KSJ ANS

Konin Language Studies
Faculty of Humanities and Social Sciences, University of Applied Sciences in Konin, Poland
KSJ 10 (3). 2022. 213-227
http://ksj.konin.edu.pl
*doi: 10.30438/ksj.2022.10.3.1*

# *A simplified taxonomy of offensive language (SOL) for computational applications*

Barbara Lewandowska-Tomaszczyk ✉
University of Applied Sciences in Konin, Poland
https://orcid.org/0000-0002-6836-3321
*barbara.lewandowska-tomaszczyk@konin.edu.pl*

Abstract

The focus of the paper is a discussion of problems connected with the analysis and semantic categorization of concepts reflecting meanings of types of linguistic offense in English. The first part of the paper is related to general definitional and cross-boundary categorization problems of classes of words using the example of English offensive lexicon with the identification of main approaches to it and solutions proposed in linguistic and computational literature. The issue of computational identification of classes of offensive language is particularly emphasized in the further sections with the discussion of extended models of offensive language. Insights from relevant lexical embeddings and further annotation exercise lead in conclusion to positing a simplified semantic taxonomy of offensive language (SOL) in the final section of the paper.

*Keywords:* annotation; computational applications; conceptual fuzziness; conceptual indeterminacy; offensive language; simplified taxonomy

## 1. Introduction

The main aim of the paper is to discuss some categorization problems connected with semantic categorization generally and with a taxonomy of offensive language in particular. The data are acquired from corpora of English hate speech and their categorization models are discussed. They are based on rich linguistic analysis

rooted in current linguistic literature and scrutinized by means of the Sketch Engine collocational, thesaurus and synonymic sets revolving around the offensive terms in question. The need to introduce some simplification in terms of a semantic upper-dimension taxonomy is undertaken as a result of computational verification and an annotation campaign (Lewandowska-Tomaszczyk et al., submitted b), which was conducted in the winter of 2021 and spring 2022 by members of Working Group 4.1.1. *Incivility in media and social media* at the COST Action Nexus Linguarum.

## 2. Conceptual indeterminacy and fuzziness in offensive language categorization

Problems with the definition and categorization of linguistic meaning have been discussed in the later 20th century linguistic literature at least since Ludwig Wittgenstein's (1953) presentation of semantic and definitional questions in linguistic semantics. With his discussion of the concept of *game* and attempts of its necessary and sufficient set of features identification, he persuasively presented the problems with a common set of features for its definition, which would cover such diverse types of games as, for example, board games, card games, ball games, etc. Wittgenstein's (1953) proposal of the inter-categorial concept of *family resemblance* adequately captures the ontological relations within one category. Since then researchers of different background have been developing ideas concerning the leaking lexical category boundaries. Mathematician and philosopher Lotfi Zadeh (1965) developed his approach to what he called *fuzzy sets* in his proposal to solve the categorial boundary problems. Psychologist Elinor Rosch's seminal papers on categorization and prototypes, Charles Fillmore's (1977) concept of *frame* and George Lakoff's (1987) influential study of categorization, radial concepts, etc., and their reflection in linguistic meanings, gave an impetus to thousands of studies of these phenomena in diverse languages. On the one hand, the research confirmed the problems with the identification of strict category boundaries and questions of sets of necessary and sufficient conditions to define lexical meanings and, on the other, it uncovered at the same time the weaknesses in positing rigid hyper-/hyponymic semantic categorization structures to model lexical semantics. Although it is not needed here to discuss particular cognitive linguistic cases reflecting the problems and their attempted solutions, it is necessary to mention this segment of linguistic research that has its repercussions for computational modelling of linguistic meanings. In the present case, the taxonomically problematic case of the category of *offense* will be briefly discussed.

## 3. Languages of offense

*Offense* is defined in pragmatic literature as intentional face-attacks, typically accompanied by offense perception by the hearer/addressee (Culpeper, 2005; Haugh & Sinkeviciute, 2019). Behavioral offense is usually accompanied by the use of derogatory language, addressed at an individual or a group target. Occasionally, offense is used to address an individual through some discriminating group stereotypes, on grounds of religion, abilities, gender, etc. This type of verbal offense is defined as one type of offensive language, that is, *hate speech*. A more complete analysis of hate speech by Victoria Guillén-Nieto (2023) within wider linguistic and legal ramifications is forthcoming.

The present study discusses an attempt to build a taxonomy of the categories of offensive language using the example of English, as one of best, if not the best, studied language at present to be applied to a computational application of offensive language identification in social media for English, and through its more universal schema, to other language systems. The taxonomies which have been proposed by researchers typically aim at the universal status and are frequently tested on large unannotated language corpora in various languages. The discussed variants of our offensive language taxonomy have precisely a similar objective – to be successfully used in a number of languages for offensive language identification in untrained, non-annotated texts.

## 4. Computational approaches to offensive language

Exploration of the language of offense by means of linguistic and computational linguistic methods has been conducted by numerous researchers. Such attempts have frequently been targeted towards both offensive language analysis and recognition as well as offensive language identification in untrained textual data.

The development of systems for automatic language identification has been progressing by reference to current linguistic theorizing and computational capacity. It includes identification of offensive language *feature-based* linear classifiers (Waseem & Hovy, 2016), via *corpus-related* neural network architectures (Birkeneder et al., 2018; Mitrović et al., 2019), and, at present, the development of fine-tuned *pre-trained language models* as, for example, BERT and RoBERTA system (e.g., Liu et al., 2019; Swamy et al., 2019). Results in the first linear classifier architectures proved quite efficient, although it is the pre-trained language models such as HateBERT (Caselli et al., 2020) that reach the best performance.

## 5. Cognitive-pragmatic offensive language taxonomy

First attempts at building taxonomies of languages of offense, also called variably *abusive*, *extremist*, *aggressive*, *radical,* etc. language, have been based on simple identification of three categories: offensive, non-offensive, or neutral, or their naming variants. Although such categorization has been successfully used, for example, for the elimination of some coarse offensive content, especially in social media, it may not be sufficient to identify the gravity of the language inadequacy for particular contexts.

The initial offensive language taxonomy work presented in Lewandowska-Tomaszczyk et al. (2021, submitted a) proposes a model of offensive language taxonomy along cognitive-pragmatic lines and verified by computational methodology. The first annotation campaign brought tangible results, which allowed its simplification in the shape to be discussed in the next section of the present paper.

An earlier more linguistically profiled taxonomy was proposed in Lewandowska-Tomaszczyk at al. (2021), developed later into an extended model in Lewandowska-Tomaszczyk et al. (Lewandowska-Tomaszczyk et al., submitted a). The model uses Zampieri et al's (2019a, 2019b) idea with more than one level of offensive language categorization. The concept of offensive language is used in our models as a superordinate category, with 17 subcategories, arranged into four levels. This complex semantic ontology is then applied to a corpus built of 25 publicly available web-based hate speech datasets, pre-tagged by their compilers according to their respective tagging schemas.[1] The results of our proposed taxonomy were verified by resorting to non-contextual, neural-based word embedding tools (i.e., Word2Vec, fastText, Glove). Together with the pre-trained transformer models such as HateBERT they presented a more advanced computational linguistics method, assessing the particular categories cosine distance and degree of their semantic similarity. These measures are dimensions needed in determining inter-categorial closeness and distance. Figure 1 presents an expanded offensive language taxonomic model (Lewandowska-Tomaszczyk et al., submitted a), extended in comparison with the initial proposal put forward in Lewandowska-Tomaszczyk et al. (2021).

[1] 25 selected hate speech English datasets are listed and referred to with regard to their accessibility in Lewandowska-Tomaszczyk et al. (submitted b).

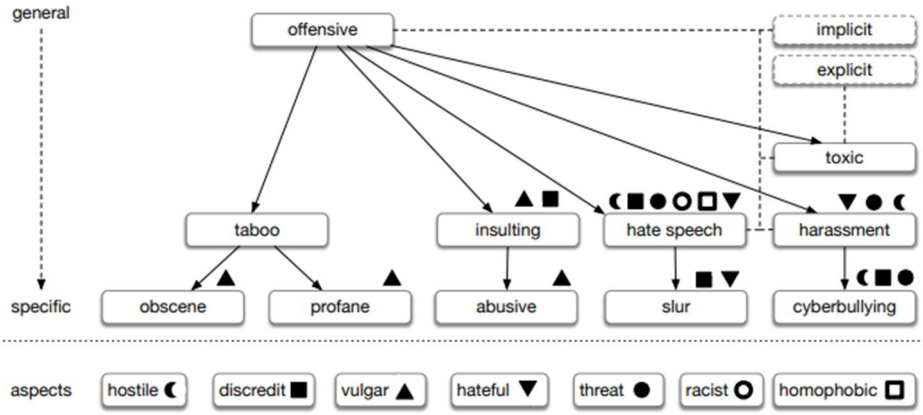


Figure 1 Offensive language taxonomy (Lewandowska-Tomaszczyk et al., submitted a)

The pragmatically oriented ontology of offensive language, visualized in Figure 1 was established by considering results of thesaurus, synonymity and collocational profiles for each of the taxonomic keywords in large English databases on the corpus-management platform Sketch Engine (Lewandowska-Tomaszczyk et al., submitted a). As mentioned before, it was originally inspired by the three-level hierarchy of offensive language put forward by Zampieri et al. (2019a, 2019b). However, unlike Zampieri et al.'s models, offensive language in our taxonomy was further refined and divided into two basic levels of analysis (Level I and Level II), and four sublevels (A, B, C, D) within Level I. Level I distinguishes lexical items that are offensive from those that are not (Level A: offensive vs. non-offensive). Secondly, in Level B (targeted vs. non-targeted), the question whether the selected items are targeted at some addressee should be answered. If there is no identifiable addressee then the use of offensive language is an example of self-expression, which has an exclamatory function, such as, for example, the use of swear words to express anger, frustration, pain etc., the sample is considered non-offensive. Targeted offensive items are further divided into either *implicit* or *explicit* cases of offensive language at Level C. While implicitness may be encoded by, for example, hyperbole and irony, whereby offense is veiled, explicitness entails more linguistically straightforward forms of verbal attack. Classes of explicit targeted categories of *offence* are further subcategorized into types characterized by varying kinds of internal or external targets as well as partly distinct characterization of the lexicon. Some of the identified classes overlap between social media such as online space use and outside world (offline) linguistic and behavioral content, such as, for example, *cyberbullying,* in fact falling outside of our instruments in current linguistic analysis. Some additional insight, in particular concerning offensive language meaning distances has been gained comparing Sketch Engine results and relevant word embeddings calculations.

The models proposed in Lewandowska-Tomaszczyk et al. (2021) and Lewandowska-Tomaszczyk et al. (submitted a) are hierarchical ones. However, although showing the basic features of offensive language and targeted at an individual or a group, the hyperonym does not always include all the instances lower in the hierarchy, the phenomenon referred to above in the first section, very well-known in linguistic analysis and typical of most language categorization schemata. This leads to problems with the uniformity of the annotation results, which will be mentioned in the section to follow. The schema with 4-tier levels in Figure 1, also visualizes hierarchical steps, reflecting the annotation levels planned as the next analytic and verification tool.

The concept of *offense* is the main category, immediately dominating either linguistically explicit or else linguistically implicit language types to be further analyzed into variable figurative language categories, irony, etc. In the social media texts, which are the basis of our investigation, linguistically explicit expressiveness was additionally marked by capitalization, punctuation or visual symbols, not considered in the presented taxonomy. The two basic levels: Level 2 – taboo, insulting, hate speech and the lowest Level 3 – *aspects* – are the categories which can be identified not by sets of necessary and sufficient properties but rather by some characteristic, or typical features, of the Wittgensteinian family resemblance character. The categorial labels – keywords in our system – were scrutinized by means of the Sketch Engine instruments, more precisely by their particular thesaurus, synonymity and collocational profiles (see Figure 2 for examples), which were the basis of our offense intensity and specificity judgments. A comparison of the scope of reference and the offensive intensity judgment of the meanings of offensive and insulting in Figure 2, evaluated by respective lists of the surfacing forms, makes it possible to identify a larger, more extensive range of reference, and a more distributed offensive force of the former.

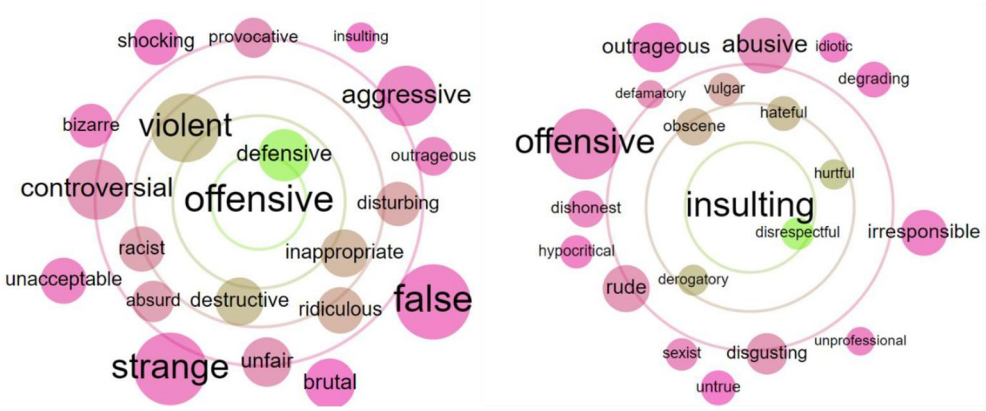


Figure 2 Thesaurus synonymy networks of Eng. *offensive* and *insulting* (Lewandowska-Tomaszczyk, 2022)

Special types of offense in terms of *toxicity*, *harassment* and *(cyber-)bullying*, identified in the right part of Figure 1, are considered distinct from the others as their manifestation in the corpus data would require identification of their more persistent character of use, possibly in a number of consecutive posts, not considered in that study.

The extended model was verified by three word embedding instruments: Word2Vec, fastText, and Glove. A vector space model for word embeddings was first proposed by Mikolov et al. (2013). It shows words with related vectors according to their sharing a similar context in a corpus. In other words, word embeddings present the results of clustering of similar words in similar contexts. One of the word embeddings performed by us for the proposed offensive language keywords (Word2Vec [t-SNE]) is illustrated in Figure 3.

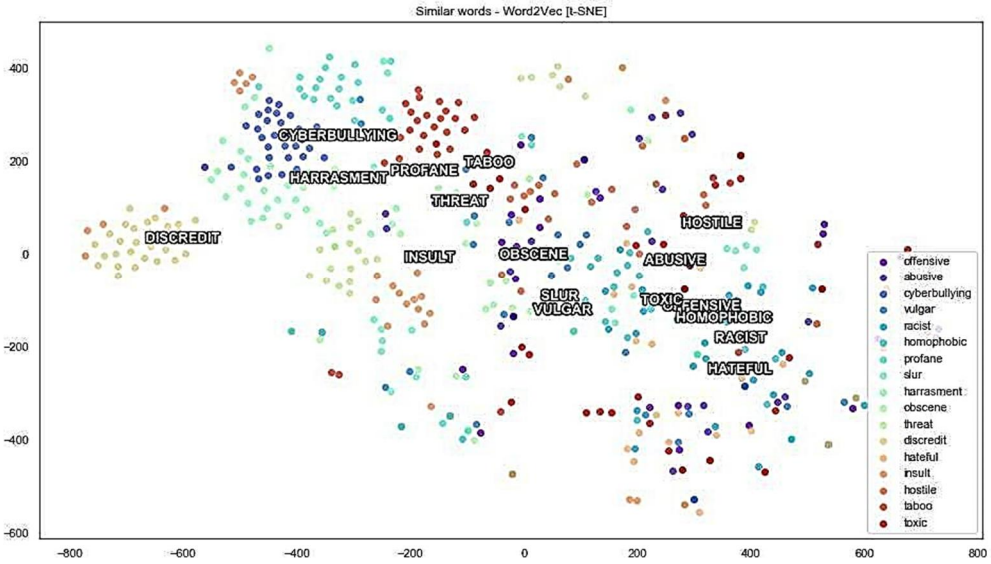


Figure 3 Word2Vec top 30 neighboring vectors visualization using t-SNE. The categories used in the Extended Model are presented in their semantic similarity clusters (see Lewandowska-Tomaszczyk, submitted a)

## 6. Extended taxonomy verification: The first annotation campaign

Following the proposal concerning the extended taxonomy of offensive language (Lewandowska-Tomaszczyk et al., submitted a), the first annotation campaign was carried out on a semantic annotation INCEpTION platform (https://inception-project.github.io/). The campaign conducted by members of WG 4.1.1. of Nexus Linguarum in early 2022 brought about important results (Lewandowska-Tomaszczyk et al., submitted b). Some of the categories were well recognized

by both annotators and confirmed by a curator, whereas some other distinctions brought about lower inter-annotator agreement results.

A particularly problematic point was presented by the dichotomy between explicit and implicit offensive language, or *insult* versus *abuse*, as this distinction proved to be difficult for some of the annotators to judge. Some other categories too showed relatively strong interconnections, reflecting their rather high semantic similarity (in the range of 4-5 out of the 10-point scale, with 10 points representing the maximum similarity value) and problems for category distinction in the annotation, such as, for example, between *hostile*, *threat* and *hateful*. This contrasts with the lower co-occurrences between these and the other aspect items, namely *racist*, *homophobic*, *vulgar* and *discredit*, with the significantly lower similarity values (in the range between 2.1-3.2), thus causing less problem in the annotation campaign.

The primary reasons for the sense discrimination problems – as argued in the Lewandowska et al.'s (submitted b) paper – is the non-crisp character of the category keywords, hard or impossible to define in terms of exceptionless sets of necessary and sufficient conditions. Additionally, some external reasons like the lack of incentives to support the annotators' work led to lower engagement on their part. Moreover, the problems involved in recruiting native English annotators and possibly insufficient intensity of the annotation training sessions might have also played some role.

The simplified offensive language (SOL) taxonomy I propose here in the following section of the present paper is meant to de-complexify the annotation system, particularly for computational purposes of efficient offensive language identification in naturally occurring data. For example, the distinction between *explicit* vs. *implicit* in the *expressiveness* tag, the categorial differentiation between *insult* and *abuse*, or the clusters of *toxic*, *abusive*, *hostile* were not convincingly supported in the embedding results (Figure 3). The re-consideration of the tags in the aspects compartment, as performed by the annotators with a varying inter-rater success, led to the keyword repertory modifications and some of the aspects distinctions were dispensed with. The simplified taxonomy contains fewer offensive language categories than our previous models but, as a whole, this approach is predicted to be a more effectively discriminating offensive category system for computational applications than the models proposed before.

## 7. A simplified taxonomy of offensive language

Taking into consideration the linguistic and computational limitations discussed in the sections above, what is proposed in this section is the simplification of the Extended Model both as far as the number as well as types of the key categories

are concerned. The simplified offensive language taxonomy is presented in terms of a step-by-step hierarchical procedure presented in Figure 4 below. This taxonomy prepares the ground for the second annotation campaign, which aims to include also languages other than English (e.g., Hebrew, Polish and Lithuanian).

SIMPLIFIED OFFENSIVE LANGUAGE TAXONOMY

1. OFFENSIVE [YES or NO]
2. Target 1
   Individual // Group // Ind wrt Gr/Gr wrt Ind [by reference to group stereotypes]
3. Target 2
   present//absent
4. Vulgar [YES or NO]
5. Choose either (i) or (ii); Then select (iii) or (iv) or both (iii) and (iv)
   (i) INSULT [addressed to: individual or group – varied offense types but not by group stereotypes]
   (ii) HATE SPEECH [individual or group; offense by reference to group stereotypes]
   (iii) DISCREDIT [individual or group//on various grounds – lying-cheating, immorality, unprofessionalism, unfairness]
   (iv) THREAT [individual or group, inducing fear]
6. Aspects – [Choose one or more]
   *[racist] [xenophobic] [homophobic] [sexist] [profane (religion)] [ageism] [physical/mental disabilities [ableism]] [social class [classism]] [ideologism] [other]*
7. Select categories below – [Choose one or more]
   RHETORICAL QUESTIONS
   METAPHOR
   SIMILE
   IRONY
   EXAGGERATION
   OTHER

Figure 4 Step-by-step hierarchical procedure of the simplified offensive language taxonomy

Starting from the highest ontological level, the question concerning the overall offensiveness status of the selected sample is crucial to establish. The answer *yes* or *no*, similarly to the other *yes-no* categorization proposed in the categories below, is not a reflection of actual conceptual-linguistic reality but, rather, indicates annotation requirements adjusted to a computer program to distinguish between dichotomic judgments (e.g., *offensive vulgar* or *not*). Furthermore, the first decision as to the offensive status of the examined sample affects all the following choices, keeping in the area of our interest only those instances which are judged as offensive by the annotator. Thus, a non-offensive vulgarism used not to offend but employed,

for example, for unaddressed emotion expression, will be considered non-targeted and hence will not be further categorized in the present taxonomy.

The next levels – 2 and 3 – refer to individuals or group, that is, to the *targets* of an offensive act. Target 1 tag distinguishes among the target which denotes an individual (*target a*), a group (*target b*), or else a *target c*, addressed at a group through a particular individual or else an individual meant to be a group representative. The main criterial property of the latter (*target c*) is the use of a gender, race, etc., *stereotype* in the offensive language sample and paves the way to the category of *hate speech* as one of the offense types in the hierarchy. *Target* 2 is a tag which represents a circumstantial property of presence or absence of the offensive language target at the *locus* of the interactional encounter.

The next selection falls between *vulgar* and *non-vulgar language* (i.e., words, phrases). Realizing the problems here connected with the varying senses of vulgarisms among annotators, let alone, in general language contexts, the first-level selection between *offensive* or *non-offensive* type, is instrumental in further judgements of the vulgarity of a particular sample provided in a larger linguistic context, as in the *Yep usual bullshit* response (Lewandowska-Tomaszczyk, 2017).

The lower distinctions excerpted from Figure 3 are definitional with respect to the character of the used offense. I propose a category of *insult* to determine an individual or group offense, *not* by reference to any *group stereotypes* (e.g., *its the state of your own mental health you should be VERY concerned about presents*, identified as an INSULT with the Aspect of ablism[2]), as juxtaposed to the concept of *hate speech*, whose discriminating property is precisely the reference to a group or individual *via* discriminatory group stereotypes.

The *discredit* tag signifies an offensive act addressed at an individual or a group on grounds of accusation of lie, immorality, unprofessionalism, and unfairness. The category of *threat* presents a special type. It is considered in this model as long as it is identified as *offensive* on the first categorization level, as accompanying any of the previous three labels on the same level. In the domain of law, threat is a statement intended to frighten or intimidate a person or a group into believing in prospective harm they will experience (Brenner, 2002). In our datasets *threat* can originate either from the offender or else be part of the commentator's *warning* against a third party not necessarily present in the interaction, or against an event, either offensive or not. Threats can contain language elements which accompany the subclasses of *aspects*, as well as *hate speech* or *insult*. The level of *aspects* as defined in this model is considered a type of *a discriminating act*, and can be addressed at the offendee's race, gender, age, religion, ethnicity, ability, etc., or else a combination of these (see Figure 5).

[2] The example quoted from Lewandowska-Tomaszczyk's (2017) Internet radical language dataset.

Choose either (i) or (ii); Then select (iii) or (iv) or both (iii) and (iv)
- (v) INSULT [addressed to: individual or group – varied offense types but not by group stereotypes]
- (vi) HATE SPEECH [individual or group; offense by reference to group stereotypes]
- (vii) DISCREDIT [individual or group//on various grounds – lying-cheating, immorality, unprofessionalism, unfairness]
- (viii) THREAT [individual or group, inducing fear]

Aspects – [Choose one or more]
*[racist] [xenophobic] [homophobic] [sexist] [profane (religion)] [ageism] [physical/mental disabilities [ableism]] [social class [classism]] [ideologism] [other]*

Figure 5 Lower-level taxonomic types

The last of the categorial distinctions refers to a crucial differentiation between linguistically explicit versus implicit types of utterances. Although this distinction is a basic differentiation in the languages of offense, bearing in mind the problems with its definitions the annotators experienced in the first annotation campaign (Lewandowska-Tomaszczyk et al., submitted b), it is not labelled as such in the simplified taxonomy. Instead, we propose a selection of one or more of the following linguistically implicit (in some cases *indirect*) categories (cf. Bączkowska et al., 2022; see Figure 6):

Select categories below – [Choose one or more]
RHETORICAL QUESTIONS
METAPHOR
SIMILE
IRONY
EXAGGERATION
OTHER

Figure 6 Implicit categories of offence

## 8. Conclusions

Although the offensive language taxonomic model (Lewandowska-Tomaszczyk et al., submitted a), based on the ontology proposed in Lewandowska-Tomaszczyk et al. (2021), used a set of scrutinized linguistic criteria for the offensive category identification, their conceptual definitions cannot be considered to contain all necessary and sufficient properties needed for their exception-free category detection in discourse. The categories remain fuzzy by their linguistic nature constraints. Nevertheless, they are coherent and combined into one larger, though complex, conceptual category system by sharing the feature of offensiveness and, consecutively, going further down the hierarchy by sharing at least one, if not more, properties between the higher and lower ranks.

The SOL taxonomy proposed in this paper can be considered a possible workable solution to the offensive language detection problem not only for English but possibly for a number of other languages. This certainly would not imply a full-scale inter-annotator agreement. The nature of linguistic categories, and these are no exception, is their non-crisp character and fuzzy boundaries. However, the extent of this inherent fuzziness should be reduced, while being directly related to the annotation success rate, and it will be more precisely determined after the results of the next annotation cycle are known.

The SOL taxonomy is proposed as the next annotation model for the 2nd Offensive Language Annotation Campaign in WG 4.1.1. (2023) to be tested on Polish and some other languages.

The foundation of the taxonomy is its complementarity to general use of offensive language ontologies and tagset systems, while the ultimate aim is its integration with the public In our taxonomy, the Linguistic Linked Open Data (LLOD) resources.

Acknowledgment

The study was conducted in the context of research carried within COST Action CA 18209 NEXUS LINGUARUM *European network for Web-centred linguistic data science*.

References

Bączkowska, A., Lewandowska-Tomaszczyk, B., Žitnik, S., Liebeskind, Ch.,Valunaite Oleskeviciene, G., & Trojszczak, M. (2022). *Implicit offensive language taxonomy and its application for automatic extraction and ontology*. Presentation at *LLOD Approaches to language data research and management*, Vilnius, 21-22 September 2022, Lithuania.

Birkeneder, B., Mitrović, J., Niemeier, J., Teubert, L. & Handschuh S. (2018). Offensive language detection in German tweets. In M. Ruppenhofer, M. Siegel, & M. Wiegand (Eds.), *Proceedings of the GermEval 2018 workshop*, (pp. 71-78). Austrian Academy of Sciences.

Brenne, J. L. (2002). True threats: A more appropriate standard for analyzing FirstAmendment protection and free speech when violence is perpetrated over the Internet. *North Dakota Law Review,* 78(4), 753-784.

Caselli T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th international conference on language resources and evaluation (LREC 2020)* (pp. 6193-6202). European Language Resources Association.

Culpeper, J. (2005). Impoliteness and entertainment in the television quiz show: The weakest link. *Journal of Politeness Research, 1*(1), 35-72.

Fillmore, C. J. (1977). The case for case reopened. In P. Cole & J. Saddock (Eds.), *Grammatical relations* (pp. 59-81). Academic Press.

Guillén-Nieto, V. (forthcoming, 2023). *Hate speech – Linguistic perspectives.* De Gruyter Mouton.

Haugh, M., & Sinkeviciute, V. (2019). Offence and conflict talk. In E. Mathew (Ed.), *The Routledge handbook of language in conflict* (pp. 196-214). Routledge.

Lakoff, G. (1987). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 63-100). Cambridge University Press.

Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrović, J., & Valunaite Oleskeviciene, G. (2021). Lod-connected offensive language ontology and tagset enrichment. In R. Carvalho & R. Rocha Souza, R. (Eds.), *Proceedings of the workshops and tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference* (pp. 135-150). CEUR Workshop Proceedings.

Lewandowska-Tomaszczyk, B. (2022). Meaning – significado – znaczenie. In J. R. Aixelá & M. Muñoz Martín (Eds.) *ENTI Encyclopedia of translation and*

*interpreting* (online edition). Iberian Association for Translation and Interpreting Studies. www. https://www.aieti.eu/en/encyclopaedia/home

Lewandowska-Tomaszczyk, B. (2017). Conflict radicalization and emotions in English and Polish online discourses on immigration and refugees. In S. Croucher, B. Lewandowska-Tomaszczyk, & P. A. Wilson (Eds.), *Conflict, mediated message and group dynamics: intersections of communication* (pp. 1-24). Rowman & Littlefield.

Lewandowska-Tomaszczyk B., Bączkowska, A., Liebeskind, Ch., Valunaite Oleskeviciene, G., & Žitnik, S. (submitted a). An integrated explicit and implicit offensive language taxonomy.

Lewandowska-Tomaszczyk, B., Žitnik, S., Liebeskind, Ch., Valunaite Oleskevicienė, G., Bączkowska, A., Wilson, P. A., Trojszczak, M., Brač, I., Filipić, L., Ostroški Anić, A., Dontcheva-Navratilova, O., Borowiak, A., Despot, K., & Mitrović, J. (submitted b) Annotation scheme and evaluation: The case of OFFENSIVE language.

Liu, P., Li, W., & Zou, L. (2019). nlpUP at SemEval-2019 Task 6: Transfer learning for offensive language detection using bidirectional transformers. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th international workshop on semantic evaluation* (pp. 87-91). Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, K., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* 26 (NIPS 2013) (pp. 3111-3119). Association for Computing Machinery (ACM) Digital Library.

Mitrović, J., Birkeneder, B., & Granitzer, M. (2019). nlpUP at SemEval-2019 Task 6: A deep neural language model for offensive language detection. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, & S. M. Mohammad (Eds.), *Proceedings of the 13th international workshop on semantic evaluation (SemEval)* (pp. 722-726). Association for Computational Linguistics.

Rosch, E. (1973). Natural categories, *Cognitive Psychology, 4,* 328-350.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019). Studying generalizability across abusive language detection datasets. In M. Bansal & A. Villavicencio (Eds.), *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 940-950). Association for Computational Linguistics.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictivefeatures for hate speech detection on Twitter. In J. Andreas, E. Choi, & A. Lazaridou (Eds.), *Proceedings of the NAACL Student Research Workshop* (pp.88-93). Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophical investigations.* Macmillan.

Zadeh, L. (1964). Fuzzy sets. *Information and Control, 8*(3), 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X. ISSN 0019-9958.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. In J. Burstein, Ch. Doran, & Th. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (Vol 1, pp. 1415-1420). Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.

Datasets and tools

25 English hate speech corpora (for the itemized list cf. Lewandowska-Tomaszczyk et al. 2021)

Annotation INCEpTION platform https://inception-project.github.io/

Sketch Engine webcorpus of English https://www.sketchengine.eu/ententen-english-corpus/