

Beyond Translation: Cross-Linguistic Expression Strategies of Online Incivility on YouTube: A Comparison of English and Russian

Mursal Dawodi¹ Albina Fatykhova¹ Yingping Sun^{1,2} Juergen Pfeffer¹

¹Computational Social Science and Big Data, Technical University of Munich, Germany

²School of Business, Shandong Normal University, Jinan, China

^mursal.dawodi@tum.de, cherblosni@gmail.com, yingping.sun@tum.de, juergen.pfeffer@tum.de

Abstract

Multilingual moderation systems often need to account for the possibility that online incivility is expressed differently across languages, beyond differences in vocabulary or other surface-level linguistic features. To examine this issue, we analyze a manually annotated bilingual corpus of 800 YouTube comments (400 English and 400 Russian) labeled across four dimensions: incivility type, severity, target, and clarity of expression. While severity does not differ significantly between the two corpora (Cramér's $V = .06$), significant differences emerge for type, target, and clarity ($V = .13$, $.22$, and $.10$, respectively). English comments contain more abusive and explicit forms of incivility, whereas Russian comments more frequently target individuals and employ implicit expression, including sarcasm and irony. User engagement analyses further show that the forms of incivility associated with greater visibility differ across the two corpora. Together, these findings suggest that, within the corpora examined, severity is comparatively stable across languages, whereas communicative dimensions exhibit greater cross-linguistic variation. The results suggest that multilingual moderation may benefit from distinguishing comparatively stable dimensions from language-sensitive communicative dimensions when developing annotation guidelines and moderation resources.

1 Introduction

Multilingual content moderation systems often assume that online incivility is expressed similarly across languages and that annotation schemes and detection models transfer once lexical differences are addressed (Davidson et al., 2017; Zampieri et al., 2019; Röttger et al., 2022). However, abusive language datasets often differ substantially in annotation guidelines and category definitions, limiting comparability and transferability across languages

(Seemann et al., 2023). Although multilingual moderation has received increasing attention, whether the underlying dimensions of incivility themselves transfer equally across languages remains largely unexplored.

Research in cross-cultural communication suggests that norms governing disagreement, criticism, and hostility vary across linguistic communities (Sifianou, 1992; Wierzbicka, 1991). Likewise, online incivility has been described as a violation of communicative norms rather than merely the use of offensive words (Papacharissi, 2004), while communicative context has been shown to influence how hostility is expressed online (Frischlich et al., 2021). Together, these findings motivate the central question of this paper: do cross-linguistic differences in online incivility primarily reflect differences in hostility intensity, differences in communicative expression, or both?

Previous multilingual studies have primarily investigated how well abusive language detection models transfer across languages, often reporting substantial performance degradation under cross-lingual transfer (Pavlopoulos et al., 2017; Röttger et al., 2022). In contrast, we examine whether the annotation dimensions themselves exhibit comparable cross-linguistic stability. If languages systematically differ in how hostility is expressed, multilingual moderation may require language-specific annotation strategies in addition to cross-lingual lexical transfer.

To investigate this question, we compare manually annotated English and Russian YouTube comments. English serves as the reference language because it dominates existing online incivility research, whereas Russian provides a linguistically and culturally distinct comparison with documented use of more indirect forms of public criticism in some communicative settings (Oates, 2013; Gorham, 2017). We analyze four annotation dimensions (type, severity, target, and clarity) together

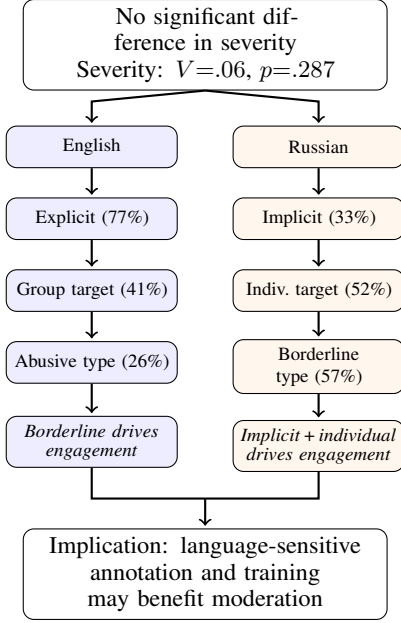


Figure 1: Conceptual summary of the distinction between comparatively stable and language-sensitive dimensions of online incivility observed in the present corpora.

with user engagement patterns to determine which aspects of online incivility remain comparatively stable across languages.

Our results reveal a consistent pattern. Severity does not differ significantly between the two corpora, whereas type, target, and clarity do. In addition, the forms of incivility receiving the greatest user engagement differ across the two language communities. Together, these findings suggest that multilingual moderation may benefit from distinguishing between comparatively stable and language-sensitive annotation dimensions. Figure 1 summarizes this conceptual distinction.

2 Related Work

2.1 Online Incivility

Online incivility is widely understood as a violation of communicative norms rather than merely the use of offensive language. Papacharissi (2004) distinguished incivility from simple impoliteness, while Coe et al. (2014) showed that uncivil comments influence readers beyond immediate interaction. More recently, Frischlich et al. (2021) demonstrated that social and platform contexts shape how incivility is expressed. Together, these studies suggest that incivility is context-dependent and concerns how disagreement, criticism, disrespect, or hostility are communicated.

In this study, online incivility is used as a broader communicative concept than hate speech. Hate speech generally focuses on hostile or derogatory expression directed at people because of their membership in a social or identity group, whereas incivility can also involve insults, contempt, sarcasm, dismissive remarks, or hostile criticism directed at individuals, groups, or ideas. Thus, not every uncivil comment in our corpus constitutes hate speech, and the annotation scheme is not intended to classify comments exclusively as hate speech. Instead, it captures multiple dimensions of uncivil expression, including its form, perceived hostility intensity, target, and explicitness.

This broader definition is important for the present comparison because the study asks not only whether hostile content occurs, but also how hostility is communicated across languages. Communicative norms can vary across contexts and linguistic communities, providing a theoretical basis for examining whether the dimensions used to characterize incivility transfer equally across languages.

2.2 Annotation Frameworks

Most annotation frameworks have been developed using English-language data. Davidson et al. (2017) showed that lexical cues alone are insufficient for detecting hate speech, Zampieri et al. (2019) introduced target-based annotation, and ElSherief et al. (2018) distinguished implicit from explicit hate speech. More recently, Seemann et al. (2023) highlighted that differences in annotation guidelines and taxonomy definitions remain a major source of disagreement across abusive language datasets, while recent multilingual resources advocate multidimensional annotation schemes that jointly represent complementary aspects of offensive language, including target and severity (Liebeskind et al., 2024; Damo et al., 2025). Building on this literature, we adopt a multidimensional annotation scheme comprising type, severity, target, and clarity. These dimensions draw on complementary strands of prior annotation research: type captures the form of offensive or uncivil expression, target captures the intended recipient of the hostility, and clarity captures the distinction between explicit and implicit expression. Severity is used to represent the perceived intensity of hostility independently of the particular form through which it is expressed. The four dimensions therefore serve different analytical purposes rather than treating incivility as a single binary category. This distinction allows us

to examine whether cross-linguistic differences are concentrated in what form hostility takes, whom it targets, and how explicitly it is expressed, or instead in the overall perceived intensity of hostility. Our framework is consequently an adaptation of existing multidimensional annotation principles to the specific question of cross-linguistic stability, rather than a claim that these dimensions are individually new.

2.3 Cross-Linguistic Work

Previous multilingual studies have shown that markers of online incivility vary across languages and that cross-lingual abusive language detection models often experience substantial performance degradation (Sanguinetti et al., 2018; Ousidhoum et al., 2019; Pavlopoulos et al., 2017; Röttger et al., 2022). However, this research has focused primarily on model transfer rather than on whether the underlying annotation dimensions themselves exhibit comparable cross-linguistic stability.

2.4 User Engagement

User engagement provides behavioral evidence of which forms of incivility receive the greatest social visibility. Oz and Greeves (2024) found that moderately uncivil content can attract greater engagement than overt hostility, while Thelwall (2018) showed that likes and replies capture different forms of audience response. This motivates our analysis of user engagement alongside annotation distributions.

2.5 Research Gap

Existing work has largely examined either language-specific annotation resources or cross-lingual model performance. Much less attention has been paid to whether the annotation dimensions used to characterize online incivility are themselves equally transferable across languages. As a result, it remains unclear which dimensions reflect broadly comparable patterns across languages and which are more sensitive to language-specific communicative norms. Addressing this question is important because successful multilingual moderation implicitly assumes that at least some annotation dimensions generalize across linguistic contexts. The present study addresses this gap by comparing the cross-linguistic behavior of four annotation dimensions—type, severity, target, and clarity—using manually annotated English and Russian YouTube comments.

3 Data and Methodology

3.1 Platform and Language Selection

YouTube was selected because it hosts large-scale public discussions across diverse topical domains and provides engagement indicators (likes and replies) through the YouTube Data API, making it a widely used platform for research on online discourse and abusive language (Frischlich et al., 2021). English was chosen as the reference language because it dominates the online incivility and abusive language literature (Davidson et al., 2017; Zampieri et al., 2019), whereas Russian provides a linguistically and culturally distinct comparison with documented use of more indirect forms of public criticism in some communicative settings (Oates, 2013; Gorham, 2017). One of the authors is a native Russian speaker, allowing annotation to be performed directly on the original comments without translation.

3.2 Data Collection

We collected English and Russian comments from YouTube videos covering a range of topical categories, including news, politics, interviews, science and culture, crime, psychology and society, lifestyle and entertainment, music, movies and TV, and sports. For each language, we sampled 75 videos and collected up to 100 comments per video, resulting in initial pools of 6,901 English and 7,283 Russian comments. To reduce the influence of temporal changes in platform discourse, we excluded comments posted before January 2022. We also removed comments that were not primarily linguistic content, such as link-only or emoji-only comments.

The video selection aimed to provide broad topical coverage in both languages rather than to reproduce the overall distribution of YouTube content. Table 1 summarizes the distribution of the 75 sampled videos in each language. The specific videos and channels were not matched across languages; instead, the sampling sought comparable topical breadth while allowing the two corpora to reflect naturally occurring content within each language.

3.3 Subset Construction

Candidate comments were first identified using language-specific lexicons of negative terms. The lexicons were initially compiled manually and subsequently expanded with additional candidate terms suggested by an AI-assisted tool. The resulting lexicons were used only for initial candidate

Category	RU <i>n</i>	EN <i>n</i>	RU%	EN%
News	14	8	18.7	10.7
Politics	10	7	13.3	9.3
Interviews	15	6	20.0	8.0
Science & Culture	9	6	12.0	8.0
Crime	7	4	9.3	5.3
Psychology & Society	6	6	8.0	8.0
Lifestyle & Entertainment	6	10	8.0	13.3
Music	4	8	5.3	10.7
Movies & TV	3	12	4.0	16.0
Sports	1	8	1.3	10.7
Total	75	75	100.0	100.0

Table 1: Distribution of sampled YouTube videos across topical categories by language.

retrieval and were not used to assign annotation labels.

The same two-stage procedure was applied to both languages: keyword-based candidate retrieval followed by manual contextual review. The initial keyword filtering identified 773 English and 358 Russian candidate comments from the original comment pools. Manual review was then used to assess the comments in context and remove false positives, including cases in which negative terms were used in neutral or otherwise non-uncivil contexts. The review also allowed implicit, sarcastic, ironic, or otherwise indirectly expressed forms of incivility to be identified.

Because lexical filtering cannot reliably capture all forms of implicit or indirect incivility, additional comments were manually identified from the broader unfiltered comment pools. After manual review and selection, 400 comments per language were retained for the final corpus.

The final selection was intended to provide comparable coverage of incivility types, targets, and communicative forms across the two languages. Near-duplicate or highly repetitive comments were excluded, and the same selection procedure was applied to both language corpora. The resulting corpus contained 800 comments in total, comprising 400 English and 400 Russian comments, which were then manually annotated across the four dimensions described in Section 3.4.

3.4 Annotation Schema

Annotation guidelines were developed by adapting definitions from prior work (Papacharissi, 2004; Davidson et al., 2017; Zampieri et al., 2019) and iteratively refined through pilot annotation. Each comment was annotated on four dimensions:

- **Type:** *abusive*, explicit attacks, insults, slurs, threats; *borderline*, sarcastic, dismissive, or contemptuous without overt slurs; *non-abusive*, criticism within civil norms.
- **Severity:** perceived hostility intensity, coded as high, medium, or low. High severity includes dehumanization, threats, or severe slurs; medium severity indicates clear or substantial disrespect without the most extreme forms of hostility; low severity indicates mild rudeness or relatively limited hostility. Severity is evaluated independently of type so that a comment’s perceived hostility intensity is not determined solely by whether its form is abusive, borderline, or non-abusive. This dimension is intended to capture perceived hostility intensity rather than culturally specific linguistic form. We recognize that judgments of severity can nevertheless involve contextual and cultural interpretation, particularly when comments rely on stereotypes, irony, or culturally specific references. The cross-linguistic comparison therefore concerns the distribution of these annotation judgments within the two corpora and should not be interpreted as evidence that severity is culturally invariant.
- **Target:** *individual*, *group*, *idea/policy*, *unclear*.
- **Clarity:** *explicit*, plainly stated; *implicit*, conveyed through irony, sarcasm, or coded subtext.

Our annotation design follows recent work advocating multidimensional representations of offensive language that distinguish target, severity, and other communicative properties, allowing richer analyses than binary offensive/non-offensive classification alone (Liebeskind et al., 2024; Damo et al., 2025). Together, these four dimensions distinguish communicative form (type), hostility intensity (severity), target, and explicitness of expression (clarity), enabling us to examine how these dimensions vary across languages.

Table 2 provides illustrative examples. All 800 comments were annotated manually by one of the authors. Following pilot annotation, a stratified subset of 50 comments per language (12.5% of each corpus) was independently annotated by a second annotator to estimate inter-annotator agreement. This proportion is comparable to reliability assessments reported in previous annotation studies

Comment (paraphrased)	Tp	Sv	Tg	Cl
<i>English</i>				
“Garbage human being.”	ab	hi	in	ex
“Another ‘expert’ who never did this.”	bl	lo	in	im
“Americans are so entitled.”	bl	me	gr	ex
“This policy is poorly designed.”	na	lo	id	ex
<i>Russian (translated)</i>				
<i>Genius move, bravo.</i>	bl	lo	in	im
<i>Russia keeps building mountains of scrap metal.</i>	bl	me	gr	im
<i>Poor girl, fell for an addict.</i>	bl	me	in	ex
<i>This person is a complete idiot.</i>	ab	hi	in	ex

Table 2: Examples. Tp=Type (ab=abusive, bl=borderline, na=non-abusive); Sv=Severity (hi/me/lo); Tg=Target (in=individual, gr=group, id=idea); Cl=Clarity (ex=explicit, im=implicit).

(Artstein and Poesio, 2008).

3.5 Analysis

Distributional differences between the English and Russian corpora were evaluated using chi-square tests together with Cramér’s V effect sizes (Lakens, 2013). User engagement was analyzed separately for each corpus using Kruskal-Wallis tests across annotation dimensions, while Mann-Whitney U tests compared engagement between the two language communities. Statistical analyses were implemented in Python.

4 Results

4.1 Annotation Distributions

Table 3 summarizes the annotation distributions, while Figure 2 visualizes the same patterns. Significant differences between the two corpora were observed for type, target, and clarity. The largest difference was found for target ($\chi^2 = 39.41$, $V = .22$, $p < .001$), followed by type ($\chi^2 = 14.40$, $V = .13$, $p < .01$) and clarity ($\chi^2 = 8.54$, $V = .10$, $p < .01$). In contrast, severity did not differ significantly between the English and Russian corpora ($\chi^2 = 2.50$, $V = .06$, $p = .287$).

The distributions therefore differ more clearly in the type, target, and clarity of incivility than in perceived hostility intensity. Although the effect sizes are modest, the pattern is consistent across these three dimensions. These findings support distinguishing between the intensity of hostility and the ways in which it is expressed across languages.

Dim.	Label	EN	nEN%	RU	nRU%	Total	nTotal%
Type	Abusive	102	25.5	59	14.8	161	20.1
	Borderline	199	49.8	226	56.5	425	53.1
	Non-abusive	99	24.8	115	28.7	214	26.8
Severity	High	97	24.2	81	20.2	178	22.2
	Medium	193	48.2	213	53.2	406	50.7
	Low	110	27.5	106	26.5	216	27.0
Target	Individual	162	40.5	207	51.7	369	46.1
	Group	164	41.0	96	24.0	260	32.5
	Idea/policy	70	17.5	71	17.8	141	17.6
	Unclear	4	1.0	26	6.5	30	3.8
Clarity	Explicit	308	77.0	270	67.5	578	72.2
	Implicit	92	23.0	130	32.5	222	27.8

Table 3: Annotation distributions across dimensions by language and total count.

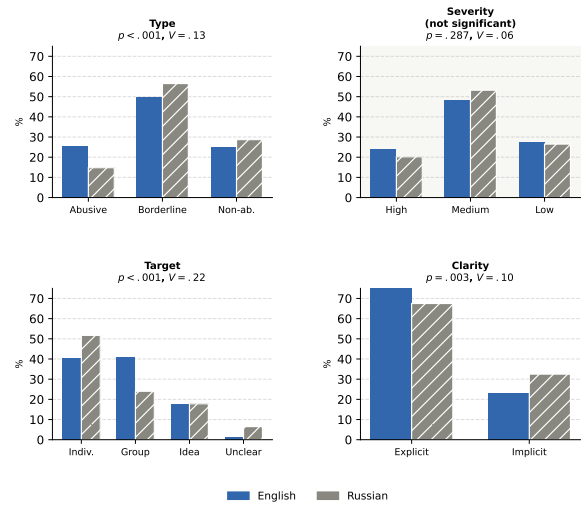


Figure 2: Annotation distributions by language. Severity is the only dimension without a significant cross-linguistic difference ($p = .287$, $V = .06$); target shows the largest effect ($V = .22$).

4.2 Engagement Analysis

Across the sampled videos, Russian comments had a higher median number of likes than English comments (139 vs. 60; Mann-Whitney $U = 71494$, $p = .009$). This cross-language comparison should be interpreted cautiously because the videos and channels were not directly matched and audience size, channel popularity, and other platform-level factors may influence raw engagement. We therefore focus primarily on within-language differences in engagement across annotation dimensions. Table 4 summarizes the Kruskal-Wallis tests examining whether engagement differed across annotation dimensions within each language. The overall pattern differs between the two corpora. In the English corpus, only incivility type is significantly asso-

Dimension	EN $H(p)$	RU $H(p)$
Type	7.86 (.020)	0.23 (.892)
Severity	2.75 (.253)	0.62 (.734)
Target	3.33 (.344)	14.23 (.003)
Clarity	0.09 (.767)	13.77 (< .001)

Table 4: Kruskal-Wallis test results by annotation parameter and corpus.

Dimension	EN κ	RU κ	Interp.
Type	.674	.669	Substantial
Severity	.633	.904	Subst. / Almost perf.
Target	.798	.832	Subst.–Almost perf.
Clarity	.830	.712	Almost perf. / Subst.

Table 5: Cohen’s κ by dimension and language. All $\kappa \geq .63$.

ciated with engagement, whereas in the Russian corpus significant differences emerge for target and clarity but not for type or severity. In the English corpus, borderline comments attracted the highest median number of likes (137) and replies (3), whereas abusive comments received the lowest median engagement (33 likes and 1 reply). In contrast, engagement did not differ significantly by incivility type in the Russian corpus. Instead, implicit comments received substantially higher engagement than explicit comments (median 257 vs. 94 likes and 6 vs. 2 replies), while comments targeting individuals received considerably more likes than those targeting groups (median 228 vs. 56). Taken together, these findings indicate that the forms of incivility receiving the greatest community attention differ across the two language communities, suggesting that engagement with uncivil content may be associated with communicative dimensions beyond severity.

4.3 Inter-Annotator Agreement

5 Discussion

The results reveal a consistent pattern across annotation and engagement analyses. Differences between the English and Russian corpora emerge primarily in the communicative dimensions of incivility (type, target, and clarity), whereas severity remains comparatively stable. Figure 1 summarizes this pattern. In the discussion below, we interpret these findings in relation to existing theories of online incivility, cross-cultural communication, and multilingual content moderation.

5.1 The Assumption Does Not Hold Uniformly

A common implicit assumption in multilingual moderation is that incivility is expressed similarly across languages, with the same communicative phenomena appearing in different surface forms. Our findings suggest that this assumption does not hold uniformly across the two corpora examined. The most notable result is that severity does not differ significantly between English and Russian ($V = .06$, $p = .287$), whereas type, target, and clarity show significant differences. Thus, within these corpora, the observed cross-linguistic differences are more pronounced in dimensions describing how incivility is communicated than in the distribution of perceived hostility intensity. This finding should not be interpreted as evidence that English and Russian have inherently different or fixed communicative norms. Rather, it indicates that the forms of incivility represented in the two samples differ across several communicative dimensions while producing comparatively similar severity distributions. This interpretation is consistent with recent studies suggesting that offensive-language categories are shaped by linguistic and cultural context and cannot always be transferred directly across languages (Liebeskind et al., 2024).

These findings also have implications for multilingual moderation systems. The dimensions that vary across the two corpora are type, target, and clarity, particularly the distinction between explicit and implicit expression. A moderation model trained on English data may therefore encounter a different distribution of communicative signals from a model trained on Russian data, even when the perceived severity of the annotated content is comparable. This difference is not necessarily a tokenization or vocabulary problem; it may also reflect differences in the communicative phenomena represented in the training data. Such differences could contribute to systematic rather than purely random moderation errors, but this possibility requires direct evaluation using multilingual moderation models.

5.2 Implicitness and Individual Targets

The higher rate of implicit incivility in the Russian corpus (32.5% vs. 23.0%) is consistent with previous accounts of indirect expression and coded criticism in Russian communicative contexts (Gorham, 2017; Oates, 2013). However, the present data do not establish that cultural norms are the cause

of this difference. What the data directly show is that implicit incivility is more prevalent in the Russian sample and that clarity agreement is lower for Russian than for English ($\kappa = .712$ vs. $.830$). This suggests that implicitness may require greater language-specific attention in annotation and moderation, including guidelines that explicitly address sarcasm, irony, and other forms of indirect expression in the target language.

The concentration of individual targets in Russian (51.7% vs. 40.5%) and group targets in English (41.0% vs. 24.0%) also points to differences in the distribution of targeted incivility across the two samples. In the Russian corpus, comments targeting individuals received a substantially higher median number of likes than comments targeting groups (228 vs. 56). This association suggests that individually targeted incivility can also receive substantial social visibility. Consequently, moderation approaches that rely primarily on group-based detection signals (Zampieri et al., 2019) could potentially overlook some individually targeted incivility with high social visibility. This implication remains provisional, however, and should be evaluated directly in future moderation studies.

5.3 Annotation Difficulty and Engagement Patterns

The inter-annotator agreement results provide a more nuanced view of the cross-linguistic differences. Severity agreement was lower for English ($\kappa = .633$) than for Russian ($\kappa = .904$), whereas clarity agreement was lower for Russian ($\kappa = .712$) than for English ($\kappa = .830$). These differences suggest that annotation difficulty is dimension-specific rather than uniformly higher in one language. In particular, the lower agreement for Russian clarity is consistent with the difficulty of distinguishing explicit from implicit expressions when sarcasm, irony, or context-dependent hostility is involved. These results should not be taken as evidence that such ambiguity is inherently characteristic of Russian; rather, they indicate that the explicit–implicit distinction presented greater annotation difficulty in the Russian sample.

The engagement analysis provides a complementary perspective. Borderline content attracts the most engagement in English (median 137 likes vs. 33 for abusive content), consistent with Oz and Greeves (2024), while in Russian, engagement does not differ significantly by type and implicit comments receive a median of 257 likes versus 94

for explicit comments. These patterns suggest that the forms of incivility associated with greater engagement are not identical across the two samples. However, the engagement comparison should be interpreted cautiously because the videos and channels were not directly matched across languages, and audience size, channel popularity, topic, and other platform-level factors may influence raw engagement. Accordingly, these results are best understood as associations within the sample rather than evidence that particular linguistic forms inherently produce greater social visibility.

5.4 Toward a Two-Level Framework for Multilingual Moderation

The pattern across annotation distributions, annotation difficulty, and user engagement suggests a practical perspective for multilingual moderation systems.

Comparatively transferable dimensions. Severity remains comparatively stable across the two corpora ($V = .06$). Because severity represents perceived hostility intensity rather than a specific communicative realization, it may provide a comparatively stable signal for cross-lingual moderation. However, this interpretation should be treated as provisional, since the present study examines only two languages and the severity judgments may themselves be influenced by context and annotation practices.

Language-sensitive dimensions. In contrast, type, target, and clarity differ significantly across the two corpora. These dimensions capture how hostility is communicated—its directness, intended recipient, and surface realization—and appear more sensitive to the communicative context represented in the two samples. Applying annotation guidelines developed for one language directly to another may therefore introduce additional annotation uncertainty, as reflected by the lower agreement for clarity in the Russian corpus ($\kappa = .712$ vs. $.830$). These findings suggest that multilingual moderation may benefit from language-specific annotation guidelines and training data for communicative dimensions, rather than relying solely on shared lexical representations or language-specific tokenization. This interpretation is also consistent with recent work advocating richer multidimensional annotation resources for offensive language (Damo et al., 2025).

Taken together, these findings motivate a two-level framework in which comparatively stable dimensions are distinguished from dimensions that appear more dependent on communicative context. Whether this perspective generalizes across additional language pairs, platforms, and cultural settings remains an open empirical question that warrants further investigation.

5.5 Synthesis

Taken together, the findings suggest that, within the English and Russian YouTube corpora examined here, cross-linguistic differences are more pronounced in dimensions describing how incivility is communicated than in the distribution of perceived hostility intensity. Severity remains comparatively stable, while type, target, and clarity vary significantly. This pattern suggests that similar levels of perceived hostility can be associated with different communicative realizations across the two samples.

The contribution of this study therefore extends beyond documenting differences between English and Russian. It identifies dimensions of incivility that appear comparatively stable in these corpora and dimensions that appear more sensitive to language and communicative context. This distinction has implications for multilingual moderation: rather than assuming that all annotation categories generalize equally across languages, moderation resources may need to distinguish between comparatively stable dimensions and language-sensitive communicative dimensions.

The convergence of evidence across annotation distributions, inter-annotator agreement, and engagement patterns provides support for this interpretation. No single result is decisive, and the relatively small, manually annotated corpora and unmatched video contexts limit the strength of the conclusions. Nevertheless, the findings indicate that cross-linguistic variation in incivility cannot be adequately characterized only through shared lexical representations. Further research with larger datasets, additional languages, and direct multilingual moderation experiments is needed to determine the extent to which the observed pattern generalizes.

5.6 Limitations

The present study has several limitations. First, YouTube’s automated moderation removes much of the most extreme content before collection, mean-

ing that the observed severity distributions likely underrepresent the full range of online incivility. Second, comment selection relied in part on lexical filtering to identify candidate comments for manual review. Although the language-specific lexicons were manually compiled and expanded with additional candidate terms suggested by an AI-assisted tool, keyword filtering may favor comments containing recognizable negative or offensive terms and may underrepresent incivility expressed through sarcasm, irony, understatement, or culturally specific indirect language. Manual review of the filtered candidates and additional selection from the broader unfiltered corpus helped mitigate this limitation, but it could not eliminate the possibility of systematic selection bias. The final selection also involved researcher judgment to maximize diversity across topics and communicative forms. Because the same procedure was applied to both language corpora, the approach provides procedural consistency, but some sampling subjectivity may remain. Third, all 800 comments were manually annotated by a single annotator, while inter-annotator agreement was estimated on a stratified subset of 50 comments per language. Although agreement was consistently substantial across all dimensions, a larger double-annotated sample would provide a more comprehensive assessment of annotation consistency. Because primary annotation was conducted separately across the two language corpora, annotator-specific interpretation cannot be completely separated from language-specific differences; the agreement subset provides evidence of consistency but does not eliminate this potential confound.

Fourth, the annotators shared similar cultural backgrounds, which may have influenced interpretations of Russian-language content. In addition, severity represents perceived hostility intensity and may itself be influenced by cultural and contextual interpretation. Therefore, the absence of a significant difference in severity should be understood as a finding about the distributions produced by the present annotation procedure, rather than evidence that hostility intensity is universally perceived in the same way across languages. Fifth, the corpus comprises 400 comments per language collected exclusively from YouTube, limiting generalizability to other platforms, languages, and online communities. Finally, although comparable topical categories were sampled across languages, exact matching of videos and channels was not

feasible, and residual topical differences may have influenced some findings.

6 Conclusion

Across the English and Russian corpora examined in this study, severity remains comparatively stable, whereas type, target, and clarity differ significantly. The findings therefore suggest that, within these two corpora, cross-linguistic differences are more pronounced in dimensions describing the communicative realization of incivility than in the distribution of perceived hostility intensity. This should not be interpreted as evidence that severity is culturally invariant or that English and Russian have fixed communicative styles. Rather, the results identify dimensions that appear comparatively stable in the present samples and dimensions that may require greater language-specific attention.

The study suggests that multilingual moderation should not assume that all annotation dimensions transfer equally across languages. Comparatively stable dimensions such as severity may provide a useful basis for cross-lingual generalization, while communicative dimensions including type, target, and clarity may benefit from language-specific annotation guidelines and training data. The engagement and annotation-agreement results further indicate that these language-sensitive dimensions can involve different patterns of visibility and annotation difficulty.

These conclusions remain provisional given the use of two languages, a relatively small manually annotated corpus, separate primary annotators, and unmatched videos and channels. Future work should evaluate the proposed framework using larger independently annotated datasets, more language pairs and platforms, and direct experiments with multilingual moderation models.

Data Availability

Raw comment texts cannot be redistributed due to YouTube's Terms of Service. Annotation labels and non-textual metadata will be released where permitted.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

Greta Damo, Alessandra Teresa Cignarella, Tommaso Caselli, Viviana Patti, and Debora Nozza. 2025. Hodiati: A dataset for detecting homotransphobic hate speech in Italian with aggressiveness and target annotation. In *Proceedings of the 9th Workshop on Online Abuse and Harms (WOAH)*, pages 124–135. Association for Computational Linguistics.

Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

Mai ElSherief, Vidya Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, volume 12.

Lena Frischlich, Tim Schatto-Eckrodt, Svenja Boberg, and Florian Winterlin. 2021. Roots of incivility: How personality, media use, and online experiences shape uncivil participation. *Media and Communication*, 9(1):195–208.

Michael S. Gorham. 2017. *After Newspeak: Language Culture and Politics in Russia from Gorbachev to Putin*. Cornell University Press.

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4:863.

Chaya Liebeskind, Marina Litvak, and Natalia Vanetik. 2024. From linguistics to practice: A case study of offensive language taxonomy in Hebrew. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*, pages 110–117. Association for Computational Linguistics.

Sarah Oates. 2013. *Revolution Stalled: The Political Limits of the Internet in the Post-Soviet Sphere*. Oxford University Press.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4674–4683. Association for Computational Linguistics.

Mert Oz and Stephanie Greeves. 2024. Examining the impact of incivility and intolerance on participation in social media discussions: An experimental investigation. *First Monday*, 29(11).

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Scott A. Hale, and Janet B. Pierrehumbert. 2022. Two contrasting data annotation paradigms for offensive language: Move the model or move the goalposts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2042–2054. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association.

Nina Seemann, Yeong Su Lee, Julian Höllig, and Michaela Geierhos. 2023. [The problem of varying annotations to identify abusive language in social media content](#). *Natural Language Engineering*, 29(6):1561–1585.

Maria Sifianou. 1992. *Politeness Phenomena in England and Greece*. Oxford University Press.

Mike Thelwall. 2018. Social media analytics for Youtube comments: Potential and limitations. *International Journal of Social Research Methodology*, 21(3):303–316.

Anna Wierzbicka. 1991. *Cross-Cultural Pragmatics: The Semantics of Human Interaction*. Mouton de Gruyter.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420. Association for Computational Linguistics.

A Negative-Word Lexicons

The language-specific negative-word lexicons used for initial candidate retrieval are provided below for transparency and reproducibility. The lexicons were used solely to identify comments for subsequent manual contextual review and were not used to determine the final annotation labels. Terms in the lexicons therefore should not be interpreted as operational definitions of incivility or as the complete vocabulary of uncivil expression in either language.

A.1 English Lexicon

bad, awful, terrible, horrible, worst, poor, disappointing, boring, dull, lame, weak, messy, sloppy, lazy, pointless, meaningless, useless, worthless, ridiculous, absurd, nonsensical, illogical, stupid, dumb, idiotic, cringe, cringey, embarrassing, pathetic, frustrating, annoying, irritating, obnoxious, unbearable, trash, garbage, junk, nonsense, bullshit, crap, shit, mess, disaster, trainwreck, joke, chaos, catastrophe, nightmare, horror, devastation, destruction, collapse, crisis, outrage, scandal, shame, disgrace, failure, breakdown, catastrophic, horrific, shocking, disturbing, unacceptable, ridiculous situation, complete mess, total disaster, absolute joke, what a mess, what a disaster, what a joke, this is a joke, embarrassing to watch, total mess, complete failure, epic fail, big fail, massive fail, low effort, zero effort, no effort, poorly done, badly made, trash content, garbage content, overrated, clickbait, forced, tryhard, hate, despise, can't stand, cannot stand, sick of, tired of, fed up, done with, disgusted, disgusting, disturbing, awful to watch, painful to watch, hard to watch, makes no sense, doesn't make sense, waste of time, not worth it, absurd situation, no way, this is horrible, this is terrible, this is awful, horribly wrong, terribly wrong, deeply wrong, completely wrong, infuriating, outrageous, disgraceful, shameful, unacceptable behavior, unacceptable situation, what is this, how is this possible, this is messed up, so messed up, completely messed up, totally messed up, broken system, failed system, corrupt, corruption, biased, unfair, injustice, criminal, evil, sick, twisted, madness, chaos everywhere, world is broken, everything is wrong, nothing makes sense, going downhill, getting worse, worse and worse, beyond bad, beyond terrible, beyond horrible, unacceptable outcome, unacceptable result, failure of system, system failure, huge problem, serious problem, major problem, big problem, disaster waiting to happen, already a disaster, complete disaster, total failure, absolute failure, epic disaster, massive disaster, horrible outcome, terrible outcome, awful outcome, idiot, moron, dumbass, asshole, bastard, bitch, piece of shit, piece of trash, fucking, fuck, fucked up, shitshow, clusterfuck