Literature and Aligned Texts

Duško Vitas[1] and **Cvetana Krstev**[2]
[1]Faculty of Mathematics, Studentski trg 16
[2]Faculty of Philology, Studentski trg 3
University of Belgrade
Belgrade 11000, Serbia
{vitas,cvetana}@matf.bg.ac.yu

Abstract

In this paper we describe the structure of the aligned French-Serbian literary corpus and some results obtained in analyzing it. The alignment was performed on the subsentential level, which means that each text has been segmented into the `<seg>` elements in order to obtain in all cases the 1:1 alignment of the original and its translation. This means that in some cases the sentences were further segmented in two or more parts. The insertion of `<seg>` tags was initially done automatically, and obtained results were then manually validated. Some results obtained by analyzing this corpus are presented.

1 Introduction

One concise outline of the state of the art in the field of the aligned corpora was recently presented in (Tufis 06)[1]. Most of the presented corpora represent certain sublanguages, particularly of legal or administrative texts. Such choice is motivated either by the aims of the corpus development – terminology extraction, establishment of the translation equivalents, etc – or by the fact that the large amounts of this type of texts can be easily accessed in electronic form and the copyright issues for them can be easily settled.

The aligned corpora of literary texts are relatively rare, especially those where one of the languages is a Slavic language. One early attempt to develop the multilingual aligned text was the production of the aligned Plato's *Republic*, in the scope of TELRI project (Erjavec *et al.* 98). The produced multilingual aligned text included several bitexts where one of the languages in the pair was a Slavic language. This experience showed that the alignment task was by no means the easy one, especially if for a text with complex logical layout it was difficult to establish the original version. Namely, the differences of translations on some of the languages were significant, so the alignment process did not yield the convincing results. The multilingual aligned text was later produced in the scope of the same project for Orwell's *1984*, which proved to be a more suitable text and was since used in many applications.

Probably the largest aligned corpus of literary Slavic texts consists primarily of the translations from English to different Slavic languages (Barentsen 05)[2]. After some modifications were done to the logical layout of the texts, in order to satisfy, for instance, the upper limit of the paragraph length, the texts were aligned on the level of paragraphs. The need of developing the aligned corpora of Slavic languages, especially South Slavic languages, was recently stressed in (Paskaleva & Pacovski 06) and illustrated by the aligned text of the novel *The Master and Margarita* by Mikhail Afanasievich Bulgakov (Russian/Bulgarian/Macedonian). An interesting discussion on problems of development of the corpus of literary texts aligned at the paragraph level were discussed in (Gelbukh *et al.* 06) and illustrated in the example of the English-Spanish corpus of novels. In our opinion, the context in the corpora aligned at the paragraph level is usually too wide for their effective and precise exploitation.

On the other hand, literary texts illustrate more precisely then any other texts translator's strategies, doubts and solutions. Also, aligned corpora of literary texts, to the contrary of corpora of some sublanguage, offer variety of possibilities for fruitful usage. Language learning, lexicography, and contrastive studies, are just a few examples of possible applications. More specific application would be to use one of the languages in the aligned text as a meta-language for the verification of the equivalent meanings in the other langauge. Such approach was proposed for the refinement of synsets (sets of synonymous literals) in Wordnets (Krstev *et al.* 04).

It should be noted that literary texts are, as a rule, precisely translated and offer the insight into the language and grammatical phenomena that

[1]see also: http://www.cs.unt.edu/∼rada/wa/

[2]see also: http://home.medewerker.uva.nl/a.a.barentsen/

cannot be observed in other registers. Some of the examples are illustrated in this paper. The newspaper texts are often translated with greater freedom. One example is given in the fragment from the translation to Serbian of *Le Monde diplomatique* (see Appendix A). The third sentence in paragraph is omitted, and the adjective *tentaculaire* (engl. *tentacular*) from the French text is translated by the phrase *poput ogromne hobotnice sa dugim pipcima* (engl. *like the giant octopus with the long tentacles*). This metaphoric translation is further distributed through the whole paragraph (these occurrences are given in italic).

2 Corpus description

2.1 Sources

The corpus consists mostly of the classical works of French literature. The choice of texts was guided by their literary value, but also by their accessibility in electronic form. The largest part of French texts was obtained in electronic form from some web sites[3], the other texts were scanned and corrected. The French originals span the time period from the end of the XVIII century to the present time, while all Serbian translation were done after 1926. They were all done by the most prominent translators from French to Serbian, and they are in accordance with the contemporary Serbian norm. The corpus also contains one text that is translated from Serbian to French (text no. 6 in Appendix B). Some Serbian texts were obtained by OCR, the others were retyped, and one was obtained from the translator. The edited versions of Plato's *Republic* and Orwell's *1984*, according to (Erjavec *et al.* 98), were added to the corpus.

The format of the source French texts varied from the plain ASCII without any mark-up, to pdf, including some texts from *Gallica* that represented just the scanned images of the original texts. Most of the Serbian texts were in MS-Word format while others were ASCII texts with paragraph mark-up.

Corpus contains two Serbian translations of some texts. The inclusion of multiple translations was motivated by the possibility to investigate different variations in contemporary Serbian, in the first place on morphological level, by using French as a meta-language. The multiple translations exist for three texts:

- The corpus contains the translations of Voltaire's *Candide* from 1934 and 1964. The analysis of these translations showed that the more recent translation is the correction of the older translation from 1934.
- *The manuscript found in Saragosa* from Jan Potocki is translated in Serbian from the French abridged edition[4] and from the Polish integral edition[5] .
- Jules Verne's *The journey around the world in 80 days* is represented in corpus by two independent translations in Serbo-Croatian[6].

These translations were independently aligned with French original, as well as between themselves. The text of Potocki's novel was aligned to the Polish electronic version of text[7], that was corrected according to the paper edition, while several parts of the Verne's novels were aligned with a number of other languages as Greek, English, Spanish, and Slovene.

2.2 The size of corpus

The corpus contains the integral texts of Serbian translations and their French equivalents. In some cases, for instance for the *Anthology of French Fantastic Novels*, the Serbian translation of good quality was chosen first and then the corresponding original was sought for. The corpus contains the total of 14 texts written by 21 authors (see Appendix B for the full description). The precise data on the size of the corpus are given in Table 1. It can be noted that the French texts are longer then Serbian texts (approximately by 25%, if measured by number of tokens). This can be explained by the usage of articles in French, the omission of subject in Serbian, etc. On the other hand, the number of types is much bigger in Serbian due to the Serbian rich morphology.

2.3 Corpus preparation

Logical layout of texts was marked using the minimal set of tags: `<head>`, `<p>` and, `<seg>`. Wherever possible, the pagination of the source text was retained as a comment. The bibliographic

[3] see http://abu.cnam.fr/ and http://gallica.bnf.fr/

[4] days from 1st to 14th, edited by Roger Caillois, Gallimard, 1958. see no. 12 in Appendix B.

[5] Jan Potocki: *Rękopis znaleziony w Saragossie*, Czytelnyk, Warszawa, 1965

[6] Belgrade 1962; Zagreb, 1961

[7] see http://univ.gda.pl/∼literat/sarag/index.htm

No. of novel	Tokens in French	Types in French	Tokens in Serbian	Types in Serbian
1.	68843	10975	53666	13712
2.	27223	5191	22547	6021
3.	72253	9496	58661	11903
4.	93550	14662	78120	19385
5.	4710	1437	3743	1668
6.	39124	5132	31043	5774
7.	7158	1780	6083	2086
8.	33001	5457	29282	7617
9.	119489	10900	103127	17380
10.	960	472	769	509
11.	3654	1156	2957	1311
12.	65029	8178	51640	12024
13.	113419	11576	90026	16832
14.	154000	9266	108595	14184
Σ	802413	38247	640259	65162

Table 1: Corpus size (number of novel correspond to the numbers given in Appendix C)

description of the source text was given according to *TEI Guidelines*[8] in the TEI header element.

Tags `<head>` for the titles were inserted manually, while tags `<p>` for the paragraphs were inserted automatically or semi-automatically, depending on the format of the source text.

The recognition of sentence boundaries and insertion of `<seg>` tags was done by the finite-state transducer *Sentence.grf* that was implemented in the programming system *Intex*[9]. This transducer identifies the problems caused by the ambiguous usage of punctuation marks and with the high precision inserts a `<seg>`-tag at the beginning of the sentence. It enables, for instance, the distinction between the usage of the ASCII character 2E as a full stop from the other possible usages, like *Prof. X. Savary* in French or *7. 5. 2007.* in Serbian. This transducer is language dependent: for instance, a semi-colon in French is always at the end of the sentence, while this is not the case for Serbian.

All texts were aligned using the program *Xalign*[10]. The associated program *Concordancier* enables visualization and relinking of the concordances of the aligned texts and detection of the alignments that are not 1 : 1. Many-to-one correspondences can have various origins. The most common of them are:

- *Differences between an original and its translation.* For instance, the Serbian translation of Flaubert's *Bouvard et Pécuchet* contains a draft of the end of the novel that is not attached to the French text that served as a source. Thus, before the alignment process the Serbian text had 40 paragraphs more than the French original.

- *The omission of some parts of the text in translation.* For instance, the publisher obtained the text no. 9 (Appendix C) by scanning the previous edition. In this process one page was omitted by mistake, so the translation is shorter from the original by 14 sentences. The translators can make the similar mistake: in the same text seven sentences were not translated, in *Tours du monde en 80 jours* eight sentences, etc.

- *Differences in the paragraph marking.* The original and its translation often have different numbers of paragraphs. Although the paragraph represents the unit of the logical layout of one document, translators sometimes separate it into two or more paragraphs. For instance, if the original text contains sequence *...X said:* `</seg><seg>`*"..."* it can be transformed into *...X reče:* `</seg></p><p><seg>` – ... in the translation. Different document formats and various ways of their acquisitions can also lead to discrepancies between the number of paragraphs in the original text and its translation.

- *Differences in the sentence segmentation.* This type of many-to one correspondences is the most frequent one and it happens mostly because of the various orthographic issues that influence the segmentation process. For

[8] see http://www.tei-c.org/P5/
[9] see http://intex.univ-fcomte.fr/
[10] see http://led.loria.fr/outils/ALIGN/align.html

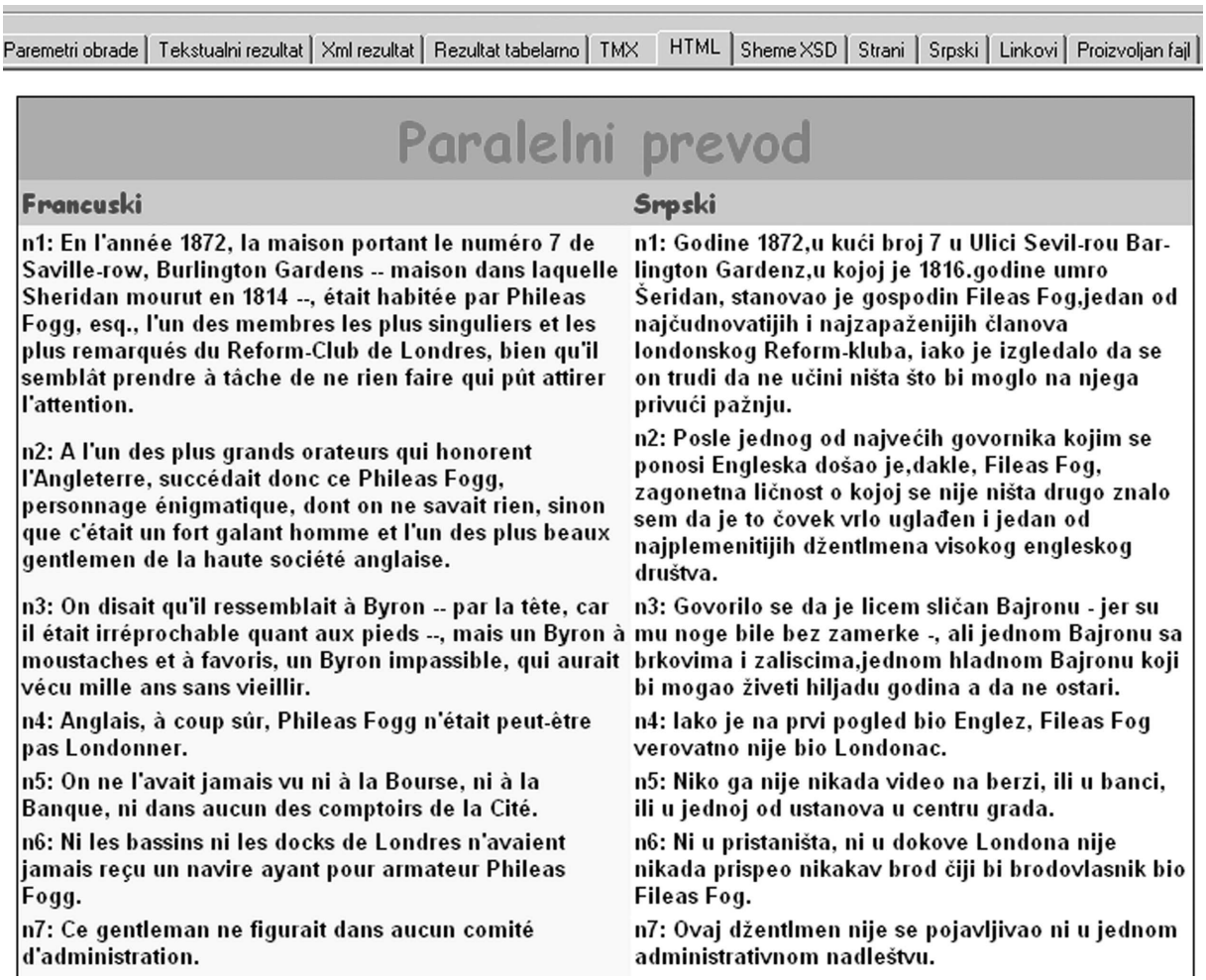

Figure 1: The HTML visualization of the French-Serbian aligned text of *Tours du monde en 80 jours*

instance, in the following segment from the novel *Bouvard et Pécuchet*:

> <seg> – "Mon Dieu, oui ! </seg>
> <seg> On pourrait prendre le mien à mon bureau!" </seg>
> <seg> – Eh, bože, naravno, mogao bi mi je ko uzeti u kancelariji! </seg>

the exclamation mark is used in the original, which is omitted in the translation, so there is no sentence ending at that point and the 2:1 correspondence is obtained. The opposite case can be found in the next segment:

> <seg> – "Messieurs, je vous écoute ! quel est votre mal ? " </seg>
> <seg> – Gospodo, slušam vas! </seg>
> <seg> Na šta se žalite? </seg>

where the alignment is 1 : 2. Texts were manually edited in all such cases by insertion or omission of one pair of tags </seg>...<seg> in order to obtain the alignment 1 : 1 in all cases. At the same time, the validation of the whole alignment process was performed.

The results obtained by the described alternations in document's logical layout can be illustrated by some figures related to the novel *Bouvard et Pécuchet*: at the beginning, the initial French text had 3258 paragraphs (tags) and 7163 sentences (<seg> tags), while the Serbian text had 3672 paragraphs and 6488 sentences. After the editing procedure both texts had 3108 paragraphs and 7075 sentences. Thus, it can be said that the alignment procedure aligns subsentential segments, that is, the <seg> tag is used to mark the equivalent segments that are whole sentences or parts of the sentence.

2.4 The corpus processing

The processing was performed in three steps. In the first phase all texts were independently processed using *Intex* as the processing tool which enabled their segmentation into sentences. Besides that, French and Serbian lexical resources integrated into the *Intex* system were applied to the

particular texts in order to obtain the insight into their lexical structure. As a byproduct, this procedure also enabled the correction of some spelling errors.

In the second phase, texts prepared in the first phase were separately aligned using the system Xalign. For each of the obtained bitexts the TMX format was produced using the system WS4LR (Krstev *et al.* 06), as well as HTML format that enables the visualization (see Figure 1).

In the third phase all texts were assembled and processed with the system IMS/CQP[11] that enabled the exploitation of the corpus as a whole.

3 Some examples

3.1 The interjections

One of the peculiarities of the dialogs in the literary texts is the usage of interjections. We will consider only three French interjections: *ah!, eh! oh!* and their translations. Their role is interesting for two reasons. First, the interjections potentially mark the end of the sentences. Second, as all three interjections exist in Serbian as well, they could be regarded as cognates. The frequencies of their usage, however, does not support the validity of these suppositions: there are only 297 occurrences of these interjections in Serbian texts in comparison to 749 occurrences in French texts:

French	**Serbian**
ah 220	*ah* 173
eh 352	*eh* 73
oh 177	*oh* 51

From 352 occurrences of *eh*, 302 are in the expression *eh bien!*.

According to the French-Serbian (Serbo-Croatian) dictionary (Putanec 89), the suggested translations of these interjections are:

French	**Serbian**
ah!	*o(h)! jao! ah!*
eh!	*eh! ah! uh! aj!*
eh bien!	*no! dakle!*
oh!	*o! oh! ah!*

The concordances of the aligned corpus reveal the following equivalents:

French	**Serbian**
ah!	*ah! e! aha! ali! no! o! oh! a!*
eh!	*eh! he! o!*
oh!	*oh! eh! ah! ooo!*

[11] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

In particular, the interjection *et bien!* has been translated freely:

French	**Serbian**
eh bien!	*dakle! pa lepo! i onda? pa dobro!* *da, pa? pa onda? pa šta!* *eto vidiš! šta veliš! e pa lepo!...*

This example shows that the translation of interjections is often determined by translator's understanding of the context.

3.2 The expression *sans doute*

Among the French expressions containing *doute*, dictionary (Robert 04) lists the expressions *sans doute* as the adverbial expression with description *selon toutes les apparences* (engl. *according to all the appearances*) and adds the comment that today this expression bears some doubt: *peut-être (maybe), probablement (probably).* On the contrary, the adverbial expression *sans nul (aucun) doute* has the meaning *certainement (without any doubt).* The Serbian equivalent *bez sumnje* according to the Serbian explanatory dictionary (Stevanović 76) has the meaning of French *sans aucun doute*. The bilingual French-Serbian dictionaries rarely indicate this subtle difference.

The analysis of the occurrences of the form *doute* in corpus gives the following results. In the French part of the corpus *doute* occurs 366 times, and almost all of them, except 58, are part of the phrases *sans (aucun) doute*. The French phrase *sans doute* occurs 270 times, while phrase *sans aucun doute* occurs 38 times. In the Serbian part of the corpus the noun *sumnja* occurs 143 times, and it is part of the phrase *bez sumnje* only 69 times.

The translator has to decide between two possibilities relying on his/her own intuition. The corpus search with the phrase *sans doute* retrieves the following equivalents in Serbian:

French	**Serbian**
sans doute	*verovatno, nesumnjivo, jamačno,* *van svake sumnje, sigurno, očigledno,* *bez sve sumnje, naravno, svakako,...*

Translation depends on a context, so the following equivalences can be spotted *aveugle sans doute = slepa za sve* (engl. *blind for everything*).

The phrase *sans aucun doute* is translated as *sigurno, bez sumnje, bez ikakve sumnje.*

If the equivalences are looked for in the only text that is translated from Serbian to French,

then the different set of possible equivalences is retrieved:

French	Serbian
sans doute	*valjda, verovatno, naročito, jesam, baš*
sans aucun doute	*bez sumnje*

Translation of the phrase *sans doute* shows that even the top-class translators, in this case to Serbian, can be misled to use the literal translation of the phrase that has the diachronic meaning. It is interesting to notice that in this case the dictionary can not be of much help since it is not clear when did the meaning of *sans doute* move from denoting certainty to probability.

3.3 Diminutives

In Serbian language diminutives and augmentatives can be produced in regular manner from most of the nouns (Vitas 04). Dictionaries, like the explanatory dictionary (Stevanović 76) record them only exceptionally, while in the bilingual lexicography there is no systematic way to describe this phenomenon that is on the border of the inflection and the derivation (Vitas & Krstev 04). Some of the diminutives that illustrate this phenomenon in the Serbian part of the corpus are *baroničica* besides *baronica* (engl. *baronette*), *Cigančica* but not *Ciganka* (engl. *Gipsy woman*), *divljačić* and also *divljak* (engl. *savage*), *kajganica* but not *kajgana* (engl. *scrambled eggs*) etc. In the Serbian part of corpus approximately 200 forms of diminutives occur. Corpus reveals two strategies for the usage of diminutives. The first one translates French sequence *petit N*, where *N* is a noun, with a form of diminutive, and the second one tries to fill the lexical gap for the concept that is not lexicalized in Serbian. The examples from the corpus for the first strategy are:

French	Serbian
une petite veine	*žilica*
une petite bouteille	*flašica*
la petite bête	*životinjica*
les petites bricoles	*stvarčice*
le petit ruban	*mašnica*
les petites bouilles	*čorbice*
un petit banc	*klupica*

Translators resort to diminutives also in cases when French lexeme contains some other modification of the basic lexeme, as in cases:

French	Serbian
quelques gouttes	*kapljice*
faible portion	*delić*
la cordelette (< *le corde fine*)	*konopčić*
le corbeille (< *le panier léger*)	*korpica*

The examples excerpted from the corpus that illustrate the second strategy are given in Table 2.

The listed examples show that the translators often resort to the usage of diminutives when they cannot find the more appropriate translation equivalent in Serbian.

3.4 The proper names

The proper names can be regarded as potential cognates that can be used for the correction of the results of the alignment (Krstev *et al.* 05) and (Vitas & Krstev 04). However, the translation of proper names from French to any language with rich morphological system generates three types of correspondence problems:

- Some problems arise from the possible use of two different alphabets (Latin and Cyrillic) and from the fact that in Serbian translation, the proper names are transcribed according to the phonetic orthography. For instance, *Bouvard* is transcribed as *Buvar*, *Passepartout* as *Paspartu*, etc., and that disables the direct usage of algorithms that are used by alignment software, as LCSR (Longest Common Subsequence Ratio) (Melamed 01) or Levenshtein distance.

- French personal names can occur in Serbian translation as a personal name and its possessive adjective in some of the inflective forms. The translation of a toponym is a toponym and its relational adjective. This phenomenon is illustrated by the following examples of the translations of proper names:

French	Serbian
Bouvard	*Buvar* realized as **N:** *Buvar, Buvara, Buvaru, Buvaru, Buvarom* and as **AdjPoss:** *Buvarov, Buvarova Buvarove, Buvarovi, Buvarovih, Buvarovim, Buvarovo, Buvarovoj, Buvarovom, Buvarovu*
Thérèse	*Tereza* realized as **N:** *Tereza, Tereze, Terezi, Terezu, Terezo, Terezom* and as **AdjPoss:** *Terezine, Terezinih*
Paris	*Pariz* realized as **N:** *Pariz, Pariza, Parizom, Parizu* and as **AdjPoss:** *pariska, pariske, pariski, pariskim, pariskom*

	N	"PETIT" N	N+DEM
Sr	*brod*	*mali brod*	*brodić*
Fr	*le navire, le bateau, le paquebot*	*une goélette*	*une goélette*
Sr	*soba*	*mala soba*	*sobica*
Fr	*la chambre*	*une petite chambre*	*le cabinet, une petite pièce, l'arrière boutique*
Sr	*vrata*	*mala vrata*	*vratanca*
Fr	*la porte*	*une petite porte*	*les battants, une petite porte*
Sr	*čovek*	*mali čovek*	*čovečuljak*
Fr	*l'homme*	*un petit homme*	*le bonhomme, le petit homme*
Sr	*sto*	*mali sto*	*stočić*
Fr	*la table*	*la petite table*	*une console, un guéridon, une table de nuit*
Sr	*grad*	*mali grad*	*gradić*
Fr	*la ville*	*la petite ville*	*le petit bourg = un gros village!*

Table 2: In the column N+Dem, the examples of diminutives are given that are used to fill the lexical gap.

- In many cases, the frequencies of proper names are not equal in the original and its translation. The frequency data for the corresponding proper names in our corpus are:

French	**Serbian**
Bouvard 618	*Buvar.***N** 617
Phileas 316	*Fileas.***N** 314
Fogg 655	*Fog.***N** 673
Passepartout 423	*Paspartu.***N** 433
Thérèse 238	*Tereza.***N** 238
Paris 106	*Pariz.***N** 109

The difference in the number of occurrences do not necessarily mean that there actually exists the $1-1$ correspondence in the number of cases that is equal to the smaller number of two frequencies. One example of the complex anaphora is given by the next segment:

<seg id="n2049"> Enfin, Bouvard et Pécuchet s'adressèrent à **Larsonneur.** </seg>
 ...
 ...
<seg id="n2053"> Par Gorju, ils s'en procurèrent une douzaine, **lui** expédièrent la moins grande – les autres enrichirent le muséum. </seg>

<seg id="n2049"> Najzad se Buvar i Pekiše obratiše **Larsoneru.** </seg>
 ...
 ...
<seg id="n2053"> Preko Goržija nabaviše jedno tuce tih sekira, **Larsoneru** poslaše najmanju, a ostalima obogatiše svoj muzej.</seg>

The pronoun *lui* from the French segment n2053 refers to the personal name *Larsonneur* that was mention in the segment n2049. The translator, however, chose not to use the pronoun, believing perhaps that there would be a too big distance between the pronoun and its referent.

4 The conclusions

The first results in exploitation of the aligned French-Serbian corpus of literary texts indicate that in translations plenty of solutions can be found that are not recorded in bilingual French-Serbian dictionaries. Also, this corpus enables the analysis of translation strategies in resolving the lexical gaps or ambiguities in the original text, as well as discovering the sources of the inconsistencies in translation.

The further work in the aligned corpus exploitation might encompass the identification of the corresponding structures in the source and target language in the sense in which it has been suggested in (Blanco 01):

Un système de TA doit donc être basé non sur des dictionnaires bilingues (ni, à plus forte raison, multilingues) mais sur [...] des descriptions lexicales de différentes langues effectuées d'après les mêmes principes.

The existence of tools and resources in the same format for both languages, French and Serbian, makes the continuation of the experiments performed on this corpus feasible.

5 Appendix A: Non-literal translation – *Le Monde diplomatique* (May 2001)

5.1 French:

<head>La pieuvre publicitaire</head>
<seg id="n1"> Tentaculaire, étouffante, oppressive, la publicité ne cesse d'étendre ses domaines d'intervention. </seg> <seg id="n2"> Elle a récemment conquis de nouveaux territoires, en particulier ceux de la galaxie Internet. </seg> <seg id="n3"> Le chiffre d'affaires publicitaire sur la Toile, en France l'an dernier, avant la crise actuelle, a dépassé le milliard de

francs, soit plus que les recettes publicitaires des salles de cinéma. </seg> <seg id="n4"> Sous la forme discrète du parrainage, son champ d'intrusion ne connaît pratiquement plus de limites. </seg> <seg id="n5"> Par ce biais quasi clandestin, elle est parvenue à investir, ces dernières années, l'art, la culture, la science, l'éducation, et même la religion. </seg> .

5.2 Serbian:

<head>Čudovišna reklama</head>
<seg id="n1"> *Poput ogromne hobotnice sa dugim pipcima, kojima će stegnuti i ugušiti* sve što joj se nadje na putu, reklamna delatnost neumorno širi polje svog dejstva. </seg> <seg id="n2"> Nedavno je osvojila nova područja, medju kojima je posebno značajna Internet galaksija. </seg> <!-- missing segment --> <seg id="n3"> *Njeni kraci dopiru svuda;* najzad, sa ulogom sponzora koju je preuzela, reklama stiče neograničene mogućnosti. </seg> <seg id="n4"> *Uvlači se* poslednjih godina, neupadljivo i gotovo kradomice, u umetnost, kulturu, nauku, obrazovanje, čak i u religiju. </seg>

6 Appendix B: Corpus composition

1. *Anthology of French Fantastique* (edited by Zoran Mišić, Orfej 37, NOLIT, Beograd, 1964), including extracts from: J. Cazotte: *Le Diable amoureux*, D.A.F. de Sade: *Juliette*, Ch. Nodier: *Smarra*, G. de Nerval: *Aurelia*, H. de Balzac: *L'élixir de longue vie* and *Séraphta*, P. Mérimée: *La Vénus d'Ille*, G. de Maupassant: *Horla*, *Qui sait?* and *La nuit*, Villiers de L'Isle-Adam: *L'Éve future*, *Véra* and *Le tueur de cygnes - Tribulat Bonhomet*, Comte de Lautramont: *Les Chants de Maldoror*, G. Appolinaire: *Le Roi-Lune*;
2. P. Louys: *La femme et le pantin*;
3. J. Verne: *Tours du monde en 80 jours*;
4. G. Flaubert: *Bouvard et Pécuchet*;
5. P. Besson: *Zodiaque amoreux*;
6. B. Blagojević: *Sve zveri što su sa tobom* (French translation: *L'arche de Boba*);
7. Voltaire: *Micromégas*;
8. Voltaire: *Candide ou l'optimisme*;
9. D.A.F. de Sade: *Justine*;
10. J.-P. Proudhon: *Idée générale de la Révolution au XIXe siècle* (extract);
11. F. Arrabal: *L'enterrement de la sardine*;
12. J. Potocki: *Manuscrit trouvé à Saragosse*;
13. G. Orwel: *1984*;
14. Platon: *La République: du régime politique*

References

(Barentsen 05) A. A. Barentsen. Aspac – amsterdam slavic parallel aligned corpus. Personal communication, Novi Sad, 2005.

(Blanco 01) X. Blanco. Dictionnaires électroniques et traduction automatique espagnol-français. *Langages*, 143:143–166, 2001.

(Erjavec *et al.* 98) T. Erjavec, A. Lawson, and L. Romary (Eds.). *East meets West A compendium of Multilingual Resources*. TELRI Association e.V., IdS, Mannheim, 1998.

(Gelbukh *et al.* 06) A. Gelbukh, G. Sidorov, and J.A. Vera-Felix. A bilingual corpus of novels aligned at paragraph level. In T. Salakoski and. F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing, FINTAL 2006*, LNAI, 4139, pages 16–23. Springer Verlag, 2006.

(Krstev *et al.* 04) C. Krstev, D. Vitas, G. Pavlović-Lažeti 'c, and I. Obradović. Using textual and lexical resources in developing serbian wordnet. *Romanian Journal of Information Science and Technology*, 7:147–161, 2004.

(Krstev *et al.* 05) C. Krstev, D. Vitas, D. Maurel, and M. Tran. Multilingual ontology of porper names. In Z. Vetulani, editor, *Proceedings of 2nd Language and Technology Conference*, pages 116–119. Wydawnictwo Poznańskie Sp. z o.o., Poznań, 2005.

(Krstev *et al.* 06) C. Krstev, R. Stanković, D. Vitas, and I. Obradović. Ws4lr - a worksation for lexical resources. In *Proceedings of the Fifth Interantional Conference on Language Resources and Evaluation, LREC'06*, pages 1692–1697. Genoa, Italy, 2006.

(Melamed 01) D. Melamed. *Empirical Methods for Exploiting Parallel Texts*. The MIT Press, Cambridge, MA, London, 2001.

(Paskaleva & Pacovski 06) E. Paskaleva and I. Pacovski. Aligning the translation – a ppossible strategy for creation of aligned corpora (for south-slavic languages). In S. Koeva and M. Dimitrova-Vulchanova, editors, *The Fifth International Conference "Formal Approaches to South Slavic and Balkan Languages", FASSBL V*, pages 113–117. Institute of Bulgarian Language, Bulgarian Academy of Sciences, Sofia, 2006.

(Putanec 89) V. Putanec. *Dictionnaire français-croate ou serbe*. Školska knjiga, Zagreb, 1989.

(Robert 04) P. Robert. *Dictionnaire Le Petit Robert*. Le Robert, Paris, 2004.

(Stevanović 76) M. Stevanović. *Rečnik srpskohrvatskoga književnog jezika*. Matica srpska : Matica hrvatska, Novi Sad, Zagreb, 1967-1976.

(Tufis 06) D. Tufis. Cross-lingual knowledge induction from parallel corpora. In S. Koeva and M. Dimitrova-Vulchanova, editors, *The Fifth International Conference "Formal Approaches to South Slavic and Balkan Languages", FASSBL V*, pages 15–18. Institute of Bulgarian Language, Bulgarian Academy of Sciences, Sofia, 2006.

(Vitas & Krstev 04) D. Vitas and C. Krstev. Intex and aligned texts. In C. Muller, J. Royaut, and M. Silberztein, editors, *Intex pour la linguistique et le traitement automatique des langues*, pages 255–270. Presses Universitaires de Franche Comté, Besançon, 2004.

(Vitas 04) D. Vitas. Morphologie dérivationnelle et mots simples: Le cas du serbo-croate. In C. Leclere, E. Laporte, M. Piot, and M. Silberztein, editors, *Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis and Lexicon-Grammar (Papers in honour of Maurice Gross)*, volume 24 of *Lingvisticae Investigationes Supplementa*, pages 629–640. John Benjamin, Amsterdam/Philadelphia, 2004.