A Sample French-Serbian Dictionary Entry based on the ParCoLab Parallel Corpus

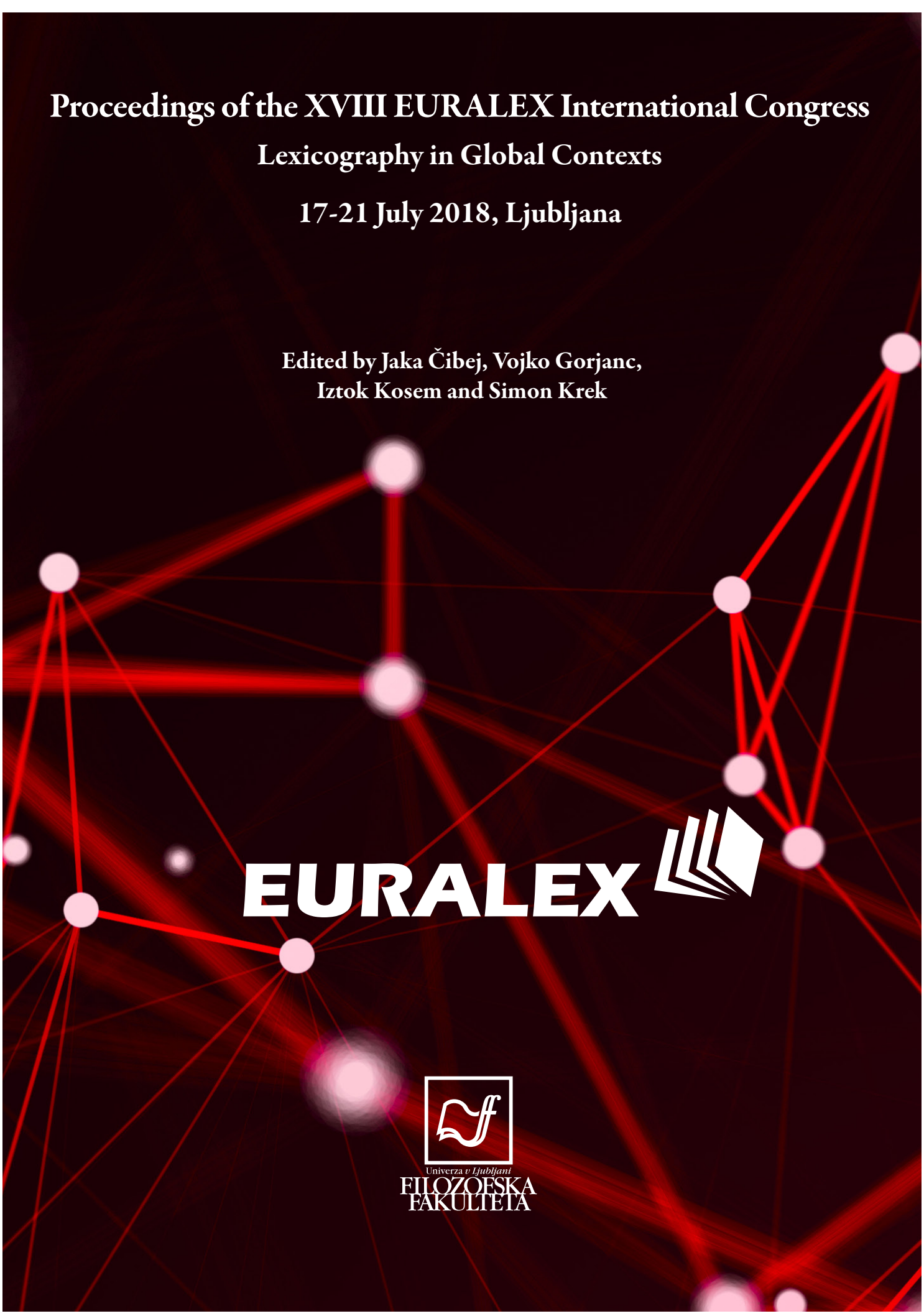
Saša Marjanović, Dejan Stosic, Aleksandra Miletic

Proceedings of the XVIII EURALEX International Congress

Lexicography in Global Contexts

17-21 July 2018, Ljubljana

Edited by Jaka Čibej, Vojko Gorjanc,
Iztok Kosem and Simon Krek



Univerza v Ljubljani
FILOZOFSKA FAKULTETA

A Sample French-Serbian Dictionary Entry based on the *ParCoLab* Parallel Corpus

***Saša Marjanović*[1], *Dejan Stosic*[2], *Aleksandra Miletic*[2]**
[1] *Faculty of Philology, University of Belgrade, Serbia*
[2] *CLLE, Université de Toulouse, CNRS, Toulouse, France*
E-mails: sasa.marjanovic@fil.bg.ac.rs, dejan.stosic@univ-tlse2.fr, aleksandra.miletic@univ-tlse2.fr

Abstract

It has already been shown in the state- of the art in lexicography that the bilingual dictionary making process can be improved by relying on parallel corpora. The aim of this paper is to present such an application of the *ParCoLab* parallel corpus, a searchable trilingual 11 million token electronic database of aligned texts in French, Serbian and English, developed at the University of Toulouse (France) in cooperation with the University of Belgrade (Serbia). In this paper, we first point out the shortcomings of the leading general French-Serbian dictionaries, which were made using traditional lexicographic methods. We pay special attention to the treatment of the equivalents offered. Taking the case of the French adjective *sale* 'dirty' as an example, we show that the *ParCoLab* parallel corpus makes it possible to: 1) have quick and easy access to meanings missing from the existing dictionaries and to corresponding equivalents; 2) find new equivalents that are not included in any of the existing dictionaries, and which are in some cases the most common translation solutions; 3) order equivalents by their relative corpus frequency; and 4) disambiguate different usages through adequate contextual examples. The solutions we offer are shaped into a sample dictionary entry.

Keywords: parallel corpus, lexicography, dictionary, bilingual, French, Serbian

1 Introduction

General and phraseological Serbian-French dictionaries have already been examined by Marjanović (2013a, b) taking the example of animal metaphors and similes processing as an example. In both cases, the author underscored the incoherent selection of content, the unsystematic processing and inadequate sense representation, poor or lack of illustrative material and the lack of guidelines on the use of entries and their equivalents. However, only one paper has been published so far on the existing French-Serbian dictionaries (Stanojević-Knežević 2005), which is in fact a bibliography of French-Serbian lexicography. The paper suggests that French-Serbian dictionaries are outdated, incomplete and full of mistakes. However, there is no mention of whether the effectiveness of existing French-Serbian dictionaries has been tested on real texts, as Bujas (1975) did in Yugoslav lexicography for the English-Serbo-Croatian dictionary that he was editing. He gave his students the task of examining how well the equivalents from the latest edition could be applied to translating selected newspaper and magazine articles. The analyzers read the assigned texts thoroughly, checked every word in the dictionary and carefully annotated their findings. After taking all the students' proposals into consideration, Bujas introduced 2,200 new items into the new edition of the dictionary and concluded that the verification of the effectiveness of a bilingual dictionary on contemporary texts can significantly improve the quality of the dictionary (cf. Bujas 1975: 204). Luckily, such manual verification of effectiveness is not necessary anymore: owing to the advent of information technologies, this can be accomplished using electronic parallel corpora (cf. Hartmann 1994: 291-292).

2 Parallel Corpora and Bilingual Lexicography

Two types of electronic corpora fall under the term 'parallel corpus' – translation corpora and comparable corpora (Atkins & Rundell 2008: 476). However, this term is mostly used to signify translation corpora in the metalexicographic literature, and it will also be used in this sense throughout this paper. A translation parallel corpus is an electronic database that contains texts written in one language and their translations into one or more languages. The corresponding texts are paired at the sentence level. Such a corpus, among other things, makes it possible to find translation equivalents and contexts in which they are used swiftly and easily by comparing aligned pairs of texts (cf. Atkins & Rundell 2008: 478). This is why one of the applications of parallel corpora is in the bilingual dictionary making process. It has been unambiguously proved on specific examples that parallel corpora, regardless of the language pair, can lead to a set of good equivalents, usable in lexicography, which are not listed in existing dictionaries (see, among others, Hartmann 1994; Roberts 1996; Roberts & Montgomery 1996; Dickens & Salkie 1996; Teubert 2002; Citron & Widmann 2006; Salkie 2008; Goossens 2012; Perdek 2012; Perko & Mezeg 2012; Zavaglia & Galafacci 2014). Furthermore, it has been shown that parallel corpora are useful when paired bilingual texts, which both represent the translation of the same source text in a third (so-called *pivot*) language, are compared with each other (e.g. the French-German corpus of Plato's *Republic*, in Teubert 2002). Considering that the use of a parallel corpus can help the lexicographer gain insight into which equivalents are used in translation, and in which contexts, these resources can largely compensate for the lexicographer's intuition and increase the objectivity of lexicographic work.

However, while the usefulness of parallel corpora is not called into question in the metalexicographic community, for a long time there has been no mention of dictionaries based on parallel corpora in applied lexicography. Atkins and Rundell's (2008: 477) survey on the EURALEX discussion list as to whether there is a single dictionary publisher who uses a parallel corpus brought no new findings. The only dictionary systematically based on a parallel corpus was the *Bilingual Canadian Dictionary* (Roberts 1996; Roberts & Montgomery 1996), unfortunately unfinished. Furthermore, the parallel corpus was only "used at the end of the translation stage to ensure that no good equivalents have been missed" (Roberts & Montgomery 1996: 460).

Some of the reasons for this situation include the poor availability of parallel corpora for most language pairs, the high cost of creating a new parallel corpus, the inadequate size of existing parallel corpora, and the unreliability of the translations they include (cf. Salkie 2008). Whereas little can be done about the first three problems, when it comes to the last reason, it has been shown that even bad translations can serve a purpose in the dictionary making process (cf. Marjanović 2017: 487-492): the frequency of incorrect translations in certain language units indicates problematic points, to which lexicographers should pay more attention. Atkins and Rundell (2008: 478) also stress that parallel corpora offer too much evidence, which has then to be carefully considered by the lexicographer. A detailed examination of all corpus findings slows the lexicographer down, which is not profitable in commercial lexicography. This obstacle, however, is considerably alleviated by new technologies that contribute to the automation of lexicographic work. For all these reasons parallel corpora are gaining steam in modern e-lexicography (see Héja 2010; Lindemann 2013; Lindemann et al. 2014; Škrabal & Vavřín 2017).

The goal of this paper is to show how existing lexicographic processing, and especially the processing of equivalents in French-Serbian dictionaries, can be improved by relying on a new language resource, the *ParCoLab* parallel corpus. First, we provide an overview of existing traditional French-Serbian dictionaries on a specific example and point to their shortcomings (Section 3); next, in Section 4, we describe the *ParCoLab* corpus, and finally, in Section 5, we present the results of the analysis of the

corpus-based equivalents and their lexicographic usability. Section 6 then summarizes the findings and offers a sample dictionary entry based on the results of the analysis.

3 The Big Four: French-Serbian Dictionaries and Their Shortcomings

In order to check the quality and to address the shortcomings of the existing French-Serbian (Serbo-Croatian, Croatian) dictionaries, in this section, we will illustrate and discuss in detail the processing of equivalents on the example of a common French entry *sale* 'dirty' in the following four leading general bilingual dictionaries: Marković (1980), Jovanović (1991), Putanec (1995) and Točanac et al. (2017). We chose this entry because we believe it to be a representative sample: it is polysemous, its correspondent in the primary sense (*prljav* 'dirty') belongs to a wide synonym set at the systemic level, and, depending on the context, it can be translated with a set of equivalents. By analyzing this entry we can gain insight into how lexicographers order different senses, how they discriminate them and how they list multiple equivalents. The excerpt from the four aforementioned dictionaries looks as follows:

> **sale** [sal] *adj.* prljav, nečist; gadan, odvratan; pokvaren || *c'est une affaire sale* to je prljav posao; *c'est un caractère sale* odvratan karakter; *linge sale* prljavo rublje; *avoir les mains sales* imati prljave ruke; *être sale comme un porc, comme un peigne* biti vrlo prljav (Marković 1980)
>
> **sale** [sal] a. (posle imenice) prljav, nečist, prašnjav, blatnjav, neuredan, odvratan; neprijatan, nepristojan; (pre imenice) fam. pokvaren; gadan (Jovanović 1991)
>
> **sale** [sal] a. zamazan, nečist, prljav, gadan; osiromašen (o rudi); mutan (o boji); nesiguran (o obali); ružan, nepristojan, pokvaren, gadan (Putanec 1995)
>
> **sale** [sal] *adj.* 1. prljav, nečist, neuredan • *mains ~s* prljave ruke • *histoire ~* prljava priča • *argent ~* prljav novac (od šverca, ...) 2. *fam.* grozan, užasan, odvratan, gadan • *~ temps* grozno vreme, *fig.* loši dani • *~ coup* nizak, težak udarac • *~ type* odvratan tip (Točanac et al. 2017)

All four dictionaries are made using traditional lexicographic methods, mostly via bilingual adjustment of French monolingual dictionaries. In other words, lexicographers read the definitions and examples in French dictionaries and noted equivalents they could recall. Therefore, we base our analysis on the senses of the French word *sale* represented in the *Le Petit Robert* dictionary, which was the one that all lexicographers used, according to their respective bibliographies. In order to conduct our analysis, we established eight different senses: (1) 'covered or marked with an unclean substance', (2) '(of a colour) not bright or pure', (3) 'immoral, or dishonest', (4) 'concerned with sex in a lewd or obscene way', (5) '(of money) obtained through illegal or disreputable means', (6) 'used to emphasize how bad something is', (7) 'used to emphasize one's disgust for something', and (8) 'used to emphasize one's disgust for someone'. Table 1 shows which of these senses are present in the four dictionaries:

Table 1: Distribution of senses in the analyzed French-Serbian dictionaries

Dictionary / Sense	1	2	3/4	5	6	7	8
Marković 1980	+	-	+	-	+	-	-
Jovanović 1991	+	-	+	-	+	-	+
Putanec 1995	+	+	+	-	-	-	-
Točanac et al. 2017	+	-	+	+	+	+	+

Two facts can be noticed immediately: only two senses are present in all four dictionaries, and the largest dictionary (Putanec 1995) processes only the first three senses. However, two other senses were introduced into Putanec's dictionary, with the equivalents *osiromašen* and *nesiguran*, but these

two senses are too technical, so they should not have been processed in a general dictionary, especially not when common senses under items 5-8 from general language were omitted. The most comprehensive dictionary is Točanac et al. (2017), but the senses are not well discriminated there, which is something we will address later on.

By reviewing the excerpts, we can also note that senses in the first three dictionaries are listed in a linear manner and that they are separated by a semi-colon, while in the fourth one, the senses are explicitly separated by numbers. Marković introduces senses 3 and 6 with "examples"[1] (*c'est une affaire sale* and *c'est un caractère sale*), while Točanac et al. (2017) do the same for senses 4, 5, 7 and 8: *histoire sale*, *argent sale*, *sale coup* and *sale type*. In the first dictionary, the examples are always given in a separate block, while in the second, they are listed within the senses they refer to. The latter implies: 1) that the sense introduced by non-contextual equivalents must be represented in the examples, or 2) if more senses are grouped in one because they share the same equivalent, the equivalent can be used in all examples. This means that the equivalents "grozan, užasan, odvratan, gadan" from the second sense are potentially interchangeable in the "examples" (e.g. *grozno* for *užasno, odvratno*, *gadno vreme*), which justifies grouping several equivalents. However, such replacements are not possible in all "examples" in the first sense: (*prljave / nečiste / neuredne ruke*, but **nečista* / **neuredna priča*, **nečist* / **neuredan novac*). This inconsistent approach and overlapping are justified to a certain extent by the target group, that is, Serbian users who intuitively know that the equivalent *prljav* in examples *prljave ruke, prljava priča, prljav novac* – and, consequently, the adjective *sale* in source collocations – does not have the same meaning. However, native French speakers use these dictionaries as well, and this would not be obvious to them.

When it comes to explicit sense indicators, we note that only Putanec (1995) – with three secondary senses, but not the fourth – and Točanac et al. (2017) – with the "example" *argent sale* – state the sense discriminator. However, it is unclear why the sense discriminator was provided only for that collocation, but not for *histoire sale*. In both cases, the adjective *prljav* is not used in its primary sense, which was introduced through non-contextual equivalents. There is no sense discriminator either in the case when the equivalent is polysemous (cf. *pokvaren* in Marković 1980 and Jovanović 1991). Consequently the user cannot know which sense is the right one.

Furthermore, within a single sense, there are equivalents that introduce a completely different meaning. For instance, in Jovanović (1991), the equivalent *odvratan* 'repulsive' and *gadan* 'disgusting' in Putanec (1995) are unjustifiably among the set of equivalents that denote the primary sense of the entry *sale*. These mistakes occur in other entries of those dictionaries too. We can also note that close, but different secondary meanings were grouped into the same sense (e.g. in Putanac 1995: *ružan*, *nepristojan*, *pokvaren*, *gadan*, where the first two denote sense 4, while the others denote sense 3 and 6-8 respectively). Although lumping in general is a justified procedure in bilingual lexicography (Adamska-Sałaciak 2006: 76-79), it can only be applied when two or more senses share one equivalent or a set of the same equivalents, which is not the case in our example. Due to the vagueness of meaning and overlapping of senses, we could not separate equivalents for our third and fourth senses with certainty in the dictionaries analyzed, so they were presented together in Table 1.

Interestingly, the order of equivalents in these four dictionaries is not the same, not even with the primary sense: three dictionaries give the equivalent *prljav* in first place, which is also the formal correspondent of the French entry, while in Putanec (1995), it is listed in third place. What is more, it is surprising that the interlingual hyponym *zamazan* is in first place. With regard to formal correspondence, it should be pointed out that this does not allow users to perceive explicitly that the equivalent

1 The term *example* is placed in quotation marks when it is not used in the metalexicographic sense, but in the sense the authors of the four dictionaries used in their forewords.

prljav has a polysemantic structure and that it mostly corresponds to the French one, as we will see in sections 5 and 6. Because of that, these dictionaries would not be able to serve as an aid to language acquisition (cf. Tarp 2008: 195-198).

Finally, equivalence itself is understood quite broadly. Often, interlingual hyponyms, co-hyponyms and even interlingual hypernyms are listed, even when it is unjustified. Examples of this can be found in the excerpts from three dictionaries (Jovanović 1991: *prašnjav* 'dusty', *blatnjav* 'muddy', *neuredan* 'untidy'; Točanac et al. 2017: *neuredan* 'untidy'; Putanec 1995: *zamazan* 'daubed').

This detailed analysis of the four dictionaries confirms that there are plenty of shortcomings in traditional French-Serbian lexicography with regard to sense representation and processing of equivalents: not all of the common senses were processed; those that were processed are often intertwined and incoherently listed, and they are not systemically discriminated with indicators or examples; equivalence is understood broadly and polysemous equivalents are not repeated even when they are used in different senses. Since we consider the processing of the entry *sale* to be a representative sample, the findings should be understood as a general image of equivalent processing in the analyzed dictionaries, as well as in French-Serbian lexicography in general. This stance is supported by the results of the analysis of similes (Marjanović 2017: 493-551) and prepositions (Stosic et al., forth.) in French-Serbian dictionaries. In Section 5, we will examine if and how the *ParCoLab* parallel corpus, briefly presented in Section 4, can contribute to improving French-Serbian lexicography.

4 A new resource: the PARCOLAB Parallel Corpus

ParCoLab is a searchable trilingual database of aligned texts in French, Serbian and English, which has been developed by the research unit CLLE-ERSS (CNRS and the University of Toulouse-Jean Jaurès, France) in cooperation with the Romance Department of the Faculty of Philology, University of Belgrade (Serbia). The corpus is freely available at the following address: http://parcolab.univ-tlse2.fr/. *ParCoLab* contains original texts in one of the three aforementioned languages, and their translations in one or both remaining languages. In other words, it is based on three distinct subcorpora, each having a different pivot language. To date, the entire corpus contains 11.1 million tokens. In the past year, the number of tokens has increased by four million (7.1 million mid-2017, see Miletic et al. 2017: 158).

Aside from quantitatively enriching the corpus, it has also been diversified by introducing many different text types. So far, literary texts and their translations make up the largest share of the corpus, because they are the most readily available. However, there are also a certain number of legal and political texts, film subtitles, web content and biology texts. The detailed distribution of tokens according to genre is listed in Table 2:

Table 2: Distribution of tokens per text type and language

Text type	English	French	Serbian	Total tokens
Literary texts	1,919,428	4,257,773	3,495,363	9,672,564
Legal texts	233,556	291,996	79,679	605,231
Web content	229,006	186,256	63,018	478,280
Film subtitles	48,383	125,919	104,935	279,237
Biology	0	40,759	35,113	75,872
Politics	0	9,529	8,576	18,105
TOTAL	2,430,373	4,912,232	3,786,684	11,129,289

The corpus data are stored in an XML format based on the TEI P5 Guidelines. The alignment of the original texts with their translations was performed using an algorithm integrated in *ParCoLab*; no external resources were used for this task. The algorithm proceeded in descending order, creating one-to-one alignments, first at chapter level, then at paragraph level, and finally at sentence level. Errors were signaled by the tool and corrected manually, which guarantees the reliability of the corpus alignments. As of yet, only a small part of the corpus includes morphosyntactic and syntactic annotations, but our current short-term efforts are focused on this task (for further information, see Miletic et al. 2017: 160-162; Miletic et al. 2016; Miletic & Urieli 2017). Queries are carried out via the ElasticSearch search engine, which is well adapted to querying data in NoSQL databases, and a query form offers quiet very good search possibilities.

Regardless of the fact that *ParCoLab* is unbalanced and small compared to large-scale web corpora, it already contains a large number of words pertaining to the core vocabulary, which enables its usefulness for lexicographic purposes in the process of making new general French-Serbian dictionaries to be tested. Our conviction is supported by the authors referred to in Section 2, who proved the lexicographic usefulness of parallel corpora that are even smaller than *ParCoLab*.

5 The ParCoLab Corpus Evidence *vs* Dictionary Evidence

We examined the same French word as in Section 3 in order to compare corpus equivalents with the ones found in the dictionaries. First, we limited the search to original French texts with their Serbian translations. We found 80 occurrences, out of which only two were not from literary texts. They encompassed six out of eight senses, and we found the following equivalents (listed alphabetically): *bezobrazan, gadan, nečist, odvratan, podao, pokvaren, prljav, zamrljan*. However, such a short list was unexpected. In order to increase the number of occurrences, and potential equivalents, we expanded the search to French translations of Serbian and English original texts. By doing this, we were able first verify whether, as stated by Citron and Widmann (2006: 255), better translation solutions are found in the source language (L1) when searching through the translated language (L2), that is, in our case, when Serbian equivalents of the French word *sale* (L2) are searched in original Serbian texts (L1). Second, we were able to verify whether, in our case, we can reach interesting translation solutions through a third *pivot language* (cf. Teubert 2002).

The content we searched through included the French-Serbian, the Serbian-French and the English-French-Serbian subcorpora, and contained approximately four million tokens, from which we extracted 277 occurrences. In Table 3, we list a detailed distribution of occurrences according to sense, text type and original language. In five cases, we could not ascribe a single sense to the analyzed lexeme (due to the unclear translation), so we listed those cases in a special column, represented by "?":

Table 3: Distribution of occurrences per sense, text type and language

Original Language	Text type	Senses								
		1	2	3	4	5	6	7	8	?
English	Literary texts	27	1	0	0	0	4	0	1	3
	Film subtitles	2	0	0	1	0	0	0	0	0
French	Literary texts	49	3	1	0	0	4	5	15	1
	Film subtitles	1	0	1	0	0	0	0	0	0
Serbian	Literary texts	119	5	1	0	0	5	2	7	1
	Film subtitles	0	0	0	0	0	0	0	18	0
TOTAL		198	9	3	1	0	13	7	41	5

The numbers from Table 3 show that seven out of eight isolated senses of the lexeme *sale* are attested in the corpus and that, in its current version, the database can represent a polysemantic structure of the lexeme *sale*. Also, the number of different equivalents has been increased. We evaluated the relevance of their inclusion in a French-Serbian dictionary by annotating them according to the level of *lexicographic potential* (LP) (Perdek 2012: 382-386)[2]. Following Perdek (2012), we established four levels of LP – *high*, *medium*, *low* and *zero* LP. *Zero* LP includes incorrect equivalents, while *high* LP includes ideal equivalents which can be used in most common contexts of a certain sense. *Medium* (cf. *average*, in Perdek 2012) encompasses good equivalents which may enter the dictionary, but need to be indicated in a proper way, because they are limited to specific contexts or require translation transformations. *Low* refers to good equivalents which correspond to a single context.

Out of the total number of equivalents, 230 (83%) are characterized by a high and medium level of LP, which is a quite encouraging result for using *ParCoLab* for lexicographic purposes. Through English, we reached good equivalents such as *mastan, neopran, običan, zamašćen*. Since there are far more equivalents from Serbian original text, they will be presented later, especially under the last sense (cf. below). In the following section, we will thoroughly analyze all the translation equivalents in order to pinpoint the specific features of each sense. This will allow us to more easily compare corpus equivalents with existing lexicographic equivalents.

5.1 Corpus Findings Per Sense

The primary sense of the lexeme *sale* is, as one would expect, the most common in *ParCoLab*: out of 198 tokens, the equivalent *prljav* is listed in 161 cases, which means that its LP is undoubtedly high. This corroborates our observation that the solution from Putanec (1995) is poor (cf. Section 3). Furthermore, if we disregard 10 cases when there was no translation, or when it was wrong, there are a dozen more equivalents for the primary sense in *ParCoLab* (see Table 4). The lexicographically marked equivalents are in italics in the table. However, in the corpus, there are no occurrences of the lexicographic equivalents *blatnjav* and *neuredan*.

Table 4: Distribution of the first sense equivalents per LP

LP	Equivalents
High LP	*prljav* (161), *nečist* (4x), neopran (3x), mastan (2)
Medium LP	zaprljan (3x), uprljan, umašćen, izmašćen, *zamazan*, zamrljan
Low LP	garav, *prašnjav*, pun prljavštine, zagađen (2x)

Such a rich set of translation equivalents can help the lexicographer who is working on a French-Serbian dictionary to select the corresponding equivalents much more swiftly and easily, depending on the target group, its needs and the situations in which the dictionary will be used.

Only Putanec (1995) lists the second meaning (‘of a color, not bright or pure’) and processes it with the equivalent *mutan*, which is a co-hyponym of the French entry. For this meaning, we found the equivalent *prljav* seven times in *ParCoLab* (e.g. *prljavobela*), while the equivalents *siv* and *gadan* appeared once each (e.g. *žutosiva*). The first equivalent (*prljav*) has a high LP, the second (*siv*) a medium one, because it is used only with certain color adjectives (e.g. white, yellow, green, blue). Both equivalents are part of compounds according to orthography norms, and their use needs to be carefully illustrated in the dictionary. The third equivalent (*gadan*) is an example of a bad equivalent with zero LP.

2 See also *OK*, *fuzzy* and *false*, in Lindemann *et al.* (2014), as well as the notion of Basic *vs* Rich Translation Equivalence, in Dickens and Salkie (1996).

The third meaning ('immoral' or 'dishonest') is illustrated by three citations; in two of them, the equivalent is *prljav* (1)-(2), and in one, it is *kvaran* (3):

(1) Pour elle, le peuple est quelque chose de *sale* et de rusé, et Prerovo, un repaire de loups.	Narod — to je njoj nešto *prljavo* i lukavo. Prerovo — vučja jazbina!
(2) C'est lâche, c'est *sale*, et petit, et commun de calomnier une femme !	To je kukavički, to je *prljavo*, i sitničarski, i nisko, klevetati jednu ženu!
(3) C'était *sale*, mais bien joué.	Bilo je *kvarno*, ali se dobro izvukao.

In our opinion, the equivalents are adequate only in the first and third citations. In the second, it would have been better if the equivalent *pokvaren* had been used, as suggested by Marković (1980), Jovanović (1991) and Putanec (1995) in their dictionaries. It would also correspond to the third translation situation: in fact, since he cheated the girl by lying to her, the young man agrees that the means was immoral, but that the goal has been achieved. Neither of the corpus equivalents is listed in the dictionaries analyzed, which is disappointing since the first equivalent has the same polysemantic structure as the entry (cf. Section 3). Putanec (1995) offers the equivalent *gadan* with *pokvaren*, which is unjustified, since these two lexemes are not interchangeable.

For the fourth sense ('concerned with sex in a lewd or obscene way'), we found one citation in *ParCoLab*:

(4) Ce mec m'envoyait des trucs très sexuels et *sales* et il y avait une fille dans son avatar.	Jedan momak mi je slao hiperseksualne, *gadne* stvari, a na avataru mu je bila devojka.

The translation equivalent proposed (*gadan*) has a medium LP. In other words, out of context, it would not refer to the adequate meaning, so it should be contextualized in the dictionary. What is more, the best option would be to structurally transform *gadna stvar* into the morphologically related noun *gadost*. When it comes to dictionaries, only Točanac et al. (2017) processes this sense and lists the equivalent *prljav* in the collocation *histoire sale*. This equivalent has a high LP, but, aside from it, with the same collocation, the adjective *mastan* is more natural. The equivalents *nepristojan* (Jovanović 1991; Putanec 1995), as well as *ružan* (Putanec 1995) correspond to the same collocation. Jovanović's equivalent *neprijatan* is only an interpretation of the sense, so it can be considered that its LP is zero.

The citation for the fifth sense ('of money, obtained through illegal or disreputable means') was not found in *ParCoLab*. We assume that it could be found in newspaper articles, which have yet to be included in the *ParCoLab*. Among the four dictionaries, only Točanac et al. (2017) lists the collocation *argent sale* and its corresponding equivalent *prljav novac*.

For the sixth meaning ('used to emphasize how bad something is'), we have found 13 occurrences in the corpus. In one case, the translation is not adequate, while in six cases, the equivalent *gadan* was used (e.g. 5, 6), and in one case – *prljav* (7):

(5) Oh, il a vécu un *sale* moment.	O, taj je preživeo *gadan* trenutak.
(6) Alors Athénaïs vomit les plus *sales* injures, les invectives les plus obscènes sur les magistrats et les grenadiers [...].	Tada Atenaida stade da izbacuje *najgadnije* psovke i najbestidnije pogrde na činovnike i grenadire [...].
(7) Pour ne pas ébruiter une si *sale* affaire, car je suis dans l'impossibilité de justifier la conduite de mon père, je vous écris au dernier moment.	Da se ne bi raščula jedna tako *prljava* stvar, jer ja sam u nemogućnosti da opravdam postupak svog oca, pišem vam u poslednjem trenutku.

A larger number of occurrences of the first equivalent can be ascribed to the text type from which the citations stem. In ordinary Serbian language, especially in example 6, the synonymous lexemes *odvratan*, *grozan*, *ružan*, *užasan* would also be used in that sense: Marković (1980) lists two equivalents

(*gadan* and *odvratan*), Točanac et al. (2017) offers four (*grozan*, *užasan*, *odvratan*, *gadan*). We cannot deduce further guidelines on the order of the equivalents based on corpus information, because the number of occurrences is small.[3]

There are seven occurrences of the seventh sense ('used to emphasize one's disgust for something'). In three occurrences, the equivalent is *prljav* (in one case, the translation is incorrect), and *lopovski*, *podao* and *gadan* occur once (e.g. 8, 9, 10 and 11, respectively):

(8)	Je ne veux même pas prononcer son *sale* nom de Shiptar.	Neću ni da izgovorim njegovo *prljavo* šiptarsko ime.
(9)	Vous l'avez érigé dans votre *sale* patelin.	Podigli ste ga u vašem *lopovskom* mestu.
(10)	[...] et méprisant la guillotine de 89 comme une *sale* vengeance.	[...] a puno preziranja prema gijotini iz 89 kao prema jednoj *podloj* osveti.
(11)	[...] regretter de [...] n'avoir à lui offrir qu'une *sale* soutane de prêtre dont elle aura peur et dégoût!	[...] žaliti što [...] moći joj ponuditi samo *gadnu* svešteničku mantiju koje se ona boji i gnuša!

Alongside these, we excerpted one quite interesting translation solution: in one citation, the adjective *sale* has been left out from the translation, but an expressive lexeme was used for the noun with which its stands in original:

(12)	j'ai quitté la *sale baraque* à Deneulin, je descends demain au Voreux avec douze Belges	Ostavio sam onu Denelenovu *straćaru*, silazim sutra u Vore sa dvanaestoricom Belgijanaca

None of the corpus equivalents are lexicographically processed, since the seventh sense was treated only in Točanac et al. (2017) within the example where the equivalents *nizak* and *težak* were provided for *sale*, which is a justifiable lexicographic solution. Unlike the previous ones, it is more difficult to present the equivalents for this sense in the dictionary, because the choice of equivalents depends on the context: it is, therefore, necessary to present it with examples.

The final sense ('used to emphasize one's disgust for someone') was only processed in Točanac et al. (2017) as the example of *sale type* with the equivalent *odvratan tip*, and there are 41 citations in *ParCoLab* for it. If we put aside the two citations with a non-existent translation and the one with an incorrect translation, there are numerous equivalents left. More than half of the citations belong to the literary text type; in them, we have identified the following equivalents: *gadan* (8x), *prljav* (4x), *odvratan* (2x), *smrdljiv*, *pogan*, *bezobrazan*, *običan*. By analyzing each individual case, we come to the conclusion that the equivalents that occur more than once have either medium or low LP, while others have either a high or medium LP. We will list the examples of good translation solutions.

(13)	Te voilà collé au mur, *sale* crapule !	Sad si ti sabijen uza zid, *gadna* huljo!
(14)	*Sales* youpins, [...] vous avez crucifié mon Dieu et vous voulez ma peau ;	*Prljavi* gadovi, [...] razapeli ste moga boga i sad hoćete moju kožu;
(15)	J'ai vu Mouquet, tu vas encore au Volcan, où il y a ces *sales* femmes de chanteuses.	Videla sam Mukea, ideš opet u »Vulkan«, gde su one *odvratne* pevačice.
(16)	Jusqu'à présent, c'était du gâteau, *sales* Youpins, mais c'est fini.	Dosad vam je bilo lepo, *smrdljivi* Čivuti, ali sad je tome kraj.
(17)	« Tu chantes, *sale* petite souris ! » Il lui serre le cou, le secoue et cherche à lui briser la tête contre le mur.	Pevaš, mišiću *pogani*! — steže mu vrat, trese ga i hoće glavu o zid da mu slupa.

3 With regard to this sense, we have to point out that in three corpus citations, the idioms *être/se trouver dans une sale passe* and *être dans un sale pétrin* occur with good equivalents. In one case, *sale* is used in the translation of the Serbian idiom: *Nisu mi čista posla → Ça m'a l'air d'une sale affaire*. Since phraseology is not the topic of this paper, we will not expand on this any further.

(18) Partie devant eux, la Mouquette s'exclamait dans l'escalier noir, en les traitant de *sales* mioches et en menaçant de les gifler, s'ils la pinçaient.

Muketa, koja je pošla ispred njih, vikala je niz mračne stepenice, nazivala ih *bezobraznom* dečurlijom i pretila da će ih išamarati ako je budu štipali.

We also found equivalents where structural transformations occurred: *jarac*, *poganija* (both with low LP) and *gad*. The lexeme *gad* is interesting because it is precisely the equivalent that could have been used for the example *sale type* in Točanac et al. (2017):

(19) il [...] cria qu'il ferait se repentir un jour le *sale monde* qui manquait de reconnaissance [...].

vikao [je] da će se jednoga dana pokajati ti *gadovi* koji ne znaju za zahvalnost [...].

In film subtitles, whose source language is Serbian, we found six different equivalents, which all are good, but have a medium LP. Those are *najobičniji* (2x) and *prljav*, and then equivalents where structural transformations of the type Adj + N → N + Adj occurred: *pička* (4x), *đubre [jedno]* (3x) and *stoka*. Aside from that, in two cases from films, and two cases from literary texts, there are no adjectives in Serbian, we only find the noun with the pejorative suffix (cf. example 25). We will list a few examples of good translation solutions:

(20) Lâche-moi, *sale* skinhead ! Pusti me, *pičko* ćelava!
(21) Tu ferais quoi ? *Sale* menteur ! Šta bi mi pokazao, *đubre jedno* lažljivo!
(22) T'es mort, *sale* Tchetnik ! Krvavu ti nedjelju jebem, *stoko* četnička!
(23) Violeta est une *sale* pute. Violeta je *najobičnija* kurva.
(24) De *sales* capitalistes. *Prljavi* kapitalisti.
(25) Qu'est-ce qu'il y a ? *Sale* pute ! Šta je, šta je, *kurvetino*?

Considering that this common sense has not been processed in French-Serbian dictionaries, all the listed solutions from *ParCoLab* represent valuable content for identifying new equivalents. With this sense, we also see a drastic difference between text types, which justifies the introduction of film subtitles in the *ParCoLab* database. Furthermore, all the equivalents from films are located in original Serbian texts, so we can confirm the aforementioned claims by Citron and Widmann (2006) that equivalent processing can be improved by searching equivalents in source language texts.

Seeing as *sale* is part of the core vocabulary and that it is relatively frequent, we can assume that the result will be approximately the same with other frequent and common words with a similar profile (cf. Marjanović et al., forth.). The same can also be expected with a majority of highly frequent grammatical items (cf. Stosic et al., forth.). Let us recall once again that it has been demonstrated in the metalexicographic literature (see references in Section 2) that corpora of a narrower scope than *ParCoLab* can make a significant contribution to the dictionary making process. Therefore, the *ParCoLab* parallel corpus can already help lexicographers to verify the effectiveness of listed lexicographic equivalents and/or extraction of translation equivalents in general in their work on medium-size dictionaries, such as the dictionaries we analyzed in section 3.

6 The Sample French-Serbian PARCOLAB-based Dictionary Entry

Based on the results of the analysis of existing dictionaries (cf. Section 3) and bearing in mind the equivalents with a high and medium LP that we extracted from the *ParCoLab* parallel corpus, we offer a sample corpus-based dictionary entry, that satisfies both the reception and L2→L1 translation needs of native speakers of Serbian and the production, and L1→L2 translation needs of French users. The sample is given in Figure 1:

SALE /sal/ *adjectif*

1. (pas propre) **prljav**, **nečist**, [personne] **musav**; **mains sales** prljave ruke; **vaisselle sale** prljavi ou neoprani sudovi; **il a les cheveux sales** prljava ou masna mu je kosa
2. [couleur] **prljav**; **blanc sale** prljavob[ij]elo; **jaune sale** sivožut, žutosiv
3. (immoral) **prljav**, **pokvaren**, **kvaran**
3. (obscène) **prljav**, **nepristojan**, **bezobrazan**; **histoires sales** masne priče; **dire des choses sales** govoriti gadosti ou perverzije
4. [argent] **prljav**
5. (désagréable) [temps, habitude, affaire etc.] **odvratan**, **gadan**, **grozan**, **užasan**; **quel sale temps !** kako je vr[ij]eme odvratno ou ružno!
6. PÉJ. **proklet**, **smrdljiv**, VULG. **jeben**; **cette sale voiture** ta prokleta kola; **sale capitaliste** prljavi kapitalist[a], ⇄ đubre kapitalističko; **sale menteur !** ⇄ đubre lažljivo!, ⇄ VULG. pičko lažljiva!, *(souvent traduit par un nom péjoratif de sens augmentatif)* lažovčino!

Figure 1: A sample French-Serbian corpus-based dictionary entry

The sample carefully presents the polysemy of the entry, discerns its senses based on contextual information found in *ParCoLab*, and offers sets of corpus-based equivalents, which are ordered according to the number of occurrences. We provide the information on the use of equivalents and illustrate them contextually. In several cases, especially within the last sense, we also added equivalents found through association by reading the citations from *ParCoLab* and their translations (e.g. *proklet* 'damned', *jeben* 'fucking'). However, this should not be seen as a deviation of the corpus approach to lexicography because, as stressed in the literature (cf. Roberts & Montgomery 1996; Lindemann 2013: 252), corpus content is the raw material that lexicographers need to adjust to the type of dictionary they are working on.

7 Concluding remarks

The existing French-Serbian lexicography, based exclusively on traditional methods, suffers from a number of shortcomings, as demonstrated through the analysis of the entry *sale* in the four leading French-Serbian dictionaries. The most prominent issue concerns the unsystematic equivalent processing, as well as the lack of authentic illustrative material and sense discriminators. In this paper, we have shown that by relying on the 11.1M French-Serbian-English parallel corpus *ParCoLab*, these shortcomings can be remedied or at least lessened to a large extent. Based on the subcorpus of aligned French and Serbian original and translated texts of approximately four million words, we extracted 277 occurrences of the French adjective *sale* and their Serbian translations. We classified these occurrences based on their sense. This allowed us to establish that *ParCoLab* currently contains seven out of the eight senses. Based on the extracted material, we listed the equivalents for every sense. The results of this paper indicate that using *ParCoLab* can lead to a set of equivalents, a large number of which are not included in the existing dictionaries. In some cases, those equivalents were the most common translation solutions. We have also shown that the overwhelming majority of equivalents (around 83%) have a high or medium LP. Based on the results of the analysis, we have offered a sample entry of the future French-Serbian *ParCoLab* based dictionary. Thus, we have shown that the content of the parallel corpus *ParCoLab* in its current scope can contribute to improving the existing French-Serbian dictionaries. That is the main purpose of this paper.

However, we must bear in mind that the number of extracted equivalents would have been higher if the corpus had been larger and more diversified in terms of genre, which is something the *ParCoLab*

research team is working on. At the same time, we need to mention that the methodology applied in this paper is based on manual equivalent extraction from *ParCoLab*. While quite reliable, such a method is very time-consuming; therefore, we are aware that the extraction process should be automated in the work on specific commercial dictionaries. For this reason, it is necessary to develop tools for automatic equivalent extraction, such as *The Translation Equivalents Database* (Treq), developed at the Institute of the Czech National Corpus (cf. Škrabal & Vavřín 2017), or the *Bilingual Word Sketches*, developed within *The Sketch Engine* tool (cf. Baisa et al. 2014), which quantify pairs of extracted equivalents in terms of their relative and absolute frequency respectively. Considering that texts in *ParCoLab* will be lemmatized, and morphosyntactically and syntactically annotated in the near future (cf. also Miletic 2018), the next phase is to align texts at word-level, which will enable us to develop the application for automatic *ParCoLab* equivalent extraction. Such a tool would contribute to making the described methodology completely applicable in the work on good commercial French-Serbian corpus-based dictionaries.

References

Adamska-Sałaciak, A (2006). Meaning and the Bilingual Dictionary: The Case of English and Polish. Frankfurt am Main: Peter Lang.

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baisa, V., Jakubíček, M., Kilgarriff, A., Kovář, V. & Rychlý, P (2014). Bilingual word sketches: the translate button. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 505-513.

Bujas, Ž. (1975). Testing the Performance of a Bilingual Dictionary on Topical Current Texts. In *Studia Romanica et Anglica Zagrabiensia*, 39, pp. 193-204.

Citron S., Widmann T. (2006). A Bilingual Corpus for Lexicographers. In E. Corino et al. (eds.) *Proceedings of the 12th EURALEX International Congress*. Torino: Edizioni dell'Orso, pp. 251-255.

Dickens, A., Salkie, R. (1996). Comparing Bilingual Dictionaries with a Parallel Corpus. In M. Gellerstam et al. (eds.) *Proceedings of the 7th EURALEX International Congress*, Göteborg: Göteborg University, pp. 551–559.

Goossens, D. (2012). Translation equivalents in translation corpora and bilingual dictionaries: the case of approximators in English and French. In: R. Vatvedt Fjeld, J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, pp. 514-522.

Hartmann, R.R.K. (1994). The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography. In W. Martin et al. (eds.) *Proceedings of the 5th EURALEX International Congress*. Amsterdam: Vrije Universiteit, pp. 291-297.

Héja, E. (2010). The Role of Parallel Corpora in Bilingual Lexicography. In: N. Calzolari et al. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valetta: European Language Resources Association (ELRA), pp. 2798–2805.

Jovanović, S.A. (1991). *Srpskohrvatsko-francuski rečnik.* Beograd: Prosveta.

Le Petit Robert de la langue française. Accessed at: https://www.lerobert.com. [05/03/2018]

Lindemann, D. (2013). Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair. In *Procedia – Social and Behavioral Sciences*, 95, pp. 249-257.

Lindemann, D., Manterola, I., Nazar, R., San Vicente, I. & Saralegi, X. (2014). Bilingual Dictionary Drafting. The example of German-Basque, a medium-density language pair. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 563-576.

Marjanović, S. (2013a). Le traitement des métaphores animalières dans les dictionnaires serbe-français. In *Godišnjak Filozofskog fakulteta*, 38(3), pp. 117–128.

Marjanović, S. (2013b). Obrada poredbenih frazema u srpsko-francuskim rečnicima. In S. Gudurić, M. Stefanović (eds) *Jezici i kulture u vremenu i prostoru*, 2 (2). Novi Sad: Filozofski fakultet, pp. 255–268.

Marjanović, S. (2017). *Poredbene frazeme s komponentom* comme/kao *u francuskom i srpskom jeziku.* PhD thesis. Faculty of Philology, University of Belgrade, Belgrade, Serbia.

Marjanović, S., Stošić, D. & Miletić, A. (forth.). Paralelni korpus *ParCoLab* u službi srpsko-francuske leksikografije. In J. Novaković, M. Srebro (eds.) *Srpsko-francuske književne i kulturne veze u evropskom kontekstu*. Novi Sad: Matica srpska.

Marković, R. (1980). *Francusko-srpskohrvatski rečnik*. Beograd: BIGZ.

Miletic A., Stosic D. & Marjanović, S. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštein, V. Matoušek (eds) *Text, Speech, and Dialogue. TSD 2017.* Lecture Notes in Computer Science, vol. 10415. Cham: Springer, pp. 156-164.

Miletic A., Urieli, A. (2017). Non-projectivity in Serbian: Analysis of Formal and Linguistic Properties. In S. Montemagni, J. Nivre (eds.) *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, Linköping University Electronic Press, pp. 135–144.

Miletic, A. (2018). *Un treebank pour le serbe : constitution et exploitations.* PhD thesis. University of Toulouse - Jean Jaurès, Toulouse, France.

Miletic, A., Fabre, C. & Stosic, D. (2016). Mise au point d'une méthode d'annotation morphosyntaxique fine du serbe. *Conférence conjointe JEP-TALN-RECITAL 2016*, Paris, pp. 506-513. Accessed at: https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/V02-TALN.pdf. [05/11/2017]

Perdek, M. (2012). Lexicographic potential of corpus-derived equivalents. The case of English phrasal verbs and their Polish equivalents. In: R. Vatvedt Fjeld and J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, pp. 376-388.

Perko G., Mezeg, A. (2012). Uporaba francosko-slovenskega vzporednega korpusa pri slovarski analizi nekaterih mejnih področij idiomatike. In M. Šorli (ed.) *Dvojezična korpusna leksikografija : slovenščina v kontrastu: novi izzivi, novi obeti*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp 12-34.

Putanec, V. (1995). *Francusko-hrvatski rječnik*. Zagreb: Školska knjiga.

Roberts, R., Montgomery, C. (1996). The Use of Corpora in Bilingual Lexicography. In M. Gellerstam et al. (eds.) *Proceedings of the 7th EURALEX International Congress*, Göteborg: Göteborg University, pp. 457-464.

Roberts, R. (1996). Parallel-Text Analysis and Bilingual Lexicography. Accessed at: http://www.dico.uottawa.ca/articles-fr.htm. [01/12/2017]

Salkie, R. (2008). How can lexicographers use a translation corpus?. In X. Richard et al. (eds.) *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008)*. Hangzhou: Zhejiang University. Accessed at: http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf. [01/12/2017]

Stosic, D., Marjanović, S., Miletic, A. (forth.). Corpus parallèle *ParCoLab* et lexicographie bilingue français-serbe : recherches et applications. In M. Srebro, J. Novaković (eds.) *Serbica*.

Škrabal, M., Vavřín M. (2017). The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In. I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno: Lexical Computing CZ, pp. 124-137.

Stanojević-Knežević, M. (2005). Rečnici francuskog jezika u Srbiji s Bibliografijom od 1904. do 2004. In M. Pavlović, J. Novaković (eds.) *Srpsko-francuski odnosi 1904-2004*. Beograd: Društvo za kulturnu saradnju Srbija-Francuska, Arhiv Srbije, pp. 219-233.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and the Non-Knowledge.* Tübingen: Max Neiemeyer Verlag.

Teubert W. (2002). The role of parallel corpora in translation and multilingual lexicography. In B. Altenberg, S. Granger (eds.) *Lexis in contrast: corpus-based approaches*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 189-214.

Točanac, D., Dinić, T. & Vidić, J. (2017). *Francusko-srpski rečnik*. Beograd: Zavod za udžbenike.

Zavaglia, A., Galafacci, G. (2014). Corpus, Parallélisme et Lexicographie Bilingue. In A. Abel et al. (eds.) *Proceedings of the 16th EURALEX International Congress*. Bolzano: EURAC research, pp. 587-597.