Morpho-Syntactic Descriptions in MULTEXT-East — the Case of Serbian

Cvetana Krstev
Faculty of Philology, University of Belgrade
Studentski trg 3, 11000 Begrade, Serbia and Montenegro
cvetana@matf.bg.ac.yu

Duško Vitas
Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Begrade, Serbia and Montenegro
vitas@matf.bg.ac.yu

Tomaž Erjavec
Department of Knowledge Technologies, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Keywords: natural language processing, language resources, Serbian language, multilinguality

Received: July 21, 2004

MULTEXT-East is a multilingual dataset for language engineering research and development. This standardised and linked set of resources covers a large number of mainly Central and Eastern European languages and includes the EAGLES-based morphosyntactic specifications, defining the features that describe word-level syntactic annotations; medium scale morphosyntactic lexica; and annotated parallel, comparable, and speech corpora. The most important component is the linguistically annotated corpus consisting of Orwell's novel "1984" in the English original and translations. MULTEXT-East has already seen several editions, with the latest one being Version 3, where the most important addition are the Serbian language resources, including the structurally annotated "1984", the morphosyntactic specifications, the morphosyntactic lexicon and the linguistically annotated "1984". The complete dataset, unique in terms of languages and the wealth of encoding, is extensively documented, and freely available for research purposes.

Povzetek: članek opisuje uporabo MULTEXT-East v srbščini.

1 Introduction

The mid-nineties saw – to a large extent via EU projects – the rapid development of multilingual language resources and standards for human language technologies [6, 2]. However, while the development of resources, tools, and standards was well on its way for EU languages, there had been no comparable efforts for the languages of Central and Eastern Europe. The MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) was a spin-off of the EU MULTEXT project [6]; MULTEXT-East ran from '95 to '97 and developed standardised language resources for six CEE languages [1], as well as for English, the 'hub' language of the project. The project also adapted existing tools and standards to these languages. The main results of the project were lexical resources and an annotated multilingual corpus. The most important resource turned out to be the parallel corpus — heavily annotated with structural and linguistic information — which consists of Orwell's novel "1984" in the English original and translations.

One of the objectives of MULTEXT-East has been to make its resources freely available for research purposes. In the scope of the TELRI concerted action (Trans European Language Resources Infrastructure), the results of MULTEXT-East had been extended with several new languages and first released on a CD-ROM, and later through Web download via TRACTOR, the TELRI Research Archive of Computational Tools and Resources.

The Serbian language did not have its representative in the MULTEXT-East project. The researchers from the Faculty of Mathematics, however, participated in the TELRI concerted action. One of the results of this participation was the Serbian "1984" structurally annotated corpus, but the morphosyntactic specification, lexicon and linguistically tagged "1984" were not produced.

Following the TELRI release, the MULTEXT-East resources were used in a number of studies and experiments. In the course of such work, errors and inconsistencies were discovered in the MULTEXT-East specifications and data, most of which were subsequently corrected. But because this work was done at different sites and in different manners, the encodings of the resources had begun to drift

apart.

The '98–'00 EU Copernicus project CONCEDE (Consortium for Central European Dictionary Encoding) offered the possibility to bring the versions back on a common footing. Although CONCEDE was primarily devoted to machine readable dictionaries and lexical databases, one of its workpackages did consider the integration of the dictionary data with the MULTEXT-East corpus. The CONCEDE release contained the revised and expanded morphosyntactic specifications, the revised lexica, and the significantly corrected and re-encoded linguistically annotated "1984" corpus.

In addition to delivering resources per-se, a focus of the MULTEXT-East, TELRI and CONCEDE projects was also the adoption and promotion of encoding standardisation. On the one hand, the morpholexical annotations and lexica were developed in the formalism of the (EAGLES-based) specifications for six Western European languages of the MULTEXT project [6]. On the other, in the TELRI edition, all the corpus resources were encoded in SGML, in CES, the Corpus Encoding Standard [5]. For the corpus taken forward into the second edition, the Text Encoding Initiative Guidelines were adopted, in particular TEI P3 [9].

Finally, in 2004 the third version of the MULTEXT-East resources was released [3]. This release offers several contributions: it brings together the first two, i.e., offers both the TELRI and CONCEDE versions in one package; all the resources have been recoded in XML, according to TEI P4 [10], thus enabling them for processing with XML-based tools; and resources for new languages have been added, in particular the morphosyntactic specification for Resian, a dialect of Slovene, and, crucially the morphosyntactic specification and the annotated Orwell for Serbian.

Version 3 also contains extensive documentation, e.g., navigational HTML pages, which serve to structure and link the resources, and which include the list of participants and indexes to the resource by type and language. While the TEI headers give the most precise and up-to-date information on the corpus components, the documentation also contains a bibliography with copies of the MULTEXT-East project reports (giving details of the resources, e.g., the corpus markup process), published papers, a mirror of the TEI P4 and CES documentation and certain related MULTEXT and EAGLES reports.

A complete description of the Version 3 resources is given in [3] and in the on-line documentation, while this paper concentrates on the Serbian resources. In the next section we introduce the structurally annotated Serbian "1984" (already a part of the TELRI release), in Section 3 we describe INTEX, the system that has for a long time served as the infrastructure for developing LR resources for Serbian, Section 4 explains the MULTEXT-East (Serbian) morphosyntactic specification, Section 5 the linguistically annotated "1984", Section 6 the Serbian lexicon and the last section gives some conclusions and direction for further work.

```
<text id="mteo-sr." lang="sr">
 <body id="Osr" lang="sh">
  <div id="Osr.1" n="1" type="part">
   <head>Prvi deo</head>
   <div id="Osr.1.2" n="1" type="chapter">
    <head>1.</head>
    <p id="Osr.1.2.2">
      <s id="Osr.1.2.2.1">Bio je vedar i
hladan aprilski dan; na &#x10D;asovnicima
je izbijalo trinaest.</s>
      <s id="Osr.1.2.2.2"><name>Vinston
Smit</name>, brade zabijene u nedra da
izbegne ljuti vetar, hitro zama&#x10D;e u
staklenu kapiju stambene zgrade
<hi rend="it">Pobeda</hi>, no nedovoljno
hitro da bi spre&#x10D;io jednu spiralu
o&#x161;tre pra&#x161;ine da u&#x111;e
zajedno s njim.</s>
    </p>
```

Figure 1: The structurally annotated Orwell

2 Structural "1984"

The MULTEXT-East multilingual parallel corpus consists of the novel "1984", about 100,000 words in length. The corpus contains extensive headers and markup for document structure, sentences, and various sub-sentence annotations, which have been harmonised over languages. As an example, the start of the text from the Serbian Orwell is given in Figure 1.

The translations of "1984" have been automatically sentence aligned with the English original, and the alignments hand-validated. The bilingual alignments are stand-off, i.e., they are stored not with the primary data but in separate documents, as references to sentence IDs.

The cesDoc encoded novel served as the basis for producing the linguistically annotated version. The link between the two is maintained via sentence identifiers.

The Serbian version was produced already in the scope of TELRI. The digital source was the same as for the English and Slovene versions, namely the Oxford Text Archive, via the ECI multilingual CD-ROM. This version was plain ASCII, so it was first marked up, similar to other versions in SGML, and then sentence segmented and aligned with the English original. Also, many typographical errors were corrected.

3 Serbian INTEX resources

Before discussing the MULTEXT-East Serbian morphosyntactic resources (the specifications, lexicon and linguistically annotated "1984") we first describe the basis for these resources, which had been developed independently of European projects, namely the Serbian morphological lexicon in the format of the INTEX system, which is based on the technology of finite-state transducers [8].

In this dictionary a lemma is of the form $W_t, W_l.Cn +$

$SSD : (Codes)*$ where W_t represents the textual word, W_l the corresponding lexical word, C is the part of speech, n is the code of inflective class, SSD is the set of syntactic and semantic attributes of the lemma that is classified as Cn, and $codes$ describe the values of morphological categories that realized with the form W_t. For instance, the dictionary entry `prozorom,prozor.N01+Com:ms6q` descibes the form *prozorom* as the form of *prozor* (Engl. window), which is common (+Com) musculine (m) inanimate (q) noun (N) from the inflective class 01 in instrumental (6) of singular (s). It can be seen that this format is not as compact as MSD, as the relevant information is distributed among the inflective code, syntactic and semantic information, both associated to lexical word, and the grammatical codes which are assigned to the textual word. The $codes$ are not positional — for a part of speech one alphanumeric character represents a value of one and only one of its attributes.

The present size of the Serbian morphological dictionary is 74,000 lemmas and more than 1 million word forms, which enables morphological text analysis with a high percentage of success, around 92% for literary texts. Some of the word-forms that are not covered by the dictionary itself can be successfully morphologically tagged by additional tools (lexical transducers) incorporated in Intex. The use of this specifically constructed set of lexical transducers enables the recognition of various derived word-forms, such as several classes of compounds, possessive adjectives, diminutives, augmentatives, etc. [7].

The team from the University of Belgrade plans to convert its full INTEX lexicon to a MSD-type lexicon. It is also planned to tag with MSDs the corpus of contemporary Serbian that is being developed at the Faculty of Mathematics [12], where it plans to use INTEX and the lexica incorporated in it as a preprocessor.

4 Morphosyntactic Specifications

The MULTEXT-East morphosyntactic specifications give the syntax and semantics of the morphosyntactic descriptions (MSDs) used in the lexica and corpora. The MSDs, are structured and more detailed than is commonly the case for part-of-speech tags; they are compact string representations of a simplified kind of feature structures. The first letter of a MSD encodes the part of speech, e.g., `Noun` or `Adjective`. The letters following the PoS give the values of the position determined attributes. The specifications define, for each part of speech, its appropriate attributes, their values and one-letter codes. So, for example, `Ncmpi` expands to `PoS:Noun, Type:common, Gender:masculine, Number:plural, Case:instrumental`. It should be noted that in case a certain attribute is not appropriate (1) for a language, (2) for the particular combination of features, or (3) for the word in question, this is marked by a hyphen in the attribute's position. Slovene verbs in the indicative, for example, are not marked for gender or voice, hence the two hyphens in `Vcip3s--n`.

The specifications have been developed in the formalism and on the basis of specifications of the EU MULTEXT project [6] and in cooperation with EAGLES, the Expert Advisory Group on Language Engineering Standards. Originally, these specifications were released as a report of the MULTEXT-East project but have been revised for both subsequent releases, and have become, if not a standard, then at least a reference for comparison [4].

The MULTEXT-East morphosyntactic specifications have the following structure: (1) introductory matter; (2) the common specification; and (3) a language particular section for each language.

The common part of the specifications first defines the parts of speech and their codes; MULTEXT-East distinguishes the following, where not all PoS are used for all languages - we mark in italic those that are not used for Serbian: Noun (N), Verb (V), Adjective (A), Pronoun (P), *Determiner (D)*, *Article (T)*, Adverb (R), Adposition (S), Conjunction (C), Numeral (M), Interjection (I), *Residual (X)*, Abbreviation (Y), and Particle (Q).

The formal core of the specifications resides in the common tables; they define the features, their codes for MSD representation, and their appropriateness for each language — an example is given in Figure 2.

Technically, the complete specifications are a LaTeX document (with derived HTML and PDF renderings), where the common tables are plain ASCII in a strictly defined format. This format is suitable for a printed version, tolerable for one in HTML, and reasonably manageable for modification and addition of new languages. However, it is not suitable for processing needs, in particular to enable smooth manipulation and linking to an XML encoded corpus using the MSDs.

We have therefore implemented a (Perl) conversion of the common tables into XML, using the TEI.fs module, a tagset devoted to encoding feature-structures. This tagset is currently being used as the basis of an evolving ISO standard (currently a Draft International Standard), as part of work of ISO/TC 37/SC4 Language Resource Management.

The XML version of the common tables has one feature library for each category, e.g., ⟨*fLib type="Noun"*⟩. Each feature in such a library is comprised of the identifier, which enables the linkage to corpus MSDs, the name of the attribute, the languages the feature is appropriate for, and the symbol that is its value; examples are given in Figure 3.

The Serbian specifications was produced on the basis of the Croatian one (which was added in the scope of CONCEDE), with some modifications stemming less from the differences between languages and more by the set of morphosyntactic attributes already incorporated in the Intex e-dictionaries for Serbian. For instance, in the verb table the value gerund for the attribute VForm is the most appropriate to account for present and past gerund active in Serbian. Also, the attribute Clitic is applicable to Serbian

3.2 Verb (V)

P	ATT	VAL	C	EN	RO	SL	CS	BG	ET	HU	HR	SR	SL-ROZAJ
1	Type	main	m	x	x	x	x	x	x	x	x	x	x
		auxiliary	a	x	x	x	x	x	x	x	x	x	x
		modal	o	x	x	x	x		x		x	x	x
		copula	c		x	x	x				x	x	x
		base	b	x									
2	VForm	indicative	i	x	x	x	x	x	x	x	x	x	x
		subjunctive	s		x								x
		imperative	m		x	x	x	x	x	x	x	x	x
		conditional	c	x		x	x		x	x	x	x	
		infinitive	n	x	x	x	x		x	x	x	x	x
		participle	p	x	x	x	x	x	x		x	x	x
		gerund	g		x			x	x			x	
		supine	u			x			x				x
		transgressive	t				x						
		quotative	q						x				

Figure 2: Start of Common Table for Verbs

```
<fLib type="Verb">
<f id="V0."
   select="en ro sl cs bg et hu hr sr sl-rozaj"
   name="PoS"><sym value="Verb"/></f>
<f id="V1.m"
   select="en ro sl cs bg et hu hr sr sl-rozaj"
   name="Type"><sym value="main"/></f>
<f id="V1.a"
   select="en ro sl cs bg et hu hr sr sl-rozaj"
   name="Type"><sym value="auxiliary"/></f>
<f id="V1.o"
   select="en ro sl cs et hr sr sl-rozaj"
   name="Type"><sym value="modal"/></f>
```

Figure 3: Morphosyntactic specifications as TEI features

copula verbs, as well as the attribute Aspect to the most of the other verbs. Both these attributes are already encoded for all verbs in Serbian e-dictionary. One of the other differences between Croatian and Serbian tables is the recognition of the value ‘paukal’ for the attribute ‘Number’ for several PoS in Serbian.

5 Lexicons

The MULTEXT-East morphosyntactic lexicons have a simple structure, where each lexical entry is composed of three fields:

1. the *word-form*, which is the inflected form of the word, as it appears in the text, modulo sentence-initial capitalisation;
2. the *lemma*, which is the base-form of the word; where the entry is itself the base-form, the lemma is given as the equal sign; and
3. the *MSD*, the morphosyntactic description.

To produce the lexica, the token lists of the MULTEXT-East corpus were first fed through morphological analysers in order to produce the lemma list; this list was further extended from the comparable corpus, to arrive at at least 15,000 lemmas – some languages have further extended this, e.g., Romanian to 41,000 lemmas. In the next step, these lemmas were fed back to morphological generators (except for the agglutinative languages) in order to produce the complete inflected lists, i.e., the full paradigms of the lemmas, which constituted the final lexica of the project.

The MULTEXT-East lexica serve as medium sized morphological lexica for the languages. In addition to explicating the inflectional behaviour of the most common (and, typically, morphologically the most complex) words of the languages, the lexica also serve to establish the definitive set of valid MSDs for the languages.

For Serbian, currently, only a minimal lexicon was produced, which contains just the word-forms that in fact appear in the annotated “1984” corpus. This lexicon has 20,294 entries, 16,907 different word-forms, 8,392 lemmas and 906 MSDs.

To serve as a standard registry of MSDs, we converted the lexical MSDs to TEI feature structure libraries, ⟨*fsLib*⟩, one for each category. Here each MSD is expressed as a feature structure specifying its *id*, the language(s) it is appropriate for, and its decomposition into features. Some examples are given in Figure 4.

Both ⟨*fsLib*⟩s and ⟨*fLib*⟩s are stored in dedicated ⟨*TEI.2*⟩ element, complete with its TEI header; this document also constitutes a part of the linguistically annotated MULTEXT-East corpus.

```
<fsLib type='Verb'>
<fs id="Van" select="en et"
    feats="V0. V1.a V2.n"/>
<fs id="Van----an----n" select="cs"
    feats="V0. V1.a V2.n V7.a V8.n V13.n"/>
<fs id="Van----an-n---p" select="sr"
    feats="V0. V1.a V2.n V7.a V8.n V10.n
           V14.p"/>
<fs id="Van----ay----n" select="cs"
    feats="V0. V1.a V2.n V7.a V8.y V13.n"/>
<fs id="Vanp" select="ro"
    feats="V0. V1.a V2.n V3.p"/>
```

Figure 4: MSDs as TEI feature structures

6 Linguistically annotated "1984"

The centrepiece of the MULTEXT-East resources is the linguistically annotated "1984"; it contains word level markup, namely context disambiguated lemmas and MSDs. Because it was the first such resources for many of the MULTEXT-East languages, also Serbian, it was the most difficult and time-consuming to produce as the work had to proceed mostly manually. The annotated novel is useful as a dataset for tagger and lemmatiser induction and testing, and has already been used for this purpose in a number of experiments; c.f. the bibliography section on the MULTEXT-East web site.

The work on the Serbian annotation proceeded in the following steps. First, using Intex as the tool and all of the Serbian lexical resources "1984" was morphologically tagged. As a result, a textual file is obtained that contains the finite automaton of the text represented in the form of a regular expression.

In the second step the text annotated in this way was manually checked and disambiguated. It means that not only have right lemmas and morphological categories been chosen for ambiguous word-forms and added for those words that had not been recognized, but non-ambiguous forms have also been checked in case that they had been incorrectly recognized.

As a result, a non-ambiguous representation of a text is obtained in the same format. This step had been done iteratively, which enabled both the correction of the used dictionaries and other lexical resources, and their enhancement.

In the third step, a Perl script was written and used to convert the Intex annotated text with to the MULTEXT-East annotation. The conversion is not, however, a straightforward task, not only because of the different encoding systems, as described in section 3, but also because of the differently chosen attributes. This difference can be most easily described in the case of verbs. For verbs, in MULTEXT-East the second attribute specifies a verb form, and the third a tense. However, due to the composite tenses, some verb forms are used for the construction of different tenses. For instance, in Serbian, verb form *imao* is the active past participle of the verb *imati* (Engl. *to have*), and is used to produce both perfect tense if used with the indicative form of the present tense of the copula verb *biti* . In Intex, however, only the verb forms are recognized, which in the case of simple tenses enables the recognition of a tense as well (for instance, for present or aorist). For analytical tenses, the word-form recognition is not enough and more complex tools have to be used, as described in [11]. These tools, as not yet being fully developed, were not used in the first step of the annotation process, and thus the precise mapping from Intex to MULTEXT-East tags was not possible. As a consequence, despite having different functions, the active past participle is always given the same value in the third attribute, that is `tense=past`, as being the most frequent.

The TEI P4 markup of the linguistically annotated Serbian "1984" obtained through this process is exemplified in Figure 5 by the same two sentences.

```
<s id="Osr.1.2.2.1">
<w lemma="biti" ana="Vmps-sman-n---p">Bio</w>
<w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
<w lemma="vedar" ana="Afpmsnn">vedar</w>
<w lemma="i" ana="C-s">i</w>
<w lemma="hladan" ana="Afpmsnn">hladan</w>
<w lemma="aprilski" ana="Aopmpn">aprilski</w>
<w lemma="dan" ana="Ncmsn--n">dan</w>
<c>;</c>
<w lemma="na" ana="Spsa">na</w>
<w lemma="&#x10D;asovnik"
   ana="Ncmsa--n">&#x10D;asovnicima</w>
<w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
<w lemma="izbijati"
   ana="Vmps-snan-n---e">izbijalo</w>
<w lemma="trinaest" ana="Mc---l">trinaest</w>
<c>.</c>
</s>
```

Figure 5: The linguistic annotation of "1984"

7 Conclusions

The paper presented Version 3 of the MULTEXT-East resources, and, in particular, its Serbian language portion. As the resources cover a number of inflectionally rich languages, are interlinked, harmonised, have a standardised encoding, and have been manually validated and tested in practice, they can serve as a "gold standard" dataset for language technology research and development.

While portions of the resources are distributed without any restrictions, the resources as a whole are available free of charge for research purposes only, as this was the condition imposed by some copyright holders of the sources.

Version 3 of the resources can be downloaded from the MULTEXT-East home page, *http://nl.ijs.si/ME/*. Access is enabled by filling out and submitting a Web based agreement, which is modelled after the one used by Edinburgh's Language Technology Group.

Currently, there are no plans to start working on Version 4; rather, the focus will be on the utility of V3, in our own

research, and in enabling others to use the resources, by providing maintenance, continuing to support their accessibility and correcting errors.

Acknowledgments

The work presented in this paper was, in part, supported by the bi-lateral project on scientific and technological cooperation between Slovenia and Serbia “The development of language resources for machine translation between the Slovene and Serbian languages”.

References

[1] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.J., Petkevič, V., and Tufiş, D., 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL ’98*. Montréal, Québec, Canada.

[2] EAGLES, 1996. Expert advisory group on language engineering standards. Http://www.ilc.pi.cnr.it/EAGLES/home.html.

[3] Erjavec, T., 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC’04*. Paris: ELRA. Http://nl.ijs.si/et/Bib/LREC04/.

[4] Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., and Vitas, D., 2003b. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*. Budapest.

[5] Ide, N., 1998. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC’98*. Granada: ELRA. Http://www.cs.vassar.edu/CES/.

[6] Ide, N., and Véronis, J., 1994. 1994. Multext (multilingual tools and corpora). In *Proceedings of the 15th International Conference on Computational Linguistics*. Kyoto.

[7] Pavlović-Lažetić, G., Vitas, D., and Krstev, C., 2004. Towards full lexical recognition. In *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag, pp.179–186.

[8] Silberztein, M., 2000. *INTEX*. Masson.

[9] Sperberg-McQueen, C. M. and Burnard, L. (eds.), 1999. *Guidelines for Electronic Text Encoding and Interchange, Revised Reprint*. The TEI Consortium.

[10] Sperberg-McQueen, C. M. and Burnard, L. (eds.), 2002. *Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines*. The TEI Consortium.

[11] Vitas, D., 2003. Composite tense recognition and tagging in serbian. In *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*. Budapest.

[12] Vitas, D., Krstev, C., Pavlović-Lažetić, G., and Obradović, I., 2003. An Overview of Resources and Basic Tools for Processing of Serbian Written Texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*.