Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data

Dušica Terzić[1(✉)], Saša Marjanović[1], Dejan Stosic[2], and Aleksandra Miletic[2]

[1] Faculty of Philology, University of Belgrade, Studentski trg 3, 11000 Belgrade, Serbia
{dusica.terzic,sasa.marjanovic}@fil.bg.ac.rs
[2] CNRS and University of Toulouse, 5, Allées Antonio Machado, 31058 Toulouse, France
{dejan.stosic,aleksandra.miletic}@univ-tlse2.fr

Abstract. In this paper we present the efforts to diversify Serbian-French-English-Spanish corpus ParCoLab. ParCoLab is the project led by CLLE research unit (UMR 5263 CNRS) at the University of Toulouse, France, and the Romance Department at the University of Belgrade, Serbia. The main goal of the project is to create a freely searchable and widely applicable multilingual resource with Serbian as the pivot language. Initially, the majority of the corpus texts represented written language. Since diversity of text types contributes to the usefulness and applicability of a parallel corpus, a great deal of effort has been made to include spoken language data in the ParCoLab database. Transcripts and translations of TED talks, films and cartoons have been included so far, along with transcripts of original Serbian films. Thus, the 17.6M-word database of mainly literary texts has been extended with spoken language data and it now contains 32.9M words.

Keywords: Parallel corpus · Serbian · French · English · Spanish

1 Introduction

ParCoLab[1] is a Serbian-French-English-Spanish corpus developed by CLLE research unit (UMR 5263 CNRS) at the University of Toulouse, France, and the Department of Romance Studies at the University of Belgrade, Serbia. The primary goal of the ParCoLab project is to create a multilingual resource for the Serbian language, searchable via a user-friendly interface that can be used not only in NLP and contrastive linguistic research but also in comparative literature studies, second language learning and teaching, and applied lexicography [14,17].

[1] http://parcolab.univ-tlse2.fr. Last access to URLs in the paper: 20 Apr 2020.

P. Sojka et al. (Eds.): TSD 2020, LNAI 12284, pp. 61–70, 2020.
https://doi.org/10.1007/978-3-030-58323-1_6

Another goal of the ParCoLab project is to add several layers of annotation to the corpus text, such as lemmas, morphosyntactic descriptions (MSDs) and syntactic relations [10,13,14,17]. Currently, two portions of the Serbian subcorpus are annotated – a 150K-token literary subcorpus, ParCoTrain-Synt [12], and a 30K-token journalistic subcorpus, ParCoJour [18].[2]

In the composition of the ParCoLab corpus, quality of the collected data and the processing of the texts is prioritized over quantity, which requires a significant implication of the human factor in the process [17]. The creation of the ParCoLab corpus started with written literary texts which, in general, come with high quality translations. The result, a useful, high-quality corpus was created based on literary classics and a careful selection of good translations. However, uniformity of the corpus has an important impact on NLP applications. For instance, the annotation models trained on a single domain corpus are not particularly robust when used to process the texts of another domain [1,2,6,15]. This was confirmed in a parsing experiment in which a parsing model was trained on the ParCoTrain-Synt literary treebank and used to parse the ParCoJour journalistic corpus (see [18]).

It is not only the uniformity of the data that has an impact on the NLP applications but also the type of that data. It was shown that the differences between spoken and written language have a significant impact on machine translation. Ruiz and Federico [16] compared 2M words from 2 English-German corpora, one of which contained TED talks and the other newspaper articles. They found that TED talks consisted of shorter sentences with less reordering behavior and stronger predictability through language model perplexity and lexical translation entropy. Moreover, there were over three times as many pronouns in TED corpus than in news corpus and twice as many third person occurrences, as well as a considerable amount of polysemy through common verbs and nouns [16].

It is therefore necessary to diversify corpus data in order to make them useful for the development of good and robust NLP models. The expansion and diversification of the ParCoLab database represents an important task for Serbian corpus linguistics, considering that Serbian is one of the under-resourced European languages in terms of both NLP resources and corpora for other specialists (teachers, translators, lexicographers, etc.). In order to accomplish the goals of the ParCoLab project, the corpus should be diversified especially by adding spoken language data.

However, collecting, transcribing, and translating an authentic spontaneous speech corpus requires considerable financial and human resources. We were therefore constrained to search for the data closest to the spontaneous speech that could be collected more efficiently. It was decided to introduce TED talks and film and cartoon transcripts and subtitles and the term "spoken language data" is used to refer to this type of data. We are aware that TED talks are written and edited to be spoken in a limited time frame and thus do not represent spontaneous speech. Film and cartoon transcripts, on the other hand, are more

[2] Both corpora can be queried via the ParCoLab search engine and are available for the download at http://parcolab.univ-tlse2.fr/about/ressources.

likely to resemble transcribed natural speech although they are also written and edited beforehand. Another possible downside to using this type of documents is the questionable quality of the available transcripts and translations of TED talks and films, which may compromise the quality of the corpus material and its usefulness (cf. [7]). The method used to include transcripts and translations of films in the ParCoLab corpus tries to palliate the shortcomings of massive inclusion of unverified data and we present it in this paper. In Sect. 2, we introduce similar corpora in order to demonstrate the position of the ParCoLab corpus amongst other parallel resources containing Serbian. In Sect. 3, we describe the state of the ParCoLab database before the inclusion of the spoken data. The ongoing work on including spoken data in the ParCoLab corpus is detailed in Sect. 4. Finally, we draw conclusions in Sect. 5, and present plans for future work.

2 Related Work

In this section, we present other corpora containing the Serbian language and one of three other languages of the project – French, English or Spanish. We also discuss the share of spoken data in those corpora. There are two bilingual parallel corpora[3] developed at the Faculty of Mathematics, University of Belgrade – SrpEngKor and SrpFranKor. SrpEngKor [8] is a 4.4M token Serbian-English corpus consisting of legal and literary texts, news articles, and film subtitles. There are subtitles of only three English films containing approximately 20 K tokens. SrpFranKor [21] is a Serbian-French corpus of 1.7M tokens from literary works and general news with no spoken data. Texts in both corpora are automatically aligned on the sentence level and alignment was manually verified.

Texts in the Serbian language also appear in multilingual corpora. "1984" [9] of MULTEXT-East project contains George Orwell's *1984* and its translation into several languages including 150K-token Serbian translation. SETimes is a parallel corpus of news articles in eight Balkan languages, including Serbian, and English [20]. Its English-Serbian subcorpus contains 9.1M tokens. ParaSol (Parallel Corpus of Slavic and Other Languages), a corpus originally developed under the name RPC as a parallel corpus of Slavic languages [22], was subsequently extended with texts in other languages [23]. The Serbian part of the corpus contains 1.3M tokens of literary texts, of which only one novel is originally written in Serbian. These corpora either do not include spoken data in Serbian language or the film subtitles they contain are neither relevant in size nor originally produced in the Serbian language.

There are, however, two multilingual corpora, each containing a Serbian subcorpus with film subtitles – InterCorp and OPUS. InterCorp[4] [5], contains 31M tokens in Serbian. Texts from literary domain contain 11M tokens, whereas another 20M tokens come from film subtitles. Given that the pivot language is Czech, sentences in Serbian are paired with their Czech counterparts. It is

[3] Consultable at: http://www.korpus.matf.bg.ac.rs/korpus/login.php. It is necessary to demand authorization to access the interface.

[4] The official website of the project is: https://intercorp.korpus.cz.

unclear which portion of the Serbian subcorpus can be paired with the subcorpora in languages of the ParCoLab project. According to the information[5] on the official website, subtitles are downloaded from the OpenSubtitles[6] database. OPUS[7] [19] also contains subtitles from this database. The Serbian subcorpus contains 572.1M tokens. Neither alignment nor the quality of the translations are manually verified in these two corpora, leading to a significant amount of misaligned sentences and questionable quality of the translations. It is highly unlikely that these corpora contain films originally produced in Serbian.

Serbian spoken data can also be found in several multilingual corpora of TED talks. TED talks are lectures presented at non-profit events in more than 130 countries [24]. They are filmed and stored in a free online database at https://www.ted.com/talks. TED provides English transcripts which are translated by volunteer translators. The translation is then reviewed by another TED translator, who has subtitled more than 90 min of talk content. Finally, the reviewed translation is approved by a TED Language Coordinator or staff member [24]. Hence, the TED talks are supposed to be of higher quality than the subtitles from OpenSubtitles database, which are not verified. Free access to hours of spoken data translated into more than 100 languages has generated works on collecting corpora based on TED talks. WIT[8] [4], is an inventory that offers access to a collection of TED talks in 109 languages. All the texts for one language are stored in a single XML file. There are 5.3M tokens in the Serbian file. In order to obtain parallel corpus, it is necessary to extract TEDs by their ID and to use alignment tools since the subcorpus for each language is stored separately [4]. MulTed [24] is a parallel corpus of TED talks which contains an important amount of material in under-resourced languages such as Serbian. The Serbian subcorpus comprises 871 talks containing 1.4M tokens. All the translations are sentence-aligned automatically. Only the English-Arabic alignment was manually verified [24]. According to the official website[9] of the project, the corpus will be available for download soon.

As already mentioned in Introduction, the goal of the ParCoLab project is to create a parallel corpus of high quality. Even though it is clear that ParCoLab is not the largest available parallel corpus containing the Serbian language, an important effort is devoted to ensuring the quality of the alignment. Besides prioritizing quality over quantity, we pay special attention to including original Serbian documents. This is also true for film subtitles, whose translation we improve. Another advantage of the ParCoLab corpus is that it contains transcripts of Serbian films, providing original Serbian content. In comparison to other corpora destined to NLP users, ParCoLab is accessible and freely available to general public via the user-friendly interface, which widens its applicability.

[5] https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze12.
[6] https://www.opensubtitles.org/en/search/subs.
[7] http://opus.nlpl.eu/OpenSubtitles2016.php.
[8] https://wit3.fbk.eu/#releases.
[9] http://oujda-nlp-team.net/en/corpora/multed-corpus.

Since 2018, it has been possible to use ParCoLab search engine directly online without creating an account.

3 ParCoLab Content

The texts included in ParCoLab database are aligned with their translations using an algorithm integrated in the corpus platform. The alignment process starts with 1:1 pairing of chapters. It then continues on the level of paragraphs and, finally, of sentences. Possible errors are pointed out by the algorithm and corrected manually afterwards [10,11,17]. Corpus material is stored in XML format in compliance with TEI P5 (https://tei-c.org/guidelines/p5). XML files include standardized metadata – title, subtitle, author, translator, publisher, publication place and date, creation date, source, language of the text, language of the original work, domain, genre, number of tokens, etc. [17].

ParCoLab has been growing steadily since its inception. Initially, it contained 2M tokens [17]. Before the work on diversification presented in this paper, it contained 17.6M tokens, with 5.9M tokens in Serbian, 7.4M in French, 3.9M in English and 286K in Spanish. All the languages except for Spanish were represented through both original works and translations. In Spanish, there were only fiction translations. Its low representation is due to the fact that it has been incorporated recently in order to palliate the lack of existing Serbian-Spanish corpora. There is ongoing work on including more Spanish texts, both original and translated.

Regarding the type of texts, the corpus content came from predominantly literary works [3]. A small portion of the corpus was characterized as web content, legal and political texts and spoken data, but they were not significant in size – ∼30K tokens of film and TV show subtitles and ∼60K tokens from TED talks [14]. There were some efforts to diversify the corpus by including domain specific texts from biology, politics, and cinematography, but this material remained secondary. The original number of tokens per type of data and per language is shown in Table 1.

Table 1. Token distribution per language and text type before including spoken data.

Text type	Serbian	French	English	Spanish	Total	%
Literary texts	5,535,926	6,542,014	3,301,397	286,948	15,666,285	88.77
Non-literary written texts	340,060	761,595	566,656	0	1,668,311	9.45
Spoken data	104,935	125,919	82,504	0	313,358	1.78
Total	5,980,921	7,429,528	3,950,557	286,948	17,647,954	
% of corpus	33.89	42.10	22.39	1.63		

Even though there were some diversification efforts, the literary works remained dominant and represented 88.7% of the corpus. ParCoLab corpus consisted mainly of written texts, apart from only 1.78% of spoken data [11]. As mentioned in Introduction, linguistic differences between written and spoken corpus

influence the performance of NLP tools. Therefore, we put in a great deal of effort to overcome the main shortcoming of the corpus, which we discuss in the next section.

4 Spoken Language Data in ParCoLab

As we have already discussed in Introduction, one of the easiest way to diversify a corpus by adding spoken language data is to include TED talks and film subtitles even though this material is written and edited before oral production. This method presents a number of other shortcomings. For instance, some of the subtitles are translated automatically or by amateur translators without subsequent verification by professional translators. In addition, transcripts and translations are influenced by the number of characters that can appear on the screen. Moreover, the subtitles usually do not represent the translation of the speech in the film, but the translation of the transcripts of that speech, which are edited to fit the character number limit (see [7]). In what follows, we describe how these downsides were overcome in the present work.

Although the quality of TED talks translations cannot be guaranteed, they are reviewed by experienced translators and are supposed to be of higher quality then subtitle translations downloaded from the OpenSubtitles database. Therefore, we downloaded TED talks from the official TED site in a batch. We did not use the transcripts existing in other corpora (cf. Sect. 2). Transcripts of original TED talks are included in the database alongside their translations into three languages of the project – Serbian, French, Spanish. At the time of writing this paper, 2000 TED talks have been included in the ParCoLab database for a total of 13,458,193 tokens. A TED talk in ParCoLab corpus contains 1,652 words on average. The shortest TEDs contain only brief introductions or explanations of musical or art performances of about 200 words, whereas the longest contain around 8,000 words. They date from 1984 to 2019.

As for the film subtitles, the methodology is slightly different. Original English and French transcripts are downloaded from the OpenSubtitles database. The Serbian films were manually transcribed since it was not possible to download original transcripts or to find open source speech-to-text tools for Serbian. The inclusion of the Serbian film transcripts makes the ParCoLab corpus unique. The film subtitles translations are downloaded from the OpenSubtitles database and then improved by students who are translators in training and by the members of the ParCoLab team who work as professional translators as well. Moreover, the subtitles are compared to the actual speech in the film and corrected accordingly. That way, the limit on the number of characters to appear on screen does not affect the quality of the transcript and translation.

Apart from film transcripts, the transcripts of a large collection of cartoons are being included in the ParCoLab corpus. The data is collected from the Smurfs official Youtube channels[10] in all four languages of the corpus. The transcripts

[10] https://www.youtube.com/channel/UCeY4C8Sbx8B4bIyREPSvORQ/videos.

of popular children's stories produced by Jetlag Productions[11] are also included in the corpus in all four languages. One of the advantages of this approach is the fact that the cartoons are dubbed. That way, transcripts in all languages are transcripts of the speech in that language and not the translations of the edited transcripts of that speech. There are currently 19 The Smurfs cartoons and 19 children stories from the Jetlag productions in all four languages.

All the spoken language data is stored in XML files in compliance with the TEI P5 guidelines and included in the ParCoLab database using the same methodology as for the rest of the corpus (see Sect. 3). Apart from standardized metadata, the name of the TED editor is included. Time spans are omitted. Additional metadata for film and cartoon transcripts represent names of characters, gender, and age in order to make it useful for linguistic analysis.

There are now 32.9M tokens in ParCoLab database. The Serbian subcorpus currently contains 9.6M tokens, French 11.5M, English 7.7M, whereas the Spanish portion contains 4.06M. The current percentage of spoken data is listed in Table 2.

Table 2. Token distribution per language after adding spoken data.

Text type	Serbian	French	English	Spanish	Total	% of corpus
TED talks	3,215,129	3,592,230	3,304,572	3,346,261	13,458,193	40.88
Films	292,916	356,749	252,355	0	902,020	2.74
Cartoons	110,865	68,516	173,865	23,324	376,570	1.14
Spoken data	3,618,910	4,017,495	3,730,792	3,369,585	**14,736,783**	44.77
Written data	5,989,500	7,475,463	4,030,951	687,127	**18,183,041**	55.23
Total	9,608,410	11,492,958	7,761,743	4,056,712	**32,919,824**	
% of corpus	29.19	34.91	23.58	12.32		

The percentage of literary works dropped from 88.7% to 55.23% whereas the spoken data represent 44.77% instead of 1.78% of the corpus before the diversification. We can conclude that the inclusion of, what is called here, spoken data has already demonstrated a substantial progress in diversifying ParCoLab corpus. All the spoken material can be queried via the user-friendly interface which makes this corpus accessible not only to researchers but also to the translators, lexicographers, teachers, etc. The Spanish section of the corpus rose from 1.63% to 12.32%.

When it comes to the qualitative evaluation of the corpus, this diversification helped to cover certain senses and contexts of specific words. For instance, the Serbian adjective *domaći* (Eng. domestic) mostly occurred with the sense 'related to the home' in the original corpus [11]. Currently, its dominant sense is 'not foreign', which is in accordance with the monolingual Serbian corpora. Furthermore, as was supposed previously [10], film transcripts contributed to

[11] https://en.wikipedia.org/wiki/Jetlag_Productions.

augmenting the number of the examples in which French adjective *sale* (Eng. dirty) is 'used to emphasize one's disgust for someone or something'.

5 Conclusion and Future Work

The quadrilingual corpus ParCoLab is one of rare parallel resources containing a Serbian subcorpus, especially when it comes to original Serbian texts. In the expansion of the corpus, priority was given to quality over quantity. In addition to continuing work on enlarging the corpus, a great deal of effort has also been devoted to the diversification of the predominantly literary content. This paper describes the method that allowed us to include transcripts and translations of 2000 TED talks containing 13.5M tokens in ParCoLab. Apart from TED talks, there are film subtitles, among which are those originally produced in Serbian, as well as the transcripts of dubbed cartoons that are included in the ParCoLab database. By including additional 73 film and cartoon transcripts alongside the aforementioned TED talks, ParCoLab corpus database surpasses 32.9M. Thus we created the material not only for the development of NLP tools (especially machine translation) but also for teaching and learning French, English, Serbian, and Spanish as foreign languages and for lexicography.

While the ParCoLab content is being diversified more and more, the annotated portion of the corpus still comes from written documents. Given that the training corpus for the annotation tools needs to be built on the in-domain data to perform well, it is necessary to improve the training corpus. A new spoken language data subcorpus provides us with material to pursue this goal. Therefore, our next steps in annotating the corpus would be to tag, lemmatize, and parse added spoken subcorpus.

References

1. Agić, Ž., Ljubešić, N.: Universal dependencies for Croatian (that work for Serbian, too). In: Piskorski, J. (ed.) Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015), pp. 1–8. INCOMA, Hissar (2015)
2. Agić, Ž., Ljubešić, N., Merkler, D.: Lemmatization and morphosyntactic tagging of Croatian and Serbian. In: Piskorski, J. (ed.) Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013), pp. 48–57. Association for Computational Linguistics, Sofia (2013)
3. Balvet, A., Stosic, D., Miletic, A.: TALC-sef a manually-revised POS-tagged literary corpus in Serbian, English and French. In: LREC 2014, pp. 4105–4110. European Language Resources Association, Reykjavik (2014)
4. Cettolo, M., Girardi, C., Federico, M.: WIT3: web inventory of transcribed and translated talks. In: Proceedings of the 16th EAMT Conference, pp. 261–268 (2012)
5. Čermák, F., Rosen, A.: The case of interCorp, a multilingual parallel corpus. Int. J. Corpus Linguist. **13**(3), 411–427 (2012)
6. Gildea, D.: Corpus variation and parser performance. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (2001). https://www.aclweb.org/anthology/W01-0521

7. van der Klis, M., Le Bruyn, B., de Swart, H.: Temporal reference in discourse and dialogue (Forth)
8. Krstev, C., Vitas, D.: An aligned English-Serbian corpus. In: Tomović, N., Vujić, J. (eds.) ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), vol. 1, pp. 495–508. Faculty of Philology, Belgrade (2011)
9. Krstev, C., Vitas, D., Erjavec, T.: MULTEXT-East resources for Serbian. In: Erjavec, T., Gros, J.Z. (eds.) Zbornik 7. mednarodne multikonference "Informacijska druzba IS 2004", Jezikovne tehnologije, Ljubljana, Slovenija, 9–15 Oktober 2004. Institut "Jožef Stefan", Ljubljana (2004)
10. Marjanović, S., Stosic, D., Miletic, A.: A sample French-Serbian dictionary entry based on the ParCoLab parallel corpus. In: Krek, S., et al. (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pp. 423–435. Faculty of Arts, Ljubljana (2018)
11. Marjanović, S., Stošić, D., Miletić, A.: Paralelni korpus ParCoLab u službi srpsko-francuske leksikografije. In: Novaković, J., Srebro, M. (eds.) Srpsko-francuske književne i kulturne veze u evropskom kontekstu I, pp. 279–307. Matica srpska, Novi Sad (2019)
12. Miletic, A.: Un treebank pour le serbe: constitution et exploitations. Ph.D. thesis. Université Toulouse Jean Jaurès, Toulouse (2018)
13. Miletic, A., Fabre, C., Stosic, D.: De la constitution d'un corpus arboré á l'analyse syntaxique du serbe. Traitement Automatique des Langues **59**(3), 15–39 (2018)
14. Miletic, A., Stosic, D., Marjanović, S.: ParCoLab: a parallel corpus for Serbian, French and English. In: Ekštein, K., Matoušek, V. (eds.) TSD 2017. LNCS (LNAI), vol. 10415, pp. 156–164. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64206-2_18
15. Nivre, J., et al.: The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, pp. 915–932. Association for Computational Linguistics, Prague (2007)
16. Ruiz, N., Federico, M.: Complexity of spoken versus written language for machine translation. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT), pp. 173–180. Hrvatsko društvo za jezične tehnologije, Zagreb (2014)
17. Stosic, D., Marjanović, S., Miletic, A.: Corpus parallèle ParCoLab et lexicographie bilingue français-serbe: recherches et applications. In: Srebro, M., Novaković, J. (eds.) Serbica (2019). https://serbica.u-bordeaux-montaigne.fr/index.php/revues
18. Terzic, D.: Parsing des textes journalistiques en serbe par le logiciel Talismane. In: Proceedings of TALN-RECITAL, PFIA 2019, pp. 591–604. AfIA, Toulouse (2019)
19. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Calzolari, N. (eds.) Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association, Istanbul (2014)
20. Tyers, F.M., Alperen, M.S.: South-East European times: a parallel corpus of Balkan languages. In: Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages, pp. 49–53 (2010)
21. Vitas, D., Krstev, C.: Literature and aligned texts. In: Slavcheva, M., et al. (eds.) Readings in Multilinguality, pp. 148–155. Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia (2006)
22. von Waldenfels, R.: Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B., Zdanova, V., Zimny, R. (eds.) Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9, pp. 123–138. Verlag Otto Sagner, München (2006)

23. von Waldenfels, R.: Recent developments in ParaSol: breadth for depth and XSLT based web concordancing with CWB. In: Daniela, M., Garabík, R. (eds.) Natural Language Processing, Multilinguality, Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011, pp. 156–162. Tribun EU, Bratislava (2011)
24. Zeroual, I., Lakhouaja, A.: MulTed: a multilingual aligned and tagged parallel corpus. Appl. Comput. Inform. (2018). https://doi.org/10.1016/j.aci.2018.12.003