Annotated Lexicon for Sentiment Analysis in Bosnian Language

Sead Jahić[1], Jernej Vičič[1]

[1]Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska, Koper, Slovenia

Abstract

The paper presents the first sentiment annotated lexicon of the Bosnian language. The language coverage of the lexicon was evaluated using two reference corpora. The usability of the lexicon was already proven on a Twitter-based comparison. Two approaches were observed in this experiment, the first method used a frequency list of all lemmas extracted from two relevant Bosnian language corpora, and the second method used all lemmas occurrences without using frequency as the main factor in counting. The results of the study suggest usable language coverage. The computed coverage for the first corpus was 27.25%, while the second corpus yields 24.34%. The second method yields 1.899% coverage for the first corpus and 6.05% for the second corpus.

Keywords

Bosnian lexicon, corpus, sentiment analysis, AnAwords, stopwords

1. Introduction

Sentiment analysis (or opinion mining) is a technique used to determine whether data, often performed as text, is positive, negative, or neutral. The growing interest in the efficient analysis of web texts has led to remarkable developments in the field of sentiment analysis. Sentiment analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities within a sentence or phrase. Social networks enable users to express their thoughts and feelings more openly than ever before; sentiment analysis is becoming an essential tool to monitor and understand that sentiment.

In this paper, we present coverage of the first Bosnian sentiment annotated lexicon using two reference corpora. The results of the study suggest usable language coverage. We applied two different approaches, where the first yielded 27.25% of the coverage and the second around 1.9%. The main reason of the difference in results lies in the fact that for the first approach we used lemmas together with their frequencies, while in the second approach the frequencies of the given lemmas were neglected.

Section 2 presents State of the Art, where we have gave an insight about what have been done in area of NLP in Bosnian language, sentiment analysis as well as lexicon and corpus construction. The methodology; process of cleaning corpora, covering corpora by the lexicon, and also usage of the stop-word lists and intensifiers in all were described in section 3. Stages of the annotated process have been also explained in section 3, while in section 4, which is an extension of section 3, we have presented all results of the experiment. Last section is reserved for conclusion and further work.

2. State of the Art

There has been quite extensive research in the area of Sentiment analysis and many types of models and algorithms have been proposed depending on the final goal of the analysis of the interpretation of user's feedback and queries, such as Fine-grained Sentiment Analysis (based on polarity precision), Emotion detection, Aspect-based Sentiment Analysis, Multilingual Sentiment Analysis. All those algorithms and models can be divided into one of three basic classes: rule-based systems (relying on long used linguistic methods, rules and annotated linguistic materials such as annotated lexicons), automatic (corpus-based) systems and hybrid systems that combine properties from both previous types. In such a manner, hybrid systems use machine learning techniques together with NLP techniques developed in computational linguistics such as stemming, tokenization, part-of-speech tagging, parsing and lexicons.

The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), June 7-8, Koper, Slovenia

jahic.sead@famnit.upr.si (S. Jahić); jernej.vicic@upr.si (J. Vičič)

Lexicons have been widely used for sentiment analysis. One of the first-known, human-annotated lexicons for sentiment analysis is the General Inquirer lexicon (Stone et al. [1]), which contains 11,788 English words (2291 labelled as negative and 1,915 as positive, with the rest, labelled as objective). Sentiment lexicons exist for most Slavic languages; examples are lexicons for Bulgarian (Kapukaranov and Nakov et al. [2]), Croatian (Glavaš et al. [**?**]), Czech (Veselovská [3]), Macedonian (Jovanoski et al. [4]), Polish (Wawer [5]), Slovak (Okruhlica [6]), Slovenian (Kadunc [7]) and Bosnian (Jahić, Vičič [8]).

The usability of the Bosnian lexicon, on a sentiment tagging task, was already proven on a Twitter annotation task (Jahić, Vičič [8]). It is loosely based on the Slovenian lexicon (Kadunc [7]), which consists of words and lemmas. The lexicon creation process comprised of taking words from Slovenian lexicon and translating them into the Bosnian language. The sentiment of the Bosnian translation was manually checked during the translation process. This lexicon was used for measuring of coverage and it contains 1279 entries labelled as positive and 3116 as negative.

An important question for natural language researchers, general linguists, and even teachers and students is how much text coverage can be achieved with a certain number of lemmas from the lexicon in a given language since the number of terms in the lexicon is by a few magnitudes smaller than the number of terms in the corpus. Studies of vocabulary coverage have been carried out for many languages such as the German language (J. Randall et. al [9]), where a study based on the BYU/Leipzig Corpus of Contemporary German has shown that a basic vocabulary of 3,000 high-frequency words can account for between 75% and 90% of the words in the text; moreover in Spanish language (Davies M. et. al [10]) is stated that for the language learner it is enough to know basic 4000 words in order to cover/recognize more than 90% of the words in the native text. Ortiz, Hernández et. al [11] presented Lingmotiflex: a wide-coverage, domain-neutral lexicon for sentiment analysis in English, stated that it achieves significantly better performance than the other lexicons for English, where coverage goes up to 75% and 84% (F1-score) for two data-sets.

Bučar, J., Žnidaršič, M. & Povh, J. [12] introduce new language resources (corpora, annotations and lexicon) for sentiment analysis in Slovene. They retrieved more than 250,000 news items from five Slovene web media resources. Five different measures of correlation were used to evaluate the process of annotation. In general, all the measures indicate good internal consistency at all levels of granularity; however, their values decrease steadily when applied to the paragraph and sentence levels.

At least to the author's knowledge, there were almost no attempts carried in the NLP for the Bosnian language, so this paper presents one of the first steps in NLP for the Bosnian language and puts this language side by side with all other world languages. Although the authors are aware that the method is not perfect, the reference corpus is used as a normative representing the live language to the best possible extent.

Corpus-based model methods and lexicon-based model methods have been increasingly used to compare language usage. A comparison of hundreds of thousands or millions of words/lemmas from the corpus with a few thousand words/lemmas from the lexicon presents one of the main types of corpus comparison. In this case, we refer to the corpus as 'normative' since it provides a standard against which we can compare. As it is stated in [13] it is possible to compare three or more corpora at the same time, but it will only make the results more difficult to interpret.

3. Methodology and work

Bosnian sense annotated lexicon is presented and analyzed in this paper. Since also words that were not classified as positive or negative, the notation 'sense lexicon' was used to declare lexicon propose. Moreover, 'sense' is used to declare that the lexicon consists of the "core" lexicon (positive and negative words), a list of stop-words, and a list of AnAwords (Affirmative and Non-affirmative words) which is clarified in Fig 1.

Two corpora were used to test language coverage:

- the Bosnian web corpus bsWaC 1.1 [14] which is a collection of web pages crawled in 2016. The corpus consists of texts in three languages (Bosnian, Croatian and Serbian), each text belongs (is tagged) to one of the languages. The corpus is also morph-syntactically annotated and lemmatized. It consists of more than 285 million words. At the moment of writing, this corpus was the de-facto reference corpus for Bosnian language
- Bosnian news corpus 2021 bsNews 1.0 [15], which is a collection of web news articles crawled at the start of 2021. The corpus contains a balanced set of at most 2000 most recent news articles from

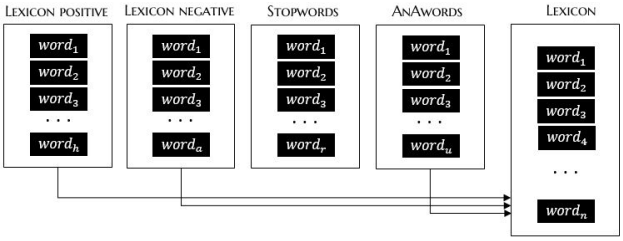

Figure 1: Construction of the Lexicon

each identified web news portal in Bosnia and Herzegovina. The list of portals is maintained by Press Council in Bosnia and Herzegovina.
The corpus contains news articles from 46 portals. This corpus was used as a contemporary and balanced source. The sentence tokens are morph-syntactically annotated with MULTEXT-East morph-syntactic annotations for Croatian, Version 6 (http://nl.ijs.si/ME/V6/). The corpus was morpho-syntactically annotated and lemmatized with ToTaLe [16]. It consists of more than 36 million words.

Two different approaches are applied:

- First, all lemmas with their frequencies were considered,
- Second, the frequencies for lemmas were ignored.

A list of lemmas with frequencies was extracted from each corpus and cut off at 5 occurrences to avoid clutter. The list of lemmas extracted from the first corpus Ljubešić N. and Klubička F. [14] consisted of 371385 different lemmas with frequency. The lemmas are ordered in increasing order by frequency, where the lowest value is 5 (cutoff) ("batkovi - drumsticks" ...) and the highest value is 16652046 for lemma "biti - to be". The list of lemmas extracted from the second corpus (BsNews 1.0 corpus [15]) consisted of 101773 lemmas ordered in decreasing order, the most frequent lemma again "biti – to be" with a frequency of 2350487 and with the lowest frequency of 5 are several lemmas such as "polegnuti – lay down".

Not all lemmas can be included in the analysis. Symbols, equitation marks, and numbers, even if being a part of the corpus, cannot be part of the lexicon, especially the sentiment annotated lexicon. These items were removed from both corpora in the cleaning: some special characters that appear in Bosnian (not usual for non-southwest Slavic group of languages); emoticons; punctuation; numbers; hyperlinks.

- First approach was to include lemmas with their frequency in analysis (all appearance of lemmas was used for each corpora).

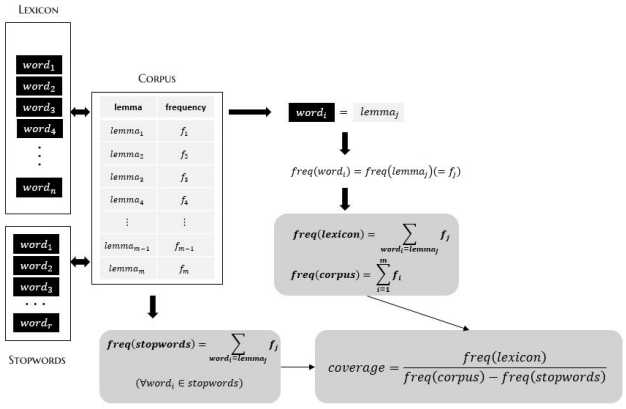

Figure 2: Process of matching lemmas from corpus with words from Lexicon

The Figure 2 shows the procedure for checking the existence of given words from the lexicon in the corpus. If this statement is true, if the word exists in the corpus, the value of $freq(lexicon)$ is accumulated for the value of the frequency of each word, otherwise the value 0 is added to $freq(lexicon)$.

Sum of all word's frequencies from corpus is given as $freq(corpus)$ and $freq(stopwords)$ presents the sum of all frequencies of the stop-words that appears in corpus. The coverage is counted as:

$$coverage = \frac{freq(lexicon)}{freq(corpus) - freq(stopwords)}, \tag{1}$$

where all stop-words were excluded from.

- Second approach was by using accuracy of lemmas without influence of frequency.

After pre-processing, 286095 lemmas from first and 92268 lemmas from second corpus were included in further analysis (see Table 1).

Table 1
Number of lemmas left after pre-processing corpora

	Corpus1	Corpus2
Overall number of lemmas	371385	101773
Cleared lemmas	285963	92183
Percent (%)	76.999	90.58

In order to be able to compare words from the lexicon and corpus, the letters typical of the South-Slavic languages such as "č, ć, đ, ž" have been replaced with "c, c, dj, z".

Given that sentimental value is not at the forefront at this stage of research (we are looking for language coverage), 1279 positive and 3116 negative words were united in a unique lexicon.

In addition to lexicons, two other groups of words: stopwords (343 in our collection) and AnAwords (Affirmative and Non-affirmative words), play a significant role in this process. Jahić and Vičič in [8] pointed out that stop-words usually refer to the most common words in a language and that there is no single universal list of stop-words used.

Besides that, 102 words from the AnAwords list were created by Jahić and Vičič [8], and it has been proven by Osmankadic et. al [17] that most of those words are intensifiers.

The influence of words from the AnAwords list was also considered in the coverage of the corpus.

The process of annotating lexicons went through several stages, and they were all based on an equation:

$$\frac{FOUND}{NOT\_FOUND} \tag{2}$$

where FOUND presented the list of all words in corpus that were matched with words from the lexicon, NOT_FOUND opposite.

Those stages are:

- Simple coverage of corpus by lexicon as shown in the first stage. The stop-words were part of the corpus at this stage.
- While in 1^{st} stage stop-words were an integral part of the corpus, in the process of coverage, in 2^{nd} stage covering of the corpus was made without them since the number of stop-words is almost negligible in relation to the number of elements in a corpus, a large difference in coverage in this stage was not expected.
- Guided by the results of research conducted for corpus-based lexical analysis of subject-specific university textbooks in English by Hajiyeva K et. al [18], in the 3^{rd} stage, coverage was observed by the frequency distribution of lemmas.
- In the 4^{th} stage, the question arises whether it is possible to group similar words (such as "andjeo" and "andjel" (angel)) and view them as a single word?! The solution to the problem Davies, M. et al [10] suggested grouping words according to word families. Given this possibility of grouping, matching functions were applied between corpus words and lexicon words.

4. Results

This section presents the results achieved by the two approaches described in Section 3.

In the first approach (Coverage of corpora by using accuracy of lemmas with influence of frequency), $freq(corpus)$, the sum of stop-words frequencies $freq(stopwords)$ and the overall sum of all frequencies of the words from Lexicon is $freq(lexicon)$ were computed.
By using equation (1) coverage of the corpus1 is 27.25%, and coverage of the corpus2 is 24.34% (see table 2).

Table 2
Coverage of corpora's lemmas with word from sentiment lexicon

	freq(corpus)	freq(lexicon)	freq(stopwords)	COVERAGE
CORPUS1	197245460	43555699	37414905	27.25%
CORPUS2	30599375	6238242	4971734	24.34%

The second approach (Coverage of corpora by using the accuracy of lemmas without the influence of frequency) was to compute the overall coverage of the corpora without using word frequencies.
The motivation behind this approach was to count how many different lemmas from corpus are already present in the sentiment lexicon.
There have been few stages in this approach.

First stage: In this first stage, 1.199% coverage of 1^{st} corpus and 3.21% for 2^{nd} corpus was achieved.

Table 3
Coverage of corpora's lemmas with word from sentiment lexicon (without additional changes made)

	Corpus1	Corpus2
FOUND	3389	2866
NOT_FOUND	282574	89317
Coverage (%)	1.199	3.21

In Table 3 are presented lemmas that were matched with words from lexicon (FOUND) and that were absent from lexicon (NOT_FOUND).
Maximum coverage of corpora is possible if all words from the lexicon are included in corpora. It means that the maximum coverage for the First corpus is 1.54% and the Second corpus is 4.76%.
On the other side, coverage of lexicon by corpora is 77.11% and 65%. It means that of 4395 words from the lexicon, 3389 were presented in corpus1 which produced use of 77.11% of the lexicon, and 2866 were presented in corpus2, which led to 65.21% use of lexicon.

Second stage basically decreases the number of lemmas in corpora since all lemmas that are stop-words or AnAwords were removed. In this case, coverage of corpora has been increased to 1.2% and 3.22%.

Table 4
Coverage of corpora's lemmas with word from sentiment Lexicon

	Corpus1	Corpus2
FOUND	3389	2866
NOT_FOUND	282330	88994
Coverage (%)	1.2	3.22

Third stage was covering by distributing lemmas by frequency, and counting number of lemmas that were or were not covered by words from lexicon.

From 50000 last lemmas from corpus1, there were about 1962 words from lexicon, which means that from 3389 overall words from lexicon that have been annotated by corpus1, 57.89% were included

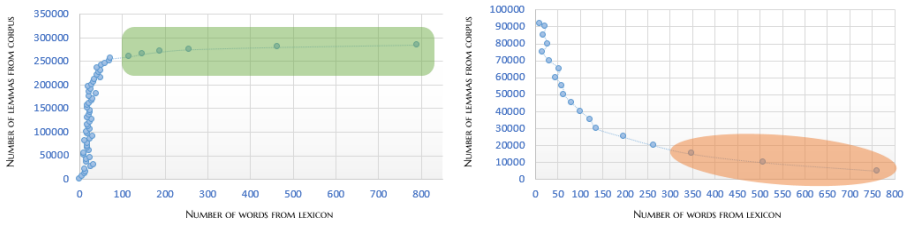

Figure 3: Annotated lexicon by distributed lemmas from corpus1 (left) and corpus2 (right)

in 50000 most frequent lemmas of corpus1 (see Figure 3 (left)).
More, of 15000 most frequent lemmas 1616 were presented in lexicon. Bear in mind that overall number of words from lexicon that are annotated is 2866, it gave that 56.39% all annotated words from lexicon are presented in 15000 most frequent lemmas from corpus2 (see Figure 3 (right)).

Forth stage: In 4^{th} stage lexicon annotation was increased to more than 1.899% for 1st corpus. Even it looks like it is contradiction with explanation that maximum coverage for the first corpus is about 1.54%, it is not. The reason for that is in the fact that get_close_matches (imported from difflib in Python) function was applied with cutoff 80% and $n = 1$ (one possibility).

$$get\_close\_matches(word, possibilities[, n][, cutoff])$$

The function works in such way that all words that are almost similar (80% matching in this case) are considered as one word. For example: andjel (engl. angel), andjelko (engl. little angel), andjela ("I saw an angel"), all three words were replaced with andjeo. Number of lemmas (for corpus1) in corpora have decreased (see Table 5.) to number 229210, and annotation increases to 1.899%, which means that from 229210 lemmas from corpus1, 4271 was founded in lexicon.
Same thing happens for corpus2, where annotation of 4101 words from lexicon in corpus2 were detected.

Table 5
Coverage of corpora's lemmas with word from sentiment Lexicon

	Corpus1	Corpus2
No.of lemmas	229210	71857
FOUND	4271	4101
NOT_FOUND	224939	67756
Coverage (%)	1.899	6.05

Although the 3^{rd} stage presents an insight into the annotation of most frequent lemmas, for overall annotation most important stages were first, second, and fourth, since they produce overall coverage of the corpora by lexicon (see table 6).

Table 6
Annotation of corpora

Approach:		Coverage of corpora	
		CORPUS1 (bsWaC)	CORPUS2 (bsNews)
First		by using accuracy of lemmas with influence of frequency	
		27.25%	24.34%
Second		by using accuracy of lemmas without influence of frequency	
Stage	First	1.199%	3.21%
	Second	1.2%	3.22%
	Forth	1.899%	6.05%

5. Conclusion

Sentiment annotation of a lexicon and working in the field of Sentiment analysis and corpus-lexicon based methods present new and first results for the Bosnian language. Although arguably Bosnian language is closely related to Serbian and Croatian languages, there are subtle differences in these three languages that are more evident from the Sentiment analysis point of view.

This paper presents the annotation of the first Bosnian sentiment lexicon that has been earlier proven on sentimental basis. The lexicon comprises approximately 4400 words and covers more than 27% of the lemmas in the first observed corpus (corpus1), Ljubešić N. and Klubička F. [14], and more then 24% of lemmas in the second observed corpus (corpus2), (BsNews 1.0 corpus [15]). If the emphasis is on coverage of different lemmas from corpus by lexicon, then coverage is 1.2% for corpus1 and 3.2% for corpus2. This coverage will increase by applying some matching functions between corpora's lemmas and lexicon's words (which was described in forth stage of second approach). In that case, the coverage raises to 1.9% for corpus1 and 6% for corpus2.

It means that almost 97% of the lexicon was used to annotate corpus1 (132 words from the lexicon were not found in corpus1) and more than 93% to annotate corpus2, which means that only 308 words from the lexicon were not present in corpus2.

The results show that about a quarter of the lemmas from corpora have their sentimental value annotated in the lexicon, which greatly helps in the sentimental annotation of the sentences (tweets or regular text). Stop-words and AnAwords were also included in the analysis which leads to the possibility that LSAnA group becomes a representative group for emotional words, stop-words, and intensifiers (all written in Bosnian). The language coverage of the lexicon is comparable with State Of The Art, the values can be compared in [11].

The focus in our future work will be on developing and improving LSAnA group. All members of the group should be extended, which means that our expectation is to have more items/words labelled as positive or negative in our 'core' lexicon, as well as extending of lists of stop-words and AnA words. To increase coverage, we will try to create a lexicon with all possible lemmas and in doing so we will contain all the grammatical rules found by the Bosnian language itself (declination, conjugation, change of river by gender, number and so on)

Although the process of annotation, as well as improvement of the first Bosnian lexicon, is still under development, the results shown are comparable with results shown for other languages [10] [12].

Acknowledgments

The authors gratefully acknowledge the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund).

References

[1] P. Stone, D. Dunphy, M. Smith, D. Ogilvie, The General Inquirer: A Computer Approach to Content Analysis, volume 4, 1966. doi:10.2307/1161774.

[2] B. Kapukaranov, P. Nakov, Fine-grained sentiment analysis for movie reviews in Bulgarian, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 266–274. URL: https://aclanthology.org/R15-1036.

[3] K. Veselovská, Czech subjectivity lexicon : A lexical resource for czech polarity classification, in: In Proceedings of the 7th international conference Slovko, Bratislava, 2013, pp. 279–284.

[4] D. Jovanoski, V. Pachovski, P. Nakov, Sentiment analysis in Twitter for Macedonian, in: Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 249–257. URL: https://aclanthology.org/R15-1034.

[5] A. Wawer, Extracting emotive patterns for languages with rich morphology, International Journal of Computational Linguistics and Applications 3 (2012) 11–24.

[6] A. Okruhlica, Slovak sentiment lexicon induction in absence of labeled data, Master's thesis, Comenius University Bratislava, 2013.
[7] K. Kadunc, Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega učenja (2016). URL: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=eng&id=91182.
[8] S. Jahić, J. Vičič, Determining Sentiment of Tweets Using First Bosnian Lexicon and (AnA)-Affirmative and Non-affirmative Words, Springer International Publishing, Cham, 2021, pp. 361–373. URL: https://doi.org/10.1007/978-3-030-54765-3_25. doi:10.1007/978-3-030-54765-3_25.
[9] R. L. Jones, An analysis of lexical text coverage in contemporary German, Brill, Leiden, The Netherlands, 2006, pp. 115 – 120. URL: https://brill.com/view/book/edcoll/9789401202213/B9789401202213-s010.xml. doi:https://doi.org/10.1163/9789401202213_010.
[10] M. Davies, Vocabulary range and text coverage. insights from the forthcoming routledge frequency dictionary of spanish, in: Selected Proceedings of the 7th Hispanic Linguistics Symposium, 2005, pp. 106–115.
[11] A. Moreno-Ortiz, C. Pérez-Hernández, Lingmotif-lex: a wide-coverage, state-of-the-art lexicon for sentiment analysis, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1420.
[12] J. Bučar, M. Žnidaršič, J. Povh, Annotated news corpora and a lexicon for sentiment analysis in slovene, Language Resources and Evaluation 52 (2018) 895–919. doi:10.1007/s10579-018-9413-3.
[13] P. Rayson, R. Garside, Comparing corpora using frequency profiling, in: Proceedings of the Workshop on Comparing Corpora - Volume 9, WCC '00, Association for Computational Linguistics, USA, 2000, p. 1–6. URL: https://doi.org/10.3115/1117729.1117730. doi:10.3115/1117729.1117730.
[14] N. Ljubešić, F. Klubička, bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian, in: Proceedings of the 9th Web as Corpus Workshop (WaC-9), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 29–35. URL: https://aclanthology.org/W14-0405. doi:10.3115/v1/W14-0405.
[15] J. Vičič, Bosnian news corpus 2021, 2021. URL: http://hdl.handle.net/11356/1406, slovenian language resource repository CLARIN.SI.
[16] T. Erjavec, C. Ignat, B. Pouliquen, R. Steinberger, Massive multi lingual corpus compilation: Acquis communautaire and totale, Archives of Control Sciences 15 (2005).
[17] M. Osmankadić, A Contribution to the Classification of Intensifiers in English and Bosnian, Institut za jezik, 2003.
[18] K. Hajiyeva, A corpus-based lexical analysis of subject-specific university textbooks for english majors, Ampersand 2 (2015) 136–144. doi:10.1016/j.amper.2015.10.001.