

Opus-MontenegrinSubs 1.0: First electronic corpus of the Montenegrin language

Petar Božović,^{*} Tomaž Erjavec,[†] Jörg Tiedemann,^{*} Nikola Ljubešić,[§] Vojko Gorjanc[‡]

^{*} English Language and Literature Department, Faculty of Philology, University of Montenegro
Danila Bojovića bb, 81400 Nikšić
petarb@ac.me

[†] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
tomaz.erjavec@ijs.si

^{*} Department of Linguistics and Philology, Uppsala University
BOX 635, SE-751 26 Uppsala, Sweden
jorg.tiedemann@lingfil.uu.se

[§] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ffzg.hr

[‡] Department of Translation Studies, Faculty of Arts, University of Ljubljana Aškerčeva 2, 1000 Ljubljana
vojko.gorjanc@ff.uni-lj.si

Abstract

Although recent years have witnessed a growth in the number of computational language resources and tools, a lot still needs to be done, especially with low-density languages. This is the case with all South Slavic languages and especially Montenegrin, the fourth standard of the once Serbo-Croatian language that has been re-codified only recently. Even though it became the official language of Montenegro in 2007, there still isn't any publicly available electronic corpus that would be available for empirical research of linguistic, translational or any other inquiry. This paper introduces the first publicly available English – Montenegrin parallel corpus of subtitles. It describes the process of corpus compilation, presents linguistic annotation and accessibility of the corpus through web concordancers. Furthermore, it gives a brief overview of linguistic situation in Montenegro with some of the most important recent developments especially in the light of the recent official international recognition of the language which took place in December 2017.

1. Introduction

Recent years have witnessed a growth in the number of machine-readable corpora and language tools for a number of world languages. It is currently estimated that there are 7,097 languages in the world, an updated number of officially recognized languages listed by Ethnologue¹ which is to be taken arbitrarily. Out of this number, only in 2006 there were corpora available for less than 1% of all world languages, and 20-30 of these fall into the category of high-density and medium-density languages, where “density” is understood to represent the number of computational resources available (Maxwell and Hughes, 2006). The first group would include a handful of languages only, including English, German, Arabic, etc. Today, the number of available resources has increased to app. 90 languages, which means app. 1.2% of all world languages having any kind of publicly available computational resources. The majority of these, however, are lower-density languages as resources are rather scarce.

Being a multilingual and integrative society, Europe is estimated to cover more than 80 languages, of which 23 are official and the rest are either minority or immigration languages. However, a number of these languages is technologically not supported sufficiently and run the risk of being marginalized or digitally extinct. Thus, a number of initiatives have been introduced, such as the META-NET Strategic Research Agenda for Multilingual Europe 2020, with the aim of using various language technologies

for overcoming language barriers, enabling free flow of information, goods, and innovation, thus creating a single digital space and marketplace².

When it comes to electronic language resources and corpora for some of the major official languages of former Socialist Federal Republic of Yugoslavia, Bosnian, Croatian, Montenegrin, Serbian, Slovenian (BCMSS), the majority of resources are available for Slovenian, followed by Croatian, Serbian, and Bosnian. The last language standard based of the once Serbo-Croatian or Croato-Serbian language that has recently been re-codified and internationally recognized in December 2017 is Montenegrin. So far, there has not been any electronic public corpora of any kind available for the study of this standard. This all testifies to the fact that there is still a lot of work to be done.

This paper presents Opus-MontenegrinSubs 1.0, the first parallel English-Montenegrin electronic corpus developed as a joint effort of researchers from the University of Montenegro, Jožef Stefan Institute, University of Helsinki, and University of Ljubljana. First we will briefly discuss the potentials and possible applications of parallel corpora, the specifics of subtitle corpora as a sub-type of parallel corpora, followed by an outline of the current available parallel corpora for BCMSS. Since this is the first electronic resource of Montenegrin language that is being presented, we will give a brief overview of the linguistic situation in Montenegro followed by the description of the corpus itself and the first study based on it.

¹ <http://www.ethnologue.com>, accessed: March 15th, 2018, at 11:30 pm.

² <http://www.meta-net.eu/sra/key-messages>, accessed: March 30th, 2018, at 1:00 pm.

2. Parallel subtitle corpora for BCMSS: the potential, current state and specifics

Parallel corpora have found a number of applications in linguistics, translatology, translation practice, and beyond. They proved to be an indispensable tool for a number of contrastive linguistic studies, word sense disambiguation and construction of lexicons, as an input for parallel concordancing systems. Furthermore, the use of parallel corpora has especially become a trend in translation and interpreting studies for developing and training statistical machine translation systems (which require large amounts of parallel language data), for the study of regularities of translations and translators, translation teaching and learning, translation practice including terminology extraction, identifying translation equivalents and correspondents, translation quality assessment, etc. (Bywood et al 2013; Hu, 2016). However, as noted by Tiedemann (2007), most of the existing parallel corpora cover the high-density languages and the domains of legislation, administration and technical documentation.

With regards the parallel subtitle corpora for BCMSS, most of them were developed as subcorpora within the Opus2 project (Tiedemann, 2009) and for each of the western South Slavic languages they include:

- Bosnian: subcorpus OpenSubtitles 2011 (tokens 26,491,099, words ~ 20,906,596),
- Croatian: subcorpora OpenSubtitles 2011 (tokens 111,981,881, words ~ 86,600,021), TedTalks (tokens 1,285,011, words ~ 993,749),
- Serbian: subcorpus OpenSubtitles 2011 (tokens 154,063,822, words ~ 119,149,120),
- Slovenian: subcorpus OpenSubtitles 2011 (tokens 109,690,961, words ~ 81,500,854).

Another project that involved the creation of subtitle corpora for Slovenian and Serbian was SUMAT (tokens 1,250,000/ 1,500,000) (Bywood et al., 2013, Fishel et al., 2012).

Apart from the above mentioned possible applications of corpora of this type, subtitles can be used for the study of text compression and summarization. The reason for this are unique features of subtitles that make them a specific language resource in many ways. Subtitles are usually transcriptions of spontaneous speech with a diversified language (genre, slang, colloquialisms) and they can be classified into several categories: interlingual and intralingual (depending whether they represent a translation from a source to a target (foreign) language, or they are in the same language as the source audiovisual text); monolingual or multilingual (depending on the number of translations into different languages which are shown on the screen); pre-recorded and live. Interlingual subtitling, which we refer to when we use the term in this paper, is a specific form of translation practice since subtitles per se are “a vulnerable modality” for various reasons (Diaz Cintas and Rameal, 2007). This is primarily the case because viewers are exposed both to the source and target text, and there are specific time and space constraints: they are usually shown in one or two lines with 30 – 40 characters, cca. 3 – 7 seconds only with no room for annotation. This calls for specific translation strategies among which condensation (it is estimated that subtitles are 40 – 75% shorter than spoken version), omission is seen as a legitimate strategy (especially in cases of redundancies and spoken discourse markers such as exclamations, false

starts, repetitions, hesitations, question tags, etc.), cultural substitution, generalization and specification. Moreover, standardization is also used frequently (especially in cases of slang, regionalisms, grammar mistakes, etc.), and occasional censorship. This shows that subtitles should be approached as a specific, yet an important and unique, resource of translated language.

3. Linguistic situation in Montenegro: a brief overview of recent history

As previously mentioned, Montenegrin language is the last out of four re-codified standards that stem from the same linguistic base of the polycentric Serbo-Croatian language. The remaining three include Bosnian, Croatian and Serbian. Similarly to these standards, it is based on the Eastern Herzegovian Shtokavian dialect. As it has been the case with most neighboring countries, the situation in language policy in Montenegro has reflected a rather turbulent political situation in the Socialist Federal Republic of Yugoslavia (SFRY), and later the Socialist Republic of Yugoslavia (SRJ) and the union of Serbia and Montenegro (SM), only to reach its current state in the post-2006 period when Montenegro became independent.

During the pre-1991 period, Serbo-Croatian was one of the official languages of the SFRY, together with Slovenian, Macedonian and other languages which were constitutionally of equal status, but the reality was somewhat different as they seem to have been in a position of a “competitive coexistence” (Gorjanc, 2013; Požgaj Hadži et al., 2013). In the light of the above mentioned historical events, the constitutions of the Socialist Republic of Montenegro of 1963 and 1974 define Serbo-Croatian as the language in official use in Montenegro. This polycentric language, as its very name suggests, stemmed from two main standards, the eastern (with its center in Belgrade) and the western (with its center in Zagreb), while other language forms with their center in Podgorica (i.e. Titograd, as the capital of Montenegro was called then) and Sarajevo were marginalized as regional variants, and re-codified as provincialisms, a situation which would later be seen as having significant political implications especially regarding the politics of assimilation and hegemony.

With the disintegration of Yugoslavia in 1991, Serbo-Croatian was re-codified into 4 separate standards, starting with Serbian and Croatian, and later followed by Bosnian and Montenegrin. Due to the official state policy of the day, the official language of Montenegro in the constitution of 1992 was designated as Serbian of the Ijekavian standard, and this remained the case until 2007. Shortly after the independence which took place in 2006, and upon the ratification of the new Constitution which took place on 22 October 2007, Montenegrin became the official language in Montenegro. After much controversy arising from two different approaches to the process of standardization, the first Montenegrin grammar and orthography were adopted in 2010 by the Council for General Education. The question of language standardization still remains an ongoing issue and it's highly debatable whether some of their solutions will fully integrate into language practice.

Census data from 2011 shows an increase in the number of speakers who designate their mother tongue as Montenegrin. One of the most significant events was certainly the international recognition of Montenegrin language and the assignment of the international code. This

was approved on 8 December 2017 and the ISO 639-2 and 3 code [cnr] was assigned. Needless to say, much still needs to be done. The first volume of the Dictionary of the Montenegrin Literary and Vernacular Language published by the Montenegrin Academy of Sciences and Arts in 2016 was soon withdrawn due to reactions of part of political public to some entries that seemed to be controversial. Regarding the electronic language resources, up to this date, there haven't been any officially published electronic corpora of Montenegrin language that would be available to researchers for various linguistic and translational studies. That is why the electronic corpus which we will present in this paper is of high importance as it is the first ever electronic corpus of Montenegrin language.

4. Corpus compilation and accessibility

The corpus Opus-MontenegrinSubs 1.0 contains parallel English-Montenegrin subtitles. The data and copyrights were obtained from the Radio and Television of Montenegro, the public service broadcaster of Montenegro. The corpus consists of English and Montenegrin subtitles of three series: House of Cards, Damages, and Tudors. The corpus contains 10 seasons, and 110 episodes, which are cca. 5,563 minutes in length. A detailed breakdown is given in Table 1.

Series	No. of seasons	No. of episodes	Length
House of Cards	1	13	686 mins.
Damages	5	59	2878 mins.
Tudors	4	38	1999 mins.

Table 1: Corpus breakdown

4.1. Processing the corpus

Sentence alignment and basic encoding was performed inside the OPUS project³. The original subtitle files were converted to Unicode UTF-8 using *iconv* and the Unix tool *file* for automatic detection of the character set in the original file. After that the OPUS subtitle tools (Lison & Tiedemann, 2018) were applied to convert the files to standalone XML with sentence markup; the remaining XML-well formedness problems were fixed with the program *tidy*. Finally, all translated subtitles were aligned using the time-based alignment method described in Tiedemann (2007) and the standard OPUS import pipeline was used to integrate the data in OPUS with download formats in XML, plain text and TMX.

In the second stage, the source XML data was converted to the latest version of TEI (TEI, 2018), so that the subtitles for each language are stored in a separate <text> element, with the sentence alignments being maintained by cross-links as well as separately, in a <linkGrp> element. An important part of this conversion was also the encoding of the <teiHeader> element, which contains the meta-data of the corpus, explicating its authors, license etc. but also listing all the used XML elements in the corpus, together

with a short explanation, and how the MSD annotation prefixes are to be interpreted.

Then, the English and Montenegrin texts were tokenized, sentence segmented and tagged with morphosyntactic descriptions (MSDs) and lemmas. To perform this annotation for Montenegrin, we used the ReLDI tokeniser⁴ and tagger⁵ (Ljubešić & Erjavec, 2016) with its model for Serbian. The MSD tagset used follows the MULTTEXT-East specifications (Erjavec, 2012), in particular, the version 5 specifications for Bosnian⁶.

For English, we used Tree Tagger (Schmidt 1994, 1995) with its model for English, which uses the Penn Tree Bank tagset. In order to make the English tagset harmonized with the one for Montenegrin, we converted it to the SPOOK tagset for English⁷, i.e. performed a 1-1 mapping between the original PTB tagset to MULTTEXT-East compatible SPOOK tagset.

Figure 1 illustrates the TEI encoding of the linguistically annotated corpus, giving the first translation unit (annotated as anonymous block, <ab>) for both languages. As can be seen, each language text contains the divisions marking the structure of the corpus, while the translation units are given IDs and the alignment via their @corresp attribute. Each translation unit is then divided into sentences, and these into words, punctuation symbols and whitespace. The tokens are lemmatized and MSD tagged, where the value prefix *mte* resolves to the MULTTEXT-East MSD definition (i.e. its decomposition into features), while the *spook* one resolves to the SPOOK decomposition. It should be noted that the original PTB tag is retained as the value of the @function attribute.

4.2. Corpus distribution and use

The TEI corpus was converted to the so called vertical format, used by (no)Sketch Engine and mounted on the CLARIN.SI concordancers,⁸ namely noSketch Engine and KonText, as well as Sketch Engine, so that it is available on-line for searching and exploration; both concordancers also allow displaying the aligned translation units.

The complete corpus in TEI, as well as vertical format, was also made available for download in the CLARIN.SI repository (Božović et al., 2018) under the Creative Commons - Attribution-ShareAlike license.

5. First corpus studies

The first study based on this corpus is the one conducted for the purpose of the Ph.D. studies by Petar Božović with the thesis topic *Audiovisual Translation and Elements of Culture: A Comparative Analysis of Transfer with Reception Study in Montenegro*, which is in the field of translation studies and corpus linguistics. Corpus-based translation studies are becoming increasingly relevant for the industry ever since the methodology has been introduced from an allochthonous field of corpus linguistics in the seminal paper by Baker (1993). It wasn't long after this that it became evident that using corpora in translation research was to have a great potential for scholarly empirical research, but also for terminologists and practitioners.

³ <http://opus.nlpl.eu/MontenegrinSubs.php>

⁴ <https://github.com/clarinsi/reldi-tokeniser>

⁵ <https://github.com/clarinsi/reldi-tagger>

⁶ <http://nl.ijs.si/ME/V5/msd/html/msd-bs.html>

⁷ <http://nl.ijs.si/spook/msd/html-en/msd-en.html>

⁸ <http://www.clarin.si/info/concordances/>

```

<text xmlns="http://www.tei-c.org/ns/1.0" xml:lang="cnr">
  <body>
    <div type="series" xml:id="Damages-cnr">
      <div type="season" xml:id="Damages.S1-cnr">
        <div type="episode" xml:id="Damages.S1.dam0101-cnr">
          <ab n="1" xml:id="Damages.S1.dam0101.SL1-cnr" corresp="#Damages.S1.dam0101.SL1-en">
            <s>
              <w ana="mte:Agpfprny" lemma="opasni">OPASNE</w><c> </c>
              <w ana="mte:Ncfpn" lemma="igra">IGRE</w><c> </c>
              <w ana="mte:Ncmsn" lemma="pilot">Pilot</w><c> </c>
              <w ana="mte:Ncfpg" lemma="epizoda">epizoda</w>
            </s>
          </ab>
        ...
      </div>
    </div>
  </body>
</text>

<text xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <body>
    <div type="series" xml:id="Damages-en">
      <div type="season" xml:id="Damages.S1-en">
        <div type="episode" xml:id="Damages.S1.dam0101-en">
          <ab n="1" xml:id="Damages.S1.dam0101.SL1-en" corresp="#Damages.S1.dam0101.SL1-cnr">
            <s>
              <w lemma="Season" function="NP" ana="spook:Np-s">Season</w><c> </c>
              <w lemma="1" function="CD" ana="spook:M-c">1</w><c> </c>
              <w lemma="episode" function="NN" ana="spook:Nc-s">Episode</w><c> </c>
              <w lemma="1" function="CD" ana="spook:M-c">1</w><c> </c>
              <w lemma="pilot" function="NN" ana="spook:Nc-s">Pilot</w><c> </c>
              <pc function="(" ana="spook:Z"></pc>
              <w lemma="Dimension" function="NP" ana="spook:Np-s">Dimension</w>
              <pc function=")" ana="spook:Z"></pc>
            </s>
          </ab>
        ...
      </div>
    </div>
  </body>
</text>

```

Figure 1: TEI encoding of the corpus texts

The screenshot shows the KonText concordancer interface. At the top, there's a search bar with the query 'čudo' and 11 hits. Below the search bar, there's a table with two columns: 'Opus MontenegrinSubs: Montenegrin' and 'Opus MontenegrinSubs: English'. The table lists various entries from the corpus, including titles like 'Damages.S1.dam0106...', 'Damages.S2.dam0205...', 'HoC.S1.hoc0305-cnr', 'Tudors.S1.tud0101-cnr', and 'Tudors.S2.tud0206-cnr', along with their corresponding English translations. The interface also includes a 'Line selection' dropdown set to 'simple' and a 'Display options' link.

Figure 2: Searching the corpus in the KonText concordancer

Hence, the research which is based on the corpus is focused on the highly-influential and fast-growing audiovisual translation field with the aim of mapping the different translation strategies for rendering the

extralinguistic elements of culture in subtitling. This is an issue that is at the core of some of the major challenges in the industry as transfer of elements of culture proves to be one of the “crisis points” in translation process, especially

in subtitling due to the time and spatial constraints of the modality and it can have an important influence on the reception and placement of the audiovisual product on the target market and for the target audience (Pedersen, 2011). The extralinguistic elements will be extracted as types, not tokens, by using one of the concordancers. After that, these elements will be categorized according to the level of transculturality, and translation strategy for rendering that element will be defined. The goal is to gain a better understanding and map how culture is rendered in subtitling from English into Montenegrin and to supplement this with the reception study which is also part of this research. It is hoped that this will provide an important empirical feedback for translators and broadcasting companies who could tailor the translation policy better to meet the needs and expectations of the real, not ideal or intuitive, target audience.

6. Conclusions

The Opus-MontenegrinSubs 1.0 is the first publicly available parallel electronic corpus of Montenegrin language the appearance of which is timely considering the recent sociolinguistic developments, especially constitutional and international acknowledgement that this language has received. Needless to say, a lot still remains to be done in order to provide the computational resources and tools necessary for state-of-the-art linguistic approaches and analyses. It is hoped that this corpus will encourage other researches and contribute to the affirmation and development of corpus linguistics and corpus-based translation studies in the region. Moreover, it is hoped that it will encourage the development of other corpora of Montenegrin language, first and foremost of the reference corpus, which would be of a pivotal importance for the process of restandardization and without which a modern linguistic description of Montenegrin will not be possible.

Acknowledgments

The authors acknowledge the support of the Radio and Television of Montenegro and Vuk Perović for the corpus materials and are most grateful for the permissions granted.

7. References

- Jan Pedersen. 2011. *Subtitling Norms for Television: An Exploration Focusing on Extralinguistic Cultural References*. John Benjamin Publishing.
- Jörg Tiedemann. 2007. Improved Sentence Alignment for Movie Subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*.
- Jörg Tiedemann. 2007. Building a Multilingual Parallel Subtitle Corpus. In *Proceedings of the 17th Conference on Computational Linguistics in the Netherlands (CLIN 17)*. Leuven, Belgium.
- Kaibao Hu. 2016. *Introducing Corpus-based Translation Studies*. Springer.
- Lindsay Bywood, Martin Volk, Mark Fishel, and Panayota Georgakopoulou. 2013. Parallel subtitling corpora and their applications in machine translation and translology. *Perspectives: Studies in Translatology*, 21(4): 595–610.
- Mike Maxwell and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pages 29–37, Sydney, Australia. Association for Computational Linguistics.
- Mona Baker. 1993. Corpus Linguistics and Translation Studies – Implications and Applications. In: Mona Baker et al., ed., *Text and Technology: In Honor of John Sinclair*, pages 233–252, John Benjamins Publishing.
- Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, Portorož. http://www.lrec-conf.org/proceedings/lrec2016/pdf/811_Paper.pdf
- Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018. *English-Montenegrin parallel corpus of subtitles Opus-MontenegrinSubs 1.0*, Slovenian language resource repository CLARIN.SI, <http://hdl.handle.net/11356/1176>.
- Pierr Lison and Jörg Tiedemann. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. *LREC 2018*.
- Ranko Bugarski. 2009. *Nova lica jezika*. Biblioteka XX vek.
- Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*. Manchester.
- Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin.
- Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language resources and evaluation*, 46(1): 131–142, doi: 10.1007/s10579-011-9174-8
- Vojko Gorjanc. 2013. Slovenačka jezička politika i odnosi društvene moći. In: Vesna Požgaj Hadži, ed., *Jezik između lingvistike i politike*, pp.12–36. Biblioteka XX vek.
- Vesna Požgaj Hadži, Tatjana Balažić Bulc, and Vlado Miheljak. 2013. Srpskohrvatski jezik iz slovenske perspektive. In: Vesna Požgaj Hadži, ed., *Jezik između lingvistike i politike*, pages 12–36. Biblioteka XX vek, Beograd.
- TEI Consortium, eds. 2018. *TEI P5: Guidelines for Electronic Text Encoding and Interchange 3.3.0*. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>