Orwell’s 1984 – the Case of Serbian Revisited

Cvetana Krstev¹, Duško Vitas², Aleksandra Trtovac³

¹ University of Belgrade, Faculty of Philology, Studentski trg 3, Belgrade, cvetana@matf.bg.ac.rs
² University of Belgrade, Faculty of Mathematics, Studentski trg 16, Belgrade, vitas@matf.bg.ac.rs
³ University of Belgrade, University Library, Bulevar kralja Aleksandra 71, Belgrade, aleksandra@unilib.bg.ac.rs

Abstract

In this paper we present an alternative version of the morphosyntactically annotated Serbian translation of 1984. This version follows the basic principles of MULTEXT-East version, except for one addition – the text will be annotated with multi-word units as well. We will present the resources used for annotation with multi-word units and explain how these resources were enriched with multi-word units extracted from the processed text. Finally, we will present the format of this alternative version and the benefits obtained both from preparing the new resource and from the resource itself.

Keywords: morphosyntactic annotation, multi-word units, finite-state transducers, MULTEXT-East

1. Motivation

The well known multilingual linguistic resource MULTEXT-East has a long history. Its first version was produced as part of the project that was a spin-off of MULTEXT and ran from ’95 to ’97. After the project officially ended, the resource was upgraded and enhanced as part of some other projects, but also independently as a result of the efforts of different research groups. The core of this resource were in its first, as well as in its subsequent versions: the MULTEXT-East morpho-syntactic specifications of involved languages, lexica, and aligned and annotated 1984 corpora. The latest version is v4, released in May 2010. The first version covered six languages plus English (Dimitrova et al., 1998), while the current one contains at least some of the core resources for seventeen languages (Erjavec, 2010).

The Serbian language was represented in this resource from the very beginning. Although Serbian research groups were not involved in the MULTEXT-East project, they contributed to it by providing a structurally annotated and aligned version of Orwell’s 1984 through the TELRI project (Erjavec et al., 1998). A morphosyntactic description of the linguistically annotated 1984 in Serbian and the basic morphosyntactic lexicon were added in the version 3.0 (Krstev et al., 2004). In the version 4.0, the Serbian morphosyntactic lexicon was enhanced and it now contains a superset of all lemmas and their forms that appear in 1984. The morphosyntactic description and annotation of 1984 should have been corrected and improved, but due to the lack of human resources this was not done.

The popularity of this resource is indisputable. According to Google Scholar, only the paper (Erjavec, 2004) describing its third edition has 131 citations (as of August 2011). The resource itself is still being requested on a weekly basis by members of various research communities. Its popularity can be attributed to the fact that it represents a coherent and comparable resource for many languages. Also, the existence of a comprehensive documentation cannot be overstressed.

The Serbian part of this resource was used for various research purposes, e.g. for the training of taggers (Popović, 2010), (Delić et al., 2009) and aligners (Božović, 2010) for Serbian. Its connection to Serbian morphological e-dictionaries is bidirectional: they were used to semi-automatically produce linguistically annotated 1984, and were at the same time corrected and improved in the course of this process.

On the other hand, multi-word units (MWUs) have been a topic of great interest in the NLP field for quite a long time. Much work was devoted to them and they have been analyzed and processed from a variety of different perspectives. In the computational lexicography school led by Maurice Gross in LADL, the interest in MWUs and the production of morphological dictionaries of compounds has been vivid from the very beginning (Gross, 1986). In recent years further advances have been achieved to enable effective production of such dictionaries (Savary, 2009). Also, some work has been done to produce texts annotated with MWUs rather then with simple words. In (Laporte et al., 2008), the authors present an e-version of Verne’s novel Le tour du monde en 80 jours annotated with multi-word adverbs while a corpus annotated with named entities, most of which are MWUs, is presented in (Savary et al., 2010).

Having in mind the mutual benefit that Serbian morphological e-dictionaries of simple words and the morphologically annotated Serbian version of 1984 had from each other, our idea was to try to obtain a similar positive result for MWUs. We would like to (a) collect from 1984 new candidates for Serbian morphological e-dictionaries of MWUs, which is still under development, and (b) produce, using these dictionaries, an alternative version of 1984 annotated with MWUs.

In section 2 we will briefly present Serbian morphological e-dictionaries of MWUs, in section 3 we will present our tools for collecting new candidates for these dictionaries from a text, while in section 4 we will give the results obtained for 1984, in section 5 we will present the format of this alternative version, and finally in section 6 we will give some concluding remarks.

2. Serbian tools for MWU processing

By multi-word units we mean contiguous sequences of simple words that have a constant reference which can be treated as a single unit when used in many NLP applications. Two types of such MWUs are of interest. The MWUs with a predictable structure and a potentially...
infinite number of instances, e.g. date and time expressions, can be described by finite-state automata. The other type of MWUs that are idiosyncratic in nature, have a constant reference and show a certain degree of non-compositionality can be listed in a dictionary. The idea and the basic principles of treatment of these two types of MWUs for Serbian were described in (Krstev and Vitas, 2006). Both types of MWUs are treated in a similar way as simple word units during text processing: after applying dictionaries and other lexical tools to a text, a lemma and sets of values of grammatical categories are assigned to a recognized MWU. That is, MWUs belong to same Part-of-Speech (PoS) as simple words – nouns (N), adjectives (A), adverbs (ADV), interjections (INT), conjunctions (CONJ), prepositions (PREP) – and are, consequently, described by the same morphosyntactic categories. Verb phrases are not contiguous by nature and, therefore, they are not treated in either of these ways.

Finite state transducers that recognize and annotate the following MWUs were developed:

- Numerals, both cardinal and ordinal, written as words, digits (Arabic and Roman) and their combinations (sto cetvredes i pet miliona ‘145 million’);
- Inflected and derived forms of numerals written as digits (30-ak ‘approximately 30’) and acronyms (e.g. SPPT ‘FFCC’, NATO-a ‘NATO (in the genitive)’);
- Interjections with a repetition of patterns (ki-hi-hi-ho-ho-ho).

- Compounds of numerals (written as Arabic digits or words) and nouns (32-godišnjak ‘32 years old man’) or adjectives (12-godišnji ‘12 years old’), either separated with a hyphen or not. It should be noticed that these occurrences represent simple words written as MWUs.
- Noun phrases expressing measure (tri stotine metara ‘300 metres’) and money expressions (osam funti ‘eight pounds’).
- The code following the lemma in a dictionary entry represents a human (+Hum) in the instrumental case (6).

![Figure 1: One path from a transducer that recognizes compound nouns <NB->godišnjak](image)

The role of these transducers is to assign to a recognized form its lemma and a set of values of grammatical categories, as illustrated by one path in one of these graphs in Figure 1. The result is the same as if the following were listed in an e-dictionary:

32-godišnjakom,32-godišnjak.N+Hum+C:ms6v

That is, 32-godišnjakom is the singular (s) form of a masculine (m) noun 32-godišnjak representing a human (+Hum) in the instrumental case (6). The other MWUs are actually listed in e-dictionaries. Serbian morphological e-dictionaries are being developed in accordance with the methodology and the format (known as DELAS/DELAF for simple words, and DELAC/DELACF for MWUs) presented for French in (Courtois and Silberztein, 1990). One line from DELACF for the MWU noun jednakost pred zakonom ‘equality before the law’ is:

jednakosti pred zakonom, jednakost pred zakonom.N+Comp:s7qf

The MWU lemmas are listed in a separate dictionary, DELAC, and provided with information that enables automatic generation of all of their inflected forms. For instance, the entry for the lemma jednakost pred zakonom is:

jednakostjednakost.N704:fs1q pred zakonom,NC_N4X1

The paper (Krstev and Vitas, 2006) describes how all inflected forms of a lemma can be generated form such an entry, and highlights the expressiveness and limitations of this approach. As lemma entries are complex in themselves, a tool was developed that produces such entries from a list of MWUs in their normalized (dictionary) form. This tool and its successfulness in producing lemma entries for various types of MWUs (named entities, general lexica, economy domain, etc.) were presented in (Krstev et al. 2010). We use the Unitex corpus processing system for processing text using finite-state transducers and e-dictionaries.2

The code following the lemma in a dictionary entry (e.g. NC_N4X1 in the above example) is at the same time the name of the inflectional transducer that is responsible for the generation of forms and the code that determines the PoS of an MWU lemma (NC – a compound noun) and its structure (N4X1 – it consists of a noun followed by four tokens, including separators, that are re-copied in all inflected forms).

<table>
<thead>
<tr>
<th>MWU structure</th>
<th>No. of lemmas (in %)</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>AXN</td>
<td>71.90</td>
<td>Veliki Brat ‘Big Brother’</td>
</tr>
<tr>
<td>2NX</td>
<td>10.56</td>
<td>patent-zatvarač ‘zipper’</td>
</tr>
<tr>
<td>N2X</td>
<td>5.81</td>
<td>igra rečima ‘play on words’</td>
</tr>
<tr>
<td>N4X</td>
<td>3.75</td>
<td>ram za slike ‘picture-frame’</td>
</tr>
<tr>
<td>NXX</td>
<td>3.32</td>
<td>osećanje niže vrednosti ‘sense of inferiority’</td>
</tr>
<tr>
<td>AXAXN</td>
<td>1.11</td>
<td>pištolj-igračka ‘toy pistol’</td>
</tr>
<tr>
<td>other</td>
<td>3.64</td>
<td>jeftina radna snaga ‘cheap labour’</td>
</tr>
<tr>
<td>2XA</td>
<td>46.30</td>
<td>vaspitno-popravni ‘correctional’</td>
</tr>
<tr>
<td>A4X</td>
<td>21.01</td>
<td>beo kao kreda ‘white as chalk’</td>
</tr>
<tr>
<td>A3XX</td>
<td>17.12</td>
<td>gladan kao vuk ‘hungry as a wolf’</td>
</tr>
<tr>
<td>AXA</td>
<td>9.73</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>5.83</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. The most frequent structures of Serbian MWU nouns and adjectives

Serbian MWU nouns and adjectives have various structures — some of them are very frequent, while

1. All examples in this paper (in English with Serbian translation) are from the novel 1984, if such an example occurs in the text.

others occur in just a few lemmas. The most frequent structures for lemmas that inflect (MWU nouns and adjectives) are listed in Table 1. The percentages were computed at the moment when the Serbian DELAC contained 6647 lemmas for MWU nouns and adjectives.

In addition to nouns and adjectives, Serbian dictionaries contain MWUs belonging to other PoS. These MWUs also have various structures, but their structure cannot be deduced from their PoS codes because they do not inflect. However, something can be said about their structure too. The most frequent structure of MWU prepositions is PREP_N – 94.70%, MWU conjunctions is CONJ_ADV – 32.43% and CONJ_CONJ – 48.65%, MWU adverbs is PREP_N – 52.62% and PREP_ADV – 14.24%. These percentages were computed at the moment when the Serbian DELAC contained 615 MWU lemmas that do not inflect.

3. Enhancement of Serbian DELAC dictionary

The first sources of MWUs to be included in e-dictionaries are traditional dictionaries for general lexica and various gazetteers for proper names such as personal names and geopolitical names. The initial set of MWUs included in the Serbian DELAC dictionary was partly taken from various paper sources and partly from the Serbian wordnet. However, these sources cannot be the only ones used, because they were produced from a different standpoint compared to that imposed by the NLP applications.

Namely, many compounds that occur in the contemporary texts are not listed in the already used resources either because they are underdeveloped (wordnet), new or regarded as “uninteresting” from the lexicographic point of view (paper dictionaries). Therefore, we have developed a collection of graphs that extract potential MWUs from a chosen text on the basis of their syntactic structure. The aim of these graphs is twofold: (1) to extract with high recall and precision the clauses having a certain syntactic structure; (2) to extract only the clauses that are not already in the DELAC dictionary. We consider an extraction graph to be successful if the extracted clauses have the syntactic structure we are looking for. Whether the extracted clauses are actually MWUs that we wish to include in our e-dictionaries has to be determined manually.

 Naturally, we have constructed extraction graphs that cover the most frequent structures presented in the previous sections. The structures are the following:

- For nouns, A_N, N_N, N_Ng, N_PREP_N, UNK_N (UNK stands for a MWU component that is not a simple word);
- For adjectives, UNK_A, A_A, A_kao_N;
- For prepositions and adverbs, PREP_N.

One of these graphs is presented in Figure 2. The recognition of a clause Noun/Noun in the genitive case that is not already in the DELAC is done by the sub-graph Kontekst_N_Ng_ne (the upper graph in Figure 2). In order to achieve better precision, we reject all noun words that are confusable with other PoS, as well as proper names (the sub-graph Nije_N which is not presented in Figure 2). These instances of false recognition are rejected by using positive and negative contexts in the Unitex corpus processing system (Paumier, 2008; Sec. 6.3). Despite all this, the graph Kontekst_N_Ng_ne leads to many false recognitions, because in Serbian there are some very frequent compound prepositions and adverbs with a simple preposition in the initial position, followed by one or more nouns. In order to achieve better results, the main graph accepts only those sequences accepted by Kontekst_N_Ng_ne that are either (i) preceded by a preposition which does not form a MWU preposition or adverb with the nouns that follow it or (ii) are not preceded by a preposition.

### Figure 2: A graph that recognizes syntagms NOUN/NOUN_g that are not in the dictionary of MWUs

By applying these extraction graphs to our text – the Serbian translation of 1984 – and by manually analyzing the obtained concordances, we produced a list of 415 new MWUs that we have further processed to obtain MWU lemmas that we have inflected and the obtained forms were included in our e-dictionaries. The PoS and structure of new lemmas are presented in Table 2. Most of new noun MWUs are common nouns – 1984 is not a kind of text where many named entities or proper names, other then the names of the main characters, are expected. Among the 11 proper names added are Komunistička internacionala ‘Communist International’ and Deklaracija nezavisnosti ‘Declaration of Independence’.

![Diagram](image)

<table>
<thead>
<tr>
<th>Noun</th>
<th>325</th>
<th>Adjectives</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AXN</td>
<td>259</td>
<td>AXA</td>
<td>1</td>
</tr>
<tr>
<td>N2X</td>
<td>19</td>
<td>2XA</td>
<td>3</td>
</tr>
<tr>
<td>2XN</td>
<td>7</td>
<td>A4X</td>
<td>3</td>
</tr>
<tr>
<td>NXX</td>
<td>2</td>
<td>Adverbs</td>
<td>59</td>
</tr>
<tr>
<td>N4X</td>
<td>34</td>
<td>Prepositions</td>
<td>20</td>
</tr>
<tr>
<td>AXAXN</td>
<td>1</td>
<td>Interjections</td>
<td>3</td>
</tr>
<tr>
<td>other</td>
<td>3</td>
<td>Conjunctions</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: New MWUs retrieved from Serbian 1984

In addition to these 415 MWUs that were added to our e-dictionaries of general lexica and proper names, we prepared a list of MWUs specific to this particular text. It contains 21 items like O‘Brajevov, possessiv adjective of O‘Brien and Ministarstvo ljubavi ‘the Ministry of Love’. It was not always easy to decide what is specific and what is general enough – we decided that Veliki brat ‘Big Brother’ became part of general lexica, while Policija misli ‘the Thought Police’ did not (yet).

For this type of text, there was no need to develop new or modify any of the existing finite-state transducers already in use for inflection of MWUs in the Serbian texts.
4. Application of MWU resources

After applying our lexical MWU resources – general e-dictionaries, specific e-dictionaries, and finite-state transducers – to the Serbian translation of 1984, we obtained the results presented in Table 3. The text itself has approximately 90,000 current simple words.

<table>
<thead>
<tr>
<th></th>
<th>occurrences</th>
<th>lemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FSTs</strong></td>
<td>328</td>
<td>138</td>
</tr>
<tr>
<td>numerals</td>
<td>68</td>
<td>37</td>
</tr>
<tr>
<td>measures</td>
<td>250</td>
<td>95</td>
</tr>
<tr>
<td>money</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td><strong>General dictionaries</strong></td>
<td>1787</td>
<td>625</td>
</tr>
<tr>
<td>nouns</td>
<td>739</td>
<td>429</td>
</tr>
<tr>
<td>adjectives</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>adverbs</td>
<td>494</td>
<td>102</td>
</tr>
<tr>
<td>prepositions</td>
<td>176</td>
<td>63</td>
</tr>
<tr>
<td>conjunctions</td>
<td>363</td>
<td>19</td>
</tr>
<tr>
<td>interjections</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td><strong>Specific dictionaries</strong></td>
<td>393</td>
<td>21</td>
</tr>
<tr>
<td>nouns</td>
<td>369</td>
<td>20</td>
</tr>
<tr>
<td>adjectives</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2508</td>
<td>784</td>
</tr>
</tbody>
</table>

Table 3: Application of MWU resources on 1984- results

We wanted to compare our results to the English text processed with English resources. We relied on the resources included in the standard Unitex distribution. They consist of a public e-dictionary developed using the same methodology used for Serbian and its production is described in (Savary 2000). It contains a number of MWU forms and lemmas – approximately 70,000 lemmas. For the English text that consists of approximately 105,000 simple word occurrences, this dictionary retrieved 1,172 occurrences of MWUs and 532 different lemmas. Having in mind that for English finite-state automata for numerals and other expressions – measures and money – were not supplied in the Unitex distribution and no specific dictionary was available for 1984, the results were comparable to those obtained for Serbian. It is not surprising that more MWUs were retrieved in the Serbian text, since our resources were enhanced using this particular text. The comparison of the retrieved MWUs in English and Serbian text is beyond the scope of this paper and asks for more detailed analysis. All we can say is that among them there is a number of exact equivalents, e.g. Veliki brat ‘Big Brother’, talasasti lim ‘corrugated iron’, pre ili posle ‘sooner or later’.

Although it was already established that the recognition of MWUs reduces false ambiguities (see, for instance (Alegria et al., 2004)) introduced by high-recall annotation obtained by lexical resources, they cannot be accepted without caution. In our experiment there were few false recognitions, e.g. the sequence crna kosa was recognized as the genitive case of a MWU crni kos ‘blackbird’ while in the text it represented a phrase in the nominative case ‘black hair’. The instances of ambiguous behaviour of frequent MWU prepositions and adverbs were more serious. One such case is u blizini – it can be an adverb, like in On se upita da u blizini ipak nije sakriven kakav mikrofon ‘He wondered whether after all there was a microphone hidden somewhere near’ or a preposition, like Nije bilo pametno čak ni biti viden u blizini takvih ljudi ‘It was not wise even to be seen in the neighbourhood of such people’. There are also cases where although a MWU was correctly recognized, annotating text with it could confuse further linguistic analysis. For instance, annotating text with a MWU pismeni ljudi ‘literate men’ in a sequence jedva pismeni ljudi ‘people who could barely read and write’ would mean overlooking the fact that the adverb jedva ‘barely’ modifies the adjective pismeni. All such cases have to be manually examined and some of them can be eliminated by disambiguation grammars.

5. An example from the alternative version

We will illustrate the content of the proposed alternative version with one sentence: Ispadalo je da su čak bile organizovane manifestacije u znak zahvalnosti Velikom Bratu zato što je sledovanje čokolade povećano na dvadeset grama nedeljno. ‘It appeared that there had even been demonstrations to thank Big Brother for raising the chocolate ration to twenty grams a week.’ This sentence contains four MWUs:

- A preposition u znak ‘to (thank)’ requiring the locative case;
- A noun Velikom Bratu ‘Big Brother’ in the dative case;
- A conjunction zato što ‘for’.
- A measurement expression dvadeset grama ‘twenty grams’ that inherits grammatical properties from the numeral.
Conjunctions, prepositions and interjections have the attribute ‘formation’ with values ‘simple’ and ‘compound’, while nouns, adjectives, numerals and adverbs do not, so they cannot be described as such. That does not mean that there are no deficiencies in the MULTEXT-East specification (already discussed in Przepiórkowski and Wolinski, 2003), but they did not increase with the use of MWUs. The conversion of grammatical categories and their values used in Serbian morphological dictionaries to the MULTEXT-East attributes and their values, is done using the same procedures that were used for the production of annotated text using simple words (Krstev et al., 2004).

6. Concluding remarks

With this experiment we have already achieved the following goals:

- We have enlarged our e-dictionaries of MWUs by 415 new entries;
- We have developed a set of finite-state transducers that can extract potential MWUs from processed texts. These transducers were already successfully used on texts from various domains (e.g. in the domain of library and information sciences);
- We have set the stage for the production of an alternative version of MULTEXT-East that will be morphosyntactically annotated with MWUs. This task has not been finalized yet. Before the final production stages can take place we would like to perform some more preparatory work (e.g. annotation of date and time expressions, double check the annotation of the analytical tenses, etc.).

We think that when finalized, the alternative version of 1984, will be a very successful resource. Besides tasks it is already used for, it would be a useful resource for various comparative analyses and bringing us one step closer to syntactic annotation.

Acknowledgment

This research was supported by the Serbian Ministry of Education and Science (grant ON 178006) and Cesar project funded under ICT Policy Support Programme (ICT PSP), grant agreement no 271022.

References


