UDK: 81’322
81’373.47
Izvorni znanstveni rad
Rukopis primljen 10. XII. 2022.
Prihvaćen za tisak 24. V. 2023.
https://doi.org/10.31724/rihjj.49.1.8

Barbara Lewandowska-Tomaszczyk[1], Slavko Žitnik[2], Chaya Liebeskind[3], Giedre Valunaite Oleskevicienė[4], Anna Bączkowska[5], Paul A. Wilson[6], Marcin Trojszczak[7], Ivana Brač[8], Lobel Filipić[9], Ana Ostroški Anić[10], Olga Dontcheva-Navratilova[11], Agnieszka Borowiak[12], Kristina Despot[13], Jelena Mitrović[14]

barbara.lewandowska-tomaszczyk@konin.edu.pl, slavko.zitnik@fri.uni-lj.si, cliebesk@g.jct.ac.il, gvalunaite@mruni.eu, anna.k.baczkowska@gmail.com, paul.wilson@uni.lodz.pl, marcin.trojszczak@konin.edu.pl, ibrac@ihjj.hr, lfilipic@ihjj.hr, aostrosk@ihjj.hr, navratilova@ped.muni.cz, agnieszka_borowiak@wp.pl, kdespot@ihjj.hr, jelena.mitrovic@uni-passau.de

ANNOTATION SCHEME AND EVALUATION: THE CASE OF OFFENSIVE LANGUAGE

The present paper focuses on the presentation and discussion of aspects of OFFENSIVE LANGUAGE linguistic annotation, including the creation, annotation practice, curation, and evaluation of an OFFENSIVE LANGUAGE annotation taxonomy scheme, that was first proposed in Lewandowska-Tomaszczyk et al. (2021). An extended offensive language ontology comprising 17 categories, structured in terms of 4 hierarchical levels, has been shown to represent the encoding of the defined offensive language schema, trained in terms of non-contextual word embeddings – i.e., Word2Vec and Fast Text, and eventually juxtaposed to the data acquired by using a pair wise training and testing analysis for existing categories in the HateBERT model (Lewandowska-Tomaszczyk et al. submitted). The study reports on the annotation practice in WG 4.1.1. *Incivility in media and social media* in the context of COST Action CA 18209 *European network for Web-centred linguistic data science* (Nexus Linguarum) with the INCEpTION tool (https://github.com/inception-project/inception) – a semantic *annotation* platform offering assistance in the annotation. The results

[1] University of Applied Sciences in Konin, Poland; [2]University of Ljubljana, Slovenia; [3]Jerusalem Institute of Technology, Israel; [4]Mykolas Romeris University, Vilnius, Lithuania; [5]University of Gdansk, Poland; [6]University of Lodz, Poland; [8,9,10,13]Institute for the Croatian Language, Zagreb, Croatia; [11]Masaryk University, Brno, Czech Republic; [12]University of Humanities and Economics, Lodz, Poland; [14]University of Passau, Germany; [14]Institute for AI R&D of Serbia
Corresponding author Barbara Lewandowska-Tomaszczyk, orcid.org/0000-0002-6836-3321

partly support the proposed ontology of explicit offense and positive implicitness types to provide more variance among widely recognized types of figurative language (e.g., metaphorical, metonymic, ironic, etc.). The use of the annotation system and the representation of linguistic data were also evaluated in a series of the annotators' comments, by means of a questionnaire and an open discussion. The annotation results and the questionnaire showed that for some of the categories there was low or medium inter-annotator agreement, and it was more challenging for annotators to distinguish between category items than between aspect items, with the category items *offensive*, *insulting* and *abusive* being the most difficult in this respect. The need for taxonomic simplification measureson the basis of these results has been recognized for further annotation practices.

1. Introduction

Abusive or offensive language is commonly defined as hurtful, derogatory or obscene utterances made by one person to another person or group of persons (Wiegand, Ruppenhofer and Eder 2021). Offensive discourse refers to the presence of explicit or implicit verbal attacks towards individuals or groups and has been extensively analyzed in linguistics (e.g., Culpeper 2005; Haugh and Sinkeviciute 2019) and in NLP (e.g., OffensEval (Zampieri et al. 2020), HASOC (Mandl et al. 2019)), in terms of of *hate speech*, *abusive language*, *offensive language*, etc.

2. Related work and proposal

As discussed in Lewandowska-Tomaszczyk et al. (2021), most of the available offensive datasets have been created with no consideration for valid linguistic explanations. Computational models of offensive language have been based on detection of hate speech against immigrants and women (e.g., Basile et al. 2019), using results of the OffensEval Tasks of SemEval-2019 and SemEval-2020 (Zampieri et al. 2019a, 2019b). Automatic identification systems of offensive language use several diverse approaches such as, for example feature-based linear classifiers (Waseem and Hovy 2016; Ribeiro et al. 2018), neural network architectures (Kshirsagar et al. 2018; Mishra et al., 2018; Mitrovic et al. 2019), or fine-tuned pre-trained language models, such as BERT and RoBERTa (Liu et al. 2019; Swamy et al. 2019).

This paper focuses on the presentation and discussion of aspects of linguistic annotation of OFFENSIVE LANGUAGE, including the creation, annotation practice, curation, and evaluation of an OFFENSIVE LANGUAGE annotation taxonomy scheme, first proposed in Lewandowska-Tomaszczyk et al. (2021) and extended to cover a more detailed, linguistically valid, offensive language ontology (Lewandowska-Tomszczyk et al., submitted). The latter comprises 17 categories, structured in terms of 4 hierarchical levels, and represents the encoding of the defined offensive language schema, trained in terms of non-contextual word embeddings – i.e., Word2Vec and Fast Text and, eventually juxtaposed to the data acquired by using a pairwise training and testing analysis for existing categories in the HateBERT model.

The taxonomic system referenced above, is used for the annotation campaign in WG 4.1.1. *Incivility in media and social media* in the context of COST Action CA 18209 *European network for Web-centred linguistic data science* (Nexus Linguarum) with the INCEpTION tool (https://github.com/inception-project/inception) – a semantic *annotation* platform offering assistance in the annotation. This paper presents the results of the annotation campaign on English datasets and ways to proceed further. We identify and discuss corresponding offensive category *levels* (types of offence target, etc.) and *aspects* (offensive language property clusters) as well as categories of *expressiveness* (*explicit – implicit*, figurative language types) in the data. *Aspects* are meant to group those offensive cases which are lower in the hierarchy of the taxonomy than the main category types. However, due to the category character of this semantic type of expressions, no strict hyponymic relations are expected to hold between the upper main categories and the aspects that may characterize them. This type of problem is experienced in numerous computational linguistic tasks (e.g., in linguistic annotating of Named Entities, which are not well defined (e.g., Sekine and Ranchhod 2009; Yin and Shah 2010), or Events, that are typically annotated with low or moderate inter-annotator agreement (Mowery et al. 2013). There are typically strict category identification problems in using the tool, where there are overlaps, apart from a few taxonomic levels.

The paper does not refer to or discuss visual elements of social media posts, nor the graphemic peculiarities introduced by posters in social media comments, although offensiveness at such levels has been recognized by the authors in

Lewandowska-Tomaszczyk et al. (2021) and envisaged in an overall approach to offensiveness in social media.

The results partly support the proposed ontology of explicit offense and positive implicitness types (see Bączkowska et al. 2022; Despot and Ostroški Anić 2022 for more extensive discussions). However, in view of the fact that some of the categories appeared semantically very close, the final recommendations would need to re-consider the original conditions. The consecutive parts of the paper present the use of the annotation system and its results in the representation of linguistic data evaluated in a series of the annotators' comments, by means of a questionnaire and in an open discussion.

3. Offensive language taxonomy

The repertory of offensive language classification headwords used in the available tagset systems, pose classificatory and computational problems. The model in our previous study (Lewandowska-Tomaszczyk et al. 2021) was closely related to the original 3-level category system proposed by Zampieri et al. (2020). Our modifications consist of finding additional evidence to support our linguistic judgments, and retains 2 categories and 4 sub-levels, tested by means of Sketch Engine tools on a large web-based corpus and juxtaposed to the non-contextual word embeddings fastText, Word2Vec, and Glove on the relevant datasets.

The repertory of offensive categories proposed in Lewandowska-Tomaszczyk et al. (2021) comprised 11 types of language: (1) offensive, (2) taboo, (3) insulting, (4) hate speech, (5) harassment, (6) vulgar, (7) vulgar/obscene, (8) vulgar/profane, (9) abusive, (10) vulgar/slur and (11) cyberbullying. Having examined 60 publicly available datasets based on more clearly identifiable criteria (e.g., datasets by Chung et al. (2019), Ousidhoum et al. (2019), Zampieri et al. (2019a, 2020)), an extended offensive language taxonomy was proposed (Lewandowska-Tomaszczyk et al., submitted) that covers 17 offensive language categories and subcategories as presented in Figure 1 to provide more variance among offensive language types as well as to put forward clearer linguistic and distributional criteria of their identification in actual language types.

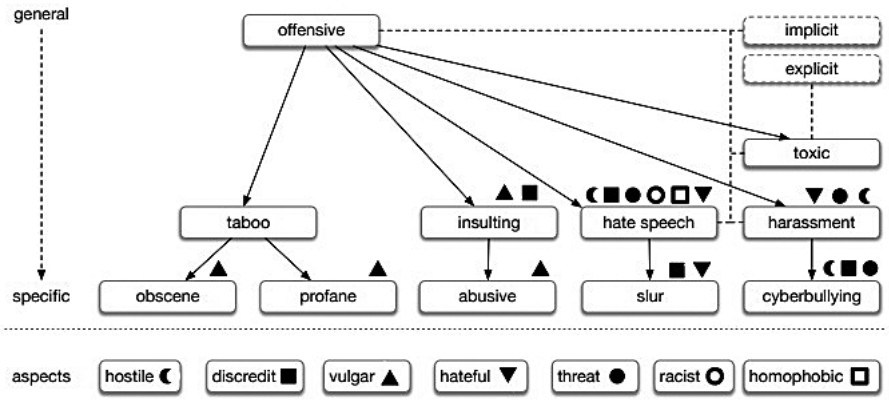

Figure 1. Offensive language taxonomy

The upper part is hierarchically organized from general to specific and some categories are annotated with contexts. Implicit and explicit offensive types are also differentiated. The bottom line identifies offensive language variants [aspects], marked by symbols (e.g., circles) to indicate semantic overlaps.

The computational instruments used for verification of offensive language types in actual texts as well as the results are presented in the forthcoming sections of the paper. While keeping the number of proposed headwords to 17 categories, which means that it has a similar tazonomic structure to that in our initial proposal in Lewandowska-Tomaszczyk et al. (2021), this paper aims to provide a practical evaluation for the categories presented in Figure 1.

Asis the case with any type of study employing a classification methodology, the identification of categories of offensive language is not easy or straightforward. We propose this extended taxonomy, bearing in mind the fuzzy categorial nature of all concepts, first signalled by mathematicians and philosophers (e.g., Zadeh 1965) and widely recognized in contemporary cognitive studies and cognitive linguistics theories (Lakoff 1987). The categories of offensive language are no exception and they are subject to natural 'leaking' as any taxonomic system referring to natural language. On the other hand, there are some reasons to assume that the proposal might present a more explicitly delineated schema than the ones currently used in offensive language identification systems.

The concept of *offence* is considered in the present study as a dominant category, which includes both more weakly and more strongly experienced offence, depending on the use of expressive means by the offender(s). Expressiveness can be realized by a wide range of semiotic means – from offending behaviour, external attributes, such as garment or hairstyle, which can be visually sive, to some addressee(s), through various types of verbal offence of different degrees of intensity. Furthermore, the messages conveyed by the vocal communication channel are most direct, while in the present study we focus on offensive language in social media texts, where expressiveness can be additionally marked by multiple repetitions, capitalisation, punctuation or visual symbols, not considered in the present study.

4. Offensive language annotation campaign evaluation

In recent years, a large number of offensive language datasets covering various facets of offensive language have been compiled as referred to in Tharindu and Zampieri (2020). A significant portion of the available datasets were generated ad hoc or as part of broader assessment initiatives, such as Kaggle competitions or shared projects. We have found approximately 60 offensive language corpora, of which about 30 are in English. Others are in Croatian, Danish, Arabic, German, French, Greek, Hindi, Indonesian, Italian, Polish, Portuguese, Spanish and Turkish. We were then able to extract 25 datasets (a complete list is provided in Lewandowska-Tomaszczyk 2021), from which we randomly sampled texts and configured the INCEpTION tool[2] to support the annotation of the proposed offensive language annotation schema. Most of the annotators were previously engaged in the schema preparation and discussions. The annotation campaing started on December 9, 2021 when we had an introductory annotators' meeting to explain the guidelines and rules for the annotation. The first introduction to the annotation guidelines was given at a workshop of the *Nexus Linguarum* general meeting in Skopje in September, 2021. The guidelines presented a step-by-step procedure for the annotatation categories (Figure 1), with carefully selected social media examples of each. Following the presentation an exercise in

[2] https://inception-project.github.io/ (Accessed March 25, 2023).

annotating authentic corpus samples took place, with a conclusive part devoted to a discussion of the tags each of the particiants in the team proposed for the samples. For the campaign we eventually attracted nine voluntary annotators (most of them with linguistic background); however, it was not possible to offer remuneration. There were three curators (the main contributors to the annotation schema and to the guidelines), and two members for technical support (data import/export and analysis).

The INCEpTION tool was configured to automatically feed new annotation documents to the annotators. Each document was annotated by two different annotators and then finally checked by a curator. According to the annotation guidelines distributed among the annotators, the annotators needed to select one or more consecutive sentences that comprised offensive language instances and then fill out nine different parameters (i.e., offensive language type, three possible aspects, expressiveness, target type, target level, and figurativeness). If there were multiple disjoint or various types of offensive language in a document, the annotators were instructed to annotate all. Also, annotations could be overlapping (e.g., if a sentence was a part of two different offensive language categories). Figure 1a shows the annotation workspace for the annotators within the INCEpTION tool. The tool was flexible enough to facilitate all needed options for annotation (e.g., overlaps, various levels, optional categories). The annotators could add, edit and remove all the annotations and features. When they finished annotating a document, they marked it as annotated. A curator could curate an annotated document after two different annotators marked a document as finished.

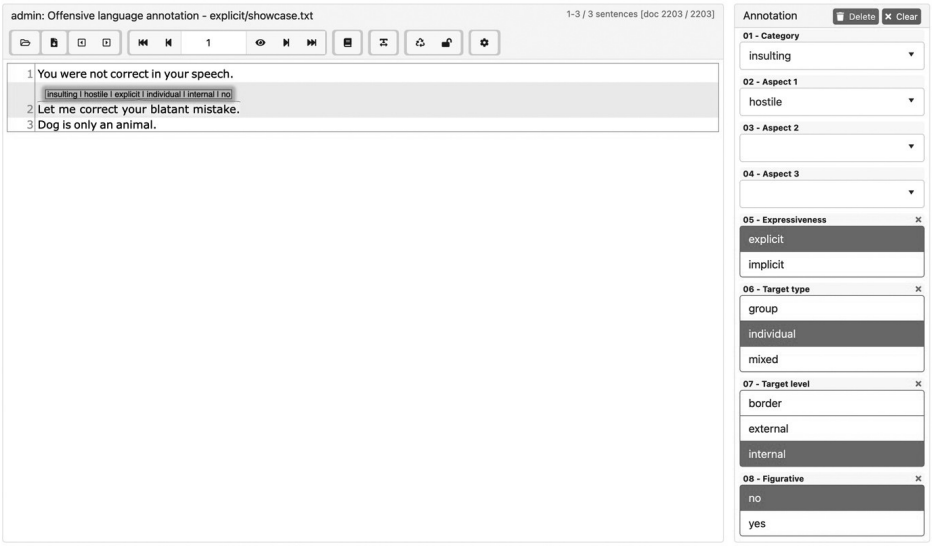

Figure 1a: Annotation workspace within the INCEpTION tool

During the annotation campaign, messages were interchanged among the annotators and curators to align their annotation problems and the main goal of the project was to validate the proposed annotation schema and continue with a larger annotation campaign. The curators checked the annotated documents and then we exported the dataset on July 28, 2022. The analysis[3] showed that we had 331 documents annotated by at least two annotators and 519 annotator-curator document pairs. We compared the annotations agreement using Cohen's Kappa which represents a statistic that tries to minimize awarding matches that might occur only by chance. The achievement of agreement scores of 0.60 or more represents substantial or perfect agreement. The aim of the questionnaire was to determine possible correspondence between ease of differentiation between aspect items and category items and inter-annotator agreement.

[3] The analysis source code repository: https://github.com/UL-FRI-Zitnik/offensive-language-organization/tree/master/NexusDataset%20agreement%20tests, see IAA calculation.ipynb (Accessed October 31, 2022).

5. Inter-annotator agreement

We first checked inter-annotator agreement between annotators. As the basic annotation unit was a sentence, we compared all the sentences (1264) from the annotated documents (331).

The agreement for each specific annotation type is represented in Table 1.

Table 1. Inter-annotator agreement per annotation type

Annotation type	**Agreement**
Category	0.322
Aspect 1	0.291
Aspect 2	0.260
Aspect 3	0.182
Expressiveness	0.487
Figurative	0.188
Target level	0.444
Target type	0.432

We observed that none of the annotation dimensions was successfully annotated to a level that would be useful to compile a corpus.

The annotators were instructed to select three aspects (the most appropriate aspect as *Aspect1*, then *Aspect2*, and lastly *Aspect3*, if applicable). Thus, we could check the matchings among all the three aspects. Based on the annotations we calculated matches in all of the three annotated aspects. Table 2 shows annotated aspects matches in the sentences jointly annotated as offensive.

Table 2. Annotated aspects in the sentences

Number of aspect annotations matches	The proportion (in %)	Number of sentences
No aspect matches	20.6%	62 sentences
Aspect 1 matches	69.4%	209 sentences
Aspect 2 matches	9.6%	29 sentences
Aspect 3 matches	0.3%	1 sentence

It can be observes that 70% of sentences that had an aspect value annotated by both annotators (we did not count other annotations) match in one aspect. Only 10% of such examples match in two identical aspects, while there are 20% mismatches and almost no complete matches.

Offensive category (*Category*) is the criterial element of the taxonomy and it is crucial for it to achieve an acceptable score. However, it was also the case that

a Kappa score of only 0.32 was achieved. Below we list the top three combinations of matches and non-matches between different values of the offensive category by the annotators. Non-matches can reflect inter-category identification problems and are the key for further improvement on annotation guidelines or the taxonoy system (see also annotator questionnaire results in the sections to follow).

Table 3. Top non-matching pairs

Pair	Proportion (in %)	Number of pairs (out of 1264)
insulting, offensive	4%	51
abusive, offensive	2%	29
offensive, slur	2%	21

Table 4. Top matching pairs

Pair	Proportion (in %)	Number of pairs (out of 1264)
hate speech, hate speech	5%	57

We can observe that the categories *insulting, offensive, abusive* and *slur* were often difficult to distinguish by the annotators, while a high degree of agreement was reached in identifying the category *hate speech*. The difficulty in differentiating between the categories *insulting, offensive* and *abusive* is consistent with the questionnaire results (see section 7).

5.1. Inter-annotator agreement between pairs of annotators

As some annotator pairs reached better agreement than others, we assessed results of specific annotator pairs that jointly annotated some of the documents. There are ten annotator pairs that jointly annotated fewer than 10 documents and these were not included. Other pairs annotated the following number of joint documents: 74, 74, 68, 42, 20 and 11.

Annotation results between pairs of specific annotators showed that there were not particularly high levels of interannotator agreement. Two pairs that annotated the highest amount of joint documents achieved around 80% of matches for one joint aspect. One of the pairs achieved more than a 0.60 Kappa score for expressiveness, the target level and the target type, while other dimensions are rather low.

6. Inter-annotator agreement between annotators and curators

There were three curators for the annotation campaign, who prepared the annotation instructions for the annotators prior to the campaign. Comparing the annotations by annotators and curators might reveal better insights into the quality of the annotations. The number of documents that were curated by specific annotators are as follows: 294, 270 (Annotator 1), 112, 98 (Annotator 2), 85, 68, 56, and 33.

The Kappa scores achieved by annotators and curators are higher compared to those achieved by inter-annotators, and in general higher than 0.40. Based on the results, the best performing annotators (i.e., those that had the highest degree of alignment with the curators) were Annotator 1 and Annotator 2. We provide their agreements scores in Table 5.

Table 5. The agreement between the best performing annotators

	Annotator 2	**Annotator 1**
Sentences	653	1066
Category	0.75	0.71
Aspect 1	0.76	0.73
Aspect 2	0.68	0.62
Aspect 3	0.46	0.55
Expressiveness	0.79	0.81
Figurative	0.67	0.65
Target level	0.77	0.78
Target type	0.79	0.79

Furthermore, results of matching of 3 types of offense *aspects* appear more persuasive than those of general category types. A significant increase in matches of two or even three aspects are observed.

Table 6. Matching of offense aspects

% of matchings (# documents)	**Annotator 2**	**Annotator 1**
No aspects match	0.8% (1)	10.5% (38)
Aspects 1 match	70.0% (91)	44.4% (161)
Aspects 2 match	26.9% (35)	24.0% (87)
Aspects 3 match	2.3% (3)	21.2% (77)

The annotation results performed by these two annotators (Annotator 1 and Annotator 2) converge to a fairly large extent. Furthermore, they also align with the

curators' annotations and this demonstrates their usefulness in the annotation campaign.

For the annotation of categories, the following category pairs (Table 7) had the highest number of mismatches.

Table 7. Category pairs according to mismatches

Pair	Proportion (in %)	Number of mismatches (out of 1066)
insulting, offensive	3%	32
hate speech, insulting	3%	28
insulting, slur	1%	8
insulting, toxic	1%	4

7. Findings

On the basis of on the annotation campaign results we conclude the following:

(a) Either the annotation guidelines were not precise enough, or the communication with the annotators was not suffciently effective. Furthermore, a more effective communication among the annotators or curators during the annotation campaign might have improved the results. Moreover, some of the category terms require knowledge of specialized contexts and/or linguistic expertize to correctly annotate the naturally occurring texts.

(b) The results show that annotator-curator agreement is better than annotator-annotator agreement. As the curators were particularly comprehensive in checking the data, they annotated all possible offensive language instances. In contrast, the annotators might have mainly detected the most obvious parts of offensive language. If two annotators identified and annotated different parts, their annotation outcomes did not match and their inter-annotator agreement could be much lower.

(c) Although the annotation was performed at a sentence-level, an annotator could mark multiple annotations for a sentence. The idea of the annotation campaing was to identify parts of the text that represent offensive language and not to annotate the whole document as offensive, which is

a procedure performed in the majority of existing datasets for offensive language classification.

(d) The data for annotation was randomly selected from existing offensive language corpora. For further campaigns, data preparation and corpus compilation might be selected on the basis of more thorough scrutiny of the offensive dataset criteria.

(e) The annotators were acting on a voluntary basis, with no remuneration, and their involvement in the task could therefore have been somewhat partial.

8. Questionnaire

In order to shed some more light on the annotation practice and identify the level of the discrimination between particular offensive categories, a questionnaire was administered among the annotators who took part in the annotation exercise presented above.

8.1. Methodology

Eight annotators rated the aspect items and nine annotators rated the category items. There were two sets of rating tasks that were performed by the participants that took part in the annotation exercise – one on the aspect items (*threat, hostile, hateful, racist, homophobic, vulgar, discredit*) and one on the category items (*offensive, insulting, abusive, toxic, harassment, hate speech, slur, bullying, taboo, obscene, profane*). The methodology for each of these sets was the same. Items in each of the sets were paired with all of the other items in their respective sets. Participants were asked to rate how easy or difficult it was to distinguish between each pair of items when performing the annotation (Table 8).

Table 8. Rating scale for a pair of category items

Abusive – Bullying										
Very easy to distinguish between	1	2	3	4	5	6	7	8	9	Not at all easy to distinguish between

In the present study we adapted the NodeXL (Smith et al. 2010) tool to provide information pertaining to the ease with which participants were able to differentiate between the use of category and aspect items while annotating. This is represented in our NodeXL graphs, with higher values representing greater difficulty in distinguishing between pairs of items.

8.2 Questionnaire results

8.2.1. Aspect Results

Figure 2 shows that the higher co-occurrence values for the aspect items are between *hostile*, *threat* and *hateful*. Examining these relationships more fully, it can be seen that *hostile* has a particularly close relationship with *threat* (5.2) and *hateful* (5.4), with a relatively weaker association between *threat* and *hateful* (4.1). These relatively strong interconnections between *hostile*, *threat* and *hateful* contrast with the lower co-occurrences between these and the other aspect items, namely *racist*, *homophobic*, *vulgar* and *discredit*. It can also be seen in Figure 3 that there are relatively low interconnection values between *racist*, *homophobic*, *vulgar* and *discredit*.

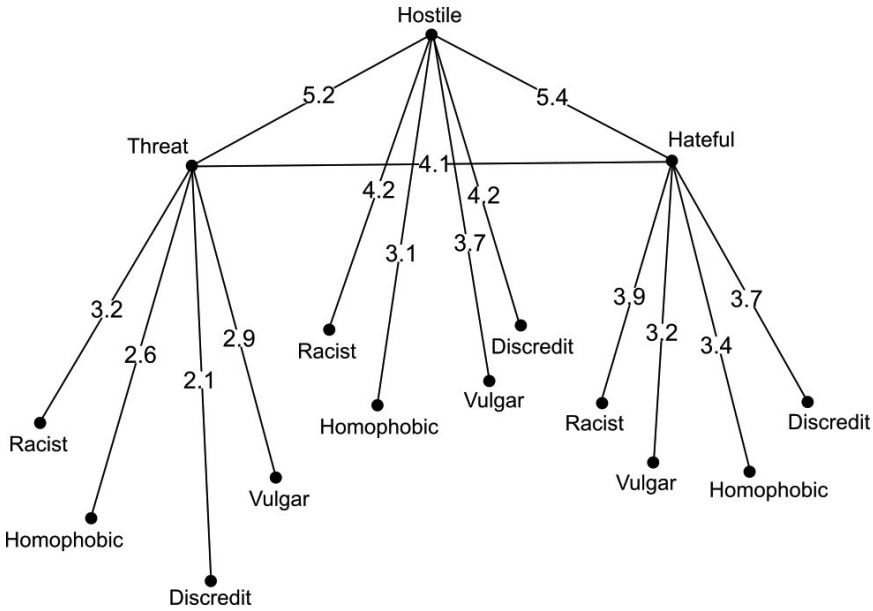

Figure 2. Interconnections between aspect items

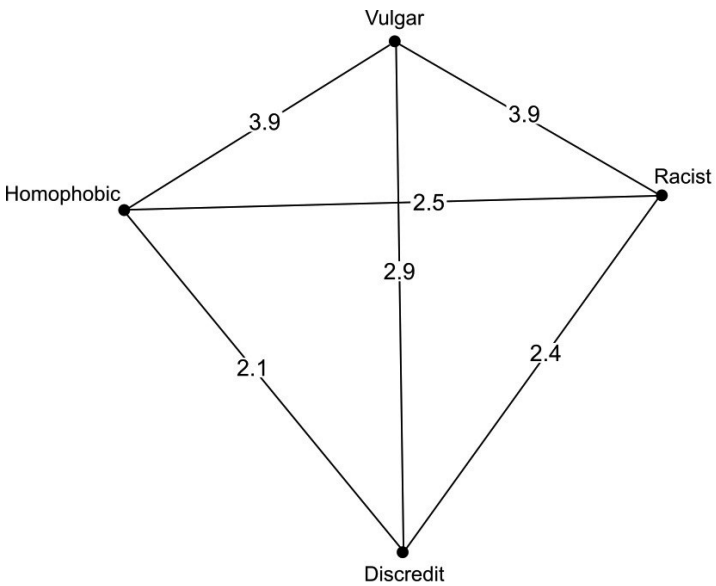

Figure 3. Interconnections between *racist*, *homophobic*, *vulgar* and *discredit*

8.2.2. Category Results

The first point to note from a general perspective is that there are higher interconnections between category items (Figures 4 and 5) than between aspect items (Figures 2 and 3). With respect to category result specifically, Figure 4 shows the interconnections between *offensive* and *abusive* (7.5), *offensive* and *insulting* (7.2), and *abusive* and *insulting* (6.9). By contrast, there are low co-occurrences between these three items and *toxic*, *harassment*, *slur* and *hate speech*. Also, there are relatively low interconnections between *harassment*, *hate speech*, *slur*, *bullying*, *taboo*, *obscene*, *profane* and *toxic* (see Figure 5).

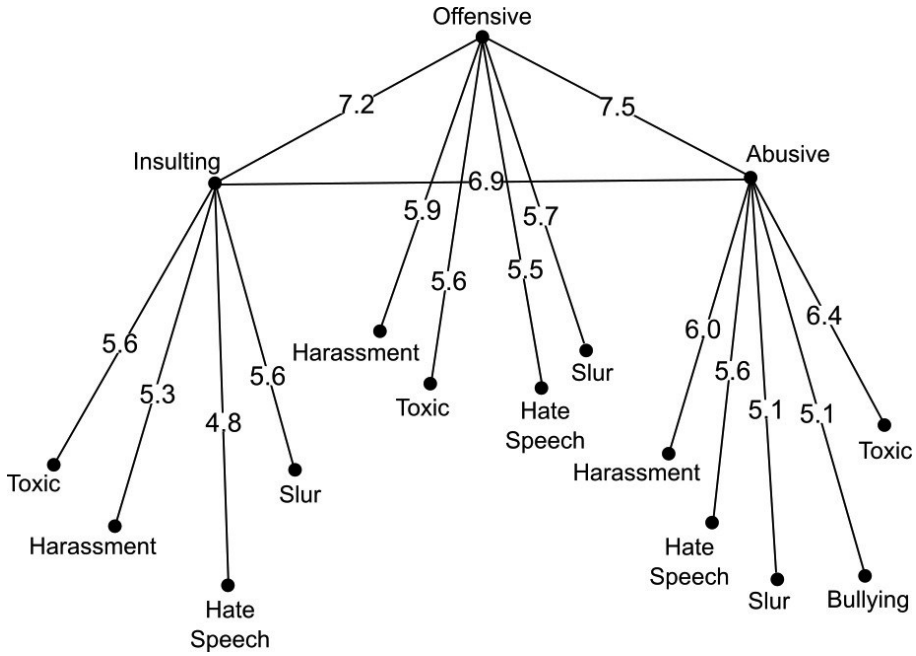

Figure 4. Interconnections between category items

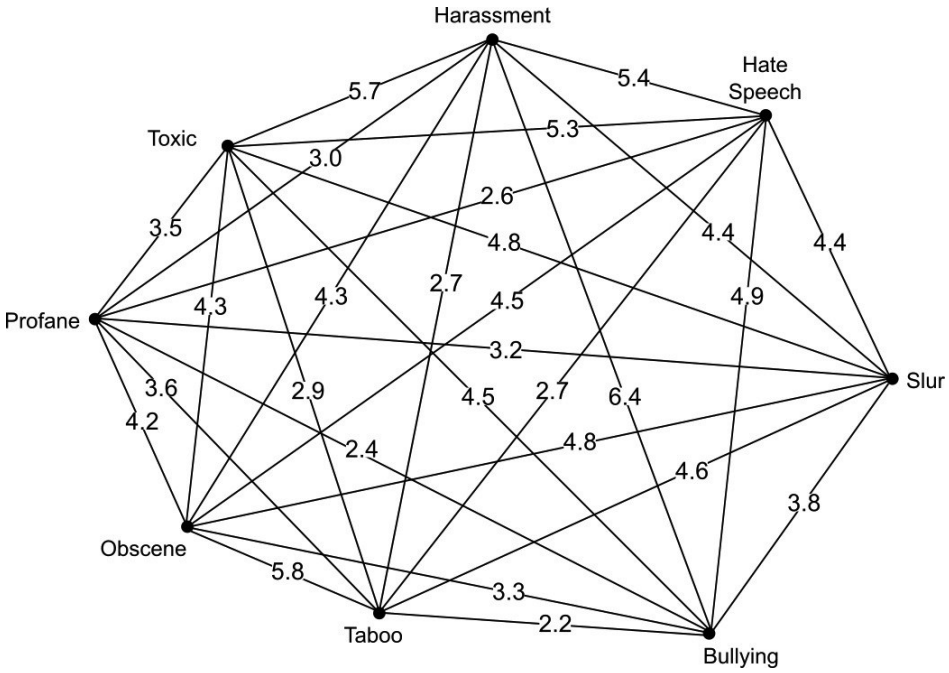

Figure 5. Interconnections between *harassment*, *hate speech*, *toxic*, *slur*, *bullying*, *profane*, *obscene* and *taboo*

8.3. Questionnaire conclusions

To conclude, the higher interconnection values between the category items than between the aspect items show that annotators found distinguishing between the category items more demanding than between the aspect items. With respect to the aspect items, the annotators reported that it was more challenging to distinguish between *hostile* and *threat* and between *hostile* and *hateful*. Taking into account the results of the aspect items and category items as a whole, it was the most difficult for annotators to differentiate between the category items of *offensive*, *insulting* and *abusive* when performing the annotation task. This corresponds with the findings of the inter-annotator agreement between annotators (see section 4).

These results might indicate a need for a simplification of the taxonomy at this point by collapsing the categories of *insulting* and *abusive* into one taxonomic

level. The category *offensive* on the other hand should be retained as it answers the basic question referring to the most general - superordinate - taxonomic level to provide information whether the annotated sample considered as a whole is regarded offensive or not. Only if it is, will it pass to the more subordinate levels of offensive language categorization.

9. General Conclusions

The paper presented the identification of Offensive Language categories in English, based on the analysis of the linguistic corpus data (collocations and synonyms in particular) of Sketch Engine. It also puts forward an elaborated Offensive Language taxonomy, based on that analysis and used as a model in the first annotation campaign of selected English hate speech and offensive language corpora. The inter-annotator agreement between annotators and the post-task questionnaire answered by the annotators showed consistent results for the category items of *offensive*, *insulting* and *abusive*, although a selection of the other categories met with a more diverse tagset selection in the annotation. It suggests that one of the main reasons for the low inter-annotator agreement for these items was the difficulty that annotators had in distinguishing between them during the annotating task. It also indicates a need to test an option of reducing the granularity of the taxonomic types (as e.g., in the Simplified Offensive Language (SOL) taxonomy model, proposed in Lewandowska-Tomaszczyk 2022), predominantly of the categories and aspects diagnosed as the most similar ones in the post-task questionnaires, however, retaining the sufficient taxonomic granularity for those distinctions which appear criterial for the inter-categorial differentiation.

Acknoweldgement

This study was performed within the COST Action CA 18209 *European network for Web-centred linguistic data science* (Nexus Linguarum).

References

Bączkowska, Anna; Lewandowska-Tomaszczyk, Barbara; Valunaite Oleškevičiene, Giedre; Žitnik, Slavko; Liebeskind, Chaya. 2022. *A taxonomy of implicit language*. Nexus Workshops days in Jerusalem. Jerusalem, May 23–24, 2022.

Chung, Yi-Ling: Kuzmenko, Elizaveta; Tekiroglu, Serra Sinem; Guerini, Marco. 2019. CONAN – COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. Florence. 2819–2829.

Culpeper, Jonathan. 2005. Impoliteness and the Weakest Link. *Journal of Politeness Research* 1/I. 35–72.

Despot, Kristina; Ostroški Anić, Ana. 2022. *Reflections on Implicitness*. Nexus Workshops days in Jerusalem. Jerusalem, May 23–24, 2022.

Haugh, Michale; Sinkeviciute, Valeria. 2019. Offence and conflict talk. *Routledge Handbook of Language in Conflict*. Eds. Evans, Matthew; Jeffries, Lesley; O'Driscoll, Jim. Routledge. London. 196–214.

Klie, Jan-Christoph; Eckart de Castilho, Richard; Gurevych, Iryna. 2020. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains, *ACL 2020,* 5. July - 10. July 2020.

Lakoff, George. 1987. Cognitive models and prototype theory. *Concepts and conceptual development: Ecological and intellectual factors in categorization.* Ed. Neisser, Ulric. Cambridge University Press. Cambridge. 63–100.

Lewandowska-Tomaszczyk, Barbara. 2022. A simplified offensive language (SOL) taxonomy for computational applications. *Konin Language Studies* 10/2. 213–227.

Lewandowska-Tomaszczyk, Barbara; Žitnik, Slavko; Bączkowska, Anna; Liebeskind, Chaya; Mitrović, Jelena; Valunaite Oleškevičiene, Giedre. 2021. LOD-connected offensive language ontology and tagset enrichment. *Proceedings of the workshops and tutorials – SALLD-3, co-located with the 3rd Language, Data and Knowledge Conference.* CEUR Workshop Proceedings. Eds. Carvalho, Sara; Rocha Souza, Renato. 135–150.

Lewandowska-Tomaszczyk, Barbara; Liebeskind, Chaya; Žitnik, Slavko; Bączkowska, Anna; Valunaite Oleškevičiene, Giedre; Trojszczak, Marcin. 2022. *An offensive language taxonomy and a webcorpus discourse analysis for automatic offensive language identification.* Presentation at 3rd International Conference: *Approaches to Digital Discourse Analysis (ADDA 3)*. St Petersburg, Florida, May 13–15, 2022.

Lewandowska-Tomaszczyk, Barbara; Bączkowska, Anna; Liebeskind, Chaya; Valunaite Oleskeviciene, Giedre; Žitnik, Slavko. (submitted). An integrated explicit and implicit offensive language taxonomy.

Modha, Sandip; Mandl, Thomas: Majumder, Prasenjit; Patel, Daksh. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. *Proceedings of the 11th Forum for Information Retrieval Evaluation*. Eds. Majumder, Prasenjit; Mitra, Mandar; Gangopadhyay, Surupendu. Association for Computing Machinery. New York. 14–17.

Mowery, Danielle L.; Jordan, Pamela; Wiebe, Janyce; Harkema, Henk; Dowling, John and Chapman, Wendy W. 2013. Semantic Annotation of Clinical Events for Generating a Problem List. *AMIA Annual Symposium Proceedings.* 1032–1041.

Ousidhoum, Nedjma; Lin, Zizheng; Zhang, Hongming; Song, Yangqiu; Yeung, Dit-Yan. 2019. Multilingual and multiaspect hate speech analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. Hong Kong. 4675–4684.

Sekine, Satoshi; Ranchhod, Elisabete (eds). 2013. *Named Entities: Recognition, classification and use.* [Benjamins Current Topics 19]. Benjamins. Amsterdam.

Smith, Mark; Ceni, Arber; Milic-Frayling, Natasa; Shneiderman, Ben; Mendes Rodrigues, Eduarda. 2010. NodeXL: A free and open network overview, discovery and exploration add-in for Excel 2007/2010/2013/2016, from the Social Media Research Foundation. http://www.smrfoundation.org (Accessed: December 1, 2022).

Tharindu, Ranasinghe; Zampieri, Marcos. 2020. Multilingual offensive language identification with cross-lingual embeddings. arXiv. Preprint. arXiv:2010.05324.

Wiegand, Michael; Ruppenhofer, Josef; Eder, Elisabeth. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Eds. Toutanova, Kristina; Rumshisky, Anna; Zettlemoyer, Luke; Hakkani-Tur, Dilek; Beltagy, Iz; Bethard, Steven; Cotterell, Ryan; Chakraborty, Tanmoy; Zhou, Yichao. Association for Computational Linguistics. 576–587.

Yin, Xuaixin; Shah, Sarthak. 2010. Building taxonomy for Web-searched intents for Name Entity queries.

https://www.researchgate.net/profile/XiaoxinYin/publication/221023101_Building_taxonomy_of_web_search_intents_for_name_entity_queries/links/54341fd80cf294006f734daa/Building-taxonomy-of-web-search-intents-for-name-entity-queries.pdf (Accessed December 1, 2022).

Zadeh, Lofti A. 1964. Fuzzy sets. *Information and control*, Vol. 8. 338–353.

Zampieri, Marcos; Malmasi, Shervin; Nakov, Preslav; Rosenthal, Sara; Farra, Noura; Kumar, Ritesh. 2019. SemEval- 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation*. Eds. May, Jonathan; Shutova, Ekaterina; Herbelot, Aurelie; Zhu, Xiaodan; Apidianaki, Marianna; Mohammad, Saif M. Association for Computational Linguistics. Minneapolis. 75–86.

Zampieri, Marcos; Nakov, Preslav; Rosenthal, Sara; Atanasova, Pepa: Karadzhov, Georgi; Mubarak, Hamdy; Derczynski, Leon; Pitenis, Zeses; Çağri Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (Offenseval 2020). arXiv:2006.07235 [cs.CL]

Anotacijska shema i njezina evaluacija: primjer uvredljivoga jezika

Sažetak

U ovome je radu predstavljen proces označavanja uvredljivoga jezika koji uključuje izradu klasifikacije toga jezika, označivačku praksu, vođenje procesa i evaluaciju. Klasifikacijska je shema prvi put predložena u Lewandowska-Tomaszczyk i dr. (2021). Proširena ontologija uvredljivoga jezika sadrži 17 kategorija posloženih u četiri hijerarhijske razine te tako predstavlja shemu uvredljivoga jezika koja je trenirana u okviru nekontekstualiziranih vektorskih prikaza riječi (engl. *word embeddings*) poput Word2Vec i Fast Text koji su naposljetku supostavljeni podatcima prikupljenima korištenjem analize parova i analize testiranja za postojeće kategorije u modelu HateBERT (Lewandowska-Tomaszczyk i dr., u postupku recenzije). U radu se izvještava o označivačkoj praksi u okviru radne grupe WG 4.1.1. *Incivility in media and social media* COST-ove akcije *CA 18209 European network for Web-centred linguistic data science* (*Nexus Linguarum*). Označavanje je provedeno u alatu INCEpTION (https://github.com/inception-project/inception) – platformi za semantičko označavanje koja ima ugrađene alate za takvu obradu podataka. Dobiveni rezultati podupiru predloženu ontologiju eksplicitnoga i implicitnoga uvredljivog jezika koja omogućuje veću raznovrsnost među već prepoznatim tipovima figurativnoga jezika (primjerice metafora, metonimija, ironija itd.). Upotreba sustava za označavanje i prikazivanje jezičnih podataka također je procijenjena u povratnim komentarima koje su pružili označivači. Komentari označivača prikupljeni su metodom upitnika te otvorenom raspravom. Na kraju je ususstavljen niz preporuka za buduće označivačke prakse.

Keywords: annotation, annotators, explicit, implicit, offensive language, word embeddings
Ključne riječi: označivanje, anotatori, eksplicitan, implicitan, uvredljivi jezik, vektorski prikaz riječi