Multi-word Expressions for Abusive Speech Detection in Serbian

Ranka Stanković
University of Belgrade
ranka@rgf.bg.ac.rs

Jelena Mitrović
University of Passau
jelena.mitrovic@uni-passau.de

Danka Jokić
University of Belgrade
dankaiv@googlemail.com

Cvetana Krstev
University of Belgrade
cvetana@matf.bg.ac.rs

Abstract

This paper presents our work on the refinement and improvement of the Serbian language part of Hurtlex, a multilingual lexicon of words to hurt. We pay special attention to adding Multi-word expressions that can be seen as abusive, as such lexical entries are very important in obtaining good results in a plethora of abusive language detection tasks. We use Serbian morphological dictionaries as a basis for data cleaning and MWE dictionary creation. A connection to other lexical and semantic resources in Serbian is outlined and building of abusive language detection systems based on that connection is foreseen.

1 Introduction

This paper presents initial results in an on-going collaboration between University of Passau and University of Belgrade aimed at improving the lexical resources that will aid abusive speech detection in Serbian. Discriminatory messages and exhortation to violence are related to offensive and hate speech, which has been gaining more attention due to the extensive use of online media and the Internet in general. The concept of abusive speech, as an umbrella term for phenomena such as offensive and hate speech, its content and forms of expression are analysed, trying to define its vocabulary, collocations, colloquial expressions, and context.

Starting from the definition of hate speech as 'any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic' (Nockleby, 2000) it is clear that hate speech is a complex social and linguistic phenomenon. Abusive language and its detection have been gaining more attention recently. Caselli et al. (2020) define abusive language as 'hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions. This might include hate speech, derogatory language, profanity, toxic comments, racist and sexist statements.' Computational processing of such language requires usage of finely-tuned, task specific language tools and resources, especially for morphologically rich and low-resource languages such as Serbian.

In the process of building a conceptual framework of abusive language, special attention is paid to Multi-word expressions (MWEs) which allow for better and more precise detection of this linguistic phenomenon. The development of the MWE lexicon also helps in reducing ambiguity. The lexical resource, consisting of words that could be used as a trigger for recognition of abusive language is built, with an idea that the Serbian system for recognition and normalization of abusive expressions will also take into consideration phrases and figurative speech as an indicator. Both explicit and implicit abusive language (hateful and offensive messages that are not apparent at first glance) will be analysed.

The remainder of the paper is structured as follows. Related work is given in Section 2 with a short overview of approaches for developing this type of resources in Subsection 2.1. One of the existing,

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 74–84
Barcelona, Spain (Online), December 13, 2020.

publicly available resources, Hurtlex, that we base our work on, is introduced in Subsection 2.2. The analysis and improvement of Hurtlex for Serbian is given in Subsection 3.1. Overview of acquiring, describing and classifying the additional dataset is given in Section 4.1. Building of the morphological dictionary and its application in graphs is given in Subsection 4.2. In conclusion, our main research questions related to whether MWEs include trigger words and what is their role, and are the abusive MWEs mainly compositional or not, are presented at the end of the paper.

2 Related work

The use of offensive and hateful language has been a concern since the early days of the Internet. It has been estimated that the number of MWEs in the lexicon of a native speaker has the same order of magnitude as the number of single words (Moreno-Ortiz et al., 2013). Consequently, their detection is of great importance for abusive language identification. In our abusive language detection system, we are giving the same relevance to MWEs as to the individual words. Without MWE identification, an expression could be marked as not abusive since it does not contain abusive words e.g. *zapržiti nekome čorbu* (eng. to spice up someone's stew) with figurative meaning of meddling with someone's life in a negative way, deliberately making things difficult.

The majority of current hate speech, offensive and abusive language detection systems in social media are based on lexicons or blacklists (Chen et al., 2012; Colla et al., 2020; Pamungkas et al., 2020). The advantage of this approach relates to a vast number of swear words and offenses that can be detected solely using lexicons. The disadvantage of using lexicons is that swear words are used in everyday speech often without offensive intent, therefore their detection may lead to false-positive results. Another disadvantage of lexicons is due to the necessity of their maintenance since they evolve with natural language changes. While Pedersen (Pedersen, 2020) reported high accuracy of hate speech detection when using a lexicon only, lexicons are not sufficient as a resource for hate speech detection. They could be used as a baseline for comparison with more advanced methods since subtle hateful messages (Kwok and Wang, 2013), and language nuance cannot be detected accurately using this method. In addition, some insults which might be unacceptable to one group might be acceptable to another, thus taking context into account is very important (Nobata et al., 2016).

Several authors reported using a multilingual online lexicon of hate speech available at hatebase.org in their research. (Wiegand et al., 2018; Silva et al., 2016; Nobata et al., 2016). Wiegand et al. (2018) built a lexicon of abusive words using the subjectivity lexicon of Therese Wilson that is in essence a sentiment lexicon. They took words with negative polarity as a baseline for creating a basic lexicon of 551 words, which was further enriched via machine learning into a lexicon of 2898 abusive words. Several authors used the Wiegand lexicon as a blacklist in their hate speech and abusive language detection systems (Wiegand et al., 2018; Pedersen, 2020; Caselli et al., 2020). As noted by Wiegand et al. (2018), lexicons that contain different part of speech give better results than those containing just nouns, therefore we employed this approach in building our first abusive words lexicon.

An approach for racial, national, and religious hate speech detection adopted by Gitari et al. (2015) was based solely on the usage of lexicon and rules. They used semantics and subjectivity features – polarity, intensity, and subjectivity level of words, using the domain corpus of hateful content and Subjectivity lexicon of Therese Wilson in combination with the SentiWordNet (Esuli and Sebastiani, 2006).For classification, they leveraged rules and achieved a result of F1 = 0.783 for strongly hateful sentences on a manually annotated domain corpus.

Razavi et al. (2010) created one of the first systems for flame speech detection that used a 3-layer classifier, rules, and dictionary of abusive and derogatory terms. They noticed that offensive language was characterized by extreme subjectivity. Hence, they used a combination of term's offensiveness and subjectivity impact to calculate and assign a weight from 1 to 5 to each entry of their dictionary. Those weights were later corrected through adaptive learning on the training data set resulting in the Insulting or Abusive Language Dictionary (IALD). The dictionary contains terms such as "chew Somebody's ass out" that become a search template by replacing somebody with * or by adding suffixes or prefixes to verbs or nouns (e.g. suffix -ing for verbs and -es to nouns). They achieved an accuracy of 96,78% in

10-fold cross-validation on the binary classifier.

2.1 Resources for offensive and hate speech

When developing a lexicon of hateful, offensive or abusive words, researches usually start from the existing resources: (i) subjectivity lexicons, since it is assumed that hate speech contains elements of extreme subjectivity (Razavi et al., 2010; Wiegand et al., 2018), (2) a lexicon of sentimental words and expressions, and SentiWordNet (Gitari et al., 2015), where it is assumed that abusive language consists of words indicating negative polarity of feelings, (3) list of offensive words and expressions (Bassignana et al., 2018) and (Hatebase.org), whether made by experts and/or obtained using crowdsourcing.

In an abusive content detection system, a lexicon could be used in one of the following ways: (i) As a classification feature, either as a binary indicator of the abusive word occurrence in the examined text (Pamungkas and Patti, 2019), or a numerical value corresponding to the number of abusive words and its level of abusiveness (Razavi et al., 2010); (2) When applying rules for classification of offensive content, the authors may decide to classify the text in a certain category based on the number of abusive expressions above a certain threshold, e.g. if 2 or more notions of high abusiveness are found in text, it is marked as very abusive (Gitari et al., 2015; Pedersen, 2020); (3) Training of classifiers for recognizing abusive speech in text using the lexicon content as the training set (Wiegand et al., 2018).

On the other hand, high quality corpora of hate speech, offensive speech, and abusive language are just as important for tackling the detection of these phenomena online (Zampieri et al., 2019; Zampieri et al., 2020; Basile et al., 2019; Caselli et al., 2020). When it comes to Serbian abusive language resources and detection, the lexicon that we are working on is the first one of its kind. Still, some resources that will facilitate abusive language detection already exist. Serbian Morphological Dictionaries are certainly a staple in processing texts in Serbian (Krstev, 2008). In order to process implicitly abusive language, we need to take into account the usage of non-literal language, the rhetorical devices that are so often a part of such utterances, as shown in (Caselli et al., 2020; Mitrovic et al., 2020). The Ontology of Rhetorical Figures for Serbian (Mladenović and Mitrović, 2013) is a valuable resource for modelling and detection of rhetorical figures that play an important part in abusive language, e.g. irony, sarcasm, simile, hyperbole, litotes etc. Initial work on detecting some of these figures has been presented in (Mladenović et al., 2017; Krstev et al., 2020).

Using a corpus of newspaper articles from 2006, Krstev et al. (2007) presented the results of an information search experiment in search of attacks which are the result of national, racial, or religious hatred and intolerance. The aim was to develop a system which would recognize the news covering these topics, annotating certain components of the text, which, viewed individually or together, indicate the required information. The authors conclude that further development of the system could go in the direction of adding weight factors to the components (neutral, less neutral and offensive content and explicit content) which could be used to calculate the overall importance of a news item for the examined topic.

2.2 Multilingual HurtLex

HurtLex is a multilingual lexicon of hateful words in over 50 languages. The words are divided into 17 categories, plus a macrocategory indicating whether there is stereotype involved (Bassignana et al., 2018). Lemmas in this dictionary belong to one of these two levels: 1) conservative: obtained by translating offensive senses of the words in the original lexicon and 2) inclusive: obtained by translating all the potentially relevant senses of the words in the original lexicon.

The basis for HurtLex was a lexicon of offensive terms prepared by the Italian linguist Tullio De-Mauro, where offensive terms were split into 3 categories (negative stereotypes, derogatory words, and negative in context), and 17 subcategories. The creators of HurLex opted for a detailed categorization in order to have the possibility to search for a specific category or group of category types. This makes HurtLex amenable to automatic usage for tasks in many languages. Koufakou et al. (2020) used HurtLex in the TRAC-2 task for aggression and misogyny detection, to facilitate retrofitting of fastText word embeddings for English, Hindu, and Bengali. In Pamungkas et al. (2018), HurtLex was used to aid automatic identification of misogyny in English and Italian tweets, while in Colla et al. (2020) HurtLex was used in a system submission at OffensEval 2020, in the process of fine-tuning offensive language models

for Danish, Turkish, and English. In the research presented in this paper, we are improving the Serbian part of HurtLex, as it can be a powerful resource for detecting abusive language in Serbian.

3 Serbian HurtLex revision

3.1 srHurtLex lexical cleaning

The initial version of HurtLex for Serbian[1] has been analysed, first from a lexical point of view, then from a grammatical point of view. The errors in srHurtLex were introduced due to the automatically generated translation. In the retrieved data set, consisting of 2518 entries, there were 1903 unique lemmas, written in both Latin and Cyrillic alphabet. After alphabet unification, 1819 unique lemmas were first analysed using the Serbian Morphological Dictionaries (Krstev, 2008). The manual check-up of unrecognised words followed, resulting in the removal of 803 entries (602 unique).

Our next task was to check each lemma and its assigned part of speech (POS): 1) in 1057 entries (678 unique) the correct lemma was used, for which 93 (64 unique) the incorrect POS was assigned; 2) 658 entries (467 unique after correction) had incorrect lemma, out of which 48 (41 unique after correction) with incorrect POS.

If we look at the percentages on unique lemmata, 34.5% were non words, 38.8% lemma forms were correct, 26.7% lemmata were wrong, but 6% had wrong POS in total. So, we had a correct lemma with a correct POS assigned in 35.1% of the cases. Statistical overview is given in Table 1. A small set of orthographic corrections, such as first upper case lemma, was also conducted. Several types of errors were detected: 1) transliteration of foreign words into Cyrillic: *diddlei, villainess, ferociousness, carcharodon*; 2) foreign (not-translated) words: *anguillidae, anguilliformes, animal*; 3) irrelevant named entities: *Istočni Goti, Abulija, Animalija, Drag kraljica*; 4) literal translations that are meaningless in Serbian: *jabuka poliranje, javni pogodnost, japanskih jedinica merenja, nevolja kafu, nestašluk odluka, novog krompira*; 5) lemma correction in order to respect agreement in gender and number: *ekstremne desničar* → *ekstremni desničar* 'extreme right-winger', or to put a lemma in its dictionary form *zaprljane* → *zaprljan* 'soiled', *zlostavljanja* → *zlostavljanje* 'abusing', *zlostavljao* → *zlostavljati* 'to abuse'.

	Entries	Unique lemma after correction	%	Entries wrong POS	Unique wrong POS	% wrong POS
Non words	803	602	34.5			
lemmaOK	1057	678	38.8	93	64	3.7
lemmaNOT	658	467	26.7	48	41	2.3
Total	2518	1747		141	105	6.0

Table 1: Statistic of lexical cleaning.

Bearing in mind that the initial version of HurLex for Serbian was mostly done automatically, without support of any tools and resources for Serbian language processing, such results were expected and certainly indicate that this phase is inevitable in the construction of similar lexicons.

After the removal of all the wrong entries, 1725 entries remained. After removing duplicates, 1402 entries remained with 1000 unique lemmata. A total of 90 candidate entries for removal were annotated as both inclusive and conservative.

3.2 srHurtLex reclassification

The focus of this research was on MWEs, where 265 entries with 198 unique lemma were retrieved. Out of 265 entries, agreement in assigned category is confirmed for 156 entries, with a few suggestions for better translation: *sveštenikov pomoćnik* 'priest's assistant' → *đakon* 'deacon', *svinjski mesar* 'pig butcher' → *kasapin* 'butcher', *ženski imitator* 'woman impersonator' → *travestit*'transvestite'. 109 entries were eliminated for various reasons: 34 entries were marked as inappropriate due to a bad translation, and 12 were marked both as inclusive and conservative, of which only one remained. Most of

[1] https://github.com/valeriobasile/hurtlex/tree/master/lexica/SR/1.2

Label	HurtLex category description	no	yes	total
ps	negative stereotypes ethnic slurs	5	14	19
pa	professions and occupations	3	5	8
ddf	physical disabilities and diversity		2	2
ddp	cognitive disabilities and diversity	7	7	14
dmc	moral and behavioral defects	4	11	15
is	words related to social and economic disadvantage		3	3
or	plants	1		1
an	animals	26	10	36
asm	male genitalia	2	1	3
asf	female genitalia	2	1	3
pr	words related to prostitution	5	5	10
om	words related to homosexuality		8	8
qas	with potential negative connotations	9	23	32
cds	derogatory words	38	45	83
re	felonies and words related to crime and immoral behavior	5	16	21
svp	words related to the seven deadly sins of the Christian tradition	2	5	7
	total	109	156	265

Table 2: Statistic of HurtLex MWE categories.

others candidate for elimination were literate translations e.g. *domaća svinja* 'domestic pig' → *krme* 'pig', *komunalni otpad* 'communal waste' → *đubre* 'trash', *životinjski svet* 'animal world' → *stoka* 'cattle'.

A few examples may further illustrate why some MWEs had to be deleted from the Serbian HurtLex. The MWE *meka na dodir* 'soft to touch' may have near synonyms with abusive meaning *mekana, ljigava, slabašna* 'soft, slimy, week' but not in categories that were assigned to this entry: animals, female genitalia, male genitalia, derogatory words, cognitive disabilities and diversity, ethnic slurs. Also, *nekompetentna osoba* 'incompetent person', *neobrazovana osoba* 'uneducated person' can not be in the category *animals*. Instead of *zmija u travi* 'snake in the grass' one would use in Serbian just *zmija* 'snake'. Table 2 shows number of MWEs that were rejected (no) and confirmed (yes) per each HurtLex category.

4 MWE - dictionary construction

4.1 Selection of new abusive MWE entris

In order to find a set of words that can be triggers for MWEs and generally for offensive speech, a set of trigger (single) words was created. The lexical database Leximirka (Stanković et al., 2018), which supports Serbian electronic dictionaries (Krstev, 2008) was analyzed and entries with one of the following semantic markers were selected: Aug (augmentative, 103), Pej (pejorative, 89), POG (derogatory, 41). The additional 602 items from srHurtLex were added to the list. The Dictionary of Serbian Language (DS) (Vujanić, 2007) was also analysed and following abbreviations from dictionary entries were used to select additional words: vulg. (vulgar, 68), ir. (ironic, 224), pej. (pejorative, 981), pogrd. (derogative, 3), podr. (elongated, 29), prezr. (scornful, 17). A set of threats and offensive chunks (tweets, posts) was processed and additional 694 words were obtained.

The final list with 2,851 trigger (single) words (lemma) was used to collect MWEs that contains at least one of selected trigger word. Various sources were used: dictionaries, collection of threats and results of online search. Finally, a list of 4,624 MWEs was compiled that were candidates for the detection of some kind of offensive or hate speech.

This list was manually checked and each MWE was put into on of three categories: YES - abusive speech (1260), MAYBE – could lead to abusive content (462), NO – not abusive (2902). The manual classification was supported by search over a Twitter corpus collected specifically for his research, Web

	A	ADV	N	PRO	V	(blank)	Total
maybe	93	12	152	0	168	37	462
no	432	142	978	17	1333		2902
yes	213	39	367	0	474	167	1260
Total	738	193	1497	17	1975	204	4624
	%	%	%	%	%	%	
maybe	12.6	6.2	10.2	0.0	8.5	18.1	10.0
no	58.5	73.6	65.3	100.0	67.5	0.0	62.8
yes	28.9	20.2	24.5	0.0	24.0	81.9	27.2

Table 3: MWEs classified as yes, no, maybe and part of speech of trigger words.

and other corpora previously compiled. The distribution of MWEs by part of speech categories of their trigger word is presented in Table 3.

Further analysis showed that 45% of trigger words yielded no MWE marked as abusive, 19% had less abusive than MWEs categorized as not abusive or potentially abusive (NO and MAYBE), 14% had more abusive MWEs, while for 22% trigger words all extracted MWE were marked as abusive. An example of a trigger word for which both abusive and not abusive MWE were extracted is *junak* 'hero'. MWEs marked as abusive are: *junak gradskih salona*, 'hero of city salon' and *junak na jeziku* 'hero on the tongue (scaramouch)', while non-abusive are *junak romana* 'a hero of the novel', *junak našeg naroda* 'hero of our people'.

Since the list has been acquired automatically, manual correction of lemmas was necessary for 285 lemmas in YES classes. For example, MWEs composed of adjectives *topovski* 'relating to a cannon' are described in the dictionary as: "~meso" (meat, gender n.), "~hrana" (food, gender f.), where ~stands for a lemma itself. The automatic substitution produces incorrect MWEs "topovski meso", "topovski hrana" that have to be corrected in order to conform to the gender agreement with a noun, obtaining finally *topovsko meso*, *topovska hrana* 'meat/food for cannons'.

The categories in the lexicon are based on hate targets similar to (Silva et al., 2016) that originated from Hatebase.org scheme, which are further enriched with additional categories: Immoral and criminal activities, slurs, curses, and offense. A certain term in the lexicon can be assigned to several categories, in case it appears in the context of several types of abusive speech. Table 4 presents examples of abusive words in each category.

4.2 Lexical Representation of Multi-Word Abusive Expressions

In order to enable the detection of abusive language in Serbian texts it is necessary to represent in a lexicon both simple- and multi-word abusive expressions. Lexical representation should address various aspects of these expressions: morphological, syntactic, semantic, and usage. Morpho-syntactic characteristics of simple abusive words are for most of them already described in the Serbian Morpho-syntactic Dictionary (SrpMD) due to its comprehensiveness (Krstev, 2008). Various classes of multi-word expressions are represented in SrpMD as well, primarily noun and adjective expressions. However, none of the dictionary entries were labeled specifically for hate speech and abusive language detection (except with general markers for derogatory or pejorative usage, as mentioned in Subsection 4.1). Our aim was to enrich SrpMD with new MWEs related to abusive language, and to provide all relevant entries, both already existing and new, with markers appropriate for its detection.

In the first step, we analysed the mopho-syntactic structure of MWEs marked as positively or potentially abusive (markers MAYBE or YES). This list, originally having 1772 items, contained after separating variations in MWEs (e.g. *neka te (mutna) voda nosi* 'let the (muddy) water carry you', *visiti o (koncu / dlaci) 'to hang on a tread/hair'*) 1832 items. The most frequent were, as expected, MWEs with 2 components (893), followed by MWE with 3 components (464), 4 components (279), 5 components (119) and 77 MWEs with more than 5 components. MWEs were tagged using Serbian tagger (Stanković et al., 2020) and separated in two groups: nominal phrases (653) and verbal phrases (1179). Among nominal

Abusive category	Examples – single word	Example - MWE
Ethnicity and nationality (ABUS=racial)	*Ciganin*/Gipsy, *fašista*/ fashist, *Kinez*/Chinese, *jevrejski*/jewish	*praviti se Kinez*/pretending to be Chinese, *ciganska posla*/ gypsy business
Physical/mental disability (ABUS=disability)	*bogalj*/disabled, *budala*/ fool, *imbecil*/imbecile	*ptičiji mozak*/bird's brain
Physical/age discrimination (ABUS=appearance)	*kicoš*/dandy, *ćumez*/schack *debeljuca*/fatty, *baba*/grandma	*ružan k'o lopov*/ugly as a thief *matora devojka*/old maid
Sexual orientation (ABUS=sexual)	*guza*/butt, *gej*/gay *homoseksualac*/homosexual *travestit*/transvestite	*pederast izgled*/gay look
Behavior (ABUS=behavior)	*cepidlaka*/stickler *danguba*/dangler *drkadžija*/wanker	*pokondirena tikva*/ conceited pumpkin
Class (social, economic) (ABUS=seclass)	*bedan*/poor *ubog*/retched *buržoazija*/bourgeoisie	*go k'o crkveni miš*/ naked as a church mouse
Immoral/criminal activities (ABUS=law)	*bandit*, *bitanga*/rascal *lagati*/to lie, *izdajnik*/traitor	*ratni profiter*/war profiteer
Religion (ABUS=religion)	*nevernik*/infidel	*islamski fundamentalista*/ islamic fundamentalist
Race (ABUS=race)	*crnja, crnčuga*/Negro *Azijat*/Asian	*crn čovek*/black man
Gender (ABUS=gender)	*kurva*/whore, *drolja*/slut *krava*/cow, *žigolo*/gigolo	*ženski petko*/feminized man *laka žena*/easy woman

Table 4: Categories of abusive words and expressions with characteristic examples.

phrases the most frequent pattern is A N (448), a noun preceded by an adjective that agrees with it in the number, the case and the gender, for instance *belosvetska kurva* 'worldwide whore'. The other frequent patterns are: N N (50), usually a noun followed by a noun in the genitive or the instrumental case, e.g. *šaka jada* 'handful of misery', or two coordinated nous, e.g. *krava muzara* 'dairy cow'; N PREP N (25), a noun followed by a prepositional phrase, e.g. *govno od čoveka* 'shit of a man', *roba s greškom* 'damaged goods'; A ADV N (11), adjectives as simile figures, e.g. *glup kao noć* 'stupid as night', N CONJ N (9), two nouns connected with a conjunction, e.g. *bruka i sramota* 'shame and disgrace'. It should be noted that since MWEs were not syntactically parsed, some expressions were incorrectly grouped, e.g. *leglo opozicije* 'opposition's lair' was incorrectly recognized as V N pattern instead of N N due to the ambiguity of the form *leglo* (*leglo* is a noun 'lair' but also a form of a verb *leći* 'to lie down'); these cases were manually corrected. It should also be stressed that the assignment of POS tags to MWEs does not define the POS, or the role, an MWE itself does. For instance, *pukla bruka* 'scandal burst' has a common verbal phrase structure V N; however, it is a frozen expression mostly used as an interjection.

As already mentioned, prior to this task SrpMD contained 79 multi-word entries (noun phrases) from the compiled list of 653 nominal MWEs, however, without any marker pointing to their usage. After reallocating those that were incorrectly put into this group, and separating those that are not used as nominal but rather as frozen expressions, e.g. *duga kosa kratka pamet* 'long hair short wit' the list of 518 new MWE nominal entries was prepared, using semi-automatic procedure for MWE lemma construction (Krstev, 2008; Stanković et al., 2016) that relies on the information about its components already represented in SrpMD. The morpho-syntactic information is automatically assigned to all forms of multi-word expressions, while specific markers that point to their abusive usage were added according to the prior classification: HRT=yes and HRT=maybe. This is certainly a very rough classification, but a systematic annotation with more granulated classes is an ongoing activity.

In Table 5 we present some examples of produced MWEs lemmas, and some of their automatically produced forms. It should be noted that this way of representing MWEs in lexicon has a drawback

noun lemma	ženski(ženski.A2:adms1g) petko(petko.N68:ms1v),NC_AXN+Hum+HRT=yes
noun form	ženskog petka,ženskog petka.N+Hum+HRT=yes:ms2v
adj lemma	glup(glup.A15:akms1g) kao klada,AC_A4X+Simile+HRT=yes
adj form	glupog kao klada,glup kao klada.A+Simile+HRT=yes:adms2g kao klada glupog,glup kao klada.A+Simile+HRT=yes:adms2g

Table 5: Two examples of MWE lemmas and their forms in the genitive case singular: *ženski petko* (abusive for a man not manly enough) and *glup kao klada* 'stupid as a log'

because it does not cover any sort of insertions that may occur in an analysed text. Thus, a lexical description of *glup kao klada* will recognize different word orders *glup kao klada* and *kao klada glup* but not even the simplest insertions *glup je kao klada* 'lit. stupid is as a log'. For that reason we have started the more elaborate description of adjective expressions – simile figures – that relies on their joint tabular and finite-state description (Krstev et al., 2020).

We have started to apply the similar approach to verbal expressions. Among the group of verbal expressions the most frequent are those with a structure V N (289), followed by V PREP N (124), V PRO N (52), V N PREP N (40). Again, this is a rough analysis because each group may contain syntactically very different expressions. For instance, the V N group contains besides frozen expressions, e.g. *ode glava* 'head gone', expressions without complements, e.g. *sejati strah* 'sow fear', expressions with complements in the dative case, e.g. *prosuti creva* NEKOME 'to spill guts (to somebody)', expressions with complements in the accusative case, e.g. *lišiti* NEKOGA *slobode* 'deprive (somebody) of freedom', expressions with prepositional phrases as complements *zabadati nos* U NEŠTO 'pierce one's nose (into something)'. So far, a number of different structures were described in tables that cover lexical variants, e.g. Ekavian od Ijekavian word form (*pogaziti reč/riječ* 'trample the word', or synonyms (*lomiti/polomiti/slomiti vrat* 'break a neck'), complements, adjuncts etc. These tables are complemented with finite-state automata (FSA) that deal with word order, model complements, etc. and that are used to retrieve verbal expressions in texts. So far three classes of V N were modelled, covering 68 verbal MWEs.[2] This approach enables formulation of elaborate retrieal queries, similar to those proposed in (Razavi et al., 2010), but more precise since instead of a joker character * a more sophisticated patterns are used for complements and other insertions, e.g. a nominal phrase in the dative case.

We used our simple and MWE dictionary entries marked as (potentially) abusive to search our Twitter corpus containing approximately 8000 tweets and obtained around 800 hits, of which 80-90% indicated the abusive language. These are, however, just preliminary results that have to be confirmed on larger and more versatile corpora.

Figure 1 presents a Leximirka panel[3] for MWE editing: the syntactic class is assigned to a MWE, components and their morphological information are described which allows automatic production of all inflected forms that can be examined. Specific markers for abusive speech, that are proposed in this paper, can be assigned to a MWE entry through this panel.

5 Conclusion

In this paper we presented initial results on the analysis, improvement and creation of the lexical resources that will aid abusive speech detection tasks in Serbian, with a special focus given to MWEs, but there is still much to be done. Options of using a hybrid approach that would merge a dictionary with machine-learning will be explored. Finally, a user-friendly interface that will enable the usage of these resources on the Web is under development. We plan to use our lexicon of abusive speech to build a corpus of abusive content similar to (Rezvan et al., 2018) who firstly created an offensive word lexicon and then collected Twitter messages that contain at least one word from the lexicon. As authors noted, presence of a word in a tweet is just an indication of its offensiveness, thus subsequent manual annotation

[2]It should be noted that more MWEs from our list are described since in our model one expression groups variations.

[3]Leximirka is an online application based on lexicographic database, covering a wide range of users (http://leximirka.jerteh.rs//) (Lazić and Škorić,).

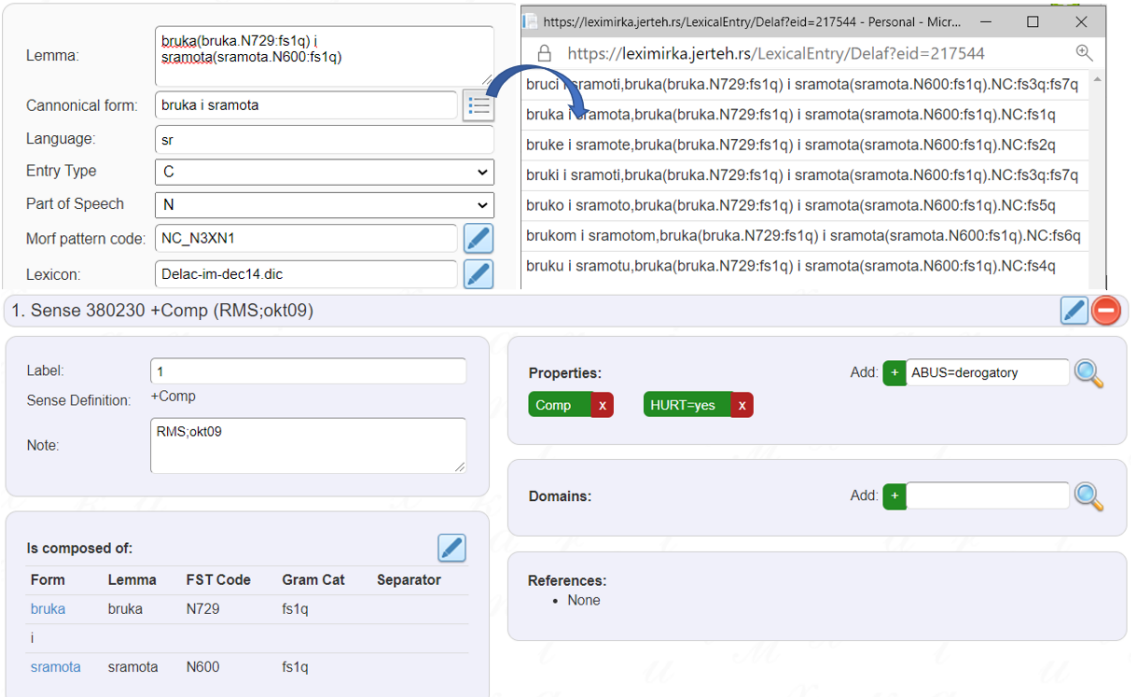

Figure 1: MWE editor in web portal Leximirka – description of a noun MWE *bruka i sramota* 'shame and disgrace'

is mandatory to assure correct classification of tweets.

In the next phases of the abusive words lexicon development, we plan to use: lists of slurs, abusive expressions, and courses built by conducting surveys and crowdsourcing (Mitrović et al., 2015), slang and dictionaries of synonyms, translation of existing lexicons in other languages, sentiment lexicon for Serbian language (Mladenović et al., 2016b), synsets from the Serbian WordNet (Mladenović et al., 2016a). We plan to use the lexicon for building a corpus of abusive content in social networks in Serbian as well as a classifier using rules and existing resources for Serbian language (Krstev et al., 2007). In addition, we plan to include the context rules and intensifiers following the approach presented in (Moreno-Ortiz et al., 2013) about the MWEs sentiment lexicon for Spanish. Additional attention will be given to the extension of the vocabulary with expressions that are not present in any existing lexicons, but evidenced in corpus as having offensive usage. The recognition of the different usages, that can be both offensive and non–offensive will be marked. The additional information about context or sense embeddings that will be useful for distinguishing between the two usages, could be added in the lexicon.

Acknowledgements

SPONSORED BY THE
Federal Ministry of Education and Research

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The author is responsible for the content of this publication. This research was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia and Deutcher Akademischer Austauschdienst - DAAD, Project years 2020-2021, Cross-Lingual Hate Speech Detection, and COST Action CA18209 - European network for Web-centred linguistic data science.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 11–16, 2020.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Davide Colla, Caselli Tommaso, Valerio Basile, Jelena Mitrović, and Granitzer Michael. 2020. Grupato at semeval-2020 task 12: Retraining mbert on social media and fine-tuned offensive language models. In *Proceedings of the 14th International Workshop on Semantic Evaluation(SemEvaleval)*.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Anna Koufakou, Valerio Basile, and Viviana Patti. 2020. Florunito@ trac-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 106–112.

Cvetana Krstev, Sandra Gucul, Duško Vitas, and Vanja Radulović. 2007. Can we make the bell ring? In *Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages*, pages 15–22.

Cvetana Krstev, Jelena Jaćimović, and Duško Vitas. 2020. Analysis of similes in serbian literary texts (1840-1920) using computational methods. In Svetla Koeva, editor, *Proceedings of the Fourth International Conference Computational Linguistics in Bulgaria (CLIB 2020)*. Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences, June.

Cvetana Krstev. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence*, pages 1621–1622.

Biljana Lazić and Mihailo Škorić. From dela based dictionary to leximirka lexical database.

Jelena Mitrović, Miljana Mladenović, and Cvetana Krstev. 2015. Adding mwes to serbian lexical resources using crowdsourcing. In *poster presented at The 5th PARSEME general meeting. Iași, Romania*, pages 23–24.

Jelena Mitrovic, Cliff O'Reilly, Randy Allen Harris, and Michael Granitzer. 2020. Cognitive modeling in computational rhetoric: Litotes, containment and the unexcluded middle. In Ana Paula Rocha, Luc Steels, and H. Jaap van den Herik, editors, *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 2, Valletta, Malta, February 22-24, 2020*, pages 806–813. SCITEPRESS.

Miljana Mladenović and Jelena Mitrović. 2013. Ontology of rhetorical figures for serbian. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech, and Dialogue*, pages 386–393, Berlin, Heidelberg. Springer Berlin Heidelberg.

Miljana Mladenović, Mitrović Jelena, and Cvetana Krstev. 2016a. A language-independent model for introducing a new semantic relation between adjectives and nouns in a wordnet. In *Proceedings of Eighth Global WordNet Conference*, pages 218–225.

Miljana Mladenović, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016b. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620.

Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, and Ranka Stanković. 2017. Using lexical resources for irony and sarcasm classification. In *Proceedings of the 8th Balkan Conference in Informatics*, BCI '17, New York, NY, USA. Association for Computing Machinery.

Antonio Moreno-Ortiz, Chantal Pérez-Hernández, and Maria Del-Olmo. 2013. Managing multiword expressions in a lexicon-based sentiment analysis system for spanish. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 1–10.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and V. Patti. 2018. Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In *EVALITA@CLiC-it*.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study.

Ted Pedersen. 2020. Duluth at semeval-2019 task 6: Lexical approaches to identify and categorize offensive tweets. *arXiv preprint arXiv:2007.12949*.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, pages 33–36.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016*, pages 507–514.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. Electronic dictionaries-from file system to lemon based lexical database. In *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018) (clocated with LREC 2018), McCrae, JP, C. Chiarcos, T. Declerck, J. Gracia and B. Klimek*, pages 48–56.

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3947–3955, Marseille, France, May. European Language Resources Association.

Milica Vujanić, editor. 2007. *Rečnik srpskoga jezika*. Matica srpska.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 75–86. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.