► Annotation Scheme and Evaluation: The Case of OFFENSIVE Language

Barbara Lewandowska-Tomaszczyk,
State University of Applied Sciences in Konin, Poland
Slavko Žitnik,
University of Ljubljana, Slovenia
Chaya Liebeskind,
Jerusalem Institute of Technology, Israel
Giedre Valūnaitė-Oleškevičienė,
Mykolas Romeris University, Vilnius, Lithuania
Anna Bączkowska,
University of Gdansk, Poland
Paul A. Wilson,
University of Lodz, Poland
Marcin Trojszczak,
State University of Applied Sciences in Konin, University of Bialystok, Poland,
Ivana Brač,
Lobel Filipić,
Ana Ostroški Anić,
Institute of Croatian Language and Linguistics, Zagreb, Croatia
Olga Dontcheva-Navratilova,
Masaryk University, Brno, Czech Republic
Agnieszka Borowiak,
State University of Applied Sciences in Konin, Poland
Kristina Despot,
Institute of Croatian Language and Linguistics, Croatia
Jelena Mitrović,
University of Passau, Germany; Institute for AI R&D of Serbia

Purpose: 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 & Sinkeviciute, 2019) and in NLP (e.g., OffensEval (Zampieri et al., 2020), HASOC (Mandl et al., 2019)), under the names of *hate speech*, *abusive language*, *offensive language*, etc. The paper focuses on the presentation and discussion of aspects of the linguistic annotation of OFFENSIVE LANGUAGE, including creation, annotation practice, curation, and evaluation of an OFFENSIVE LANGUAGE annotation taxonomy scheme first proposed in Lewandowska-Tomaszczyk et al. (2021) and Žitnik et al. (in press). An extended offensive language ontology in terms of 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 pairwise training and testing analysis for existing categories in the HateBERT model.

Approach: The system is used for 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 with annotation. The current authors are the taxonomy proposers, annotators, and curators of the annotation practice. 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.

Findings: The results support the proposed ontology of explicit offense and positive implicitness types and a preliminary typology of more refined offensive implicitness categorization criteria to provide more variance among widely recognized types of figurative (metaphorical, metonymic, ironic, etc.) and other languages. The use of the annotation system and the representation of linguistic data will be evaluated in a series of annotators' comments, by means of a questionnaire method and in an open discussion.

Value: The results will be presented, and further developments in the annotation taxonomy creation and practice will be included in a **recommendation package** to be considered in new proposals, i.e., an implicit offense annotation system (Bączkowska et al., 2022; Despot & Ostroški Anić, 2022), and its possible application to other languages represented in the research team towards a subsequent LOD use.

Keywords: *annotation, automatic detection, offensive language taxonomy*

Research type: Research paper

References

Bączkowska, A., Lewandowska-Tomaszczyk, B., Valunaite Oleškevičiene, G., Žitnik, B., & Liebeskind, C. (2022). *Jerusalem workshop: A taxonomy of implicit language*.

Culpeper, J. (2005). Impoliteness and the Weakest Link. *Journal of Politeness Research*, *1*(I), 35–72.

Despot, K., & Ostroški Anić, A. (2022) *Jerusalem workshop: Reflections on Implicitness*.

Haugh, M., & Sinkeviciute, V. (2019). Offence and conflict talk. In *Routledge Handbook of Language in Conflict* (pp. 196–214). Routledge

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (2021). LOD-connected offensive language ontology and tagset enrichment. *SALLD-2 2021 workshop LDK Proceedings*. Saragossa.

Lewandowska-Tomaszczyk, B., Liebeskind, C., Žitnik, B., Bączkowska, A., Valunaite-Oleškevičiene, G., & Trojszczak, M. (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.

Mandl, S., Modha, P., Majumder, D., & Patel, D. (2019). Overview of the HASOCFire 2019: Hate speech and offensive content identification in Indo-European languages. *Proceedings of the 11.Forum for Information Retrieval Evaluation*. India.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75–86). Minneapolis.

Žitnik, B., Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite-Oleškevičiene, G., & Mitrovic, J. (in press). *Detecting Offensive Language: A new approach to offensive language data preparation*.