CONSTRUCTION AND EXPLOITATION OF X-SERBIAN BITEXTS

1 Introduction

In this paper we will present aligned corpora in which one of the languages is Serbian and which were developed primarily for linguistic and lexicographic research. The choice of texts for these corpora was influenced by this goal – they contain primarily literary texts in which Serbian is either a source or a target language and translations were done by reliable translators. Two major corpora were developed along these lines – French/Serbian and English/Serbian literary corpora - as well as a smaller corpus in which Serbian texts were aligned with Serbian texts or texts in some other Slavonic or Balkan language. For the purpose of some specific projects other types of aligned corpora were produced that contain non-fictional texts. Technical applications of aligned corpora that include Serbian emerged only recently and were restricted to certain specific domains that include experiments with word alignment and cross-lingual information retrieval and extraction.

It should be noted that the work on the translation of EU legislation (acquis communautaire) into Serbian is in progress, and results are available in the form of concordances on the Web.[1] This will eventually lead to a new aligned resource including Serbian that already exists for many languages.

On the other hand, much of the work in the natural language processing in Serbia is devoted to the development of monolingual and multilingual lexical resources, and developed aligned corpora enabled their testing and enhancement as well as production of new software tools that can support these aims.

Aligned corpora that include Serbian can be used in two different software environments: one is IMS CWB[2] that is also used for exploitation of monolingual Serbian corpora on web, and the other is the corpora processing system Unitex[3] (Vitas and Krstev 2012) that can be used locally. In the next section we will present the functionalities of both

[1] http://prevodjenje.seio.gov.rs/evroteka/index.php?jezik=engl
[2] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
[3] http://igm.univ-mlv.fr/~unitex/

systems, with the emphasis on the second one because it enables much richer linguistic queries due to the usage of lexical resources that will be presented in section 3. Next we will present the process of producing an aligned text (section 4), and the content of the aligned corpora that include Serbian produced so far (section 5). At the end we will give some examples that illustrate the power of this approach in processing aligned corpora.

2 Software for querying aligned corpora

The software used for searching Serbian corpora on the web is IMS CQP Workbench (Christ 1994), while the online interface was developed at the Faculty of Mathematics.[4] The corpus is used by more than 300 Slavists all over the world.

The Unitex system (Paumier 2010) is completely different. It represents a corpus processor that implements the theory of finite-state automata and transducers in processing of raw texts by applying electronic dictionaries of both simple words and multi-word units.[5] The basic principle of its work is to tokenize and normalize a text in the preprocessing phase, and then to tag it using the information from the applied e-dictionaries. It is then possible to search a text (or a corpus) using not only regular expressions (as `[a-z]+` *ošću* in the previous example) but also with all lexical and grammatical tags attached to simple words and multi-word units in an analyzed text or a corpus. For instance, the query `<N+Attr:6>` would search for all occurrences in an analyzed text of attributive nouns (N+Attr) in the instrumental case (6). The finite-state transducers enable not only a search in a text but also its transformation by using transducer outputs. A graph interface that is a part of the Unitex system enables formulation of very complex queries and transformations. Each node in a graph contains a simpler regular expression whose terms are strings, lexical tags or calls of other graphs.

The system Xalign for text alignment is being, as proposed in (Vitas et al. 2006), incorporated in Unitex starting from the version 2.0 and a new interface was developed that facilitates working with aligned texts. The main advantage for users is the possibility to apply the same powerful text search capabilities to each of aligned texts (a *Locate...* button) and as a

[4] http://www.korpus.matf.bg.ac.rs/ (a password required)
[5] Unitex is distributed under the LGPL license, was developed in C/C++ and Java and it fully supports Unicode.

result aligned concordances are obtained, as represented in Figure X-1~~Figure X-1~~.

Formatted: Font: 10 pt

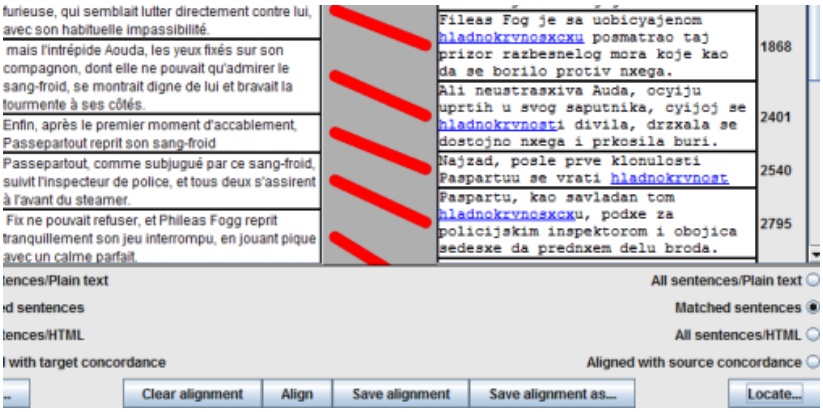

Figure X-1 An excerpt of concordances produced by <N+Attr:6> applied to the Serbian part of a French/Serbian bitext Around the World in 80 Days

3 Lexical resources for Serbian

Exploitation of corpora that include Serbian texts is connected with lexical resources, both monolingual and multi-lingual, that were developed for Serbian and other languages.

The most important monolingual lexical resource is organized as a system of morphological electronic dictionaries and their extensions in a form of local grammars. The format of these dictionaries, known as a LADL format, was initially developed for French (Gross and Perin 1989) and later for many other European languages. Each entry in these dictionaries has a form w_t`,`w_l`.K+SynSem:(mgc)*`, where w_t represents a word from a text, w_l its lemma, `K` a PoS tag, `SynSem` a set of syntactic and semantic markers assigned to a lemma, and mgc a (potentially empty) string of morphosyntactic codes that describe a relation between w_t and w_l. In the Unitex implementation the PoS tag also identifies a lemma's type of inflectional paradigm, that is, the inflectional finite-state transducer that is responsible for generating for a lemma all its forms in the above format. For instance, one entry from the English e-dictionary of simple words is: `grapes,grape.N+Conc:p` where w_t `= grapes`, w_l `= grape`, `K = N` (a noun tag), `SynSem = Conc` (a

concrete noun), and `mgc = p` (plural). Electronic dictionaries have two basic components: a dictionary of simple words (alphabetic strings between two separators) and a dictionary of multi-word units. A structure of a dictionary of simple words for a language with a rich morphology implies that in a field w_t all possible forms of a lemma from w_l can occur. One entry from the Serbian e-dictionary of simple words is: `trojicom,trojica.N+NumN+MG+Pl:fs6v`, Information in the `SynSem` field if a word `trojica` 'three men' says that it is a noun (`N`) whose natural gender is masculine (`MG`), while its grammatical gender is feminine (`f`) and it represents a plural (`Pl`) although its grammatical number is singular (`s`). Another entry from the e-dictionary of simple words is: `prozora,prozor.N1:ms2q:mp2q:mw2q:mw4q`. It demonstrates that a word form `prozora` represents various representations of the lemma `prozor` 'window', as stated by various sets of values of grammatical categories.

The dictionary of multi-word units (or compounds) that defines inflective characteristics of sequences of simple words was developed on similar principles. One entry from the Serbian e-dictionary of multi-word units is:

`Nobelovu nagradu,Nobelova nagrada.N+Comp:fs4q`.

The attribute +Comp assigned to the lemma `Nobelova nagrada` 'Nobel prize' indicates that it is a MWU. We should note that a possessive adjective `Nobelov` has to agree in the gender, the number, and the case with the noun `nagrada`. The PoS tag in the case of MWU identifies beside a lemma's part-of-speech also an inflectional finite-state transducer that takes care of agreement conditions and other specifics of MWU inflection (Krstev et al. 2006a). The syntactic and semantic properties for a MWU can be assigned independently, because due to the uncompositionality of MWUs often they cannot be inherited from properties assigned to its constituent simple words. For instance, a lemma *crna ovca* 'black sheep' contains in the `SynSem` field the attribute `+Hum` (a human) which cannot be automatically derived from the properties of lemma's constituents: `crn.A+Col` (color) and `ovca.N+Zool` (animal).

The Serbian dictionary of simple words is well developed and it can be compared in size and content with similar resources for other better-resourced languages. It contains 127,000 lemmas (91,000 belong to general lexica and 36,000 are proper names) from which near to 4.5 million word form realizations are automatically generated. The development of the Serbian dictionary of MWUs started only recently. Its

production is far more demanding (in respect to collecting candidates, producing lemmas and inflecting them) so it is far from being complete. At present it contains 8,000 lemmas (7,000 belong to general lexica and 1,000 proper names). It should be noted, however, that dictionaries are not the only way of dealing with MWUs: some of them are dealt by so-called dictionary graphs (e.g. for compound numerals) while others covered by local grammars (e.g. dates, measurement expressions, etc.).

The processing of corpora that relies on e-dictionaries differs significantly from processing of corpora that has been unambiguously tagged in advance. Processing with e-dictionaries produces high-recall tagging that need not be a disadvantage. The possibility of formulating very complex queries that involve all information recorded in applied dictionaries and their combination (morphological, syntactic, and semantic) in many cases enables selection of an appropriate tag. Not to mention that at any moment a user can apply a different (e.g. domain specific or similar) set of dictionaries to his raw corpus to obtain different tagging that better suits his/her needs.

Besides this monolingual resource, Serbian is represented in two multilingual lexical resources. One of them is WordNet in which the Serbian part of the database is aligned with other languages through and interlingual index. Serbian WordNet presently has near to 16,000 synsets (sets of synonyms) and some domains are better represented in it than others, like biology, linguistics, medical sciences, etc. (Krstev et al. 2004). The other is the multilingual lexical base Prolex (Vitas et al. 2007) in which proper names are organized at both the conceptual and the linguistic level. For instance, each concept like ‘capital of Turkey’ is connected on the conceptual level with the concept ‘Turkey’ using the relation 'capital of'. On the linguistic level the first concept is represented in Serbian with set of synonyms: *Istanbul*, *Stambol*, *Carigrad*, *Konstantinopolj*, while for some other languages the linguistic representation would be different (in English: *Istanbul*, *Byzantium*, *Constantinople*). Derivational properties of proper names are described as well which is of great importance for successful corpus processing. For instance, by using the information from e-dictionaries it is possible to analyze the sentence *On čita knjigu o minulim ratovima* 'He reads a book about past wars'. The sentence with the similar structure *Marko čita Herodotove Istorije o grčko-persijskim ratovima* 'Marko reads Herodotus History of the Greco-Persian wars’ contains four occurrences of proper names – one of them as a possessive adjective, and two of them as relational adjectives forming one MWU – and it could not be properly analyzed if these proper names were not

adequately described. Their inflection is described by e-dictionaries while their semantics is described in the Prolex database.

The interaction between these different lexical resources on one side and aligned texts on another side is supported by a tool developed specifically for this purpose. A work station for development, maintainance and exploitation of lexical and textual resources *LeXimir* was developed by the NLP group at the University of Belgrade (Krstev et al. 2006b). It enables transfer of information from one resource to another but also the use of all mentioned resources in searching aligned texts.

4 Technology of producing aligned texts

In a technical respect, all aligned corpora that include Serbian as a language were produced in the same way. Texts were collected from different sources: they were downloaded from web, scanned using OCR software or retyped. In the preprocessing phase all texts were transformed into the ASCII encoding scheme that neutralizes the effect of two different alphabets originally used (Latin and Cyrillic in Serbian). This is essential because for the linguistic processing of corpora the alphabet used for printing or display is of no consequence. It should be noted, however, that both Cyrillic and Latin representation can be reproduced form this ASCII representation without loss of information or introduction of errors.

In the next step, logical layout tags were added to all texts (divisions, headings and paragraphs) as well as a TEI heading with basic meta-data about the texts' origin. All this was done automatically and then carefully proof-read. This was particularly important for texts derived from their .pdf versions because the information on paragraph endings can be easily lost during transformation.

After that all texts obtained by scanning or retyping were corrected by using Unitex and e-dictionaries supported by it (for French, English, and Serbian). The final step before proceeding to alignment was sentence segmentation. Here again we used Unitex and sentence segmentation graphs for the above mentioned languages. The first such graph for French was developed for the Intex system (predecessor of Unitex) (Friburger et al. 2000), and later the similar graphs were produced for English and Serbian. These graphs inserted XML tags <s> and </s> into texts. The benefits of this approach is that sentence graphs can be easily modified to handle the peculiarities of some specific texts, old orthographic norms, etc, as was done when preparing texts for the Jane Austen English/Serbian corpus (Krstev and Vitas 2011).

The prepared texts were aligned by XAligne.[6] The default behavior of Xalign does not involve cognates, but one can introduce his own (numbers and proper names) to improve the alignment in the bootstrap process. Xaligne may take as input partially aligned texts and use this information to build more reliable alignment (Paumier and Dumitriu 2008). The result of the alignment process is an XML file that contains links between numbered segments of source and target texts. In the example in Figure X-2~~Figure X 2~~ segments n3 and n4 of a source text from the group l1 that is linked to the segment x3 of a target text, while segments x22 and x23 of a target text form the group l2 that is linked to the segment n23 of a source text.

Formatted: Font: 10 pt

```
<xptr id="x22" from="ID (n22)"/>
<xptr id="x23" from="ID (n23)"/>
<link targets="n3 n4" type="linking" id="l1"/>
<link targets="x22 x23" type="linking" id="l2"/>
 ...
 <link targets="n2 x2"/>
 <link targets="l1 x3"/>
 <link targets="n5 x4"/>
 ...
 <link targets="n23 l2"/>
```

Figure X-2 An excerpt form an output file produced by XAlign

The alignment can be performed in two different ways. One can use Xalign directly and its concordancer (from a command line) or one can use some software environment in which Xalign has been integrated. One such system is ACIDE described in (Utvić and al. 2008) that facilitates the alignment process and enables the production of various representations of aligned texts from an output file: formats Vanilla[7], HTML, TMX as well as different alphabets (e.g. Cyrillic or Latin for Serbian).[8]

5 Aligned corpora and texts

For aligned corpora that include Serbian consisting primarily of literary texts the precision of alignment is of great importance (Gelbukh et

[6] a tool developed by Patrice Bonhomme, Thi Minh Huyen Nguyen and Sean O'Rourke, http://led.loria.fr/outils/ALIGN/align.html
[7] http://nl.ijs.si/telri/Vanilla/
[8] http://korpus.matf.bg.ac.rs/prezentacija/paralelni.html

al. 2006). In that respect for the majority of aligned texts the links between text segments are 1:1. This distinguishes our corpora from enormous corpora like Europarl. Namely, one experiment that we performed with the aim of producing a French/Spanish corpus for which we used Europarl showed that it contained texts (e.g. ep-99-09-17) in which links of the type n:0 were quite often (more then 50 such links, which meant that 1/7 of all segments from the French text did not exist in the Spanish text). It is understandable that for literary texts such situations can not be tolerated. For this reason, all the texts included in our corpora were manually checked. As a result of manual verification all segments missing in translation were identified as well as differences between original and translated texts that occurred because a target text was not actually translated from a version that was chosen as a source text. Also, the number of links of type 2:1, 1:2 or 2:2 in our corpora is insignificant and it is usually due to translators' decisions as how to deal with direct speech in his/her translation.

As an example of differences between original and translated texts that are due to the use of different versions of the original – one for translation and other for alignment – is segment number 247 from the Verne's novel *Around the World in 80 Days* which is in the French version available on web as follows:

<s id="n247">-- Tout de suite.</s>

English, Serbian and versions in many other languages have an addition to this segment (which originates from some later French editions, for instance 63rd edition from 1884):

<s id="n247">"At once." <!-- Missing in FR: Only I warn you that I shall do it at your expense." --> </s>
<s id="n247">-- Odmah. <!-- Missing in FR: Samo upozoravam vas da ću to učiniti na vasx trošak. --></s>

The following example that illustrates links of type 2:1 is from another of Verne's novels, *A Fantasy of Dr Ox*. In the original French the adverbial phrase (in italic) ends with a colon which also marks a sentence ending. The Serbian translator has appropriately moved this segment in the middle of the sentence while the English translator has solved the problem differently (Table X-1~~Table X-1~~).

Even when the number of segments in a source and a target text are equal, which is rarely the case, it is necessary to manually check all links and perform any corrections necessary. This often means that it is necessary to correct the automatically performed segmentation. Namely,

Formatted: Font: 10 pt

the correction of links of type *n*:*m* (where $n * m \neq 1$) may entail insertion or deletion of tags and <s>, as illustrated in previous examples.

French	English	Serbian
<s id="n437">*De temps à autre*: </s> <s id="n438">« Je crois que ça mord, Suzel, disait Frantz, sans aucunement lever les yeux sur la jeune fille.</s>	<s id="n437">*From time to time* Frantz would say, without raising his eyes,-"I think I have a bite, Suzel." </s>	<s id="n437">-Mislim da je zagrizao, Suzelo, - govorio bi Franc *s vremena na vreme* i ne gledajući mladu devojku.</s>

Table X-1 An example of 2:1 link between segments

5.1 The content of Aligned Corpora that include Serbian

The interest of authors of this paper in aligned texts started long ago. Back in 1988 they prepared a small aligned corpus of instructions for use of drugs (Krstev et al. 1988). The purpose of this small experiment was lexical analysis that suggested that machine translation in this domain and for related languages of the former Yugoslavia was feasible. Later, a multilingual corpus containing the basic laws of the former Yugoslavia in all of its official languages was compiled and aligned concordances produced (Krstev and Vitas 1994). The purpose of this experiment was to investigate the correctness of translations and results showed that versions of these legislative acts differed in some important issues.

The production of aligned corpora that include Serbian got its impetus with the TELRI project[9] in the scope of which two multilingual texts were produced. The first one was an aligned multilingual version of Plato's *Republic* in which 16 languages were involved, including Serbian. The second one was, now well-known and much used, multilingual version of Orwell's *1984* (Erjavec et al. 1998). The first version of this resource involved 7 languages, while the actual version has 17 languages; However, Serbian was there from the beginning and it also entered in the English/Serbian literary Corpus.

5.2 French/Serbian aligned corpus

[9] Trans-European Language Resources Infrastructure (http://telri.nytud.hu/)

This corpus primarily contains works of French literary classics from the 19th century but also some contemporary authors and it is constantly being updated. An older version of this corpus, its content and size are described in (Vitas and Krstev 2006). Recently more works from Balzac and Jules Verne were added to it as well as works of some 20th century authors like Albert Camus and Amin Maalouf. A considerable effort has been made in the past few years to obtain and process French translations of some important Serbian authors. As a result, this corpus today includes works by Ivo Andrić, Danilo Kiš, Rastko Petrović, Bora Stanković and others. The total size of the Serbian part of this corpus is 1.5 MW, while the French part contains approximately 1.9 MW. Some parts of this corpus are available on web.[10]

5.3 English/Serbian aligned corpus

The first text in this corpus was, as already mentioned, the Serbian translation of Orwell's *1984*. Later some novels from some classical English authors were added (Jane Austen, Tomas Hardy, etc.), as well as some contemporary novels (by Hemingway, Dan Brown, J.K. Rowling, etc). As with the French/Serbian corpus a number of Serbian contemporary novels translated to English were collected (Kiš, Velikić, Basara, etc.). This corpus also contains some French classical novels translated to English (Verne, Stendhal, etc.), that were obtained as a by-product while compiling the French/Serbian corpus. A more detailed description of this corpus is given in (Krstev and Vitas, 2011).

Besides literary works this corpus also contains newspaper articles from the Southeast European Times Web site.[11] The total size of the Serbian part of this corpus is 1 MW.

5.4 Intera corpus

In the scope of the Intera project a Serbian/English aligned corpus was compiled that contains texts from law, business, education and health-care domains. It contains one million words in both languages. The Serbian part of this corpus was semi-automatically lemmatized and PoS tagged,

[10] http://www.korpus.matf.bg.ac.rs/ (a password required)
[11] http://www.setimes.com/

while the English part was tagged with TreeTagger[12] (Schmid, 1994). This corpus was used for the evaluation of different taggers for Serbian (Popović 2010), and for some experiments in machine translation and term extraction (Gavrilidou et al. 2005).

5.5 Around the World in 80 days

Jules Verne is the most translated French author, and the second most translated author in the world.[13] As a consequence, Verne's novels are available in e-form in many languages which makes them good candidate for aligned corpora. This is, however, not their only advantage. The novel *Around the World in 80 Days* is suitable for experiments with named entity recognition as can be easily seen from the very first sentence from this novel:

> En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens -- maison dans laquelle Sheridan mourut en 1814 --, était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres,...

For experiments with shallow parsing (recognition of adverbial phrases for dates) this text has also been successfully used.

On the other hand, the Unitex distribution for French includes this same novel to demonstrate the system's functionalities. It also contains as an evaluation resource a version of this text in which MWU expressions (nouns and adverbs) were manually tagged (Laporte et al., 2008).[14] These circumstances make it possible to compare the annotations used by e-dictionaries for various languages. One experiment in that direction was already done for Serbian and Bulgarian (Vitas et al. 2008), (Krstev et al. 2008). Since the Nooj corpus processing system[15] uses similar resources as Unitex for overlapping set of languages, this text appears to be suitable for comparison and possible standardization of morphosyntactic annotation systems for considerable number of languages (Stanković et al. 2011).

As we considered all the advantages of this novel, we decided to prepare its multilingual version. Today this resource contains twenty translations all aligned to the French original. All Slavic, Roman, Balkan

---

[12] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[13] http://www.unesco.org/xtrans/bsstatexp.aspx?crit1L=5&nTyp=min&topN=50
[14] http://infolingu.univ-mlv.fr/DonneesLinguistiques/Corpus/Visualisation.html
[15] http://www.nooj4nlp.net/pages/nooj.html

languages are represented, as well as two English and two German versions, Dutch, Hungarian and Chinese. Aligned versions of public domain texts will be available on web by the end of 2011 in a format suitable for processing by Unitex.

5.6 Serbian/Serbian corpus

This aligned corpus, the smallest of all, contains five novels for which two Serbian (or Serbo-Croatian) translations exist.[16] It contains Voltaire's *Candide*, Verne's novels *Around the World in 80 days* and *Dr. Ox's Experiment*, Hemingway's *The Old Man and the Sea*, and the first 14 days from Jan Potocki's novel *The Manuscript Found in Saragossa* (one translation was done from French, the other from Polish). One example form Potocki's novel aligned with French and Polish segment is given in Figure X-3~~Figure X-3~~. In the two Serbian translations only four words are identical; nevertheless, both sentences have the same meaning. This corpus is used for studies of free word order in Serbian and the use of paraphrases.

Formatted: Font: 10 pt

<tu> <tuv>Obudziłem się na głos pustelnika, który zdawał się niesłychanie cieszyć, widząc mnie zdrowego i wesołego.</tuv>
<tuv>Probudio *me* je glas isposnika koji se izgleda vrlo obradovao *što me vidi* zdravog i veselog.</tuv>
<tuv>Probudi *me* isposnik, veoma zadovoljan *što me vidi* živa i zdrava.</tuv>
<tuv>Je fus réveillé par l'ermite, qui parut très content de me voir sain et sauf.</tuv></tu>

Figure X-3 An example from Potocki's multilingual text with two Serbian translations

6 Examples

6.1. Lemmatized concordances

The Unitex software we presented in Section 2 and Serbian lexical resources presented in Section 3 enable versatile exploitation of the corpora presented in Section 5. Besides queries supported by most corpora

[16] The Croatian translation before separation of Croatian from Serbo-Croatian in 1991.

processing systems using regular expressions on string of characters, Unitex supports regular expressions not only on string characters but on lemmas as well. Moreover, all information stored in used e-dictionaries for particular lemma can be used as well. We will illustrate this with two small examples. Our first query will be very simple: it consists of one pattern <ljubav> that does not search for all occurrences of the string *ljubav* but rather for all occurrences of all the inflected forms of the noun *ljubav* 'love'. Three examples from the aligned concordances are given in Table X-2~~Table X-2~~.

Formatted: Font: 10 pt

Search patterns can be even more complex since all grammatical, syntactic and semantic codes and markers from e-dictionaries of texts can be used in them. For instance, the pattern <A+Nprop~Hum> retrieves adjectives (A) derived from proper names (+NProp) that are not used for humans (~Hum) in the Serbian text. This pattern retrieves various occurrences, such as *engleskog*, *francuskog*, *indijski*, *sevrskog*, *njufaundlendskim*, and some others, all used as translations of expressions like *Newfoundland puppy* or prepositional phrases like *clay of Sèvres*.

English	Serbian
Indeed, she had no *taste* for a garden;	U stvari, nije posedovala neku naročitu *ljubav* prema bašti.
And this address seemed to satisfy all the fondest wishes of the mother's heart, for she received him with the most delighted and exhulting *affection*.	Ovakvo njegovo oslovljavanje je, kako se činilo, potpuno zadovoljilo najtoplije želje majčinskog srca, jer ga je dočekala sa velikom radošću i *ljubavlju*.
Her *heart* and faith were alike engaged to James.	na šta mu je ona poklonila svu svoju veru i *ljubav*.

Table X-2 Pattern <ljubav> used on Jane Austen's *Northenger's Abbey*

6.2. Regular derivation

The phenomenon of regular derivation is as specific to Serbian as for other Slavonic languages. New lemmas can be derived starting from one lemma whose meaning can be predicted from it. The examples of such derivational processes for a noun *glumac* 'actor' are derivations of possessive and relational adjectives, *glumac*.N → *glumčev*.A (that belongs to an actor) and *glumački*.A (relating to or concerning an actor), gender motion *glumac*.N:m → *glumica*.N:f (a female actor) and amplifiers of meaning (diminutive *glumac* → *glumčić* and augmentative *glumac* → *glumčina*). From adjective lemmas abstract nouns can be derived, (for

instance *veseo* 'cheerful'→ *veselost* 'cheerfulness'), and form verb lemmas verbal nouns (for instance. *glumiti* 'to act'→ *glumljenje* 'acting'). The phenomenon of the regular derivation has immediate effect upon the structure of entries in monolingual dictionaries (Vitas and Krstev 2005a), (Vitas and Krstev 2005b) as well as bilingual dictionaries (Krstev and Vitas, 2004). The results of exploitation of aligned texts are also influenced by this phenomenon. Namely, if a search keyword is not a Serbian word, and its Serbian equivalent has a potential for regular derivation, they will be then retrieved and presented in concordances. Obviously, the search with a corresponding Serbian keyword would yield different results. For instance, for a French key <*mendiant*> ('beggar') results are obtained presented in Table X-3~~Table X-3~~.

Formatted: Font: 10 pt

French	Serbian
Il y a toujours un *mendiant* philosophe, un châtelain bourru,[…]	Uvek ima po jednog *prosjaka* filozofa, mrgodnog vlastelina, […]
Sa mère, une *mendiante*, l'amenait chez eux tous les matins.	Njegova majka, *prosjakinja*, dovodila ga je k njima svakog jutra.

Table X-3 The gender motion in French and Serbian text

It is so because in French the gender motion is treated as inflection. In Serbian, a key <prosjak> retrieves only the first example, whereas for the retrieval of the second example the key <*prosjakinja*> 'beggarwoman' should be added.

Regular derivation is important for proper names as well. From many proper names possessive and relational adjectives can be derived, as illustrated by examples from Table X-4~~Table X-4~~.

Formatted: Font: 10 pt

French	Serbian
Les bourgeois de *Chavignolles.N* désiraient les connaître	Građani *Šavinjola.N*, međutim, želeli su da ih upoznaju
Deux jours après l'émeute de *Chavignolles.N*,	Dva dana posle *šavinjolske.A* pobune,
Vous etes la femme de *Pipo.N*?	-- Vi ste žena *Pipova.A*?

Table X-4 Possessive and relational adjectives in French and Serbian texts

Examination of regular derivation on aligned texts points to the possibility to structure a lemma in Serbian as a meta-lemma whose forms would correspond to various forms of a corresponding lemma in the other language of a bitext.

On the other hand, examples of regular derivation reveal a frequently used translation technique. Namely, some lexical gaps in Serbian are often overcome by the use of amplifiers. For instance, French/Serbian aligned

texts show that a French noun *la table* ‘table’ is consistently translated with Serbian *sto*. Names for special kinds of tables, like *le guéridon, la console, la table tournante*, for which in Serbian specific terms do not exist are all translated with a diminutive *stočić*. Similarly, *le tonneau* is in Serbian always *bure*, while *le fût, la barrique le tonneau de faïence* are all translated by the diminutive *burence*. *L'oeil-de-boeuf, la lucarne, le soupirail, la croisée, le carreau* are as well translated with the diminutive *prozorče* as opposed to *prozoru* which is usual translation only for *la fenêtre*. Augmentatives are more rarely used and usually in pejorative meaning – for instance, French *la rosse* (‘bad horse’) is translated in Serbian as *konjina*, the augmentative of *konj* ‘horse’).

6.3 Semantic properties

Color terms in Czech, English and Dutch are analyzed in (Čermák 2011) in order to show that the way colors are perceived is influenced by a language, even when basic colors are concerned. Moreover, analysis of aligned texts shows that independently of differences in the perception of basic colors and even when a term for a color exists it can be translated in a way clear to a reader.

The E-dictionary of Serbian uses special markers for colours in the field for syntactic and semantic markers +Col: for instance, crven.A+Col ‘red’. Thus, with a very simple query <A+Col> it is possible to obtain concordances of equivalences that would confirm or reject Čermak's hypothesis. Thus, in the French/Serbian aligned corpus the following results are obtained for colors of the visible light spectrum.

The adjective *crven* ‘red’ in our corpus predominantly corresponds to the French *rouge*, but this Serbian adjective is used for the shades of red, for instance for the French *écarlate*. The adjective *žut* ‘yellow’ corresponds to the French *jaune*, but this Serbian adjective is used to translate shades like *chamois* or *nankin* as well - for instance, FR: *gilet de nankin* – SR: *prsluk od otvoreno žutog nankina*, where *nankin* entails the colour yellow. For French *fauve* the combination of two colors is used *crvenožut* ‘red-yellow’, besides *mrkožut* ‘dark-yellow’ and *mrkosmeđ* ‘dark-brown’. Similarly, the adjective *plav* ‘blue’ corresponds to the French *bleu*, but also to its shades: *gorge-de-pigeon* ‘pigeon blue’, *azur*, *livide*, *marin pincé*, while the adjective *zelen* ‘green’, corresponds not only to *vert*, but also to shades *herbacé* and *glauque*.

Colours can be vague, but in these cases correspondence in the corpus was almost absolute: *rougeâtres* - *crvenkast*, *rose* - *ružičast*, *verdâtre* - *zelenkast*, *bleuâtres* - *plavičast*, *grisâtre* - *sivkast*.

When used for shades of hair or beard the choice of colors is different. For instance, in Serbian hair can be *plav* 'blue', and in that case in French *blond* is used (almost yellow), but French *une lumière blonde* corresponds to Serbian *žuta svetlost* 'yellow light'. Similarly, a beard in Serbian can be *riđ* not *crven* 'red', therefore translation of French *la barbe rouge* is *riđa brada*, not *crvena brada*.

Colors in Serbian can be used to intensify meaning, as in *la plus basse jalousie – najcrnja zavist* 'the worst envy'. The description of colors is by no means exhausted by this analysis, as illustrated by following examples in Table X-5~~Table X-5~~:

Formatted: Font: 10 pt

French	Serbian
Rastko: vert piquant, incandescent et violet, tandis que l'autre en reçut des ombres immensément obscures	ljuto zelena, ognjena i violetna a druga dobi ogromno tamna osenčenja
Verne: teint coloré depuis les sombres nuances du cuivre jusqu'au blanc mat, mais jamais jaune	boje kože od bakarnosmeđe do svetlosmeđe, ali nikad žute

Table X-5 Colors and their translation in Serbian and French texts

6.4 Local grammars

Besides queries that use lemmas and their properties from e-dictionaries, Unitex enables the formulation of much more complex queries that represent combinations of properties from dictionaries and various additional conditions. Such queries represented in the form of graphs enable the extraction from text of complex objects, like analytical tenses or named entities. An excerpt from concordances that were produced by the Serbian graph for analytical tenses – in extracted examples most of them correspond to French simple tenses is presented in the first row of Table X-6~~Table X-6~~. Concordances in the second row of Table X-6~~Table X-6~~ are result of the recognition of dates in Serbian texts.

Formatted: Font: 10 pt

Formatted: Font: 10 pt

French	Serbian
En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens – maison dans laquelle Sheridan *mourut* en 1814 –, *était*	Godine 1872, u kući broj 7 u Ulici Sevil-rou Bar-lington Gardenz, u kojoj *je* 1816.godine *umro* Šeridan, *stanovao je* gospodin Fileas Fog, jedan od

habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il *semblât* prendre à tâche de ne rien faire qui *pût* attirer l'attention.	najčudnovatijih i najzapaženijih članova londonskog Reform-kluba, iako *je izgledalo* da se on trudi da ne učini ništa što *bi moglo* na njega privući pažnju**.**
Donc, à partir de ce moment, onze heurs vingt-neuf du matin, ce *mercredi 2 octobre 1972*, vous êtes à mon service.	Dakle, počev od ovog trenutka, jedanaest časova i dvadeset devet minuta pre podne u *sredu 2. oktobra 1872. godine*, vi ste u mojoj službi.

Table X-6 Recognition by local grammars in Verne bitext

6.5 The alignment on the word level

Some experiments with the alignment on the word level were performed on the *Intera* English/Serbian corpus (Obuljen 2009). This corpus was appropriate for the task because both monolingual parts were lemmatized and PoS tagged, as explained in section 5.4. Various measures for ranking the translation pairs were tested, and the most suitable measure was chosen (V is the set of word forms i of a target language for which $C(i|y) > 0$):

$$rank_y(x) = (C(x|y) / \Sigma_{i \in V} C(i|y)) * (C(x|y) / C(x))$$

In this formula $C(x)$ is the frequency of occurrences of a word x in the target language, while $C(x|y)$ represents the frequency of a word x from the target language occurring in the same segment with the chosen word y from the source language. Summing is done for all words of the source language. This formula represents a variant of the geometric average. For instance, for the English lemma *crime*, if using this kind of ranking the top six best candidates are:

zločin, rank=0,0143	'crime'
ratni, rank=0,0086	'war'
suđenje, rank=0,004	'trial'
počinjen, rank=0,004	'committed'
vojni, rank=0,003	'military'
civilni, rank=0,002	'civil'

In the presented results word forms are replaced by their lemma.

For source language words ranked in this manner, a manual evaluation was performed on one sample that encompassed all words with a

frequency greater than 50, and for 500 words from other frequency classes – frequency 1, 2-5, 6-20, and 21-29. This evaluation showed that the most frequent Serbian words that cover 87.92% of the whole corpus were aligned with the adequate corresponding word in 83.08% cases.

Words from the frequency range 6-20 were correctly aligned in 52.4% cases, while the result for words with frequency 21-49 was 61.8%. The results for words with lower frequency were not good. 72.4% of words occurring just once were not correctly aligned, while 59.6% of words occurring between two and five times were incorrectly aligned. The experiments in this direction will continue with testing of other measures based on harmonic average and weight-harmonic average.

7 Concluding remarks

Our work on aligned corpora with Serbian language included will continue. We will continue to collect and align new texts, as well as improve tools already developed that facilitate this procedure. However, our primary focus will be the development of tools that would enable full usage of lexical resources for both languages involved for literary and linguistic research based on aligned corpora. The examples from section 6 illustrate the power of this approach. To that end, the subsystem of Unitex that deals with aligned texts will be improved to enable the more efficient exploitation of bitexts, for instance through a web interface.

Bibliography

Christ, Oliver. 1994. A modular and flexible architecture for an integrated corpus query system. *Papers in Computational Lexicography (COMPLEX '94)*: 22–32, Budapest, Hungary.

Čermák, František. 2011. Colour Terms in Three Languages: Their Distribution and Function. *The Second Conference on Slavic Corpora SlaviCorp 2011*, (forthcoming) Dubrovnik, Croatia.

Erjavec, Tomaž, Ann Lawson and Laurent Romary, eds. 1998. *East Meets West – A Compendium of Multilingual Resources*. TELRI Association e.V.,IdS, Mannheim.

Friburger Nathalie, Anne Dister and Denis Maurel. 2000. Améliorer le découpage des phrases sous Intex, *Revue Informatique et Statistique dans les Sciences Humaines*, vol. 36, n°1-4: 181-200

Gavrilidou, Maria, P. Labropoulou, M. Monachini, S. Piperidis and C. Soria. 2005. Building Multilingual Terminological Resources, in Piperidis, S. and Paskaleva, E. eds. *Proceedings of the International Workshop "Language and Speech Infrastructure for Information Access in the Balkan Countries"*, *RANLP 2005*: Borovets, Bulgaria.

Gelbukh, Alexander F., Grigori Sidorov and José Ángel Vera-Félix. 2006. A Bilingual Corpus of Novels Aligned at Paragraph Level. *Proceedings of the 5th International Conf. FinTAL 2006*: 16-23.

Gross, Maurice and Dominique Perrin, eds. 1989. *Electronic Dictionaries and Automata in Computational Linguistics*. LNCS 377, Springer

Krstev, Cvetana, S. Jović-Puač i D. Vitas. 1988. The Analyse of the Sublanguage of Instruction of Using Drugs in Serbo-Croatian and Slovene*, ROJP IV*: 249-255, Institut Jožef Štefan, Ljubljana, Slovenia.

Krstev, Cvetana and Duško Vitas. 1994. Concordances of Aligned Texts, *Zbornik radova XXXVIII konferencije ETRAN*: 229-230, Niš, Serbia.

Krstev, Cvetana and Duško Vitas. 2004. Restructuring Lemma in a Dictionary of Serbian. *Zbornik 7. mednarodne "Informacijska druzba IS 2004", Jezikovne tehnologije*: 103-107 Ljubljana, Slovenija.

Krstev, Cvetana, Gordana Pavlović, Duško Vitas and Ivan Obradović. 2004. Using Textual and Lexical Resources in Developing Serbian Wordnet, *Romanian Journal of Information Science and Technology*, vol. 7, No. 1-2: 147-161, Publishing House of the Romanian Academy.

Krstev, Cvetana, Ranka Stanković, Duško Vitas and Ivan Obradović. 2006a. WS4LR - a Worksation for Lexical Resources. *Proceedings of the Fifth Interantional Conference on Language Resources and Evaluation*: 1692-1697, Genoa, Italy.

Krstev, Cvetana, Duško Vitas and Agata Savary. 2006b. Prerequisites for a Comprehensive Dictionary of Serbian, *Proceedings of the 5th International Conference FinTAL 2006*: 552-564, Turku, Finland.

Krstev, Cvetana, Svetla Koeva and Duško Vitas. 2008. A Dictionary-based Model for Morpho-Syntactic Annotation, *Proceedings of the 2nd LAW, in scope of the Sixth Interantional Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Krstev, Cvetana and Duško Vitas. 2011. An Aligned English-Serbian Corpus. *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I*: 495-508 *Belgrade,* Serbia.

Laporte, Éric, Takuya Nakamura and Stavroula Voyatzi. 2008. A French Corpus Annotated for Multiword Nouns. *Proceedings of the Language Resources and Evaluation Conference (LREC), Workshop Towards a Shared Task on Multiword Expressions*: 27-30, Marrakech, Morocco.

Obuljen, Aljoša. 2009. Kvantitativna metoda za poravnanje reči dvojezinkog korpusa. Internal report, Faculty of Mathematics, University of Belgrade, Serbia.

Paumier, Sébastien and Dana-Marina Dumitriu. 2008. Editable text alignments and powerful linguistic queries. *27th International Conf. on Lexis and Grammar (LGC'08)*: 117–125, L'Aquila, Italy.

Paumier, Sébastien. 2010. *Unitex 2.1 User Manual*, http://igm.univ-mlv.fr/~unitex/UnitexManual2.1.pdf

Popović, Zoran. 2010. Taggers applied on texts in Serbian, *Infotheca*, Vol. XI (2): 21-38, Belgrade, Serbia.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of the International Conference on New Methods in Language Processing*: 44-49.

Stanković, Ranka, Ivan Obradović, Duško Vitas, Cvetana Krstev and Miloš Utvić. 2011. On the compatibility of lexical resources, *NooJ 2011 Proceedings*, Dubrovnik (forthcoming)

Utvić, Miloš, Ranka Stanković and Ivan Obradović. 2008. Integrisano okruženje za pripremu paralelizovanog korpusa. In *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*: 563-578, LitVerlag, Muenster.

Vitas, Duško and Cvetana Krstev. 2005a. Structural derivation and meaning extraction: a comparative study on French-Serbo-Croatian parallel texts. In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, ed. G. Barnbrook, P. Danielsson, M. Mahlberg: 166-178, The University of Birmingham Press, Birmingham.

Vitas, Duško and Cvetana Krstev. 2005b. Regular derivation and synonymy in an e-dictionary of Serbian. *Archives of Control Sciences*, Volume 51 (3): 469-480, Polish Academy of Sciences.

Vitas, Duško, Cvetana Krstev and Eric Laporte. 2006. Preparation and exploitation of Bilingual Texts. *Lux Coreana*, No. 1: 110-132

Vitas, Duško and Cvetana Krstev. 2006. Literature and Aligned Texts. In *Readings in Multilinguality*, eds. Milena Slavcheva, Galia Angelova and Kiril Simov: 148-155, Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Vitas, Duško, Cvetana Krstev and Denis Maurel. 2007. A note on the Semantic and Morphological Properties of Proper Names in the Prolex Project. *Lingvisticae Investigationes*, *Special issue on Named Entities: Recognition, Classification and Use*, eds. Satoshi Sekine and Elisabete Ranchhod, Vol. 30(XXX), No. 1: 115-134, John Benjamins Publishing Company, Amsterdam, Philadelphia.

Vitas, Duško, Svetla Koeva, Cvetana Krstev and Ivan Obradović.2008. *Tour du monde* through the dictionaries, *Actes du 27eme Colloque International sur le Lexique et la Gammaire: 249-256,* L'Aquila, Italy.
Vitas, Duško and Cvetana Krstev. 2012. Processing of corpora of serbian using electronic dictionaries. Prace Filologiczne (fothcoming)