Czech Offensive Language: Testing a Simplified Offensive Language Taxonomy

Olga Dontcheva-Navrátilová
Renata Povolná
Masaryk University, Brno, Czech Republic
navratilova@ped.muni.cz
povolna@ped.muni.cz

Abstract

This contribution presents the results of an annotation campaign carried out on a Czech Corpus of Offensive Language (CCOL) compiled for the purposes of this study. The annotation was based on a Simplified Offensive Language (SOL) Taxonomy (Lewandowska-Tomaszczyk 2022) which has been proposed as part of the research work undertaken within COST Action NexusLinguarum WG 4.1.1. The aim of the study is to test the applicability of the SOL taxonomy to the Czech language, to identify the level of inter-rater agreement for all categories of the taxonomy and to compare the results to an earlier annotation campaign on English Offensive Language within the same research project. The findings of this study hope to support the application of the suggested SOL taxonomy as an ontology for effective detection and encoding of offensive language in Linguistic Linked Open Data (LLOD).

1 Introduction

Online newspaper and social media platforms have created virtual places where people can exchange opinions and views not limited by space constraints. Apart from speeding up the process of production, consumption and sharing information, these platforms have led to the emergence of huge amounts of data and the surge of offensive language (Kennedy et al. 2017, Casselli et al. 2020). As a result, there is a need for the development of methods for the automatic detection of offensive language applicable in LLOD.

In agreement with Lewandowska-Tomaszczyk et al. (forthcoming) offensive language is understood as hurtful, derogatory or obscene utterances produced by one person (or a group of people) to another or to a group of persons (see also Wiegand et al. 2021) with the intention to cause offence or insult. Offensive language, sometimes called abusive or toxic language, or hate speech, refers to the use of explicit language means representing verbal attacks towards individuals or groups of individuals. This paper does not consider visual means although they are natural part of social media platforms and their role in creating offensiveness is generally recognized (see e.g. Lewandowska-Tomaszczyk et al. 2021).

2 Offensive language taxonomy

Several attempts have been made to create an effective offensive language taxonomy (e.g. Basile et al. 2019, Liu et al. 2019, Fortuna et al. 2021, Kogilavani et al. 2021). The taxonomies suggested by Lewandowska-Tomaszczyk et al. (2022, 2023) developed within COST Action NexusLinguarum draws on Zampieri et al.'s (2019) three-level categorisation of offensive language, in which level one discriminates between offensive and non-offensive posts, level two identifies the offensive type (targeted vs non-targeted insult/offence) and the third level identifies the target of offence, i.e. individual, group or other. Within the SOL taxonomy approach (Lewandowska-Tomaszczyk 2022), an additional sub-level is added to the target of offence specifying whether the target is absent or present as an interaction participant. In addition, a specific level focusing on the type of lexical items is introduced differentiating between vulgar and non-vulgar expressions. The offensive type is split into four kinds of speech acts, i.e. *hate*, *insult*, *discredit* and *threat*. The offence is further specified in terms of the specific property of the target that is aimed at (e.g. *ageism*, *ideologism*, *ableism*, *racism*, *sexism*). Finally, the taxonomy considers implicit types of offence expressed via figurative means, labelled aspects, namely *exaggeration*, *irony*, *metaphor*, *rhetorical question*, *simile* or *other*.

1. Offensive
- Yes
- No

2. Target 1
- Group
- Ind. Wrt. Gr./Gr. Wrt. Ind. [by reference to group stereotypes]
- Individual
- Non-targeted

3. Target 2
- Absent
- Present

4. Vulgar
- No
- Yes

5. Speech act
- Hate speech (referring to group stereotypes)
- Insult (not referring to group stereotypes)
- Discredit (e.g. lying-cheating, immorality, unfairness)
- Threat (inducing fear)

6. Aspect (specific property of the target aimed at)
- Ageism
- Homophobic
- Ideologism
- Other
- Physical/mental disabilities (ableism)
- Prophane (religion)
- Racist
- Sexist
- Social class (classism)
- Xenophobic

7. Category of figurative language (implicit offence)
- Exaggeration
- Irony
- Metaphor
- Other
- Rhetorical question
- Simile

Table 1: Simplified offensive language taxonomy

3 Data and annotation

The Czech Corpus of Offensive Language (CCOL) comprises 400 comments, each consisting of one to three adjacent utterances, extracted from online discussions in ten Czech national newspapers and news platforms, such as SeznamZprávy, Idnes.cz, Forum24, Novinky.cz, HlídacíPes, published in the period January-February 2023. The corpus is sampled to represent discussions on a variety of topics, including home and foreign news, home and foreign politics, sport, celebrities, crime, finance, travelling, weather and health. The corpus was annotated by two annotators who are linguists and share a similar social background, age, and profession. In order to test whether the L1 of the annotator is an important variable, the L1 of one of the annotators taking part in the campaign was Czech and the other had a different L1 but had been living and working in Czechia for 30 years. Prior to annotating the corpus, the two annotators carried several training sessions, in which they discussed the offensive language taxonomy, practiced annotating samples, compared their results and resolved disagreements.

The CCOL was annotated with the assistance of INCEpTION tool (https://github.com/inception-project/inception), a semantic *annotation* platform, and classified according to the SOL

Taxonomy (Lewandowska-Tomaszczyk et al. 2021) proposed as part of the research work undertaken within COST Action NexusLinguarum WG 4.1.1, summarised in Table 1. The annotation campaign took place in the period February-March 2023.

4 Results

Annotator agreement was measured according to the Cohen's Kappa measure; drawing on Landis and Koch (1997) and Sim and Wright (2005), the strength of agreement for the kappa coefficient was established on the scale: ≤0=poor, .01–.20=slight, .21–.40=fair, .41–.60=moderate, .61–.80=substantial, and .81–1=almost perfect.

The Cohen's Kappa results for inter-rater agreement summarised in Table 2 show that the annotator agreement is high. More specifically, it is almost perfect for the categories Target 1 (0.89), Target 2 (0.93) and Vulgar (0.85), and substantial for the Offensive type categories (0.74 for both Insult and Discredit); the slight agreement for the *threat* category may be explained by its very low occurence in the annotations. During the curation campaign, it was revealed that in terms of target, most of the comments in the CCOL aimed at individuals and groups, while non-targeted comments were rare (e.g. *A Hitler dělal to, co teď Russáci* [*And Hitler did what the Russians are doing now*], CZ-OL-131). There were some ambiguous cases, where even in the case of Czech, which discrimitates T/V forms, it was impossible to decide whether the target is a group, or an individual addressed by the V-form. Occasional disagreements in the Vulgar category seem to reflect metaphorical uses of lexical items (e.g. *Člověče, vytáhněte si hlavu z řitního otvoru a možná to pochopíte* [*Man, pull your head out of your asshole, and maybe you'll understand.*], CZ-OL-292). The differences in the offensive type identification concerned the perceived intensity of offence categorised as *threat* (e.g. *Už tam zůstaň na věčné časy, šmejde* [*Stay there for eternity, scum.*], CZ-OL-22).

As to Aspects of offensive language, or properties of the target, and categories of implicit realisations (categories of figurative language), interrater agreement differs at the three sub-levels: there is substantial agreement at the first level of Aspect 05 and Category 06, i.e. 0.70 and 0.61 respectively, but only moderate agreement at the second level Aspect 05a (0.52) and Category 06a (0.53); the value 0 for the third level (Aspect 05b and Category 06b) reflects the very low occurrence of simultaneous selection of more than three categories per instance of offensive language. When coding Aspects of offensive language, or properties of the target, the annotators were expected to select up to three properties available in the set (*ageism, homophobic, ideologism, albeism, prophane,*

Annotation type	**Agreement**
Target 1 – Individual/group	0.89
Target 2 – present/absent	0.93
Vulgar	0.85
Offensive type – hate speech/ insult	0.74
Offensive type discredit	0.74
Offensive type threat	0.11
Aspect 05	0.70
Aspect 05a	0.52
Aspect 05b	0
Category 06	0.61
Category 06a	0.53
Category 06b	0

Table 2: Inter-rater agreement for the Czech Offensive Language Corpus

racist, social class, xenophobic and *other*) and mark them as Aspect 05, 05a and 05b. The annotators were instructed to select the most salient property as Aspect 05, but no further guidance was provided for assigning properties to the individual sub-types. Similarly, the instructions concerning the identification of the three sub-categories (06, 06a and 06b) of impicit realisations (categories of figurative language, i.e. *exaggeration, irony, metaphor, simile, rhetorical question, irony* and *other*) did not explain how the individual categories of figurative language should be assigned to the sub-types.

Out of the properties of the target, the most frequently appearing in the CCOL were *ideologism* (e.g. *České ošetřovatelství katastrofa, vládo, už se konečně prober* [*Czech healthcare is a disaster, government, wake up already*], CZ-OL-355), *albeism* (physical/mental) (see the example of metaphor below), and *sexism* (e.g. *některé ženy by neměly mít peníze, aspoň by nedělaly krávoviny* [*some women shouldn't have money, at least they wouldn't do shit*], CZ-OL-19). As to the categories of figurativeness, *metaphor* (e.g. *Lituji pana premiéra, že se musí až do poslední chvíle scházet s tou vypitou troskou...* [*I pity the Prime Minister for having to meet with that drunken wreck until the last minute.*], CZ-OL-323), *simile* (e.g. *Pokud se nechovají jako ruská šovinistická prasata, tak s nimi nemají sebemenší problém* [*As long as they don't act like Russian chauvinist pigs, they don't have the slightest problem*], CZ-OL-127), *irony* and *rhetorical question* appear to be most prominent. The curation campaign showed that the lower level of agreement is most likely affected by the absence of specific instructions concerning the order in which the individual properties of the target and categories of figurative language should be marked during the annotation process. In the absense of such instructions, the annotators ranked the properties and categories differently, for instance, annotator 1 classified *metaphor* as category 06a and *irony* as category 06b, while annotator 2 had *metaphor* as category 06b and *irony* as category 06a. The same concerns the properties of the target, where the annotators often listed the same properties, but in a different order. This suggests that the annotation scheme is robust, but should include a hierarchy of potential realisation of categories, in order to improve inter-rator agreement. In addition, some divergencies in the annotation of the two annotators are caused by differences in the splitting of a particular document into several consecutive parts, for instance, one annotator has identified as offensive a single expression, and the other has marked as offensive a whole clause, or one annotator has analysed a complex sentence as consisting of two clauses realising two speech acts of offense, while the other has marked the whole sentence as one speech act of offence. This could also be resolved during the training campaign by specific instructions on selection criteria.

Overall, in the case of the CCOL, the use of the SOL (Lewandowska-Tomaszczyk et al. 2022) has yielded a considerably higher degree of inter-rater agreement in comparison with annotation campaigns using a more elaborate taxonomy of offensive language, such as the English Offensive Language Corpus annotation performed earlier within COST Action NexusLinguarum

(Lewandowska-Tomaszczyk et al. 2023), where interrater agreement was fair (for Offensive type 0.32 and for Aspect between 0.29 and 0.18 for the individual subcategories). Apart from the simplification of the taxonomy, this considerably higher degree of inter-rater agreement seems to stem from the careful selection of the data included in the corpus, the extensive training campaign and the similarity in the professional and social background of the two annotators. The CCOL campaign also indicates that inter-rater agreement is not strongly affected by the L1 factor (one of the annotator's L1 was different from Czech), as what seems of primary importance is the knowledge of the cultural and social context, in which offensive language is used. A comparison of this annotation campaign with the earlier campaign using the extended offensive language taxonomy on English offensive language (Lewandowska-Tomaszczyk et al. Forthcoming) suggests that the substantially lower inter-rater agreement (moderate and fair agreement) achieved in the English offensive language campaign may be attributed, apart from the random selection of data and short training campaign, to the choice of annotators, who, not only were speakers of various L1s different from English, but also lived in various non-English speaking contexts failing to provide them with shared cultural and social knowledge for the analysis of the English data.

5 Conclusions

This study tested the applicability of the SOL taxonomy to the Czech language, seeking to identify the level of inter-rater agreement for all categories of the taxonomy in CCOL. The results showed that the SOL taxonomy can be successfully applied to the Czech language and that the level of inter-rater agreement was generally high. This suggests that the taxonomy is applicable as an ontology for detection and encoding of offensive language in Linguistic Linked Open Data (LLOD).

Acknowledgement

We would like to thank Slavko Žitnik, University of Ljubljana, Slovenia, for the preparation of the INCEpTION tool with the assistance of which we have annotated our data.

References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P. & Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics*, 54-63.

Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I. & Granitzerz, M. (2020) I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive Language. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 6193-6202.

Fortuna, P., Soler-Company, J. & Wanner, L. (2021) How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management 58*(3), 102524.

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

Kennedy, G., McCollough, A., Dixon, E., Bastidas, A., Ryan, J., Loo, C. & Sahay, S. (2017) Technology solutions to combat online harassment. *Proceedings of the First Workshop on Abusive Language Online.* 73-77.

Kogilavani, S. V., Malliga, S., Jaiabinaya, K. R., Malini, M. & Manisha Kokila, M. (2021) Characterization and mechanical properties of offensive language taxonomy and detection techniques. *Materials Today: Proceedings.*

Landis J. R. & Koch G. G. (1997) The measurement of observer agreement for categorical data. *Biometrics 33*, 159-174.

Lewandowska-Tomaszczyk, B. (2022) A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies 10* (3). 213-227.

Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrović, J. & Valunaite Oleškevičiene, G. (2021) Lod-connected offensive language ontology and tagset enrichment. In: 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.* CEUR Workshop Proceedings. 135-150.

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

Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valunaite Oleskeviciene, G. & Žitnik, S. (2023) An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics* 23.1.

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

Sim, J. & Wright, C. C. (2005) The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy 85* (3): 257-268. Doi:10.1093/ptj/85.3.257

Wiegand, M., Ruppenhofer, J. & Eder, E. (2021) Implicitly Abusive Language – What does it actually look like and why are we not getting there? In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. & Zhou, Y. (eds) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics. Stroudsburg. 576-587.

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). In: *Proceedings of the 13th International Workshop on Semantic Evaluation.* Minneapolis, Minnesota, USA: Association for Computational Linguistics. 75-86.