.
- **Tense** – future, past or present, present, past, any.
- **Voice** – active, passive.
- **Degree of Comparison** – positive, comparative, superlative.
- **Negation** – affirmative, negated.
- **Variant** – archaic, colloquial, bookish, abbreviation etc.

Some examples of PDT lemma-tag pairs are given below:

Praha_;G, NNFS1-----A---- : geographical name; noun, feminine, singular, nominative, affirmative
docházet_:T, VB-S---3P-AA--- : imperfect verb; present or future, singular, 3rd person, present, affirmative, active
navrhovaný_^(*2t), AAFS4----1A---- : derivational information: remove two letters, add t; adjective, feminine, singular, accusative, positive, affirmative
se_^(zvr._zájmeno/částice), P7-X4---------- : reflexive pronoun/particle; pronoun, reflexive, any gender, accusative
dva‘2, ClHP1---------- : reference to number 2; cardinal numeral, feminine or neutral, plural, nominative
kladně_^(*1ý), Dg-------1A---- : derivational information: remove one letter, add ý; adverb with degrees of comparison, positive, affirmative
v-1, RR--6---------- : homonym marked as 1; preposition in, takes locative
asi, TT---------- : /; particle
a-1, J^---------- : homonym marked as 1; conjunction, coordinating

4.1 MWUs in Prague Dependency Treebank

MWUs were introduced in PDT in version 2.5 ([2]). It was already decided in version 2.0 that MWUs are to be annotated on a level other than morphological. The decision was made to tag each of the MWU segments according to their original meaning, for example: in *New York* – *new* is lemmatized as a foreign word, not as a geographical name, while *York* is lemmatized as a geographical name, even if it finds itself in the name of a newspaper such as *New York Times* ([10]). From version 2.5, PDT distinguishes personal names, names of institutions and objects, locations, addresses, time, bibliographical information, foreign expressions and numbers, but on the tectogrammatical rather than morphology level.

During the work on version 2.5, a lexicon of MWUs SemLEx was extracted as a freely available product of the work of annotators ([2]).

5 Comparative View

Comparing the two MSDs, the most obvious difference is the one of organization – PDT being positional and MEDS attributive. However, the internal organization of information and division of information between the lemma and tag are in many respects similar.

5.1 Lemma

In PDT, lemma is treated separately and a separate set of tags is attributed to it, although it does not have to contain any tags. In MEDS, lemma is also treated separately, while still being inside the tag itself, and is always assigned a set of labels. While POS is a part of the tag in PDT, in MEDS, it is a part of the lemma. In PDT, POS is stated withing the lemma only in cases of homonymy. Homonyms are in PDT often enumerated the same way as in traditional dictionaries and often contain a comment additionally specifying the word. This comment is meant to be used by the human reader. In MEDS however, homonyms are distinguished by their POS alone, and in case homonyms belong to the same POS, it is counted on the lemma and SynSem values to make their meaning clear.

A number of features specifying the lemma (the SynSem part of the MEDS tag) is in PDT also given within the lemma. This means that all the terms (proper names and terms) are defined there, just like in MEDS. SubPOS is also placed inside the lemma in MEDS though, while in PDT it is given within the tag.

Both systems add derivational information here, and both systems use reference to other lemmas (e.g. in case of numerals). Another feature these two systems have in common is placement of verbal aspect information within the lemma section. Its values are in both systems perfective, imperfective or both.

Some additional information is added here in MEDS – syntactic information such as reflexivity, transitivity, agreement (for prepositions) and natural

gender and number. Even though Czech has the same type of discrepancy between natural and grammatical genders, it does not seem to address this matter separately.

5.2 Morphosyntactic Description of the Word Form

Table 2: Presence of PDT categories in MEDS

PDT	MEDS
POS	within lemma
SubPOS	within lemma
Gender	✓
Number	✓
Case	✓
PossGender	x
PossNumber	x
Person	✓
Tense	✓
Grade	✓
Negation	✓
Voice	defined differently
Var	some

As already noted, POS and what corresponds to SubPOS are defined within the lemma in MEDS. This organizational difference does not make much of a practical one.

On the other hand, two categories that exist in PDT, namely PossGender – gender of the possessor and PossNumber – number of the possessor, categories relevant for possessive adjectives and pronouns, are not taken into account in MEDS. Even though the two may not be important from the morphosyntactic point of view, they are a useful piece of information for morphology, and even more useful for syntax in case of resolution of sentences such as:

Marko-j je dao Mariji-i njenu-i knjigu. (Marko-j gave Marija-i her-i book.)

Unlike Czech, Serbian does not include dual as a number category. Even though in past Serbian had special forms for only two entities, today those forms can be seen in only a small number of nouns and some of their forms (genitive plural), where they have a plural meaning.

There is no specific tag for voice in MEDS, rather, a form used for making the passive voice – passive past participle is labeled. There are two types of passive in Serbian, both composite forms, but only one can be potentially recognized

on this level – the one with the passive past participle. Label Active is never attributed as any other form other than that one is considered active.

Var (variant) in PDT defines the style of the word form, such as standard, bookish, colloquial, archaic etc. Abbreviations are also defined here. An occasional Arh (archaic) tag is used in MEDS, but style is otherwise not commented on. However, there are other variants defined in MEDS – dialectal and language variants. Thus, there are tags added for the ekavien pronunciation whenever there is a difference in the word form between the two pronunciations of Serbian (ekavian and jekavian). There are additional tags for Croatisms in this section as well.

5.3 Syntactic Information

While syntactic information is completely left out of this layer in PDT, some of it is included in MEDS. This includes agreement information for prepositions, added for the purposes of easier annotation of MWUs and other composite forms. Tags for auxiliary verbs are also added here.

Reflexiveness and transitivity tags are added for all the verbs, signifying the potential of the verb, rather than concrete use in the particular context. Still, this piece of information will be useful for future annotation of syntax.

Table 3: Irregular gender and number phenomena in Serbian

braće,braća.N+Hum+MG+Pl:fs2v
(brothers, gram.cat: female singular, nat.cat: male plural)
petoricu,petorica.N+NumN+MG+Pl:fs4v
(five men, gram.cat: female singular, nat.cat: male plural)
deca,.N+Hum+NG+Pl+Ek:fs1v
(children, gram.cat: female singular, nat.cat: neutral plural)
budalom,budala.N+Hum+MG+FG:fs6v
(fool, gram.cat: female singular, nat.cat: male of female singular)

Nouns with irregular behavior are assigned special tags – ST and PT for singularia and pluralia tantum respectively. MG (male gender) FG (female gender) and NG (neutral gender) are added for nouns of natural gender different than the grammatical one. Tag Pl (plural) is added to singular nouns whose meaning is plural, which is not clear from the form of the word. These features influence the agreement with the verb and give important input for the future analysis of syntax. Some of the examples of discrepancies are given in Table 3.

5.4 Treatment of MWUs

As noted in section 4.1, MWUs are not tagged on the morphological level in PDT, but on the level of linguistic meaning (tectogrammatical level), according

to the methodology of PDT. A lexicon SemLEx is extracted from version 2.5, which contains lemmas of MWUs together with their separate tags.

MEDS already contains DELACF, a dictionary of MWU forms, with a unique tag for each of the MWUs. This dictionary will without doubt serve great purpose in the process of MWU detection in subsequent stages of annotation.

6 Conclusions and Further Work

After comparing the PDT and MEDS morphosyntactic descriptions, we conclude that even though they are different in some respects - organization, use and placement of some categories, there are no significant differences between them. However, we thought we could give some recommendations for making the MEDS tags more suitable for tagging a treebank:

- It is recommended that tags for possessor's gender and number are added as they can be important for the future use.
- There is a number of tags specific for the treebank purpose than need to be added, such as tags for punctuation, foreign words and unrecognized or misspelled words.
- Additional variant features (archaic, colloquial, bookish) could eventually be added.
- Tags for MWUs should probably not be used at this point, but could be adapted into composite tags on a higher level of annotation.

Several immediate steps follow the choice of the MSD. First, the decision needs to be made on the format of the treebank, which should be in accordance with other projects being developed for Serbian, and at the same time readable and recognizable by the wider community.

Second, a portion of text needs to be tagged with help of MEDS. The initial training corpus can either be a subcorpus of the Serbian corpus ([11]) or a controlled test corpus assembled specifically for this purpose.

In that process, ideally, an interface would be made which would propose choices given by MEDS to the human annotator and let them choose and approve them. The annotator should be given an option of assembling a new tag, following certain restrictions. In the process of building the software and annotating, comprehensive guidelines can be created to guide the future annotation choices.

Following that, the choice of a tagger to be trained on the annotated corpus will be made.

References

1. Eduard Bejček, Jarmila Panevová, Jan Popelka, Pavel Straňák, Magda Ševčíková, Jan Štěpánek, and Zdeněk Žabokrtský. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India, 2012.

2. Eduard Bejček and Pavel Straňák. Annotation of Multiword Expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1-2):7–21, 2010.
3. Alena Böhmová, Jan Hajič, Eva Hajičová, and B Barbora Hladká. The Prague dependency treebank: Three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.
4. Blandine Courtois and Max Silberztein. Dictionnaires électroniques du français. In *Langue Française*, n°87, Larousse, Paris, France, 1990.
5. Ranka Stanković, Cvetana Krstev and Duško Vitas. A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. In Nicoletta Calyolari et al., editor, *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, Valetta, Malta, May 2010*, pages 816–819, 2010.
6. Tomaž Erjavec, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić, and Duško Vitas. The MULTEXT-east Morphosyntactic Specifications for Slavic Languages. In Tomaž Erjavec and Duško Vitas, editors, *MorphSlav '03: Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 25–32, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
7. Radovan Garabík and Mária Šimková. Slovak Morphosyntactic Tagset. *Journal of Language Modeling*, 0(1):41—63, 2012.
8. Cvetana Krstev and Duško Vitas and Tomaž Erjavec. Morpho-Syntactic Descriptions in MULTEXT-East – the Case of Serbian. *Informatica, The Slovene Society Informatika, Ljubljana*, 28:431–436, 2004.
9. Jan Hajič. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Charles University Press, Prague, Czech Republic, 2004.
10. Hana Jiři and Daniel Zeman. Manual for Morphological Annotation. Revision for the Prague Dependency Treebank 2.0, Technical Report No. 2005-27. Technical report, Institute of Formal and Applied Linguistics (ÚFAL MFF UK), Faculty of Mathematics and Physics, Charles University, 2005.
11. Cvetana Krstev and Duško Vitas. Corpus and Lexicon – Mutual Incompletness. In Pernilla Danielsson and Martijn Wagenmakers, editors, *Proceedings of the Corpus Linguistics Conference, 14–17 July 2005, Birmingham*, 2005.
12. Cvetana Krstev, Duško Vitas, and Gordana Pavlović-Lažetić. Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian. In Gerhild Zybatow et al., editor, *Formal Description of Slavic Languages*, pages 3–17, Frankfurt am Main, Germany, 2008. Peter Lang. The Fifth Conference, Leipzig 2003.
13. Sébastien Paumier. *Unitex 3.0 User Manual*, 2011. Available at `http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf`.
14. Adam Przepiórkowski and Marcin Woliński. A Flexemic Tagset For Polish. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages*. 10th Conference EACL 2003.
15. Alexandr Rosen. Morphological Tags in Parallel Corpora. In František Čermák, Patrick Corness, and Aleš Klégr, editors, *InterCorp: Exploring a Multilingual Corpus*, pages 205–234. NLN, Praha, 2010.
16. Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. Designing and evaluating a Russian Tagset. In *LREC'08*, 2008.
17. Kiril Simov, Petya Osenova, and Milena Slavcheva. BTB-TR03: BulTreeBank morphosyntactic tagset. BTB-TS version 2.0. Bultreebank project technical report. Technical report, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences, 2004.

18. Marko Tadić. Building the Croatian Dependency Treebank: the initial stages. *Suvremena lingvistika*, 63:85–92, 2007.
19. Dan Tufiş, Nancy Ide, and Tomaž Erjavec. Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages. In *Proceedings of First International Conference on Language Resources and Evaluation*, pages 233–240, Granada, Spain, 1998.
20. Živojin Stanojčić and Ljubomir Popović. *Gramatika srpskoga jezika. Udžbenik za I, II, III i IV razred srednje škole*. Zavod za udžbenike i nastavna sredstva, Beograd, 1997.