On a Hurtlex resource for Bulgarian

Petya Osenova
Department of Bulgarian Language
Sofia University "St. Kl. Ohridski"
petya@bultreebank.org

Abstract

The paper reports on the cleaning of the Hurtlex lexicon for Bulgarian as part of the multilingual Hurtlex resource.

All the challenges during the cleaning process are presented, such as: deleting strings or lexica that are clear errors from the automatic translation, establishing criteria for keeping or discarding a lexeme based on its meaning and potential usages, contextualizing the lexeme with the meaning through an example, etc. In addition, the paper discusses the mapping of the offensive lexica to the BTB-Wordnet as well as the system that has been used.

Keywords: Bulgarian, Hurlex, mapping, wordnet, offensive language.

1 Introduction

The Hurtlex resource is a multilingual lexicon of hate/offensive words. This resource started with the Italian hate lexicon developed by Tullio De Mauro and organized in 17 categories. Later on, it was automatically extended to other languages – through links to available synset-based lexical thesauri, among them MultiWordNet (Pianta et al., 2002) and BabelNet (Navigli and Ponzetto, 2010). Also machine translation was used (Bassignana et al., 2018).

HurtLex is a lexicon of offensive, aggressive, and hate words in over 50 languages among which Bulgarian. The words were classified in 17 categories like the source area of usage (plants, animals) or the target qualities or groups (moral and behavioral defects, with potential negative connotations). Also, a category was added whether the word expresses a stereotype or not. Since stereotypes are culture specific and usually determined through specially designed questionnaires within sociolinguistic frameworks, for Bulgarian this information has not been gathered yet. Thus, for the moment the inherited stereotypes from the Hurtlex resource are available without a focused localalized justification..

The hate words belong also to two other categories: either *conservative* (where only words with offensive senses were translated from the original lexicon) or *inclusive* (where all the potentially relevant senses of the words in the original lexicon were translated). As a result from the automatic translation due to the non-exhaustive coverage of various linked multilingual wordnets and lexical resources, the translated counterparts might include ambiguous, incorrect or questionable words.

Thus, during the process of obtaining the Hurtlex lexicons in other languages, it became clear that there is noise in the resulting resources such as inappropriate, non-comprehensible or unclear in their offensive meanings words. As a result, some cleaning is needed over the obtained lexicon per language. Such a cleaning, for example, was already performed for Modern Greek by Stamou et al. (2022).

The authors detect Greek offensive language by cross-classifying words on three dimensions: context, reference, and thematic domain. They add to the categorization also the social and the cultural aspect which are language specific.

Here we present a work in progress on the cleaning and linking of the Bulgarian part of Hurtlex. Version 1.2 for Bulgarian was downloaded from the hurtlex repository[1] and then manually checked. Initially, it had altogether 2865 words but some were already initially deleted since there were repetitions of the wordforms of the same lexeme. Also, normalization was performed over the same words with varying spelling – mostly the same word with a capital and small letter.

At the moment the lexicon consists of 1370

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

words which means that slightly more than 50 % of the original list of words were discarded before the human checks. At the same time, the lexicon is further enriched with the synonyms in the appropriate synsets through the sense mappings with BTB-Wordnet (Simov and Osenova, 2023). In addition, it has to be noted that Hurtlex does not include only isolated words but also phrases. Despite this fact, the proper handling of multiword expressions remains for future work since some of them came as isolated words and other – as expressions.

The structure of the paper is as follows: the next section presents the system that has been used for cleaning and enriching the resource. Section 3 discusses in more detail the workflow and challenges when editing the data. Section 4 focuses on the mapping of the Hurtlex lexicon to the BTB-Wordnet. Section 5 outlines the conclusions.

2 The resource editing system

The Hurtlex list for Bulgarian was inserted into a customized version of the specially designed CLaDA-BG Dictionary Creation System (Angelov et al., 2022). The system presents a multifunctional editor that can be used for creating various types of lexical resources – thesauri like wordnets, or traditional types of dictionaries (specialized, explanatory, spelling, etc.). The system also provides possibilities of interlinking the available data depending on the goal.

In general, the following information is present in the customized version:

- the lexeme/phrase itself
- the part-of-speech of the lexeme or of the headword of the phrase
- the status of the entry completion (with lables *Ready*, *To check*, *Irrelevant*)
- indication of whether the word is present in the wordnet or not (with labels T(rue) or F(alse))
- the definition of the lexeme meaning (it comes from the wordnet, if it is present there, or from other available dictionaries – if the lexeme and/or its appropriate meaning is absent)
- comments on any related to the check-ups issues

A partial screenshot from the system is given in Fig. 1. In the first column from left to right, offensive words are given that relate to being lazy and being scruffy or dirty. The second column indicates the part of speech. Here the examples are mostly nouns as in the resource itself. The third column marks the status of the entry completion – in this case marked as *Ready*. In column 4 the Hurtlex ID is given. Column 5 keeps track to the word as it came from the automatically generated Hurtlex resource. This means that the word might have come with an unnecessary capital letter or in a certain wordform instead of a lemma, or within an ungrammatical expression, or as an error, etc. However, the edited final lexeme is in the first column form since there the words have been normalized to their lemmas and proper spelling. The last column that is shown here, outlines the mapping to the appropriate entry in BTB-Wordnet or, if missing there, the relation to a suggested definition.

An example of a word is the last one down left in the table мърморко, ‘marmorko-M.SG’ (a grumbler). An example for a phrase is социален аутсайдер, ‘sotsialen-M.SG autsayder-M.SG’ (a social outsider). Both get a noun as part of speech.

The status marking shows whether the checking is accomplished (label *Ready*) or it should be paid attention to later (label *To check*). It also indicates whether the word is considered non-appropriate for the lexicon (label *Irrelevant*). We are aware of the fact that such decisions are not easy to take since many words can become offensive in a given context. For that reason they are not removed but just marked as not appropriate. Thus, we start with the lexemes or phrases with offensive lexical meaning and gradually will consider also the context-bound cases.

If the word is already present in the wordnet with the offensive meaning, then its ID, category, definition, examples, etc. are copied into Hurtlex. If however, the lemma/meaning is not present there, a definition is added by the editing linguist together with an example. The new information will be added to the BTB-Wordnet as well.

The linguist can leave comments of the following types: either there is no definition in the available dictionaries and other sources, or the definition is newly constructed, or the meaning is not offensive.

Despite the labels of readiness, additional systematic checks will be necessary after this first

мързел	noun	Готово	BG8	Мързел	Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк
мързелан	noun	Готово	BG82	мързелан	Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк
мързелив	adj	Готово	BG1519	мързелив	Id: 617 adj.all LEMMA: ленив, мързелив DEF: Който не обича и не
мързеливец	noun	Готово	BG1622	мързеливец	Id: 12508 noun.person LEMMA: мързел, мързеливец, мързеливк
мърлява жена	noun	Готово	BG295	мърляв жен	Id: 718 adj.all LEMMA: мръсен, изцапан, замърсен, нечист, мърл
мърляч	noun	Готово			Несръчен човек, неумел, лош работник или занаятчия
мърляч	noun	Готово	BG400	мърляч	Мръсен и противен тип.
мърморко	noun	Готово	BG2040	мърморко	Човек, който все мърмори и е недоволен.

Figure 1: An example screenshot from the system.

phase of Hurtlex data cleaning. Also, a strategy is required on how to deal with words are not offensive through their lexical meaning but might become offensive in a certain expression or communicative situation. For example, the word птица, 'ptitsa-F.SG' is not offensive but in the expression странна птица, 'stranna-F.SG ptitsa-F.SG' (a strange person) it might become offensive.

Another observation is that many words are not offensive but they just refer to intolerable models in society or bear mostly neutral meanings. For example, such type of words are фалш, 'falsh' (falseness) or мълва, 'malva-F.SG' (a rumor). An additional obstacle is the fact that we had to work with words without senses and contextualizing examples.

Let us look more closely into the challenges of cleaning Hurtlex in the next section.

3 Cleaning Hurtlex: Workflow and Challenges

The workflow includes several steps and these are organized in the following way:

- The lexicon is available in the editing system where one keeps track to the initial variant of the resource but also manipulates the data accordingly.
- The content is checked - initially it is done by the alphabetically ordered words, and later it can be performed through various characteristics such as the level of completeness, the (non)-availability in the wordnet, the part of speech, etc.
- The words/phrases are categorized into three types: *Ready*, *To check* and *Irrelevant*. It should be noted that all of them are being checked later again, including the ones labeled *Ready*.
- The correctness of the part of speech is also checked and changed, if necessary. In case of phrases, the part of speech is equal to the headword.
- If the lexeme is currently present in the wordnet with the appropriate meaning, it is marked with the boolean value *True (T)*, and then all the wordnet information is copied. Respectively, if the lexeme or the meaning is not present in the wordnet, the boolean value *False (F)* is selected. The inclusion of the missing words and/or meanings to the wordnet is envisaged as a future step. Also, the potential of the existing SentiWordNet (Baccianella et al., 2010) will be researched with the aim to see what part of the words/expressions with negative polarity intersect with the offensive lexica.
- Examples are added to each offensive meaning of the lexeme. Thus, the it can be seen in an appropriate context.

Thus, for the word тиква, 'tikva' (pumpkin) which is an offensive-oriented name of the head as part of the human body, there is an accompanying offensive synset in the wordnet, and the information is copied here. This entry can be seen in Figure 2. It provides the category *noun.body*, the set of synonyms as well as the definition. It should be noted that at this stage the synonyms in BTB-Wordnet are not marked as colloquial, dialectal, jargon and the like. These markings are envisaged for the future. Then comes the spelling of the word, its part of speech – noun, the status of the entry – in this case *Ready*, and the inclusion in wordnet – in this case *True*. The method of having obtained this word in Bulgarian is marked as *Inclusive*, i.e. all the potentially relevant senses were translated. The Hurtlex category is abbreviated as *or* – this means a relation to Plants. The lexeme is not considered stereotypical but as it was mentioned above, we should take this inherited information with a grain of salt since this dimension of information requires a special survey.

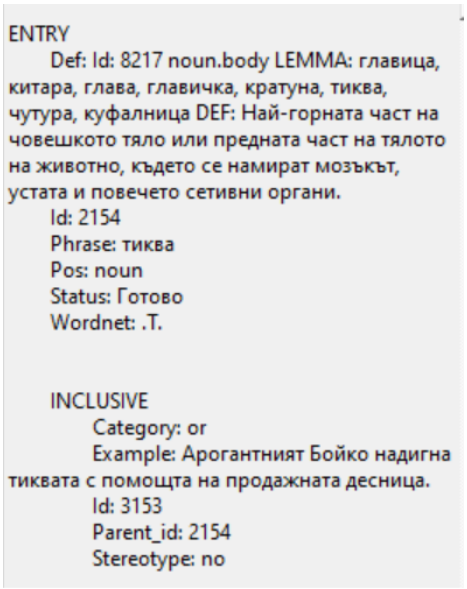

Figure 2: The information ticket for the word тиква, 'tikva' (pumpkin).

The identified challenges during the cleaning process are as follows: a) deleting strings or lexica that are clear errors resulting from the automatic translation, b) trying to establish criteria for keeping or discarding a word based on its lexical meaning and potential usages, c) contextualizing the word with an offensive meaning through an appropriate example, etc.

Let us look closer into each of these challenges:

Errors Here some errors are considered, such as collocations of arbitrary words or stop words. For the former some examples are: ябълка паркет, 'yabalka-F.SG parket-M.SG' (apple parquet) or томахавка-F.SG работа-F.SG 'tomahavka rabota' (tomahawk business); for the latter – с изключение на, 's-PREP izklyuchenie-N.SG na-PREP' (with the exception of).

Criteria Here the main problem is the presence of isolated words or phrases that might have offensive meaning in some context or from a certain perspective but without this background information it is difficult to decide.

Among the data that is not considered as part of the Bulgarian Hurtlex are the following ones: occupations like собственик на бизнес 'sobstvenik-M.SG na-PREP biznes-M.SG' (a business owner); специален пратеник, 'spetsialen-M.SG pratenik-M.SG' (a special correspondent); предприемач, 'predpriemach-M.SG' (an entrepreneur), портиер, 'portier-M.SG' (a porter); domains like растителни биотехнологии, 'rastitelni-PL biotehnologii-F.PL' (plant biotechnologies); some common language words like преобличане, 'preoblichane-N.SG' (changing clothes); пропуск, 'propusk-M.SG' (a pass); нормален, 'normalen-M.SG' (normal).

The words/expressions that have to be considered carefully, are the ones that do not have offensive connotations per se. For example, the lexeme умен, 'umen-M.SG' (smart) can become offensive only in an ironic context where the expression Много си умен,'Mnogo-ADV si-AUX.2PERS.SG umen-M.SG!' (You are very smart!) would mean the opposite statement Много си глупав!,'Mnogo-ADV si-AUX.2PERS.SG glupav-M.SG!' (You are very stupid!). Needless to say, handling irony is an important part of all the tasks related to detecting offensive language and attitude. In such tasks very often lexicons with sentiment and offensive lexica become an integral part of the complex architectures – see for example in (Hernández Farías et al., 2015). At the same time, focused annotated data, embeddings at various levels as well as Large Language Models are used for modeling contexts and pragmatic conditions.

Examples The idea of adding appropriate examples to the offensive meanings is related to the problems, described above. Adding examples to only-offensive words is by no means very important for introducing the typical context of usage. For example, the qualitative word with an offensive nuance темерут, 'temerut-M.SG' (an antisocial person) that is classified as CDS - derogatory words, is illustrated with the following example from the media: Този с държавната работа е мързелив темерут, който е постоянно ритан и бутан от по-мързеливи и корумпирани темерути като него. (*The one with the government job is a lazy antisocial person who has been constantly kicked and pushed around by even lazier and more corrupted antisocial people like him.*). It should be noted that in this example also other offensive words were used, namely the adjectives мързелив, 'marzeliv-M.SG' (lazy), and корумпиран, 'korumpiran' (corrupted).

But examples are even more important in non-trivial cases like the ones where a positive word can be turned into an offensive one, or the word is offensive with its figurative meaning only, etc. For example, the word професор, 'profesor' (a professor) might be used ironically to offend someone

who claims to be talking with competence on many topics: Тоя миндьовец (демек, селянин от с. Миндя) се изказва по всяка тема - голям "професор"се извъди. (*This man mindyovets (i.e. from Mindya village) speaks out on every subject – a great "professor" he has become.*).

For the selection of appropriate examples in Hurtlex we used internet and the Bulgarian part of the CLASSLA corpora – see more in (Ljubešić and Kuzman, 2024) – since it covers also non-standard communication in blogs, forums, etc.

4 Mappings to BTB-Wordnet

Since no mapping to synsets in Bulgarian wordnets was available in the originally compiled Hurtlex, we established our own mapping. In some cases, Hurtlex is being enriched with the synonyms of the lexeme – either also offensive ones, or from other registers. In other cases, the lexeme is not present in the wordnet and thus has to be mapped through some strategy such as a link to an appropriate hyponym and/or a hypernym.

The most common cases of not having links to wordnet are the following:

- *The word is missing.* For example, the word дивак, 'divak-M.SG' (a savage) is missing. The same holds for the word дриплъо, 'driplyo-M.SG' (a ragamuffin) and many others.

- *The word is present but the appropriate sense is missing.* This factor is the most frequent one. For example, the word свиня, 'svinya-F.SG' (a swine) has only the meaning of a domestic animal, but not the meaning of a filthy or bad person. The same holds for the word червей, 'chervey-M.SG' (a worm). It has the meanings of the animal and the computer virus but not the one of an insignificant person.

Table 1 shows some statistics on the Bulgarian Hurtlex data. The label *In Wordnet* indicates, as mentioned above, that more than half of the lemmas are not included there either as a lemma, or as an appropriate meaning.

The *Ready entry* means that there is no problem of identifying the word or phrase as an offensive one. More entries are like this but also a substantial part requires further elaboration. For example, the lexeme козел, 'kozel-M.SG' (a billy-goat) is marked as *Ready* but the link to the wordnet is to the meaning of man with a long sharp beard only. In fact, the examples show wider offensive usage towards an old man with an inappropriate behaviour.

Label	True	False
In Wordnet	641	729
Nouns	1119	251
Ready entry	817	553
No definition	72	1298

Table 1: Some statistics on the current status of the Bulgarian Hurtlex resource.

No definition refers to cases where a definition in this meaning cannot be found on a regular basis in Bulgarian dictionaries. But such cases are very rare. Most cases here refer to words that do not have an offensive meaning (like празненство, 'praznenstvo-N.SG', (celebration)) and to words that show ironic meaning in the relevant context.

5 Conclusions

This paper presents work in progress on the cleaning and enriching the Bulgarian Hurtlex resource. The main steps in this ongoing work and the related challenges have been outlined. Another challenge that was not mentioned before is the difficulty to handle the contemporary cases with rapid change of some words from positive to negative or with the emergence of an offensive meaning in a specific context like the political arena.

The next steps include: a) the handling of context-dependent offensive lexica; b) the handling of multiword expressions and bigger contexts; c) the addition of the missing lemmas and meanings into BTB-Wordnet.

As future work we also consider the annotation of a corpus with offensive senses. Only in this way more precise criteria can be set with respect to the delimitation of the offensive meanings within their context-bound process of generation.

References

Zhivko Angelov, Kiril Simov, Petya Osenova, and Zara Kancheva. 2022. The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In *CLARIN Annual Conference Proceedings*, pages 24–28, Prague, Czechia.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced

lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Italian Conference on Computational Linguistics*.

Delia Irazú Hernández Farías, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2015. ValenTo: Sentiment analysis of figurative language tweets with irony and sarcasm. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 694–698, Denver, Colorado. Association for Computational Linguistics.

Nikola Ljubešić and Taja Kuzman. 2024. CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3271–3282, Torino, Italia. ELRA and ICCL.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Kiril Simov and Petya Osenova. 2023. Recent developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 220–227, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.