CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation

Nikola Ljubešić, Taja Kuzman
Jožef Stefan Institute, University of Ljubljana, Institute of Contemporary History
Jamova cesta 39, Večna pot 113, Privoz 11, 1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si, taja.kuzman@ijs.si

Abstract

This paper presents a collection of highly comparable web corpora of Slovenian, Croatian, Bosnian, Montenegrin, Serbian, Macedonian, and Bulgarian, covering thereby the whole spectrum of official languages in the South Slavic language space. The collection of these corpora comprises a total of 13 billion tokens of texts from 26 million documents. The comparability of the corpora is ensured by a comparable crawling setup and the usage of identical crawling and post-processing technology. All the corpora were linguistically annotated with the state-of-the-art CLASSLA-Stanza linguistic processing pipeline, and enriched with document-level genre information via the Transformer-based multilingual X-GENRE classifier, which further enhances comparability at the level of linguistic annotation and metadata enrichment. The genre-focused analysis of the resulting corpora shows a rather consistent distribution of genres throughout the seven corpora, with variations in the most prominent genre categories being well-explained by the economic strength of each language community. A comparison of the distribution of genre categories across the corpora indicates that web corpora from less developed countries primarily consist of news articles. Conversely, web corpora from economically more developed countries exhibit a smaller proportion of news content, with a greater presence of promotional and opinionated texts.

Keywords: web corpora, South Slavic languages, linguistic processing, genre identification

1. Introduction

The South Slavic languages constitute one of the three major branches within the Slavic language family. They are mostly spoken in Central and Southeastern Europe, spanning from the Slovenian Alps, across the Adriatic and Dinarides regions, to Bulgaria and the Black Sea. Despite their widespread use, many languages within this group remain relatively low-resourced and underrepresented in the field of natural language processing. According to a recent report on the state of technologies for European languages (Rehm and Way, 2023), the support with the core Language Technologies (LT) – which include corpora, models and language technologies, such as text and speech processing and machine translation – was shown to be less than moderate for South Slavic languages, with Bosnian and Macedonian having weak support to no support in all LT areas. In order to facilitate the development of language technologies for these languages, it is crucial to have access to substantial amounts of textual data. Such data serve as the foundation for language technologies, including language models and machine translation systems. Additionally, general text corpora – that is, corpora that cover broad language use and are not specifically restricted to any particular subject, register or genre – enable linguists and other researchers to analyze linguistic phenomena based on statistically significant amounts of data. Web crawling, an automated method for collecting text corpora, is one of the primary approaches for quick collection of such data. Recently, the MaCoCu[1] project (Bañón et al., 2022) used this method to develop and make freely available monolingual and parallel datasets for over 10 under-resourced languages, including the South Slavic languages.

This paper introduces the CLASSLA-web corpora, which are based on the MaCoCu datasets and have been additionally enriched with linguistic annotation and genre information. This collection of corpora represents, to best of our knowledge, the first comparable corpus collection that encompasses the entire language group. The corpora are considered comparable as they were collected and processed with the same tools and within the same time frame. The paper presents a comprehensive overview of the various stages involved in creating the corpora and offers initial insights into their content based on genre information. The creation of the CLASSLA-web corpora involved three main steps: 1) web crawling based on top-level national domains, and subsequent data post-processing and filtering; 2) linguistic annotation with the latest CLASSLA-Stanza pipeline (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023), the state-of-the-art pipeline for processing South Slavic languages; and 3) genre annotation, which provides an insight into the functional content of

[1] https://macocu.eu/

LREC-COLING 2024, pages 3271–3282
20-25 May, 2024.

these corpora. Genres are text categories that are defined considering the author's purpose, common function of the text, and the text's conventional form (Orlikowski and Yates, 1994). Examples of genres are *News*, *Promotion*, *Legal*, etc. Genre annotation provides valuable insights into the functional content of the corpora, enabling genre-based corpus studies and more focused natural language processing applications in machine translation (Van der Wees et al., 2018) or summarization (Stewart and Callan, 2009).

The paper is organized as follows. Firstly, we provide an overview of the related work on construction of web corpora and the automatic genre identification in Section 2. Next, in Section 3, we present the process of creating and curating the CLASSLA-web corpora, whose content is analyzed in Section 4. Lastly, the paper concludes with Section 5, where we summarize the main findings and present future work.

2. Related Work

2.1. Web Corpora

The tradition of building web corpora can be followed back to the WaCky initiative (Baroni et al., 2009), inside which first web corpora for large European languages were built. Two other notable initiatives include the CoW corpora (Schäfer et al., 2012), developed for large European languages, and the TenTen corpora (Jakubíček et al., 2013), created for many languages, including some of South Slavic languages, namely Slovenian[2] and Bulgarian[3]. However, an important limitation of the TenTen corpora is that they are accessible only through concordancers of the Lexical Computing company, which require a paid subscription.

Recently, numerous web-based datasets have emerged from the Common Crawl[4] and the Internet Archive data collections[5]. Most prominent collections include the cc100 dataset (Conneau et al., 2020), the mC4 dataset (Xue et al., 2021), and the OSCAR dataset (Suárez et al., 2019). While these multilingual collections are highly useful for multilingual language modelling tasks, they suffer, inter alia, from blind spots, such as an unexplainable small amount of all Western South Slavic (Slovenian, Croatian, Bosnian, Montenegrin, Serbian) data in OSCAR[6] and an unexpectedly small amount of Croatian content in the mC4 dataset (Xue et al., 2021).

The previously developed open web corpora for South Slavic languages adhere to the naming convention of the WaCky (Baroni et al., 2009) initiative (e.g., slWaC, hrWaC, srWaC). However, they were actually built using the TenTen technology, more specifically, the SpiderLing crawler (Suchomel et al., 2012), the jusText text extraction tool, and the Onion near-deduplication tool (Pomikálek, 2011). The first web corpora of South Slavic languages were compiled for Slovenian (slWaC) and Croatian (hrWaC) (Ljubešić and Erjavec, 2011). Three years later, the list of languages was expanded to include Bosnian (bsWaC) and Serbian (srWaC) (Ljubešić and Klubička, 2014). Additionally, the Croatian and Slovenian crawls were later updated (Erjavec et al., 2015). However, there have been no recent activities related to open web corpora for South Slavic languages.

2.2. Genre Prediction

Until recently, the available technologies for text classification, including support vector machines and logistic regression classifiers, were insufficient to accurately identify genres across languages and datasets (Sharoff et al., 2010; Kuzman et al., 2023a). However, recent advancements in deep neural technologies led to a breakthrough in this field. Transformer-based language models (Vaswani et al., 2017), specifically those fine-tuned on manually-annotated genre datasets, demonstrated the capability to identify genres in various web corpora and languages (see Rönnqvist et al. (2021); Kuzman et al. (2022)). The main advantage of Transformer models in this task is that their representations incorporate both lexical and syntactic knowledge about a text, which is crucial for accurate genre prediction (Kuzman and Ljubešić, 2022). The advancement in these technologies has facilitated the development of multilingual Transformer-based genre classifiers, which can now already be effectively applied to large-scale datasets in various languages. Recently, Laippala et al. (2022) published the Register Oscar dataset, consisting of 351 million documents in 14 languages to which genre labels were automatically assigned. The dataset was created by applying the multilingual genre classifier (Rönnqvist et al., 2021) on the OSCAR datasets (Suárez et al., 2019). Their classifier was trained on genre datasets in multiple languages and uses a schema with 8 genre labels, such as *Narrative*, *Opinion* and *Interactive Discussion*. The unseen-language classification was evaluated based on annotated datasets in eight new languages (Arabic, Catalan, Chinese, Hindi, Indonesian, Portuguese, Spanish and Urdu). The evaluation results for these new

[2] https://www.sketchengine.eu/sltenten-slovenian-corpus/
[3] https://www.sketchengine.eu/bgtenten-bulgarian-corpus/
[4] https://commoncrawl.org
[5] https://archive.org
[6] https://oscar-project.github.io/documentation/versions/oscar-2301/

languages demonstrated promising performance, with F1 scores ranging from 0.58 to 0.82.

These endeavors were closely followed by the first experiments with automatic genre annotation on the MaCoCu data. Kuzman et al. (2023b) applied the X-GENRE classifier, which is described in more detail in Section 3.2, to the English part of the parallel MaCoCu corpora. Based on a preliminary manual evaluation, the genre classifier achieved a macro F1 score of 0.73 and a micro F1 score of 0.88 on a small sample of English texts derived from the Slovenian-English MaCoCu corpus.

3. Construction of the Corpora

3.1. Collecting and Curating Web Corpora

The CLASSLA-web corpora are derived from the web crawls that were gathered and curated inside the MaCoCu project[7] (Bañón et al., 2022). Namely, the following datasets were used: Bosnian web corpus MaCoCu-bs 1.0 (Bañón et al., 2023a), Bulgarian web corpus MaCoCu-bg 2.0 (Bañón et al., 2023b), Croatian web corpus MaCoCu-hr 2.0 (Bañón et al., 2023c), Macedonian web corpus MaCoCu-mk 2.0 (Bañón et al., 2023d), Montenegrin web corpus MaCoCu-cnr 1.0 (Bañón et al., 2023e), Serbian web corpus MaCoCu-sr 1.0 (Bañón et al., 2023f), and Slovenian web corpus MaCoCu-sl 2.0 (Bañón et al., 2023g). They comprise 7 South Slavic languages written in Latin and/or Cyrillic script.

The MaCoCu corpora were collected by crawling the national top-level domains, such as, in the case of Slovenian, the Slovenian top-level domain `.si`. The primary focus of the crawl was on the top-level domain, but texts from generic domains (`.com`, `.net` etc.) were also collected if they were 1. either in the list of seed URLs obtained from previous crawls, or 2. connected well through hyperlinks with the websites in the national top-level domain, and containing sufficient data in the target language. The MaCoCu crawler[8], which is based on the SpiderLing crawler (Suchomel et al., 2012), was used for the crawling. The crawling process was followed by a thorough post-processing to assure high quality data. Firstly, the jusText[9] tool was employed to remove boilerplate content (Pomikálek, 2011). Secondly, the onion[10] tool was used to identify near-duplicate documents (Pomikálek, 2011). Thirdly, the Monotextor[11] tool was used to enhance language identification accuracy and to evaluate the fluency of paragraphs using a language model. Lastly, we further refined the corpora by removing very short texts (shorter than 75 words or consisting only of paragraphs shorter than 70 characters) and texts originating from web domains that had been manually or automatically identified as automatically-generated[12].

Language identification for certain under-resourced languages within the MaCoCu corpora has proven to be a challenge, as these languages are frequently under-represented in the training data of wide-coverage language identification tools. This challenge is particularly pronounced when attempting to differentiate between Bosnian, Croatian, Montenegrin, and Serbian, as these South Slavic languages exhibit a significant level of mutual intelligibility. Thus, to ensure a high level of accuracy, we employed multiple language identification tools at various stages of the pipeline. Firstly, the identification process began with the Google's Compact Language Detector 2 (CLD2)[13] at the document level during the crawling phase. Secondly, the FastSpell[14] tool was used for a more refined language identification at the paragraph level during the post-processing of the corpora with the Monotextor tool. In these initial two steps, Bosnian, Croatian, Serbian and Montenegrin languages were treated as a single macro-language, and the objective in the case of these languages was to determine whether the text belongs to this macro-language or not. Finally, a specialized classifier was employed to distinguish between these four languages. Specifically, they were identified using a Naive Bayes classifier (Rupnik et al., 2023), which relied on lists of words specific to each variety, which were extracted from web texts published on national top-level domains. This approach assumed that, for instance, the Croatian top-level domain (`.hr`) is predominantly associated with the Croatian language.

3.2. Genre Annotation

One crucial part of the creation of CLASSLA-web corpora was also enrichment of the corpora with metadata on genres. This information offers valuable insights into the functional content of the corpora. Furthermore, it facilitates the creation of subcorpora based on genres, which can be used for genre-based linguistic analyses. We used the multilingual X-GENRE classi-

[7] https://macocu.eu/

[8] https://github.com/macocu/MaCoCu-crawler

[9] http://corpus.tools/wiki/Justext

[10] http://corpus.tools/wiki/Onion

[11] https://github.com/bitextor/monotextor/releases/tag/v1.1

[12] Further details regarding the preparation of corpora can be found at https://github.com/macocu/Monolingual-Curation/.

[13] https://github.com/CLD2Owners/cld2

[14] https://github.com/mbanon/fastspell

Language	Other	Mix
Slovenian (sl)	2.7%	5.0%
Croatian (hr)	1.9%	4.1%
Bosnian (bs)	1.5%	2.9%
Montenegrin (cnr)	2.0%	3.3%
Serbian (sr)	1.8%	3.7%
Macedonian (mk)	1.2%	3.3%
Bulgarian (bg)	2.1%	4.1%

Table 1: Percentage of documents annotated with genre labels that do not relate to a specific genre (labels *Other* and *Mix*).

fier[15] (Kuzman et al., 2023a) to automatically annotate the corpora with genre labels. The classifier uses the following genre categories: *Information/Explanation*, *Instruction*, *News*, *Legal*, *Promotion*, *Opinion/Argumentation*, *Prose/Lyrical*, *Forum* and *Other* (for a detailed description of labels, refer to Kuzman et al. (2023a)). When tested in the in-dataset scenario, i.e., on the test split that is derived from the same dataset as the training data, the X-GENRE classifier achieves a micro F1 score of 0.80 and a macro F1 score of 0.79. In the cross-dataset scenario, the classifier still maintains a strong performance with a micro F1 score of 0.68 and macro F1 score of 0.69 (Kuzman et al., 2023a).

We applied the genre classifier to each of the seven CLASSLA-web corpora. The processing was performed on the NVIDIA A100 40GB GPU, with approximately 2,000 predictions executed per second. This resulted in a total processing time of 300 hours.

In addition to genre specific categories, the classifier also returns the category *Other*, which denotes texts that do not fit inside any of the genres present in the schema, as would be the case with an exam, interview, etc. This phenomenon is present in around 2 percent of the documents of each corpus.

Additionally, given that the classifier returns per-class logits that can then be transformed into class probabilities via the softmax function, we introduced a new label, *Mix*. This label is used in cases where none of the categories reached a probability of 0.8, which is a relatively infrequent phenomenon, affecting between 3 and 5 percent of the documents in each corpus. We have come to the decision to use the *Mix* label by performing a manual analysis of samples of instances with lower probabilities of the most probable label. The analysis showed that the suggested upper threshold of 0.8 isolates the documents containing multiple genres very well, with both high precision and high recall.

Table 1 shows the distribution of documents in texts, annotated as *Other* or *Mix*, therefore the label not having a straightforward application. If we omit these labels, we can report that the X-GENRE classifier assigned a specific genre label to approximately 92%–96% of documents in each corpus.

After annotation, we performed a manual analysis of samples from Slovenian, Croatian and Macedonian corpora. Samples consisted of 10 instances per genre label (excluding *Other* and *Mix*), amounting to 80 instances per evaluated corpus. The evaluation showed very high performance of the model, both in Latin and Cyrillic scripts, namely 0.88 in macro F1 for Croatian, 0.93 in macro F1 for Macedonian and 0.94 in macro F1 in case of Slovenian[16].

3.3. Linguistic Annotation

The final layer of enrichment of the original MaCoCu datasets was linguistic annotation, which enables linguistic analyses and simplified querying of the corpora through concordancers. For linguistic annotation the CLASSLA-Stanza pipeline (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023) was used, which provides the state-of-the-art linguistic annotation of Slovenian, Croatian, Serbian, Bulgarian, and Macedonian.

The CLASSLA-Stanza pipeline is based on the Stanza neural pipeline (Qi et al., 2020), but was further improved for processing of South Slavic languages (Terčon and Ljubešić, 2023). Notable enhancements include the support of external inflectional lexicons, which greatly increases performance for morphologically rich languages (Ljubešić and Dobrovoljc, 2019). Additionally, the training datasets used for all models in the pipeline were expanded beyond the Universal Dependencies data. Moreover, the pipeline uses CLARIN.SI-embed word embeddings (Terčon et al., 2023; Terčon and Ljubešić, 2023a,b,c,d) which were trained on larger and more diverse datasets compared to the embeddings used by Stanza. As a result, the CLASSLA-Stanza pipeline demonstrates superior performance compared to Stanza, with error reduction between 34% and 98% on the Slovenian official benchmark SloBENCH[17] (Žitnik and Dragar, 2021) (see Terčon and Ljubešić (2023) for further details).

One highly useful feature of the CLASSLA-Stanza pipeline is that it provides non-standard linguistic processing models for Slovenian, Croa-

[15] https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

[16] More details on the evaluation of the model's cross-lingual performance will be provided in a future publication that we are currently working on.

[17] https://slobench.cjvt.si/

tian and Serbian. These models are trained on non-standard social media texts, as well as on texts from closely related languages, such as Croatian in the case of the Serbian language, and vice versa. Furthermore, the training data used to train these models had diacritics partially removed, which is a phenomenon often seen in non-curated online texts. The CLASSLA-Stanza system includes a "web" processing module that uses a standard tokenizer, but relies on non-standard models for morphosyntactic processing and lemmatization. This setup was shown to be a great solution for linguistic processing of the variety of texts that can be found online. Additionally, the flexibility of these models allows the Croatian web module to be applied to Bosnian and Montenegrin web corpora with high accuracy. Note that Bosnian and Montenegrin language can be considered, in simplified terms, a mixture of the Croatian and Serbian language (Ljubešić et al., 2018) and that the Croatian web module in CLASSLA-Stanza is capable of handling features of both.

Thus, we linguistically annotated the CLASSLA-web corpora using the web module of the CLASSLA-Stanza pipeline, which is implemented as a Python library[18]. The linguistic processing involved tokenization, morphosyntactic annotation and lemmatization. The CLASSLA-Stanza pipeline also allows annotation on the levels of dependency parsing and named entity recognition. However, these two processing stages are not supported for the Macedonian language due to the lack of required training data for that language. If the research community expresses interest for these two annotation layers in CLASSLA-web corpora for the available languages, or, even more, if appropriate training data for Macedonian are produced in the meantime, we will add these two additional annotation layers in the next version of the CLASSLA-web corpora. The final CLASSLA-web corpora have been made freely available on the CLARIN.SI repository as:

- Bosnian CLASSLA-web.bs corpus (Ljubešić et al., 2024a),
- Bulgarian CLASSLA-web.bg corpus (Ljubešić et al., 2024b),
- Montenegrin CLASSLA-web.cnr corpus (Ljubešić et al., 2024e),
- Croatian CLASSLA-web.hr corpus (Ljubešić et al., 2024c),
- Macedonian CLASSLA-web.mk corpus (Ljubešić et al., 2024d),
- Slovenian CLASSLA-web.sl corpus (Ljubešić et al., 2024g),
- Serbian CLASSLA-web.sr corpus (Ljubešić et al., 2024f).

The corpora are also available on the CLARIN.SI NoSketch Engine concordancers[19] which enable easy querying and linguistic analyses.

Corpus	# tokens	# docs
CLASSLA-web.sl	2,153M	4,063k
CLASSLA-web.hr	2,575M	5,422k
CLASSLA-web.bs	802M	1,993k
CLASSLA-web.cnr	177M	401k
CLASSLA-web.sr	2,765M	5,256k
CLASSLA-web.mk	557M	1,482k
CLASSLA-web.bg	3,917M	7,456k
Total	12,948M	26,076k

Table 2: Sizes of CLASSLA-web corpora in millions of tokens and thousands of documents. The following language codes are used: "sl" for Slovenian, "hr" for Croatian, "bs" for Bosnian, "cnr" for Montenegrin, "sr" for Serbian, "mk" for Macedonian, and "bg" for Bulgarian.

4. CLASSLA-web Corpora

In this section, we perform basic analyses of the seven newly introduced corpora. In the first part, we run some general analyses on the size of each of the corpora in terms of token and document count, as well as some basic analyses of the top-level domains from which the data originate. We proceed with a genre-based analysis of the corpora, exploring the relationship between corpora as well as the relationship of specific genres within them.

4.1. General Analysis

The sizes of the resulting corpora for all 7 official South Slavic languages are presented in Table 2. The combined corpora amount to almost 13 billion tokens of running text coming from 26 million documents. Among them, the Montenegrin web corpus is the smallest, consisting of 177 million tokens from 401 thousand documents. The largest corpus was derived from the Bulgarian web, consisting of almost 4 billion tokens, obtained from more than 7.4 million documents.

Each of the presented corpora represents the largest general corpus available for the respective language. What is more, the Macedonian CLASSLA-web corpus is the first linguistically-annotated general corpus of Macedonian. This achievement was made possible not only through the crawl performed inside the MaCoCu project,

[18] https://pypi.org/project/classla/

[19] https://www.clarin.si/ske/

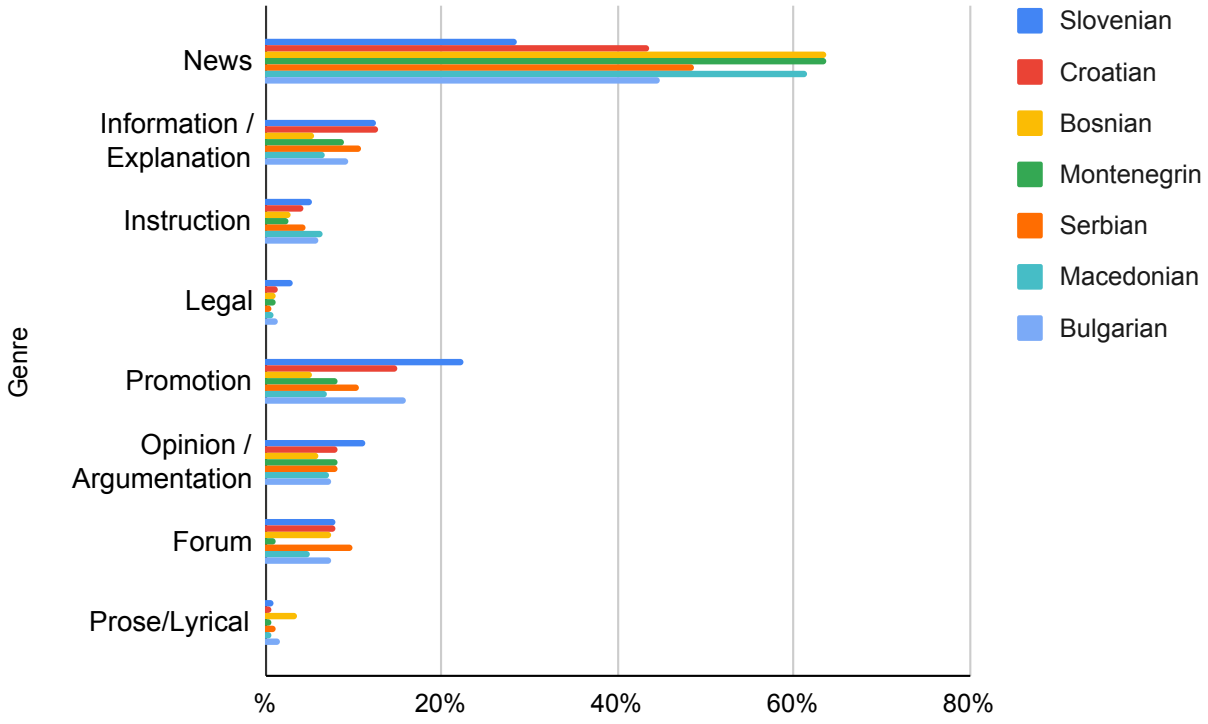

Figure 1: Genre distribution in CLASSLA-web corpora (in percentages of texts).

Corpus	National	Generic
CLASSLA-web.sl	78.2%	21.8%
CLASSLA-web.hr	75.8%	24.2%
CLASSLA-web.bs	62.5%	37.5%
CLASSLA-web.cnr	46.6%	53.4%
CLASSLA-web.sr	63.6%	36.4%
CLASSLA-web.mk	95.2%	4.8%
CLASSLA-web.bg	71.1%	28.9%

Table 3: Distribution of texts, derived from national top-level domains and other, generic domains.

but especially due to the recent inclusion of basic linguistic processing of Macedonian into the CLASSLA-Stanza linguistic processing pipeline (Terčon and Ljubešić, 2023).

Table 3 shows the differences in the distribution of texts that were derived from the national top-level domain, e.g., `.si`, and all other, generic domains, such as `.com`, `.net` etc. The percentage of texts originating from the national top-level domain varies significantly, ranging from only 47% of texts for Montenegrin to 95% for Macedonian. On average, approximately 70% of texts were sourced from the national top-level domain, highlighting the importance of crawling beyond the national top-level domain.

4.2. Genre-Based Analysis

We start our genre-based analysis of our seven new corpora by plotting the genre distribution in texts across these corpora in Figure 1. The first observation to be made is that genres in general are similarly distributed across all seven corpora. The most prevalent genre categories in the analyzed web corpora are *News*, *Information/Explanation*, *Promotion*, *Opinion/Argumentation* and *Forum*. Conversely, *Instruction*, *Legal* and *Prose/Lyrical* are the least represented, with *Prose/Lyrical* accounting for 1% to 4% of texts in the corpora.

On the other hand, the most significant disparities in genre distribution across corpora are observed in the case of *Promotion* and *News*. While the Slovenian web corpus consists only of 28% of *News* texts but 23% of *Promotion* texts, two-thirds of the Bosnian, Montenegrin and Macedonian webs are dominated by the *News* genre, with *Promotion* content constituting only 5–8% of the total texts.

The most significant differences in the occurrence of *Promotion* in web corpora appear to be observed among corpora derived from webs of countries with the greatest disparity in development levels, as measured by gross domestic product (GDP) per capita. Among the seven South Slavic nations, Slovenia has the highest GDP

Genre	Pearson r	p-value
Promotion	0.938	0.002**
Opinion	0.873	0.010*
Information	0.861	0.013*
Legal	0.861	0.013*
News	-0.900	0.006*

Table 4: Results of the Pearson correlation test of the GDP (PPP) per capita for each South Slavic country and the distribution of each genre in CLASSLA-web corpora. Asterisks denote p-values: ** for p<0.005 and * for p<0.05. Only correlations that are statistically significant, i.e., with the p-value below 0.05, are included. *Information/Explanation* is shortened to Information, and *Opinion/Argumentation* is shortened to Opinion.

per capita, corrected for purchasing power parity (GDP PPP per capita), whereas the GDP PPP per capita values of Bosnia and Herzegovina, Montenegro and Macedonia are almost half of that of Slovenia. Given this background knowledge, we decided to examine the relationship between the genre distributions across the 7 web corpora, and the GDP (PPP) per capita for each of the corresponding nations, which serves as a strong metric of development. We inspect the relationship based on the Pearson correlation test. Table 4 shows the results of the correlation test that are statistically significant with p-value below the 0.05 threshold. As anticipated, the *Promotion* genre exhibits the strongest positive correlation with the GDP PPP per capita, with an exceptionally high correlation coefficient of 0.938. Interestingly, the *Information/Explanation*, *Legal* and *Opinion/Argumentation* also demonstrate a relatively strong correlation ranging from 0.86 to 0.87. In contrast, the prevalence of *News* displays a significant negative correlation with the GDP (PPP) per capita, with a correlation coefficient of -0.90.

Given the observation that *News* appears to have a negative correlation with the other genres, as evidenced by the contrasting shapes in Figure 1, we also conducted a pairwise Pearson correlation test among all the genres. The results, presented in Table 5, reveal a nearly perfect negative correlation of -0.972 between the relative frequency of *News* and *Promotion*. This suggests that, as the country evolves from less-developed to better-developed, there is a phenomenon wherein newspaper content in the country's web is increasingly replaced by promotional material. Additionally, high, albeit lower, correlation coefficients can be observed between the *News* genre on one side and the *Information/Explanation*, *Opinion/Argumentation* and *Legal* genre. This analysis allows us to classify the genres into two distinct clusters – one consisting of

Genre pair	Pearson r	p-value
News, Promotion	-0.972	0.000**
News, Information	-0.833	0.020*
News, Opinion	-0.812	0.026*
News, Legal	-0.777	0.040*
Information, Opinion	0.783	0.037*
Information, Promotion	0.813	0.026*
Promotion, Opinion	0.831	0.021*
Legal, Promotion	0.834	0.020*
Legal, Opinion	0.851	0.015*

Table 5: Results of the Pearson correlation test over pairwise genre categories. Categories *Mix* and *Other* are not included in the analysis. Only correlations that are statistically significant, i.e., with the p-value below 0.05, are included. Asterisks denote p-values: ** for p<0.005 and * for p<0.05. *Information/Explanation* is shortened to Information, and *Opinion/Argumentation* is shortened to Opinion.

the *News* content, and the other consisting of *Information/Explanation*, *Opinion/Argumentation*, *Promotion* and *Legal* texts. Based on these findings, we can suggest a preliminary hypothesis that, as a country experiences economic development, its web becomes more diverse and incorporates a greater variety of promotional, opinionated, informational, and legal content. It is important to note that this observation does not take into account other possible factors that might have an impact on the differences between South Slavic webs. Despite being very preliminary, we put forward this hypothesis in hopes of sparking interest for further research on this phenomenon in the wider research community.

5. Conclusions

This paper introduces a collection of comparable web corpora for South Slavic languages. To the best of our knowledge, this is the first corpus collection that comprises comparable general corpora for all languages within a language group. Furthermore, these corpora represent the largest general corpora available for each respective language. The corpus collection comprises 13 billion tokens and 26 million documents in total. Additionally, for the least resourced language in this group, the Macedonian language, the Macedonian CLASSLA-web corpus is the first general linguistically-annotated corpus.

The creation of these corpora was made possible through the contribution of multiple separate endeavors. This paper aims to document all the necessary steps that were undertaken in order to create the corpora. Firstly, the MaCoCu

project (Bañón et al., 2022) played a crucial role by collecting the corpora based on a comprehensive crawl of the national top-level domains and well-interconnected general domains. Additionally, the project implemented various post-processing methods to ensure the high quality of the final datasets. Notably, that research highlighted the significant challenges associated with language identification of less prevalent languages, especially between closely-related South Slavic languages, such as the mutually intelligible Croatian, Bosnian, Montenegrin and Serbian languages.

Secondly, high-quality linguistic annotation of the CLASSLA-web corpora was made possible by recent improvements of the CLASSLA-Stanza pipeline (Terčon and Ljubešić, 2023). In addition to improving annotation accuracy with extended training datasets and embeddings, the new version of CLASSLA-Stanza pipeline provides a "web" module, specifically tailored for linguistic annotation of web corpora. This is particularly advantageous, as web corpora pose a unique challenge to linguistic processing pipelines due to their composition of standard and non-standard texts. What is even more, while Bosnian and Montenegrin do exhibit combinations of linguistic features that distinguish them from Croatian and Serbian, most of these features are covered either in the one or the other language training data, enabling for the Croatian "web" module, trained on standard and non-standard datasets of both Croatian and Serbian, to very successfully annotate the Bosnian and Montenegrin CLASSLA-web corpora. Linguistic annotation facilitates comprehensive analyses of language phenomena for corpus linguists, who perform analyses based on part-of-speech and other linguistic information.

Thirdly, we applied to the corpora a recently-developed multilingual genre classifier. This automatic annotation with genre categories allowed us to enrich the datasets with valuable information on the communicative function of documents inside the seven corpora. Our findings demonstrate that the CLASSLA-web corpora exhibit similar genre distributions. However, we observed that certain corpora, such as Bosnian, Montenegrin, and Macedonian, predominantly consist of news content. In contrast, the Slovenian corpus contains a smaller proportion of news content, with promotional texts representing a significant portion of the corpus. In future research, we plan to significantly extend explorations of genres inside the CLASSLA-web corpora. Firstly, we plan to perform manual analysis on all 7 languages to ascertain what is the overall performance of the multilingual genre classifier on each of the corpora. Secondly, we will investigate possible patterns of biases that could negatively impact downstream research relying on genre labels obtained from multilingual models. Finally, we plan to use genre information to analyze the linguistic characteristics of genres in South Slavic web corpora.

Given the substantial sizes of the CLASSLA-web corpora, they are immensely valuable for the development of language technologies for South Slavic languages and future linguistic analyses. The corpora are already being used for the development of BERT-like and generative pretrained (GPT) language models[20], specific to South Slavic languages. Furthermore, the data can be useful as the starting point for development of manually-annotated training data for numerous tasks. For instance, as part of the Slovenian EMMA project[21], focused on providing NLP solutions to the media industry, samples of the texts annotated as *News* are planned to be used for development of datasets for multilingual topic prediction in news. Moreover, the corpora are highly useful for linguistic analyses as was already shown by their predecessors slWaC, hrWaC (Ljubešić and Erjavec, 2011), srWaC (Ljubešić and Klubička, 2014) and others. Accordingly, we have made the corpora available through the CLARIN.SI concordancers[22] as well, to enable easy corpus querying. To promote their use in the wider linguistic community, which also includes language teachers and digital humanists, the use of the corpora through concordancers will be presented in CLASSLA Express[23], a series of five workshops that will take place in four South Slavic countries.

The MaCoCu approach demonstrated significant success in automatically collecting texts by crawling the top-level domains and beyond, resulting in the creation of the largest general text collections for each of the targeted South Slavic languages. Further enrichment of these text collections with linguistic and genre information resulted in a corpus collection that was not just collected, but also enriched in a highly comparable way, enabling comparable insights in the functional composition of these corpora, as well as unlocking the linguistic research potential of the corpora by performing multi-layer linguistic annotation. To ensure transparency and reproducibility, all steps involved in the process are publicly available. Our future plans involve conducting iterative crawling, following the MaCoCu web corpora collection approach, and enrichment of the South Slavic lan-

[20] See for instance the XLM-R-BERTić (https://huggingface.co/classla/xlm-r-bertic) and YugoGPT (https://huggingface.co/gordicaleksa/YugoGPT) models.

[21] https://emma.ijs.si/en/about-project/

[22] https://www.clarin.si/ske/

[23] https://www.clarin.si/info/k-centre/workshops/classla-express/

guages on a yearly or bi-yearly basis. We have set up a crawling infrastructure inside the CLARIN.SI research infrastructure[24] which is dedicated to iterative crawling of South Slavic webs and webs in other languages. We have already started performing a new run of crawling for Slovenian, Croatian, Serbian, Bosnian and Montenegrin, which will be followed by crawling of Bulgarian and Macedonian. This will allow us to further expand and update the current version of the CLASSLA-web corpora. Consistent updating the corpora will also enable research on how the texts and their distribution on the web are evolving through time, and also enable research in the field of semantic change for South Slavic languages. Additionally, by presenting this process in this paper, we aim to inspire similar initiatives to develop web corpora for other languages lacking large high-quality corpora.

6. Acknowledgments

The research presented in this paper was conducted within the research project "Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language" (J7-4642), the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070, co-funded by the Kliping d.o.o. agency) and within the research programme "Language resources and technologies for Slovene" (P6-0411), all funded by the Slovenian Research and Innovation Agency (ARIS).

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

We would like to thank Petra Bago, Virna Karlić and Lidija Milković for helping to validate and improve the content of the Croatian corpus. We would also like to extend our gratitude to Marija Runić for giving guidance through the complexity of the Bosnian web. We are finally grateful to all our collaborators in the MaCoCu project who made these corpora significantly better.

6.1. Ethical Considerations and Limitations

We are aware that using data that was collected from the web can raise questions of respecting the intellectual property and privacy rights of the original authors of the texts. The authors of the MaCoCu datasets, on which the CLASSLA-web corpora are based on, assured that no sensitive data would be included by only collecting the texts that were freely accessible. Nevertheless, we are aware that the datasets might still include some texts that the authors do not consent to be included. To mitigate this, the CLASSLA-web corpora are published with a notice, which informs the authors of the text that the texts can be taken out of the corpora upon their request.

[24] https://www.clarin.si/

7. Bibliographical References

Marta Bañón, Miquel Esplà-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation*, pages 301–302.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. 2015. The slWaC Corpus of the Slovene Web. *Informatica*, 39(1).

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, pages 125–127.

Taja Kuzman and Nikola Ljubešić. 2022. Exploring the Impact of Lexical and Grammatical Features on Automatic Genre Identification. In *Odkrivanje znanja in podatkovna skladišča - SiKDD: 10. oktober 2022, 10 October 2022, Ljubljana, Slovenia*, Odkrivanje znanja in podatkovna skladišča - SiKDD: 10. oktober 2022, 10 October 2022, Ljubljana, Slovenia. Institut "Jožef Stefan".

Taja Kuzman, Nikola Ljubešić, and Senja Pollak. 2022. Assessing Comparability of Genre

Datasets via Cross-Lingual and Cross-Dataset Experiments. In *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, Jezikovne tehnologije in digitalna humanistika: zbornik konference, page 100–107. Institute of Contemporary History.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023a. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023b. Get to Know Your Parallel Data: Performing English Variety and Genre Classification over MaCoCu Corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103.

Veronika Laippala, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, and Valtteri Skantsi. 2022. Towards better structured and less noisy Web data: Oscar with Register annotations. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 215–221.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings 14*, pages 395–402. Springer.

Nikola Ljubešić and Filip Klubička. 2014. bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, 6(2):100–124.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Georg Rehm and Andy Way. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Nature.

Samuel Rönnqvist, Valtteri Skantsi, Miika Oinonen, and Veronika Laippala. 2021. Multilingual and Zero-Shot is Closing in on Monolingual Web Register Classification. In *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 157–165.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A Benchmark for Discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120.

Roland Schäfer, Felix Bildhauer, et al. 2012. Building large corpora from the web using a new efficient tool chain. In *Lrec*, pages 486–493.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web Library of Babel: evaluating genre collections. In *LREC*. Citeseer.

Jade Goldstein Stewart and J Callan. 2009. *Genre oriented summarization*. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute, School of Computer Science.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Vít Suchomel, Jan Pomikálek, et al. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *arXiv preprint arXiv:2308.04255*.

Marlies Van der Wees, Arianna Bisazza, and Christof Monz. 2018. Evaluation of machine translation performance across multiple genres and languages. In *Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

8. Language Resource References

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Runić, Marija and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023a. *Bosnian web corpus MaCoCu-bs 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023b. *Bulgarian web corpus MaCoCu-bg 2.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023c. *Croatian web corpus MaCoCu-hr 2.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023d. *Macedonian web corpus MaCoCu-mk 2.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023e. *Montenegrin web corpus MaCoCu-cnr 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023f. *Serbian web corpus MaCoCu-sr 1.0*. [link].

Bañón, Marta and Chichirau, Malina and Esplà-Gomis, Miquel and Forcada, Mikel L. and Galiano-Jiménez, Aarón and García-Romero, Cristian and Kuzman, Taja and Ljubešić, Nikola and van Noord, Rik and Pla Sempere, Leopoldo and Ramírez-Sánchez, Gema and Rupnik, Peter and Suchomel, Vít and Toral, Antonio and Zaragoza-Bernabeu, Jaume. 2023g. *Slovene web corpus MaCoCu-sl 2.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024a. *Bosnian web corpus CLASSLA-web.bs 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024b. *Bulgarian web corpus CLASSLA-web.bg 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024c. *Croatian web corpus CLASSLA-web.hr 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024d. *Macedonian web corpus CLASSLA-web.mk 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024e. *Montenegrin web corpus CLASSLA-web.cnr 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024f. *Serbian web corpus CLASSLA-web.sr 1.0*. [link].

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024g. *Slovenian web corpus CLASSLA-web.sl 1.0*. [link].

Terčon, Luka and Ljubešić, Nikola. 2023a. *Word embeddings CLARIN.SI-embed.bg 1.0*. [link].

Terčon, Luka and Ljubešić, Nikola. 2023b. *Word embeddings CLARIN.SI-embed.hr 2.0*. [link].

Terčon, Luka and Ljubešić, Nikola. 2023c. *Word embeddings CLARIN.SI-embed.mk 2.0*. [link].

Terčon, Luka and Ljubešić, Nikola. 2023d. *Word embeddings CLARIN.SI-embed.sr 2.0*. [link].

Terčon, Luka and Ljubešić, Nikola and Erjavec, Tomaž. 2023. *Word embeddings CLARIN.SI-embed.sl 2.0*. [link].

Žitnik, Slavko and Dragar, Frenk. 2021. *SloBENCH evaluation framework*. [link].