The Growing Gains and Pains of Iterative Web Corpora Crawling: Insights from South Slavic CLASSLA-web 2.0 Corpora

Taja Kuzman Pungeršek[*], **Peter Rupnik**[*], **Vít Suchomel**[§], **Nikola Ljubešić**[*†‡]
[*]Jožef Stefan Institute; [†]Faculty of Computer and Information Science, University of Ljubljana;
[‡]Institute of Contemporary History; [§]Natural Language Processing Centre, Masaryk University
[*†‡]Ljubljana, Slovenia; [§]Brno, Czech Republic
{taja.kuzman, peter.rupnik, nikola.ljubesic}@ijs.si; [§]xsuchom2@fi.muni.cz

Abstract

Crawling national top-level domains has proven to be highly effective for collecting texts in less-resourced languages. This approach has been recently used for South Slavic languages and resulted in the largest general corpora for this language group: the CLASSLA-web 1.0 corpora. Building on this success, we established a continuous crawling infrastructure for iterative national top-level domain crawling across South Slavic and related webs. We present the first outcome of this crawling infrastructure – the CLASSLA-web 2.0 corpus collection, with substantially larger web corpora containing 17.0 billion words in 38.1 million texts in seven languages: Bosnian, Bulgarian, Croatian, Macedonian, Montenegrin, Serbian, and Slovenian. In addition to genre categories, the new version is also automatically annotated with topic labels. Comparing CLASSLA-web 2.0 with its predecessor reveals that only one-fifth of the texts overlap, showing that re-crawling after just two years yields largely new content. However, while the new web crawls bring growing gains, we also notice growing pains – a manual inspection of top domains reveals a visible degradation of web content, as machine-generated sites now contribute a significant portion of texts.

Keywords: web corpora, South Slavic languages, genre corpora, topic classification, web, crawling

1. Introduction

Building large, high-quality corpora for less-resourced languages remains a central challenge for the natural language processing field. Crawling top-level domains (TLD) has been shown to be a very efficient method to collect large numbers of texts for less-resourced languages. For South Slavic languages, this approach recently resulted in the largest general corpora for this language group – the CLASSLA-web 1.0 corpora (Ljubešić and Kuzman, 2024). The corpus collection has been widely used for pre-training BERT-like and decoder-only large language models (Ljubešić et al., 2024b; Vreš et al., 2024), for building datasets for downstream natural language processing (NLP) tasks (Kuzman and Ljubešić, 2025; Kuzman et al., 2024), and has been shown to be heavily consulted by linguists (Erjavec et al., 2024; Ljubešić et al., 2024a).

Motivated by these positive results, we set up an ongoing crawling infrastructure dedicated to iterative national top-level domain (TLD) crawling of South Slavic and other webs. In this paper, we present the first outcome of the crawling infrastructure – the CLASSLA-web 2.0 corpus collection, a new release of web corpora for seven South Slavic languages. As the crawling pipeline methodologically mirrors the crawling pipeline from the MaCoCu project (Bañón et al., 2022), with which the CLASSLA-web 1.0 South Slavic corpus collection was produced, this enables us a comparison over time and provides insights into the evolution of the web in the two years between the two crawls.

Our study makes the following contributions. First, we release a substantially larger web corpus collection containing 17.0 billion words in 38.1 million texts across Bosnian, Bulgarian, Croatian, Macedonian, Montenegrin, Serbian, and Slovenian (presented in Section 4). The CLASSLA-web 2.0 corpora can be downloaded in JSONL and VERT formats from the CLARIN.SI repository (Kuzman Pungeršek et al., 2026)[1] and queried through the CLARIN.SI concordancers[2]. Each text is linguistically annotated and labelled with genre and topic categories. Leveraging these annotation layers, we provide a detailed view of the distribution of genres and news topics across South Slavic webs (Section 4.1).

Second, we perform a comparison between the CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora (Section 5) showing that repeating TLD crawls after only two years yields much larger corpora with predominantly new content: on average, 82% of texts in CLASSLA-web 2.0 do not occur in CLASSLA-web 1.0. Analysing the correlation between the URL overlap and content over-

[1]The CLASSLA-web 2.0 corpora are freely accessible in the CLARIN.SI repository at http://hdl.handle.net/11356/2079.

[2]https://www.clarin.si/ske/, the links to each of the corpora are provided on the website https://clarinsi.github.io/classla-web/.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7545–7555
11–16 May 2026.

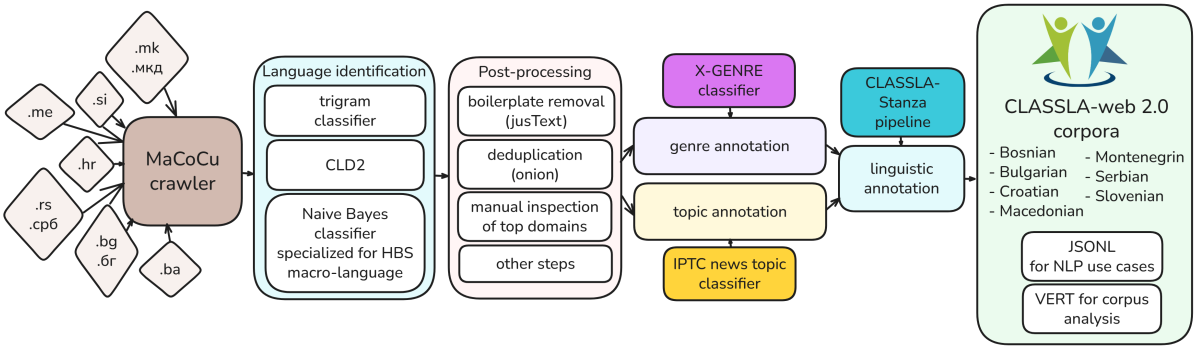

Figure 1: The CLASSLA-web construction pipeline (detailed in Section 3), consisting of several key steps: web crawling, language identification using multiple tools, post-processing to ensure high-quality corpus content, automatic genre and topic annotation, and linguistic annotation.

lap between two CLASSLA-web versions, we propose an approximate function for estimating content overlap via a weighted linear regression (Section 5.2.1). Third, we quantify web evolution in terms of change of sizes (Section 5.1), and web content quality (Section 5.3). We report a visible degradation of web content between versions, with "bad" (e.g., machine-generated or machine-translated) domains now contributing a much larger share of crawled material, underscoring the necessity of manual inspection of web domains that contribute the largest amounts of text.

2. Related Work

The practice of building large web corpora for European languages was established with the WaCky initiative (Baroni et al., 2009). The first web corpora of South Slavic languages were compiled for Slovenian (slWaC) and Croatian (hrWaC, Ljubešić and Erjavec, 2011; Erjavec et al., 2015), followed by Bosnian (bsWaC) and Serbian (srWaC, Ljubešić and Klubička, 2014). After a seven-year gap, the MaCoCu project (Bañón et al., 2022) resumed crawling activities for South Slavic languages and other less-resourced European languages. The monolingual web corpora for South Slavic languages that were produced in the MaCoCu project have later been additionally post-processed and enriched with linguistic and genre annotation, and published as CLASSLA-web 1.0 corpus collection (Ljubešić and Kuzman, 2024). The collection spans Bosnian (Ljubešić et al., 2024a), Bulgarian (Ljubešić et al., 2024b), Montenegrin (Ljubešić et al., 2024e), Croatian (Ljubešić et al., 2024c), Macedonian (Ljubešić et al., 2024d), Slovenian (Ljubešić et al., 2024g), and Serbian (Ljubešić et al., 2024f) web corpora. The CLASSLA-web corpora have received considerable attention in the South Slavic space, both from the NLP community (Ljubešić et al., 2024b; Vreš et al., 2024; Kuzman and Ljubešić, 2025; Kuzman et al., 2024) and from linguists and language teachers (Erjavec et al., 2024; Ljubešić et al., 2024a).

In parallel, numerous other web-based datasets have emerged from the Common Crawl and Internet Archive data collections, e.g., the cc100 dataset (Conneau et al., 2020), the mC4 dataset (Xue et al., 2021), the OSCAR dataset (Suárez et al., 2019), and most recent, the HPLT (de Gibert et al., 2024) and FineWeb2 datasets (Penedo et al., 2025). While these multilingual collections are highly useful for multilingual language modelling tasks, they suffer, inter alia, from lower quality of content, namely, cc100 and mc4 datasets (van Noord et al., 2024); or an unexplainable small amount of content in certain South Slavic languages, namely in OSCAR[3], mC4 dataset (Xue et al., 2021), the HPLT 3.0 datasets (de Gibert et al., 2024), and the FineWeb2 dataset (Penedo et al., 2025). More specifically, the most recent HPLT 3.0 (de Gibert et al., 2024) and FineWeb2 (Penedo et al., 2025) datasets are missing Montenegrin language. Moreover, they report unusual sizes of Serbian and Bosnian corpora where the Bosnian web corpus is significantly larger than the Serbian one, which does not correlate with the number of language speakers or web corpus sizes reported for the CLASSLA-web corpora (Ljubešić and Kuzman, 2024). It seems that these corpora that use "one pipeline to scale them all" suffer from language identification issues related to the Serbian, Bosnian and Montenegrin languages. Furthermore, since the introduction of large language models that make text generation in various languages easily available, there is a growing problem on the web that more and more content is automatically generated, as

[3] https://oscar-project.github.io/documentation/versions/oscar-2301/

we will show in Section 5.3. It is thus crucial that manual inspection of web corpora content by native speakers is included in the corpora preparation to filter out as much bad content as possible, and to assure high quality of data. Our work addresses these gaps via targeted TLD crawling, specialised language discrimination between the mutually intelligible Croatian, Serbian, Bosnian and Montenegrin languages, and manual domain validation.

3. CLASSLA-web Construction Pipeline

The goal of the crawling pipeline is to produce large, clean, and richly annotated web corpora. As shown in Figure 1, the process consists of 1) crawling based on top-level national domains and post-processing, 2) automatic text annotation with genre and topic information, and 3) linguistic annotation. In this section, we briefly present each step. For more details, refer also to the description of the CLASSLA-web crawling pipeline in Ljubešić and Kuzman (2024) and to the website that presents the CLASSLA-web collections (https://clarinsi.github.io/classla-web/).

Web Crawling Pipeline Following the methodology set up in the MaCoCu project (Bañón et al., 2022), we crawl national top-level domains (e.g., `.si` for Slovenian), as well as connected generic domains (`.com` and others). Crawling is performed with the MaCoCu crawler[4], which is based on the SpiderLing crawler (Suchomel et al., 2012) and designed for an efficient large-scale web text collection.

Language Identification Two language identification tools are used at the document and paragraph levels, namely a trigram classifier (Lui and Baldwin, 2014) and the Compact Language Detector 2 (CLD2, Sites, 2013)[5]. Language identification between Bosnian, Croatian, Montenegrin, and Serbian, comprising the HBS macro-language, has been shown to be particularly challenging, as these South Slavic languages exhibit a significant level of mutual intelligibility (Rupnik et al., 2023). We directly assign texts that come from national top-level domains to the corresponding web corpus, e.g., web domains from `.hr` are included in the Croatian web corpus. For generic top-level domains, we use an additional language identification tool specialised for distinguishing between Bosnian, Croatian, Montenegrin and Serbian, namely a Naive Bayes classifier trained on language-specific wordlists extracted from national TLDs (Rupnik et al., 2023). During the development of CLASSLA-web 2.0 corpora, we also explored language identification through zero-shot prompting with large language models, but the specialised classifier proved to be more reliable for HBS disambiguation.

Post-Processing After crawling, we remove the boilerplate with the jusText[6] tool (Pomikálek, 2011) and near-duplicates with the onion[7] tool (Pomikálek, 2011), and discard exact duplicates. We discard texts that are not written in the target language, repair encoding, clean unwanted HTML elements, and filter overly short documents (comprising only short paragraphs with less than 70 characters or consisting of less than 75 words).

Manual Inspection of Top Domains We perform a manual validation of web domains that provide the most texts in each corpus. We remove texts coming from domains that are manually identified to predominantly contain machine translation, automatically generated text, or severe encoding issues. As shown later (Section 5.3), this step has become crucial for ensuring the high quality of the web corpus.

Genre Annotation We use the multilingual X-GENRE classifier (Kuzman and Ljubešić, 2024, 2023) to automatically assign to each text one of the following labels: *Information/Explanation*, *Instruction*, *News*, *Legal*, *Promotion*, *Opinion/Argumentation*, *Prose/Lyrical*, *Forum*, or *Other* (for a detailed description of labels, refer to Kuzman et al. (2023)). Based on a multilingual AGILE genre classification benchmark[8], the model yields high performance across European languages (micro-F1 and macro-F1 of 0.85) and particularly high scores on South Slavic languages, both in Latin and Cyrillic scripts, namely macro-F1 of 0.89 for Croatian, 0.91 for Macedonian, and 0.94 for Slovenian. To ensure high annotation precision, we additionally use the label *Mix* for instances where no single label exceeds a probability of 0.8. This threshold, identified through manual analysis, effectively isolates documents containing multiple genre categories.

Topic Annotation In CLASSLA-web 2.0 corpora, a new annotation layer is introduced – in addition to genres, texts are annotated with topics to provide an additional insight into the corpora content. We

[4] https://github.com/macocu/MaCoCu-crawler
[5] https://github.com/CLD2Owners/cld2
[6] http://corpus.tools/wiki/Justext
[7] http://corpus.tools/wiki/Onion
[8] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

use a multilingual news topic classifier (Kuzman and Ljubešić, 2025) covering 17 top level labels from the IPTC Media Topic NewsCodes schema (e.g., *Politics*, *Science and Technology*, *Sport* – for a detailed description of labels, refer to Kuzman and Ljubešić (2025)). On a manually-annotated test set in Croatian, Slovenian, Catalan, and Greek, the model achieves macro-F1 of 0.75 and micro-F1 of 0.73. For instances assigned with class probabilities below 0.6, we use the *Mix* label. As with genre annotation, the confidence threshold was identified based on manual analysis. We should note that the classifier has been evaluated for now only on news texts, as it was primarily intended for news topic classification. Thus, in Section 4.1, we analyse topics only within the *News* genre.

Linguistic Annotation All texts are linguistically annotated with the CLASSLA-Stanza pipeline (Ljubešić et al., 2024c)[9], providing tokenization, lemmatization and morphosyntactic annotation with state-of-the-art accuracy for Slovenian[10], and a high performance on Croatian, Serbian, Bulgarian, and Macedonian (Ljubešić et al., 2024c). We use the "web" module of the pipeline, which has been shown to perform well on heterogeneous web texts. The Croatian module is applied to Bosnian and Montenegrin. More detailed argumentation behind the module choices is provided in Ljubešić et al. (2024c).

Differences in the Pipeline Between Versions 1.0 and 2.0 Although the overall CLASSLA-web pipeline is more or less the same between versions 1.0 and 2.0 of the CLASSLA-web corpora, three improvements were implemented during the development of the CLASSLA-web 2.0 version. First, near-duplicate removal was improved by masking numbers, punctuation, and links during paragraph-level deduplication to improve the efficiency of identifying texts that differ only in numbers. Second, we expanded the manual inspection of the top domains from only Slovenian and Croatian in CLASSLA-web 1.0 to all languages in CLASSLA-web 2.0. Third, we added topic annotation to complement genre labels.

4. CLASSLA-web 2.0 Corpora

In this section, we present the CLASSLA-web 2.0 corpus collection, collected from the South Slavic webs in 2024 and processed following the procedure detailed in Section 3.

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

[10] Based on the SloBENCH benchmark at https://slobench.cjvt.si/leaderboard/view/11.

The CLASSLA-web 2.0 collection comprises all seven South Slavic national webs and languages – Bosnian (bs), Bulgarian (bg), Croatian (hr), Macedonian (mk), Montenegrin (cnr), Serbian (sr), and Slovenian (sl). In total, it consists of 17 billion words and 38 million texts, as reported in Table 1. The biggest corpus, namely the Bulgarian (CLASSLA-web.bg 2.0) corpus, comprises 6 billion words and 15 million texts. Other large corpora are Serbian (CLASSLA-web.sr 2.0), Croatian (CLASSLA-web.hr 2.0) and Slovenian (CLASSLA-web.sl 2.0) with 2.3 to 3.7 billion words and 5 to 7 million texts. Bosnian (CLASSLA-web.bs 2.0) and Macedonian (CLASSLA-web.mk 2.0) corpora are smaller, comprising 0.7 to 1 billion words and 2.1 to 2.5 million texts. The Montenegrin (CLASSLA-web.cnr 2.0) corpus is the smallest with 0.3 billion words and 0.8 million texts. However, as shown in Section 5.1, the new Montenegrin corpus is twice the size of the Montenegrin corpus included in the CLASSLA-web 1.0 collection and, to the best of our knowledge, represents the largest general Montenegrin corpus available.

Corpus	Words (billion)	Texts (million)
CLASSLA-web.bs 2.0	1.01	2.54
CLASSLA-web.bg 2.0	5.99	14.67
CLASSLA-web.hr 2.0	3.01	5.92
CLASSLA-web.mk 2.0	0.69	2.11
CLASSLA-web.cnr 2.0	0.29	0.79
CLASSLA-web.sr 2.0	3.71	7.24
CLASSLA-web.sl 2.0	2.31	4.79
Total	17.01	38.06

Table 1: Sizes of CLASSLA-web 2.0 corpora in billion words and million texts for Bosnian (bs), Bulgarian (bg), Croatian (hr), Macedonian (mk), Montenegrin (cnr), Serbian (sr) and Slovenian (sl) corpora.

4.1. Genre and Topic Distribution

Genre and topic labels provide valuable insights into similarities as well as differences between web corpora. As shown in Figure 2, *News* content dominates in all national webs. This genre is especially dominant in Bosnian, Macedonian and Montenegrin corpora, where it exceeds 70%, whereas it represents 30–40% of texts in Croatian and Slovenian corpora. As in CLASSLA-web 1.0 corpora (Ljubešić and Kuzman, 2024), *Promotion* is more prevalent in Slovenian, Croatian and Bulgarian web corpora (12–20%) than in others. Another genre where there is a visible difference between corpora is *Forum*. Texts from forums seem to be much more frequent in Bulgarian, Serbian, and Slovenian webs (10–13%) than in Bosnian, Macedonian,

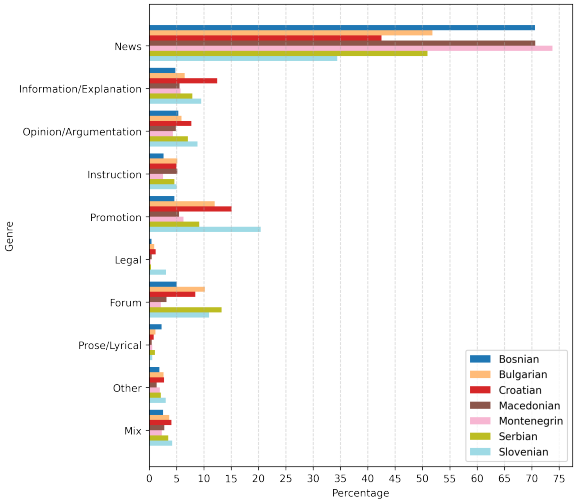

Figure 2: Distribution of genre categories across South Slavic CLASSLA-web 2.0 corpora.

and Montenegrin webs (below 5%).

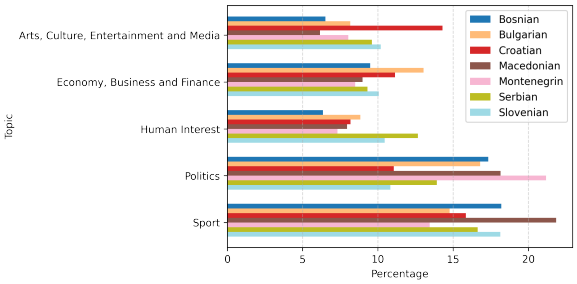

Figure 3: Most frequent topics in *News* texts in CLASSLA-web 2.0 corpora. The topics that are represented by 10% or more texts in at least one corpus are included in the figure.

As *News* is the predominant genre in all web corpora and given that the topic classifier has been primarily validated on news texts, we analyse topics within the *News* genre. In Figure 3, we show the distribution of topics that are discussed in more than 10% of news texts in at least one corpus. Despite the fact that the news topic schema comprises 17 topic labels, only five categories appear with high enough frequency in the seven corpora, namely "*Sport*", "*Politics*", "*Economy, Business and Finance*", "*Arts, Culture, Entertainment and Media*", and "*Human Interest*", in this order of average frequency across corpora. These five categories account for approximately 60% of news texts in each corpus. *Sport* is the most common topic of news texts in the majority of web corpora, and it even represents a fifth of all news in the Macedonian web corpus. In Bulgarian and Montenegrin web corpora, the most frequent topic of news texts is *Politics*, accounting for 20% of news texts within the Montenegrin corpus. The distribution of the other three most frequent topic categories is relatively consistent across the South Slavic corpora, except in the case of *Arts, Culture, Entertainment and Media* which is notably more prevalent in the Croatian corpus compared to the others, representing almost 15% of Croatian news texts.

These findings show that the information on genre and topic of web texts – aside from being very useful for data selection in various research scenarios – provides valuable insights into the variation of content in South Slavic webs. At the same time, a consistent pattern emerges across all South Slavic corpora: the predominance of news texts on the web, addressing similar topics.

5. Evolution of the Web between CLASSLA-web 1.0 and CLASSLA-web 2.0

To understand how national webs evolve over short time spans, we compare the CLASSLA-web 2.0 corpora with CLASSLA-web 1.0 (Ljubešić and Kuzman, 2024). Both web corpus collections were compiled using an identical construction pipeline (see Section 3), with the 1.0 version having been collected from the web two to three years prior to the 2.0 version, specifically in 2021 and 2022. In the following sections, we examine the size difference between the two corpora collections (Section 5.1), content overlap (Section 5.2), and the presence of automatically generated texts and other low quality content (Section 5.3).

5.1. Size Gains

As shown in Figure 4, the second crawling iteration produced larger web corpora, showing the increase of content on the web. Compared to the size of the CLASSLA-web 1.0 corpora, the web corpora in the 2.0 corpus collection comprise 57% more words and 46% more texts, that is, roughly half as many more words and texts than in the first version. Interestingly, web growth is uneven across South Slavic webs – the Slovenian and Croatian webs exhibit modest increase of size, with a 30% change in size in terms of number of words, while the size of Bulgarian and Montenegrin web corpora nearly doubled in the second crawling iteration. These differences in sizes reflect the high growth of content on the web rather than any differences in the crawling methodology, as both the CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora have been produced with practically the same pipeline.

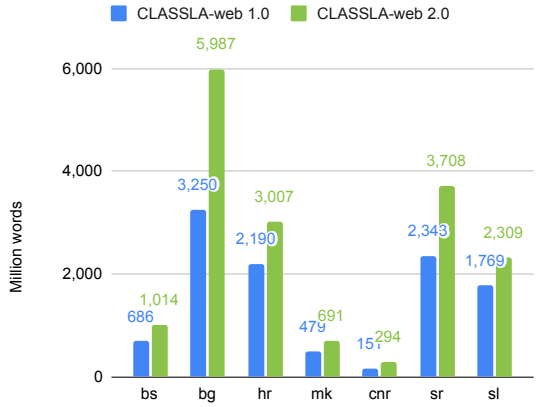

Figure 4: Comparison of sizes in millions of words between CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora for Bosnian (bs), Bulgarian (bg), Croatian (hr), Macedonian (mk), Montenegrin (cnr), Serbian (sr) and Slovenian (sl) corpora.

5.2. Small Overlap Between the Two Crawls

A comparison of CLASSLA-web 1.0 and 2.0 shows that the new edition of the web corpora contains up to twice as many texts as version 1.0. In this section, we examine the degree of overlap between the 1.0 and 2.0 versions of the CLASSLA-web corpora. This analysis aims to determine whether the growth in corpus size of the 2.0 version can be partially attributed to the inclusion of the majority of the 1.0 corpora within the 2.0 corpora. Additionally, this investigation provides us with insights into the degree of change in the content of South Slavic webs over a short period between the two observed crawls.

We measure content overlap by identifying near-duplicated texts using MinHash over word 4-grams, with a similarity threshold set to 0.7, defined via manual validation. Surprisingly, despite the fact that only two years have passed between the two crawls, there is a small content overlap – only around 20% of texts from version 1.0 are present in the CLASSLA-web 2.0 corpora.[11]

In Table 2, we report the share of texts unique to each crawling iteration. On average, 82% of texts (comprising approximately 81% of words) in CLASSLA-web 2.0 corpora are new relative to the CLASSLA-web 1.0 corpora. These findings indicate a rapid turnover of web content: most of the content present in version 1.0 has disappeared in two years, and the majority of the content in version 2.0 has appeared recently.

Taking into account also changes in corpora sizes, shown in Figure 4, Bulgarian (bg) and Montenegrin (cnr) are the fastest-changing and expanding webs. The CLASSLA-web 2.0 corpora from these two national webs nearly doubled in size compared to the 1.0 versions, while retaining only 11–14% of content that was already present in the 1.0 corpora. In contrast, Croatian web changed the least, with 24% of texts in version 2.0 already being present in version 1.0.

corpus	% unique texts in CLASSLA-web 2.0	% unique texts in CLASSLA-web 1.0
bs	79.41%	73.52%
bg	85.79%	72.08%
hr	76.08%	73.85%
mk	79.94%	71.42%
cnr	88.68%	77.32%
sr	79.67%	70.63%
sl	80.60%	77.09%
average	81.77%	73.09%

Table 2: Percentage of texts in CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora that are unique to that crawling iteration, that is, they do not appear in the other web corpus version.

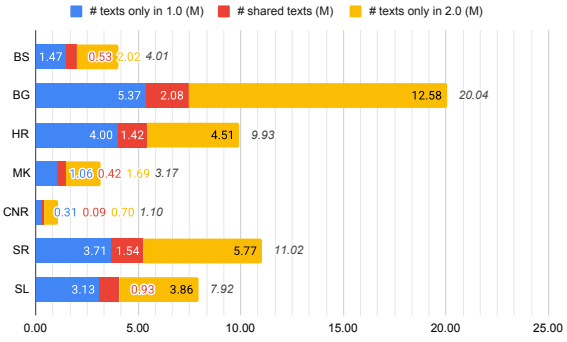

Figure 5: Number of texts that are unique in CLASSLA-web 1.0 or 2.0 versions, number of texts that are shared between the two versions, and a total number when the two corpora are merged.

The small content overlap also suggests that through iterative crawling over a two-year period, it is possible to collect large quantities of texts. Figure 5 shows the combined size of CLASSLA-web 1.0 and CLASSLA-web 2.0, with texts from version 1.0 excluded from version 2.0. The total size of the merged CLASSLA-web corpora is 57 million texts,

[11] Detailed information indicating which texts in version 2.0 are near duplicates of specific texts in version 1.0 is provided at https://github.com/clarinsi/classla-web/tree/main/more-info/duplicates-with-1.0.

with the Bulgarian corpus amounting to 20 million texts. These results demonstrate the effectiveness of iterative crawling for collecting large-scale text resources for less-resourced languages, and provide a strong motivation to continue with biennial crawling efforts.

5.2.1. URL Overlap as a Quick Insight into Content Overlap

In the previous section, we assessed the extent of overlapping content between two corpora using the MinHash algorithm to detect near-duplicate texts. Since this method is computationally intensive, we explore faster alternatives for gaining initial insights into content overlap. One potential indicator of the degree of content overlap is the overlap of URLs of web pages from which the texts were collected, as this information is included in the metadata for each text. Figure 6 shows a correlation between the percentage of shared URLs and the percentage of shared texts across both corpora. This relationship, based on seven pairs of national CLASSLA-web 1.0 and 2.0 corpora (e.g., CLASSLA-web.hr 1.0 and CLASSLA-web.hr 2.0), is strong, with a Pearson correlation coefficient of 0.908 and a p-value of 0.0047.

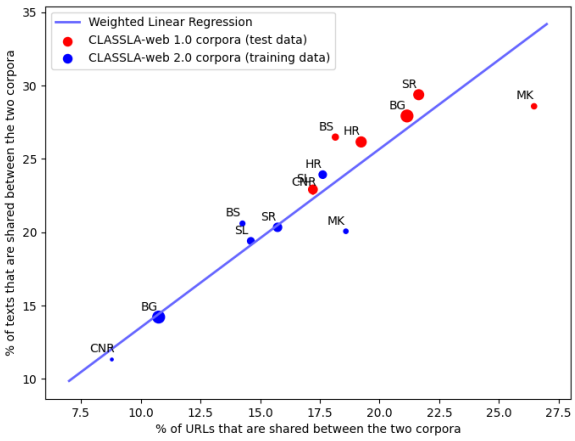

Figure 6: Correlation between percentage of shared URLs and shared texts between CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora, and a weighted linear regression predicting percentage of shared texts based on percentage of shared URLs.

As shown in Figure 6, this correlation can be modelled using a weighted linear regression, implemented in the scikit-learn Python library[12]. The model was trained on URL and text overlap data from the seven CLASSLA-web 2.0 corpora, specifically their overlap with the corresponding CLASSLA-web 1.0 corpora. The weights were based on the normalised sizes of the CLASSLA-web 2.0 corpora. When tested on CLASSLA-web 1.0 data (i.e., the percentage of overlapping texts and URLs with version 2.0), the model's prediction errors are 0.4–5 percentage points, and below 1.7 percentage points on larger corpora, namely Bulgarian, Croatian, Slovenian and Serbian. On version 2.0, the error ranges from 0.15 to 4 points, with deviations of 1 point or less for the larger corpora.

$$textOverlap = 1.347 + 1.216 \cdot URLOverlap \quad (1)$$

Equation 1 shows the function derived from the regression model, which can be used to estimate the likely percentage (0-100) of shared texts between two corpora (*textOverlap*) based on the percentage of overlapping URLs (*URLOverlap*). This function offers a rough but useful approximation of content overlap between large web corpora, enabling quick comparisons when detecting near-duplicates directly would be too time-consuming or computationally expensive.

5.3. Degradation of Web Content

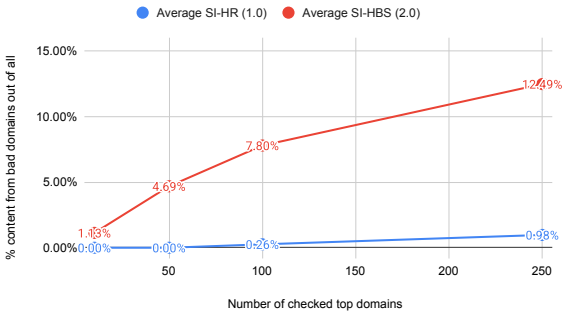

Figure 7: Percentage of content (relative to the total corpus content) originating from bad domains among the most frequent web domains in manually inspected Western South Slavic CLASSLA-web 1.0 and CLASSLA-web 2.0 corpora.

While the comparison between versions 1.0 and 2.0 of the CLASSLA-web corpora revealed substantial gains in text content, manual inspection of the most frequent domains also showed a notable increase in low-quality or automatically generated content.[13] As shown in Figure 7, a manual review of the Western South Slavic CLASSLA-web 2.0

[12] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[13] Lists of domains from the CLASSLA-web corpora that have been manually verified or identified as

corpora found that many of the top 250 domains were problematic, accounting for 15% of the total number of texts before their removal.

In contrast, this issue was much less prominent in the CLASSLA-web 1.0 corpora, crawled in 2021 and 2022. There, low-quality domains rarely appeared among the top 250 domains, and manual inspection led to the removal of only 1% of texts – 15 times less than in the 2.0 version, which was crawled in 2024. The likely cause of this surge in low-quality content is the growing availability of automated text generation tools. These findings highlight the importance of manually reviewing the most frequent domains as a necessary step in preparing high-quality web corpora.

6. Conclusion

In this paper, we present the CLASSLA-web 2.0 corpus collection for South Slavic languages, a result of an ongoing biennial web crawling infrastructure. Although the corpora were collected from the same national domains just two years after the CLASSLA-web 1.0 collection (Ljubešić and Kuzman, 2024), the new datasets are significantly larger and contain mostly new content. On average, the 2.0 versions include around 50% more words and texts than the 1.0 versions. In total, the CLASSLA-web 2.0 corpora consist of 17 billion words and 38 million texts in seven South Slavic languages. The corpora are freely available in JSONL and VERT formats from the CLARIN.SI repository (Kuzman Pungeršek et al., 2026)[14] and can be queried via the CLARIN.SI concordancers[15]. More information on corpora development and accessibility is available on a website dedicated to the CLASSLA-web collections (https://clarinsi.github.io/classla-web/).

In addition to providing metadata on the collected texts, the corpora are automatically annotated with genre, topic, and linguistic information to support various types of linguistic analyses. The genre and topic annotations offer valuable insights into the characteristics of web content and reveal interesting similarities and differences across the South Slavic web corpora. Our analysis shows that all South Slavic web corpora are predominantly composed of news content, which is largely focused on five main topics: "*Sport*", "*Politics*", "*Economy, Business and Finance*", "*Arts, Culture, Entertainment and Media*", and "*Human Interest*".

Comparison with the previous CLASSLA-web 1.0 versions showed that approximately 80% of the content in each version is unique, indicating that only about one-fifth of the CLASSLA-web 2.0 content overlaps with the CLASSLA-web 1.0 release. These results demonstrate that iterative crawling of top-level national domains is a highly effective approach for collecting large-scale text data for less-resourced languages.

Precise measurement of overlapping content requires identifying near-duplicate texts, which is computationally expensive and time-consuming for massive corpora. As an additional contribution, we propose using the percentage of shared URLs as a rough estimate of the amount of overlapping content between web corpora. Based on overlap calculations for seven pairs of 1.0 and 2.0 versions of the CLASSLA-web corpora, we find a strong correlation between URL and text overlap. Building on this, we propose a simple weighted linear model (Equation 1) that provides quick estimates of content overlap when near-duplicate detection is not feasible.

Since the CLASSLA-web 2.0 corpora were developed using the same pipeline as the previous version, CLASSLA-web 1.0, this enables longitudinal analyses of changes in South Slavic web content between the two crawls, that is, between 2021 and 2024. Most notably, there has been a significant increase in low-quality or automatically generated content. Manual inspection of the most frequent domains in CLASSLA-web 2.0 reveals that low-quality domains account for a much larger share of both texts and words compared to version 1.0 and appear far more often among the top-ranked domains. Specifically, an inspection of the top 250 domains identified a large number of bad domains, which contributed 15% of the total number of texts prior to their removal from the cleaned corpora. These findings underline the growing importance of a manual validation of high-frequency domains to maintain web corpus quality, especially in light of the widespread availability of tools for automatic large-scale text generation.

To conclude, despite the growing pains related to maintaining high-quality web corpora that would be free from automatically-generated texts, web crawling remains a highly effective method for collecting large-scale text resources for less-resourced languages such as South Slavic languages. Due to the rapid turnover of online content, repeating the crawl every two years can more than double the size of the total collected corpora and our results show small content overlap between versions collected two years apart. Thanks to their size and recency, the CLASSLA-web 2.0 corpora have already attracted interest within the South Slavic research and language technology development communities. The corpora have been included in the pretraining data for Slovenian open-source large language models (Vreš et al., 2024) and introduced

low quality are provided at https://github.com/clarinsi/classla-web/tree/main/urls.

[14] http://hdl.handle.net/11356/2079

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

to linguists through recent editions of the CLASSLA-Express workshops that focus on the use of web corpora and concordancers for linguistic analyses (Ljubešić et al., 2024a). Highly motivated by the community's uptake of this new web corpus collection, we will continue our biennial web crawling efforts to provide increasingly bigger and up-to-date web corpora for South Slavic languages.

7. Ethical Considerations and Limitations

We are aware that using web-collected data can raise questions regarding intellectual property and the privacy rights of the original authors. The web crawling pipeline is designed to avoid collecting sensitive data by only retrieving texts that are freely accessible. Nevertheless, we recognize that some texts in the datasets may still have been included without the authors' explicit consent. To address this, the CLASSLA-web corpora are published with a notice informing authors that their texts can be removed from the corpora upon request.

8. Acknowledgements

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 "Large Language Models for Digital Humanities" (GC-0002), the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070, co-funded by the Kliping d.o.o. agency), the Research Infrastructure DARIAH-SI (I0-E007), and within the research programme "Language resources and technologies for Slovene" (P6-0411), all funded by the Slovenian Research and Innovation Agency (ARIS).

This research was also supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This project is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

We would like to thank the CLASSLA knowledge centre for South Slavic languages and the Slovenian CLARIN.SI infrastructure for their valuable support.

9. 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.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

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

Tomaž Erjavec, Nikola Ljubešić, Katja Meden, Taja Kuzman, Cyprian Laskowski, Jan Jona Javoršek, Simon Krek, Mateja Jemec Tomazin, Kaja Dobrovoljc, Špela Arhar Holdt, et al. 2024. CLARIN.SI, the Slovenian node of CLARIN: ten years on. In *CLARIN Annual Conference*, pages 71–85.

Taja Kuzman and Nikola Ljubešić. 2025. LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. 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, Tanja Pavleska, Urban Rupnik, and Primož Cigoj. 2024. PandaChat-RAG: Towards the benchmark for Slovenian RAG Applications. In *Proceedings of the Slovenian Conference on Artificial Intelligence 2024, Vol. A, Ljubljana, Slovenia*, page 7–10. Institute „Jožef Stefan“.

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 and Taja Kuzman. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282.

Nikola Ljubešić, Taja Kuzman, Ivana Filipović Petrović, Jelena Parizoska, and Petya Osenova. 2024a. CLASSLA-Express: a Train of CLARIN. SI Workshops on Language Resources and Tools with Easily Expanding Route. In *CLARIN Annual Conference Proceedings 2024*, pages 31–35. Barcelona: CLARIN.

Nikola Ljubešić, Vít Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024b. Language models on a diet: Cost-efficient development of encoders for closely-related languages via additional pretraining. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 189–203, Torino, Italia. ELRA and ICCL.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024c. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Proceedings of the Conference on Language Technologies & Digital Humanities (JTDH 2024)*, pages 251–274.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora. *Disertacnı práce, Masarykova univerzita, Fakulta informatiky*.

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.

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.

Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do Language Models Care About Text Quality? Evaluating Web-Crawled Corpora Across 11 Languages. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 5221–5234. European Language Resources Association (ELRA).

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. *Jezikovne tehnologije in digitalna humanistika*, pages 485–511.

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.

10. Language Resource References

Kuzman, Taja and Ljubešić, Nikola. 2023. *Multilingual text genre classification model X-GENRE*. Hugging Face. PID 10.57967/hf/0927.

Kuzman, Taja and Ljubešić, Nikola. 2024. *Multilingual text genre classification model X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1961.

Kuzman, Taja and Ljubešić, Nikola. 2025. *Multilingual IPTC News Topic Classifier*. Hugging Face. PID 10.57967/hf/4709.

Kuzman Pungeršek, Taja and Rupnik, Peter and Ljubešić, Nikola. 2026. *South Slavic web corpus collection CLASSLA-web 2.0*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/2079.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024a. *Bosnian web corpus CLASSLA-web.bs 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1927.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024b. *Bulgarian web corpus CLASSLA-web.bg 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1928.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024c. *Croatian web corpus CLASSLA-web.hr 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1929.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024d. *Macedonian web corpus CLASSLA-web.mk 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1932.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024e. *Montenegrin web corpus CLASSLA-web.cnr 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1930.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024f. *Serbian web corpus CLASSLA-web.sr 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1931.

Ljubešić, Nikola and Rupnik, Peter and Kuzman, Taja. 2024g. *Slovenian web corpus CLASSLA-web.sl 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1882.

Sites, Dick. 2013. *Compact language detector 2*. GitHub. Software available at `https://github.com/CLD2Owners/cld2` (last updated on August 2015).