Cultural Topic Modelling over Novel Wikipedia Corpora for South-Slavic Languages

Filip Markoski[1], Elena Markoska[1], Nikola Ljubešić[2], Eftim Zdravevski[1], and Ljupcho Kocarev[3,1]

[1]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia,
[1]e-mails: filip.markoski45@gmail.com, elena.markoska@live.com, eftim.zdravevski@finki.ukim.mk
[2]Jožef Stefan Institute, Slovenia, e-mail: nikola.ljubesic@ijs.si
[3]Macedonian Academy of Sciences and Arts, Skopje, Macedonia, e-mail: lkocarev@manu.edu.mk

Abstract

There is a shortage of high-quality corpora for South-Slavic languages. Such corpora are useful to computer scientists and researchers in social sciences and humanities alike, focusing on numerous linguistic, content analysis, and natural language processing applications. This paper presents a workflow for mining Wikipedia content and processing it into linguistically-processed corpora, applied on the Bosnian, Bulgarian, Croatian, Macedonian, Serbian, Serbo-Croatian and Slovenian Wikipedia. We make the resulting seven corpora publicly available. We showcase these corpora by comparing the content of the underlying Wikipedias, our assumption being that the content of the Wikipedias reflects broadly the interests in various topics in these Balkan nations. We perform the content comparison by using topic modelling algorithms and various distribution comparisons. The results show that all Wikipedias are topically rather similar, with all of them covering art, culture, and literature, whereas they contain differences in geography, politics, history and science.

1 Introduction

Researchers studying the South-Slavic languages often face difficulties finding corpora in the respective languages. While for Slovenian there is a significant amount of corpora available both for search and download (Krek et al., 2020; Fišer et al., 2020; Erjavec et al., 2020), most other languages in this language group do not enjoy this commodity. This is the reason why web corpora with all their limitations and uncertainties are so popular in this language group[1]. This paper describes an effort to add one additional, consistent source of good-quality text for South-Slavic languages - Wikipedia corpora. To perform that, we set up a robust pipeline for preparing and linguistically processing all currently available Wikipedias of South-Slavic (macro-)languages. We use the notion of *(macro-)languages* to signify that the selected Wikipedias refer to six national languages with a respective ISO-639-1 code, and one macro-language, Serbo-Croatian, with its ISO-639-3 code *hbs*.

Processing Wikipedia corpora could seem a relatively simple task for researchers in computer science, and therefore such processing is performed on a project basis, as was already done for some South-Slavic languages (Ljubešić and Fišer, 2013; Svoboda and Beliga, 2017). However, other scientific disciplines lack the technical expertise to perform these data preparations. Furthermore, processing Wikipedia data on a per-project basis entirely disregards the questions of replicability and reproducibility of research, making the measurements or experiments performed on different dumps with different preprocessing entirely incomparable.

We are trying to break away from this poor practice by giving access to fully processed and linguistically annotated Wikipedia corpora of South-Slavic (macro-)languages whose updates we plan to publish on a yearly basis in the years to come. The current versions of the corpora are based on Wikipedia dumps[2] of the Bosnian (*bs*), Bulgarian (*bg*), Croatian (*hr*), Macedonian (*mk*), Serbian (*sr*), Serbo-Croatian (*sh*) and Slovenian (*sl*) Wikipedia, downloaded on October 17th 2020.

They are made available for download and search via the CLARIN.SI repository[3], and ad-

[1] The paper describing the Croatian National Corpus has been cited since 2017 on Google Scholar 19 times, while the paper describing the older version of the Croatian web corpus was cited in the same period 58 times. Some other languages in the language group have the web corpora as their only choice, while some do not have even that.

[2] `https://dumps.wikimedia.org`

[3] `https://www.clarin.si/repository/xmlui/`

Proceedings of Recent Advances in Natural Language Processing, pages 910–917
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_104

ditionally, they can be searched through the CLARIN.SI concordances.[4] [5]

Aside from documenting the methodology applied in preparing these corpora, we perform a topic modelling experiment on the corpora, shedding some light on the topical similarities and differences between the seven corpora. While Wikipedia as a research object and a method for performing insights into a specific group's interests and views is considered by now mainstream methodology (Niederer and Van Dijck, 2010; Callahan and Herring, 2011), there are just a handful of such inspections of Wikipedias of South-Slavic languages (Kubelka and Sostaric, 2011; Bilic and Bulian, 2014). We hope to spark additional interest in such research by making the Wikipedia corpora of South-Slavic languages standardised, versionable, and easily accessible. As the corpus data size increases, big data architectures might become handy for efficient and timely processing of it (Zdravevski et al., 2020), which in turn might require efficient algorithms for cluster-size and cost optimization (Grzegorowski et al., 2021).

The paper is structured as follows. In the following section, we overview the methods (1) applied in preparing and linguistically annotating the corpora and (2) performing topic modelling. In the third section, we perform the analysis of the topic modelling results. In the fourth section, we give a short discussion of the obtained results, wrapping up with a conclusion.

2 Methods

2.1 Preparation of Corpora

In their initial form, the seven South-Slavic corpora were obtained as wiki-dumps from `http://wikimedia.com`. WikiExtractor[6] was used to open the wiki-dumped files and extract the relevant parts of the dumped Wikipedia corpus, such as paragraphs, links, section titles, and lists. The WikiExtractor tool's output yields several enumerated folders, each containing a maximum of a hundred 1 MiB files containing HTML tags and corresponding text.

Once the preliminary files were stored in separate folders, a Python module was developed for further processing. The module was developed to be language-agnostic, and it can be applied to all of the seven South-Slavic corpora. It allows for the final output to be of significantly higher quality (both in terms of text precision and recall) than is the case with using the available Wikipedia text extractors (WikiExtractor being one of them) out-of-the-box and is also made available for free usage and adaptation. The module[7] itself contains four levels of processing of the contents of the preliminary files, outlined as follows:

1. Usage of Scrapy, the Python library to remove all relevant HTML tags from the corpora.

2. Capturing various relevant parts of the Wikipedia article, storing them temporarily in memory while other processing is conducted, and afterwards, re-injecting them into the corpus. This is relevant for cases like URLs, shortened URLs, ellipses (...), dashed or numbered lists, intralinks within the articles themselves, etc. The various elements to be captured are defined using regular expressions.

3. Substitutions of text with other pieces of text based on regular expressions.

4. Substitutions of text with other pieces of text based on the Python `.replace()` method.

The resulting files from the Python module are stored in a directory structure equal to that of the preliminary files.

In addition to the data cleaning performed on the Wikimedia dump files, we linguistically processed the data with the CLASSLA pipeline[8], which is built over the Stanza tool (Qi et al., 2020) with improvements focused on the processing of highly inflected languages. The main changes to the tool are that (1) it uses an external inflectional lexicon, (2) uses the full morphosyntactic information while predicting the lemma, and (3) named entity recognition is added. The output from the CLASSLA modules is stored in a CoNLL-U format which, in addition to the original contents of the original text, assigns annotations to each token. Thus, a subsequent set of CoNLL-U-formatted corpora was generated for each of the original seven corpora, which, equally to the previous, has the same directory structure.

[4] `https://www.clarin.si/kontext/`

[5] `https://www.clarin.si/noske/`

[6] `https://github.com/attardi/wikiextractor`

[7] `https://github.com/clarinsi/classla-wikipedia/tree/main/ling_proc`

[8] `https://pypi.org/project/classla/`

Currently, support for five South-Slavic languages is present in the CLASSLA Python module, namely, for the Macedonian, Bulgarian, Croatian, Serbian, and Slovenian languages. Notably, in lack of a corresponding Bosnian and Serbo-Croatian model, the Croatian model was used to perform the linguistic annotation of the Bosnian and the Serbo-Croatian corresponding CoNLL-U contents.

The tool allowed for the following levels of processing with the Bulgarian, Croatian, Serbian and Slovenian models:

- tokenization and sentence splitting
- part-of-speech tagging
- lemmatization
- dependency parsing
- named entity recognition

For the Macedonian language, only the first three levels of annotation are available at this point.

An example excerpt from the Croatian linguistically processed Wikipedia corpus in CoNLL-U format is given in Figure 1.

The size of the resulting corpora, measured in number of documents, number of tokens, and the final text file size, are given in Table 1.

Lang	Docs	Tokens	Sizes
bs	84,472	20,934,288	149 MiB
mk	109,276	38,792,943	435 MiB
sl	169,777	45,739,630	316 MiB
hr	205,958	56,500,881	387 MiB
sh	453,450	69,726,727	509 MiB
bg	266,415	77,701,515	856 MiB
sr	639,282	106,498,685	1.2 GiB

Table 1: Size of the resulting corpora, ordered by token size.

2.2 Topic Modelling

Latent Dirichlet allocation (LDA) is an unsupervised generative, i.e. probabilistic, statistical model that allows sets of textual observations to be explained by latent topics that explain why some parts of the data are similar. (Blei et al., 2003) Thus, each textual document can be seen as a mixture of topics, each varying in prominence.

In configuring the LDA algorithm, we chose to have ten different topics in which we classify texts for every Wikipedia article. We chose the number 10 based on an intuitive expectation of the number of topics that would be present in each article. However, it is possible to do the same analysis with a varying number of topics. To perform the topic modelling itself, we employed the multi-core LDA model provided by the Gensim[9] Python package.

Each Gensim LDA model was configured to make ten passes over the entire training set. The maximum number of iterations through the corpus per pass, when inferring a corpus's topic distribution, is 50. A configured minimum probability threshold of 0.01 discards topics with a probability lower than this threshold. Additionally, for the sake of experiment reproducibility, we trained each Gensim LDA model with an initial random state of 47.

To prepare the data set for the LDA model, we parsed a reproducible random selection of CoNLL-U generated files for a given language until we obtained 10 000 unique Wikipedia articles. As previously mentioned, the terms ”Wikipedia Article” and ”document” are used interchangeably. For each document, we collected only the lemmas tagged with a `NOUN` or `PROPN` tag, which correspond to nouns or proper nouns. From this extraction, we generated noun-documents. We do so because the essence of any text's meaning relies primarily on its nouns and proper nouns. This approach has been shown to improve topic consistency and improve model training time. (Martin and Johnson, 2015)

Each LDA model was trained on a set of 10 000 noun-documents. When choosing which noun-documents should compose the LDA training set, we added a constraint that each noun-document should be of length above a threshold of 50 nouns and below a threshold of 500 nouns. Figure 2. depicts the lower and upper noun-document length thresholds as well as the distribution of the noun-document lengths, which resembles a power-law distribution that all samples seem to have.

The noun-document length constraint was formulated because the LDA model, as a probabilistic model, measures probabilities based on the co-occurrences of the words contained in each document. A word may occur in several topics with a different probability, however, with a different set of words alongside it in each different topic. Additionally, we believe that the noun-document length constraint enables creating a more representative

[9] https://radimrehurek.com/gensim/

```
# newpar id = 2
# newsrnt id = 2.1
# text = Hrvatski jezik (ISO 639-3: hrv) skupni je naziv za nacionalni standardni jezik Hrvata, te za skup narječja i govora kojima
1       Hrvatski        hrvatski        ADJ     Agpmsny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing        2       amod
2       jezik   jezik   NOUN    Ncmsn   Case=Nom|Gender=Masc|Number=Sing        13      nsubj   _       NER=O
3       (       (       PUNCT   Z       _       4       punct   _       NER=O|SpaceAfter=No
4       ISO     ISO     PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing        2       appos   _       NER=O
5       639     639     NUM     Mdc     NumType=Card    4       nummod  _       NER=O|SpaceAfter=No
6       -       -       PUNCT   Z       _       7       punct   _       NER=O|SpaceAfter=No
7       3       3       NUM     Mdc     NumType=Card    5       conj    _       NER=O|SpaceAfter=No
8       :       :       PUNCT   Z       _       9       punct   _       NER=O
9       hrv     hrv     X       Xf      _       4       conj    _       NER=O|SpaceAfter=No
10      )       )       PUNCT   Z       _       4       punct   _       NER=O
11      skupni  skupni  ADJ     Agpmsny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing        13      amod    _       NER=O
12      je      biti    AUX     Var3s   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   13      cop     _       NER=O
13      naziv   naziv   NOUN    Ncmsn   Case=Nom|Gender=Masc|Number=Sing        0       root    _       NER=O
14      za      za      ADP     Sa      Case=Acc        17      case    _       NER=O
15      nacionalni      nacionalan      ADJ     Agpmsayn        Animacy=Inan|Case=Acc|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing
16      standardni      standardan      ADJ     Agpmsayn        Animacy=Inan|Case=Acc|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing
17      jezik   jezik   NOUN    Ncmsan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   13      nmod    _       NER=O
18      Hrvata  Hrvat   PROPN   Npmpg   Case=Gen|Gender=Masc|Number=Plur        17      nmod    _       NER=B-PER|SpaceAfter=No
19      ,       ,       PUNCT   Z       _       22      punct   _       NER=O
20      te      te      CCONJ   Cc      _       22      cc      _       NER=O
21      za      za      ADP     Sa      Case=Acc        22      case    _       NER=O
22      skup    skup    NOUN    Ncmsan  Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   17      conj    _       NER=O
23      narječja        narječje        NOUN    Ncnpg   Case=Gen|Gender=Neut|Number=Plur        22      nmod    _       NER=O
24      i       i       CCONJ   Cc      _       25      cc      _       NER=O
25      govora  govor   NOUN    Ncmsg   Case=Gen|Gender=Masc|Number=Sing        23      conj    _       NER=O
26      kojima  koji    DET     Pi-mpi  Case=Ins|Gender=Masc|Number=Plur|PronType=Int,Rel       27      obl     _       NER=O
27      govore  govoriti        VERB    Vmr3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   25      acl     _       NER=O
28      ili     ili     CCONJ   Cc      _       31      cc      _       NER=O
29      su      biti    AUX     Var3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   31      aux     _       NER=O
30      nekada  nekada  ADV     Rgp     Degree=Pos|PronType=Ind 31      advmod  _       NER=O
31      govorili        govoriti        VERB    Vmp-pm  Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part|Voice=Act      27      conj
32      Hrvati  Hrvat   PROPN   Npmpn   Case=Nom|Gender=Masc|Number=Plur        31      nsubj   _       NER=B-PER|SpaceAfter=No
33      .       .       PUNCT   Z       _       13      punct   _       NER=O
```

Figure 1: Example (partially cropped) of the final CoNLL-U-formatted encoding of the corpus. The encoding contains paragraph and sentence identifiers, the full original text, surface forms, universal part-of-speech tags and morphological features, the MulTextEast morphosyntactic description, Universal Dependencies syntactic information, and named entity annotation.

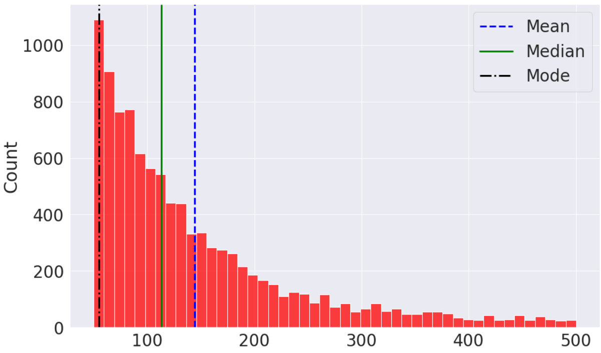

Figure 2: The noun-document length distribution for the Bulgarian sample

sample, which contains Wikipedia articles written and read by people, as opposed to auto-generated ones by the Wikipedia system. Furthermore, the LDA model aims to achieve a solution in which the topics and the words attributed to them are as disjoint as possible, and thus outliers of any form potentially may harm the results. Thus, focusing on a relevant interval of noun-documents resolves this concern as well.

Following this approach, we obtained noun-document training sets for each of the South-Slavic (macro-)languages, i.e. for the Bosnian, Bulgarian, Croatian, Macedonian, Serbian, Serbo-Croatian, and Slovenian. Subsequently, using each training set, we trained an LDA model and obtained ten topics.

3 Results

We start this section by presenting the ten topics obtained for each of the seven South-Slavic corpora in alphabetical order. Alongside each topic, in brackets, we present the percentage of documents from the 10 000 sample in which that particular topic was most prominent. Additionally, to avoid confusion, we add an enumeration index to a topic that has already been designated in the ten topics obtained for a given language sample.

An example of this occurs with the Serbo-Croatian topics, in which we obtained two topics that correspond to matters related to history. This is to be expected due to the LDA model's probabilistic machinery to produce as many topics as configured before executing the model. As we previously set 10 topics as desired output, the LDA model will strive to produce ten topics, even though naturally, there may be fewer. The detailed results per language are presented in Table 2. The languages (i.e. presented in different rows) are sorted alphabetically, and the topics within a language sample are sorted by the probability (i.e. percentage of documents in which that topic was most prominent) in descending order.

We continue this section by presenting a grouped form of the previously reported results. These consist of the ten topics obtained from each one of the seven LDA models and the percentage of documents from the sample of 10 000 in which that particular topic was most prominent. We emphasise the dominant topic because by design, the LDA model conceptualises a document as a mixture of multiple topics. Thus, as a simplifying measure, we designate a document based solely on its most dominant, i.e., most prominent topic. Furthermore, this serves also as a normalisation measure, which enables easier comparison among the South-Slavic samples.

After noticing that some topics are semantically related to each other, we grouped similar topics to form topic groups. The topic groups which we designated are the following: Art, Country, Culture, Geo-Politics, History, Science. These topic groups are formed by merging more specific topics, such as Physics, into broader topics, such as Science. Broader topics were designated because they contained contents belonging to multiple fields. For example, the Culture topic contains some of the keywords met in the Art topics, such as *painter* and *film,* or the Literature topics, which consists of keywords such as *writer*, *book* and *story*. Furthermore, the broader Country topic captures keywords related to all matters of the country. Keywords present in Country topics are *country*, *government*, *borders*, *language*, *work*, *territory*, *national holiday*, *culture* etc.

Below we list the topic groups and the topics which they encapsulate.

1. The Art group combines: Music, Literature, and Art.

2. The Country group combines: Demographics, National History, Country, and Education.

3. The Culture group combines: Language, Culture, Religion, and Sport.

4. The Geo-Politics group combines: Geography, War, and Politics.

5. The History group combines: Historical Events and History.

6. The Science group combines: Biology, Anthropology, Science, Physics, Architecture, and Astronomy.

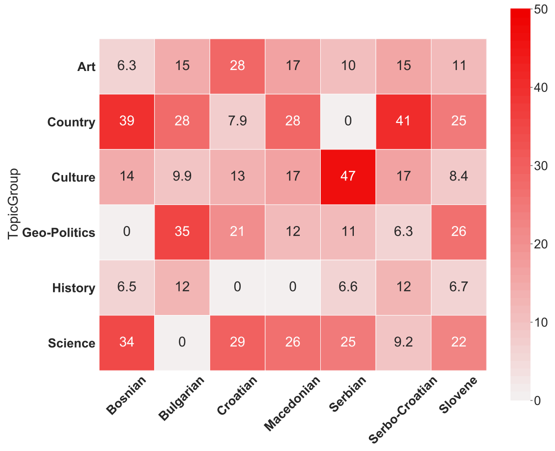

Figure 3: Dominant topic percentage distribution per topic group.

In Figure 3. we present the distribution per topic group for all seven South-Slavic corpora. A topic group's probability mass is the sum of the probability mass of its parts. For example, the Serbian exhibits a Science topic with a dominant topic probability mass of 10% and a Physics topic with a mass of 15%. Consequently, the Serbian Science topic group would have the sum of the two topics' probability mass, which equals 25%.

From the aggregated results demonstrated in Figure 3. we can see that approximately each of the seven South-Slavic Wikipedia corpora samples exhibit characteristics related to the Science and Art topic groups, which is to be expected from an encyclopedic data source such as Wikipedia.

Additionally, according to Figure 3. based on our LDA models, the Serbian sample contains the most Culture documents (47%), which is the most prominent outlier. Other outliers are the Bosnian and Serbo-Croatian samples, which contain 39% and 41% Country documents, while the Serbian sample contains none. Finally, the Bulgarian sample houses 35% Geo-Politics documents, and the Bosnian samples contain 34% Science documents, for which the Bulgarian sample contains none.

Interestingly, the Wikipedias, which have a more balanced distribution of topics, seem to be the Macedonian and Slovene, which might point towards a diverse editor structure and no agenda being pushed by the editors. On the opposite side, if we were to merge the Country and the

Table 2: Distribution of topics per language

Language	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Bosnian	Physics (20.2%)	Demographics (18.8%)	Country (10.2%)	National History (10.0%)	Culture (9.6%)	Biology (7.6%)	History (6.5%)	Music (6.3%)	Science (6.3%)	Culture (4.4%)
Bulgarian	National History (15.3%)	Politics (12.8%)	Education (12.3%)	Historical Events (12.3%)	Geography (10.1%)	Sport (9.9%)	Art (8.6%)	Geography (6.5%)	Literature (6.4%)	War (5.9%)
Croatian	Art (15.5%)	Culture (13.5%)	Art 2 (12.9%)	Geography (11.9%)	Science (10.5%)	Anthropology (9.5%)	War (8.9%)	National History (7.9%)	Architecture (4.9%)	Physics (4.6%)
Macedonian	Astronomy (20.6%)	War (12.0%)	Demographics (10.6%)	Literature (9.3%)	Language (9.2%)	Education (9.1%)	National History (8.7%)	Religion (7.6%)	Art (7.2%)	Science (5.7%)
Serbian	Physics (14.6%)	Culture (12.6%)	Sport (12.4%)	Religion (11.6%)	Art (10.3%)	Science (10.2%)	Sport (10.2%)	History (6.6%)	Geography (6.0%)	War (5.5%)
Serbo-Croatian	National History (18.9%)	Art (15.2%)	Country (12.6%)	Culture (10.1%)	Demographics (9.2%)	Anthropology (9.2%)	History (7.8%)	Religion (6.5%)	Politics (6.3%)	History 2 (4.4%)
Slovene	Education (13.3%)	Astronomy (12.6%)	Geography (12.5%)	National History (11.7%)	Literature (11.5%)	War (10.1%)	Science (9.7%)	Sport (8.4%)	History (6.7%)	War 2 (3.6%)

Geo-Politics topic groups to form a novel topic group whose probability mass distribution is the sum of the two topic groups' probability mass vectors, we can construct the following listing, namely, Bulgarian (63%), Slovene (51%), Serbo-Croatian (47.3%), Macedonian (43%), Bosnian (39%), Croatian (28.9%), and Serbian (11%). From this listing, we can measure that this new topic group's average is 47.2%, which places the Bulgarian considerably above the mean and perhaps points towards a somewhat biased editor structure. These observations are just preliminary and should be followed up by a more in-depth analysis of topics and other types of analyses, which we hope to happen in the future, especially given the improved accessibility of Wikipedia texts - the main contribution of this work.

It should be noted that each of the entries in Figure 3. are below 50% as to emphasise the point that even in the constructed samples consisting of 10 000 unique documents, there is no apparent bias.

By calculating the row-wise mean for the topic groups' entries in Figure 3. we can obtain the following ordering of topic groups and their average probability mass, namely, History (6.3%), Art (14.7%), Geo-Politics (16.0%), Culture (18.0%), Science (20.9%), Country (24.1%). This ordering is also a probability mass distribution and thus showcases that, on average, each sample's editors mostly focus on matters related to their country. However, the merger of the Art, Culture, and Science topic groups from this novel probability distribution would result in a topic group with 53.6% probability mass, which is suggestive of the notion that despite the biases that the samples' editors might have, overall, Wikipedia is still a source composed of encyclopedic knowledge.

To further quantify our comparison of Wikipedia contents, we applied pairwise Jenson-Shannon divergence (JSD) over pairs of topic distributions of Wikipedias. Jenson-Shannon divergence is a symmetrized and smoothed version of the Kullback–Leibler divergence. The JSD measure enables us to compare probability distributions, such as the discrete probability distributions housed in each one of the columns presented in Figure 3. Through this calculation, we obtain a distance or divergence estimate between each pair of Wikipedias.

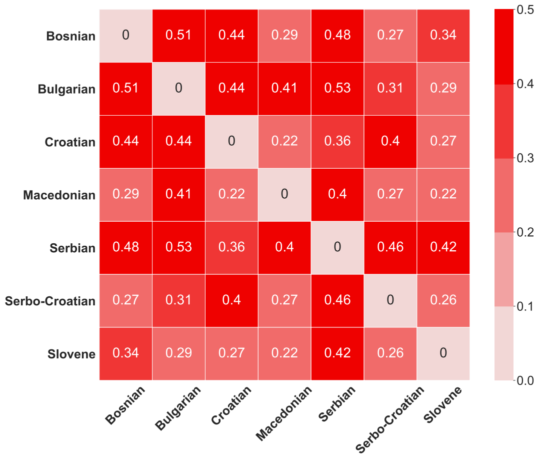

Figure 4: Pairwise Jenson-Shannon distance matrix comparing the results for every South-Slavic sample.

In Figure 4. we present a pairwise matrix that quantifies the distance between every pair of the South-Slavic Wikipedia samples. The figure shows that the Serbian sample is considerably distant from the rest of the samples, being closest to the Croatian sample. Similarly, the Bulgarian sample is notably distant from the rest except for the Serbo-Croatian and the Slovene samples. Among all, the Serbian sample and the Bulgarian sample are the most distant pair with a distance of 53%, while the least distant are the Macedonian-Croatian and the Macedonian-Slovene pairs with a distance of 22%.

Furthermore, from Figure 4., by calculating the vector component average for each row (or column, since the JSD is a symmetric matrix), we obtain the following ordering of least average distant to most average distant, namely, Slovene (25.7%), Macedonian (25.8%), Serbo-Croatian (28.2%), Croatian (30.3%), Bosnian (33.1%), Bulgarian (35.8%), Serbian (37.8%). As most average distant, the Serbian sample is reflective of the most notable outlier of Figure 3., that is, the 47% probability mass entry in the Culture topic group. Additionally, the Bulgarian sample is second most average distant due to the emphasis of the Country and the Geo-Politics topics.

It should be noted that the aforementioned average Jenson-Shannon distance ordering consists for the most part of entries below 50% which is indicative of greater likeness than dissimilarity between the South-Slavic Wikipedia samples.

4 Discussion

It is reasonable to expect that many of the Wikipedia articles' topics are recognisable scientific fields or socio-economic disciplines because Wikipedia contains various articles contributing to encyclopedic knowledge. Such is the case for Astronomy within the Macedonian corpus, Biology within the Bosnian corpus, etc.

In this work, we considered only 10 000 noun-documents out of a larger number of Wikipedia articles, each varying in size and content. Additionally, the Serbian Wikipedia corpus is considerably more comprehensive, and thus, more extensive in terms of the number of articles, while some of the other languages are half of this magnitude or less. This demonstrates that the number of documents considered, the sampling strategy, and original size of the corpora are some of the relevant factors that influence the generated topics from the LDA models.

Additionally, the presence of zero element entries in the matrix depicted in Figure 3 is most possibly related to the need for enlargement of the sample size to obtain a more comprehensive result set, in which all Wikipedias samples would contain only non-zero entries for every topic group in the corresponding topic group matrix.

The results obtained from our topic modelling endeavour are in line with our expectations. Each language describes prominent figures and historical events, which entails considering geographical notions, political influences, artistic, cultural, and ideological interpretations. This showcases the difficulty in separating the contents into distinguishable topics. To further improve our results and obtain more disjoint topics, our work could benefit from an approach that has more insight and better language comprehension abilities.

Other modelling approaches can be employed, such as a hierarchical topic modelling approach. Unlike the LDA model, the Hierarchical Dirichlet Process (HDP) model by design does not contain a configurable parameter for the number of topics. The HDP model thus outputs a varying number of topics based on the input.

Furthermore, other avenues that we may explore are different LDA model evaluations, formulating and computing model perplexity, and measuring topic coherence.

5 Conclusion

The main contribution of our work is a new collection of corpora of high-quality text for seven South-Slavic (macro-)languages, namely, Bosnian, Bulgarian, Croatian, Macedonian, Serbian, Serbo-Croatian, and Slovenian. The corpora were generated by harvesting Wikipedia dumps and post-processing them to clean the text from all unnecessary phenomena.

We linguistically processed these corpora on the levels of tokenization, morphosyntactic annotation and lemmatization. For all languages, except for Macedonian, we also performed dependency parsing and named entity recognition. The final corpora are freely available for download[10] and concordancer search[11]. We plan to generate new versions of the corpora on an annual basis, improving the availability of linguistically processed high-quality corpora for the South-Slavic language group significantly.

Using these linguistically processed corpora, we performed a content-analysis experiment via topic modelling, analysing the topics featured within the Wikipedia articles across all mentioned South-Slavic corpora. While our topic modelling results are a rather shallow and preliminary insight into the content of the mentioned Wikipedias, a trend has already emerged. Judging from the dominant topic percentage entries demonstrated in Figure 3., of which all are below 50%, and from the aver-

[10] http://www.clarin.si/repository/xmlui/

[11] http://www.clarin.si/noske/

age pair-wise Jenson-Shannon distance ordering, whose entries are to the greatest part below 50%, we may gather that the results are suggestive of the notion that the interests of the peoples are more similar than they are opposed.

The Serbian Wikipedia is surely an outlier in terms of similarity to other Wikipedias, showing the most significant topical differences to other Wikipedias, and will be the first next stop of our analysis. Similarly, a large part of the samples contained documents designated with topics attributed to matters related to the country or politics, which warrant further investigation.

While the presented corpora will be a very welcome addition to the list of resources for the South-Slavic language group, we are aware of the Wikipedia text's limitations for documenting a language. Therefore, we consider it another direction for future work to be extending these Wikipedia corpora with another source of relatively inexpensive but more diverse textual material, namely web corpora.

References

Pasko Bilic and Luka Bulian. 2014. Lost in translation: Contexts, computing, disputing on wikipedia. *iconference 2014 proceedings*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Ewa S Callahan and Susan C Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2020. The kas corpus of slovenian academic writing. *Language Resources and Evaluation*, pages 1–33.

Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. 2020. The janes project: language resources and tools for slovene user generated content. *Language resources and evaluation*, 54(1):223–246.

Marek Grzegorowski, Eftim Zdravevski, Andrzej Janusz, Petre Lameski, Cas Apanowicz, and Dominik Slezak. 2021. Cost optimization for big data workloads based on dynamic scheduling and cluster-size tuning. *Big Data Research*, 25:100203.

Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France. European Language Resources Association.

Ozren Kubelka and Petra Sostaric. 2011. Wikipedija nasuprot hrvatskoj enciklopediji, kvalitativan odnos slobodno i tradicionalno uredjenoga enciklopedijskoga sadrzaja na hrvatskom jeziku. *Studia lexicographica*, 5(9):119–133.

Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77. Association for Computational Linguistics.

Fiona Martin and Mark Johnson. 2015. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 111–115.

Sabine Niederer and José Van Dijck. 2010. Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. *New media & society*, 12(8):1368–1387.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Lukas Svoboda and Slobodan Beliga. 2017. Evaluation of croatian word embeddings. *arXiv preprint arXiv:1711.01804*.

Eftim Zdravevski, Petre Lameski, Cas Apanowicz, and Dominik Slezak. 2020. From big data to business analytics: The case study of churn prediction. *Applied Soft Computing*, 90:106164.