Annotation as Cultural Interpretation: Rethinking Data Labeling in NLP

Wajdi Zaghouani
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Human annotation is a foundational component of modern natural language processing (NLP). Labeled datasets underpin widely used benchmarks for sentiment analysis, toxicity detection, hate speech classification, and stance detection. Within standard NLP workflows, annotation is generally treated as a technical process aimed at recovering an objective ground truth according to predefined guidelines. This paper argues that such a view overlooks the inherently interpretive nature of annotation. Drawing on insights from sociolinguistics, discourse analysis, and cultural theory, and on a growing empirical literature on annotator subjectivity, we propose that annotation should be understood as a culturally situated interpretive practice. Annotators rely on culturally shaped norms, values, and communicative expectations when interpreting linguistic meaning, and labels in NLP datasets often reflect culturally specific interpretations rather than universal truths. We position this argument relative to recent work on perspectivism, annotator-aware modeling, and cross-cultural annotation, and we use published findings from large-scale cross-cultural annotation studies to illustrate the concrete consequences of treating annotation as objective. We close with a research agenda for culturally informed annotation practice that includes operational recommendations on documentation, modeling, and evaluation.

1 Introduction

Human annotation is central to the development of supervised NLP systems. Many widely used benchmarks rely on annotated datasets in which human annotators assign labels to text according to predefined task categories. These labels serve as training signals for machine learning models and as ground truth for evaluation.

Standard annotation workflows assume that annotators can reliably recover the correct label for each example when guidelines are sufficiently clear. Disagreement among annotators is typically treated as noise arising from ambiguous instructions, poorly defined categories, or annotator error. Consequently, dataset construction practices prioritize consistency and inter-annotator agreement, measured through metrics such as Cohen's kappa (Cohen, 1960) and Krippendorff's alpha (Krippendorff, 2004).

However, linguistic meaning is deeply embedded in social and cultural contexts (Duranti and Goodwin, 1992). Expressions that appear neutral in one cultural setting may carry very different connotations in another. Judgments about whether language is offensive, sarcastic, polite, or harmful depend on socially shared norms and culturally specific expectations about communication (Hymes, 1974). These are not incidental features of language use; they are constitutive of meaning itself.

This paper argues that annotation should not be understood as the identification of objective labels but as a form of cultural interpretation. Annotators inevitably interpret linguistic meaning through the lens of their cultural experiences, social identities, and normative assumptions. Annotation outcomes therefore reflect interpretive frameworks rather than universal ground truth (Aroyo and Welty, 2015).

Recognizing the interpretive nature of annotation has direct implications for how NLP datasets are constructed, documented, and evaluated. When annotation practices implicitly assume universality, culturally specific interpretations can become embedded within datasets and reproduced by machine learning models, contributing to the biases that have been documented across a range of NLP applications (Blodgett et al., 2020).

Contributions and positioning. The argument we develop builds on a line of work that has questioned the ground-truth assumption in NLP annotation, including Aroyo and Welty (2015), Plank

Proceedings of the 4th Workshop on Cross-Cultural Considerations in NLP (C3NLP 2026), pages 1–10

et al. (2014), Plank (2022), and Díaz et al. (2022). Compared with that earlier work, our contribution is threefold. First, we offer a sustained conceptual framing of annotation as cultural interpretation, drawing more explicitly on sociolinguistics and linguistic anthropology than prior NLP-internal discussions. Second, we connect this framing to a small but growing empirical literature on cross-cultural annotation, including Davani et al. (2024), and use published cross-cultural results to substantiate claims about prevalence and consequence. Third, we translate the framing into operational recommendations for documentation, modeling, and evaluation that go beyond the high-level data documentation proposed by Bender and Friedman (2018) and Gebru et al. (2021) to address culturally specific aspects of annotation design.

2 Annotation in Contemporary NLP

Annotation is typically framed as a technical process designed to produce reliable labeled data. Standard pipelines involve defining a task, developing annotation guidelines, recruiting annotators, assigning labels, and measuring inter-annotator agreement. High agreement is interpreted as evidence that the annotation scheme captures a stable underlying signal in the data.

While such procedures are essential for building large-scale datasets, they rely on an implicit assumption that the categories used for annotation correspond to objectively identifiable properties of text. Under this assumption, the role of annotation guidelines is to minimize subjectivity and ensure consistent application of labels.

Aroyo and Welty (2015) challenge this assumption directly, arguing that disagreement among annotators often reflects genuine ambiguity in language rather than annotator error. They propose the concept of *crowd truth*, which treats inter-annotator disagreement as informative signal rather than noise. Plank (2022) extends this argument by surveying the landscape of human label variation across NLP tasks, concluding that annotator disagreement is not uniformly distributed but systematically structured by social, cultural, and linguistic factors that existing annotation pipelines are not designed to capture. The associated *perspectivist* view of annotation holds that multiple valid labelings of the same item can coexist, and that collapsing them into a single gold label erases meaningful information about the range of interpretations a text supports. Davani et al. (2022) operationalize this view in modeling, showing that multi-annotator architectures that retain per-annotator labels match or exceed majority-vote baselines on subjective classification tasks while better capturing predictive uncertainty.

Not all NLP tasks are equally susceptible to cultural interpretive variation. Tasks involving factual extraction, named entity recognition, or syntactic parsing are relatively low in cultural interpretive load: the categories being assigned are constrained by linguistic form rather than by social judgment. By contrast, tasks involving offensive language detection, hate speech classification, politeness recognition, sarcasm detection, stance detection, and emotion labeling are high in cultural interpretive load because their label boundaries are defined by normative and pragmatic standards that vary across communities. These high-load tasks form the backbone of many safety-critical NLP applications, which makes the cultural situatedness of their annotation especially consequential.

Waseem (2016) demonstrated that annotator identity influences labeling outcomes in hate speech tasks. In her comparison of expert and amateur annotators, systematic differences in labeling emerged based on the annotators' awareness of feminist and anti-racist discourse. Labels in socially sensitive NLP tasks therefore reflect the interpretive frameworks of annotators rather than purely objective properties of the text. Plank et al. (2014) earlier showed, in a part-of-speech tagging setting, that annotator disagreement is not uniformly distributed but correlates with linguistic phenomena that are genuinely ambiguous. They argue for models that incorporate annotator uncertainty rather than treating all disagreement as noise.

3 Linguistic Meaning and Cultural Context

Research in sociolinguistics and linguistic anthropology has long established that meaning emerges through social interaction and shared cultural knowledge (Hymes, 1974). Language is not merely a vehicle for conveying information but a medium through which communities express values, identities, and norms. Duranti and Goodwin (1992) develop the notion of context as a central organizing principle of linguistic interaction, arguing that utterance meaning cannot be determined apart from the social and physical environments in which

communication takes place.

Cultural context shapes how linguistic expressions are interpreted in pervasive ways. Communicative norms regarding politeness, criticism, and humor vary substantially across societies. Work in cross-cultural pragmatics has documented variation in how speakers use and interpret indirect speech acts, face-threatening moves, and markers of solidarity (Grice, 1975). Indirect speech may be interpreted as respectful deference in some communities, while direct assertion may be expected and valued in others.

Sarcasm and irony depend on culturally shared assumptions about what is appropriate or expected in a given situation. Without this background knowledge, an utterance may be interpreted literally rather than ironically, or vice versa. Because sarcasm is pervasive in social media text, tasks that require its detection place substantial demands on annotators' familiarity with the communicative practices of the relevant community.

These observations extend to potentially harmful language. Sap et al. (2019) showed that NLP systems for hate speech detection perform markedly worse on text produced by African American speakers, in part because the datasets used to train such systems underrepresent African American English (AAE) and reflect annotation practices that do not account for the pragmatic and cultural specificity of that variety. Linguistic features associated with AAE are systematically misclassified as offensive.

Linguistic interpretation cannot be fully separated from the cultural frameworks through which language is understood. Annotation is therefore inherently interpretive: annotators do not simply identify objective features of text but interpret language according to culturally shaped expectations about meaning and communication.

4 Annotation as Cultural Interpretation

If annotation involves interpretation, then dataset labels should be understood as culturally situated judgments rather than objective facts. We highlight four dimensions of annotation that bring this interpretive character into focus.

Normative evaluation. Many NLP tasks require annotators to evaluate whether language violates social norms. Determining whether an expression constitutes hate speech or offensive language involves assessing whether it causes harm or violates expectations of acceptable discourse. These judgments depend on culturally specific moral frameworks and on the annotator's position within social structures of power and identity (Hovy and Spruit, 2016). Davidson et al. (2017) highlight a related problem inside a single English-language setting: hate speech annotation pipelines that rely on lexical cues conflate hate speech with offensive but non-hateful language, with downstream consequences for which utterances and which speakers get flagged.

Pragmatic inference. Tasks such as sarcasm detection and stance classification require annotators to infer speaker intent. Interpreting such cues relies on background knowledge about social practices, communicative conventions, and intertextual references that may not be shared across cultural contexts (Grice, 1975). An annotator unfamiliar with the conventions of a particular online community may consistently misread irony, understatement, or coded language.

Social positioning. Annotators bring their own identities, experiences, and cultural backgrounds to the annotation process. Waseem (2016) showed that annotators with greater awareness of structural inequalities apply labels differently from those without such awareness. Blodgett et al. (2020) argue more broadly that NLP research embeds assumptions about whose language practices are treated as normative, with consequences for how models perform across different communities of speakers. Díaz et al. (2022) formalize this point in a documentation framework, recommending that dataset releases record the lived experiences and platform conditions that shape annotator judgments.

Interpretive variability. Aroyo and Welty (2015) demonstrate that what is typically treated as annotator disagreement often reflects genuine interpretive variability in language. Preserving this variability rather than collapsing it into a single consensus label may yield a more accurate representation of how language is understood across populations (Plank, 2022; Davani et al., 2022).

These dimensions indicate that annotation outcomes reflect interpretive frameworks rather than neutral observations about the world. This reconceptualization is consistent with the broader recognition in the social sciences that classification is not a passive act of discovery but an active process of meaning-making (Duranti and Goodwin, 1992).

5 Cross-Cultural Annotation: Challenges and Evidence

The interpretive dimensions identified in the previous section become especially consequential when annotation tasks are deployed across cultural and linguistic boundaries. This section examines the empirical evidence for cultural variability in annotation outcomes, drawing on cases from multilingual NLP, offensive language detection, and sentiment analysis.

5.1 The Global Distribution of Annotation Labor

Contemporary NLP research relies extensively on crowdsourcing platforms to recruit annotators at scale. The populations accessible through major crowdsourcing platforms are far from globally representative. Joshi et al. (2020) document deep structural inequalities in the linguistic resources available for the world's languages, showing that the overwhelming majority of NLP datasets and models are developed for a small number of high-resource languages, most of them European. This skew in resource development reflects and reinforces a corresponding skew in annotation labor: the annotators whose judgments shape widely used NLP datasets are disproportionately drawn from a narrow range of cultural and linguistic backgrounds.

This concentration has direct consequences for what cultural norms get encoded as ground truth. When the annotators who label a hate speech or sentiment dataset predominantly share a particular cultural background, the labels they produce reflect the communicative expectations and normative frameworks of that background. Models trained on those labels inherit those frameworks and apply them to all users, regardless of cultural context (Hovy and Spruit, 2016). Díaz et al. (2022) argue that disclosing the platform conditions and demographic composition of annotation pools is a precondition for downstream users to assess the scope conditions of the resulting labels.

Nekoto et al. (2020) make a related argument in the context of machine translation for African languages, showing that effective language technology development for low-resource communities benefits from participatory approaches that genuinely involve members of those communities rather than relying on externally imposed annotation standards. The quality and cultural appropriateness of annotation outcomes depend not just on the technical design of the annotation task but on who is doing the annotating and under what conditions.

5.2 Empirical Evidence of Cross-Cultural Variation

The argument that annotation reflects culture rather than recovering universal truth has, in the last few years, moved from conceptual claim to empirical finding. Davani et al. (2024) present the most comprehensive evidence to date in their D3CODE study, a parallel cross-cultural annotation of more than 4,500 English sentences for offensive language by over 4,000 annotators distributed across 21 countries and grouped into eight geo-cultural regions. Annotators were also surveyed on moral foundations including care, equality, proportionality, authority, loyalty, and purity. The study reports substantial regional variation in offensiveness judgments on identical items, with that variation aligning systematically with annotators' moral value profiles rather than being attributable to noise. This is direct empirical confirmation of two claims that earlier work could only argue conceptually: that culturally shaped values predict labeling behavior on subjective tasks, and that aggregating annotators into a single gold label discards information about which interpretations correspond to which communities.

These findings situate the case of African American English documented by Sap et al. (2019) as one instance of a broader pattern. Pragmatic conventions governing insult, irony, and criticism differ across communities within a single language, and annotation schemes developed without attention to this variation produce labels that reflect the norms of whichever variety is most represented in the annotator pool. The D3CODE results extend this to the cross-national scale and supply the kind of systematic evidence that earlier critiques flagged as missing.

A natural concern is whether these findings generalize beyond offensive language. Evidence from sentiment and emotion annotation suggests they do. Cultural linguistics has long established that the conceptualization and expression of emotion vary substantially across cultures (Wierzbicka, 1997), and categories such as "anger", "sadness", or "pride" do not map uniformly onto the emotional concepts available in all languages and cultural communities. When annotators from one cultural background are asked to label the emotional con-

tent of text produced within a different cultural context, systematic mismatches can arise between the intended communicative act and the assigned label. Translation does not fully resolve the underlying conceptual differences. Multilingual sentiment datasets may impose categorical structures that fit some cultural contexts more naturally than others, disadvantaging users whose affective expression does not conform to the dominant labeling schema.

5.3 Operationalizing Culture in Annotation Research

A productive direction for future research is to move beyond intuitive appeals to cultural difference toward more precise operationalization of which cultural dimensions predict annotation variance for which tasks. Hofstede (2001) proposed dimensions such as power distance, individualism versus collectivism, and uncertainty avoidance that characterize systematic differences in values across cultural groups. While this framework has attracted critique for treating cultural groups as internally homogeneous and for its original reliance on corporate survey data, it provides testable hypotheses about annotation behavior. Annotators from cultural contexts where group harmony is highly valued may approach the labeling of interpersonal conflict differently from those operating within more confrontational communicative norms; annotators with different orientations toward authority may apply labels for respectful versus disrespectful language in systematically divergent ways.

More psychometrically robust alternatives exist. Schwartz's theory of basic human values (Schwartz, 1992), which identifies ten motivationally distinct value types validated across a wide range of cultures, offers a finer-grained account of value variation that may prove predictive of labeling behavior on tasks with explicit moral dimensions. Moral Foundations Theory, which posits distinct foundations including care, fairness, loyalty, authority, and purity, is another validated instrument suited to annotation tasks that require evaluating harm or norm violation. Davani et al. (2024) provide direct evidence for the value of this approach by showing that moral foundations measured at the annotator level explain regional variation in offensiveness judgments. A natural next step is the systematic application of such instruments across a broader range of tasks and language settings.

The practical implication is that annotator profiling for culturally sensitive tasks should go beyond national origin or first language to include validated measures of relevant value orientations. Even coarse-grained collection of such metadata, administered as a brief pre-annotation survey, would substantially improve researchers' ability to interpret label distributions and to identify cultural subgroups within annotator pools whose judgments diverge systematically.

6 Cultural Assumptions in NLP Datasets

When annotation processes assume a universal ground truth, culturally specific interpretations become embedded in datasets and in the models trained on them. This dynamic has been documented across a range of NLP tasks and has attracted growing critical attention (Blodgett et al., 2020; Bender et al., 2021).

One consequence is the standardization of particular cultural norms within datasets. Datasets constructed primarily within a single cultural or linguistic environment may implicitly encode that community's expectations about politeness, offensiveness, or acceptable discourse. When these datasets are used to train or evaluate systems intended for broader deployment, the embedded norms are applied beyond the contexts that generated them (Bender and Friedman, 2018).

A related consequence is the misinterpretation of language produced by communities with different communicative practices. Sap et al. (2019) demonstrated that this problem is not hypothetical: systems trained on hate speech datasets assembled without attention to dialectal and cultural variation perform substantially worse on text produced by minority language communities. Models inherit the interpretive frameworks of the annotation process and reproduce them at scale. The cross-national evidence from Davani et al. (2024) indicates that the same dynamic operates at the global level, not only across dialects within a single country.

These issues affect model generalization. NLP systems trained on culturally narrow datasets may perform poorly when applied to texts originating from other social contexts (Hovy and Spruit, 2016), systematically misclassifying expressions that deviate from the norms represented in the training data and producing errors that disproportionately affect already-marginalized communities. Beyond performance, there are deeper questions about what it means for a model to be evaluated as accurate when accuracy is measured against labels that themselves

reflect contested cultural judgments. Bender et al. (2021) argue that the apparent fluency of large language models can obscure the ways in which those models have absorbed and reproduce culturally specific assumptions. Dataset documentation practices have begun to address these issues, but they have not yet resolved them (Gebru et al., 2021; Díaz et al., 2022).

The problem is compounded in multilingual NLP. Cross-lingual transfer often propagates the cultural assumptions embedded in high-resource language datasets into lower-resource language settings. When annotators working in a lower-resource language are trained on guidelines developed for a high-resource language, the resulting labels may not adequately capture the pragmatic and cultural specificity of the target language community.

7 Implications for Dataset Design

Understanding annotation as cultural interpretation suggests several concrete directions for improving dataset construction practices. These directions do not require abandoning the goal of building reliable labeled datasets; they require expanding what counts as reliability to include transparency about interpretive context and representation of interpretive diversity. The recommendations below are intended as operational complements to existing documentation frameworks (Bender and Friedman, 2018; Gebru et al., 2021; Díaz et al., 2022).

Transparent documentation. Dataset documentation should include detailed information about annotator demographics, cultural contexts, and recruitment procedures, alongside the task definitions and labeling guidelines that are typically reported (Gebru et al., 2021). Bender and Friedman (2018) propose data statements as a mechanism for recording the social context of dataset construction, and Díaz et al. (2022) extend this with crowd-specific documentation covering platform conditions and labor relations. We recommend that, at minimum, dataset releases for high-interpretive-load tasks report (i) the geographic distribution of annotators by country, (ii) the recruitment platform used and any pre-screening criteria applied, (iii) the language and dialect background of annotators, and (iv) where ethically feasible, an aggregate summary of annotators' value orientations using a validated instrument such as Moral Foundations or Schwartz's basic values. This minimum imposes modest additional cost and substantially improves the interpretability of labels for downstream users.

Representation of disagreement. Rather than forcing consensus labels through adjudication or majority vote, datasets can preserve annotation distributions that reflect interpretive diversity (Aroyo and Welty, 2015; Plank et al., 2014). This treats disagreement as signal rather than noise and enables training with distributional or soft label targets, an approach for which Davani et al. (2022) report performance comparable to or better than majority-vote baselines on seven binary subjective tasks. Evaluation metrics that reward calibration to the full distribution of human judgments, rather than accuracy against a single gold label, are better suited to tasks where ground truth is culturally contested (Plank, 2022).

Culturally informed guidelines. Annotation guidelines should explicitly acknowledge the normative assumptions embedded within labeling categories. For tasks involving offensive language or hate speech, this means specifying whose standards of harm and acceptability are being applied, and why those standards were chosen. Providing culturally specific examples and counterexamples reduces the risk that annotators from different backgrounds will apply the same label to systematically different phenomena.

Annotator diversity. Recruiting annotators from diverse cultural and linguistic backgrounds is a substantive requirement for high-interpretive-load tasks. This requires investment in recruiting pipelines that reach beyond the convenience samples typically drawn from crowdsourcing platforms dominated by speakers of a small number of high-resource languages (Waseem, 2016; Joshi et al., 2020). The D3CODE protocol of stratifying annotators by geo-cultural region and gender (Davani et al., 2024) offers one concrete model. Where genuine human annotator diversity is not achievable within available budgets, the limitations this imposes on label validity should be documented rather than treated as a solved problem.

Caution regarding synthetic annotation. An emerging practice involves using large language models prompted with cultural personas as a low-cost substitute for human annotators from underrepresented communities. While this approach may have limited exploratory value, we caution against

treating synthetic persona annotations as equivalent to genuine human annotation for training or evaluation purposes. Language models reproduce the cultural assumptions embedded in their pretraining data (Bender et al., 2021), and prompting them with cultural identities does not guarantee that the resulting annotations reflect the actual communicative norms and values of the represented community. For safety-critical tasks such as hate speech detection, reliance on synthetic cultural annotations risks encoding model-derived stereotypes about cultural groups rather than genuine community perspectives.

Cross-cultural evaluation. Models should be evaluated across multiple cultural contexts to assess whether they generalize beyond the norms encoded in the training data (Hovy and Spruit, 2016). Evaluation benchmarks constructed within a single cultural context provide limited evidence about cross-cultural robustness. Where annotator metadata permits it, stratified reporting of model performance by annotator cultural subgroup gives a more informative picture of where cultural misalignment is most severe.

8 Research Agenda

The reconceptualization of annotation as cultural interpretation opens research directions spanning NLP, sociolinguistics, and cultural anthropology. We organize these around three priorities: understanding culture-annotation intersections, detecting and mitigating cultural harms, and building cross-culturally competent systems.

8.1 Perspectivist and Disagreement-Aware Modeling

A foundational direction involves developing modeling frameworks that handle interpretive plurality from the outset rather than treating disagreement as a preprocessing problem to be resolved before training. Plank (2022) surveys recent efforts in this direction, including multi-annotator models that maintain per-annotator label representations, soft-label training objectives that minimize divergence from the full annotation distribution, and uncertainty-aware evaluation metrics that reward calibration to human label distributions. Davani et al. (2022) provide a concrete instantiation, showing that multi-annotator architectures yield comparable or better performance than majority-vote baselines while better capturing predictive uncertainty.

The cultural dimension adds specificity to this agenda: the goal is not merely to preserve any disagreement but to understand which dimensions of annotator identity and cultural background predict which patterns of disagreement across which task types. Studies that cross annotator cultural background with task type and disagreement pattern, in the manner of Davani et al. (2024), would extend the empirical foundation for targeted deployment of perspectivist modeling.

8.2 Detecting and Mitigating Cultural Harms

Detection of cultural harm in datasets requires auditing tools that correlate label distributions with annotator cultural metadata. Such audits presuppose that datasets include this metadata, which in turn requires changes to current data collection practices (Gebru et al., 2021; Bender and Friedman, 2018; Díaz et al., 2022). Privacy considerations are real: collecting demographic and value-orientation data from annotators requires informed consent and careful data governance, particularly when annotators are drawn from communities that may be subject to surveillance or discrimination. Minimum viable metadata collection should balance informativeness with participant protection, prioritizing aggregate reporting over individual-level disclosure.

Mitigation strategies range from upstream interventions in annotation design to downstream debiasing in model training. Upstream approaches are preferable in principle because they address the source of cultural narrowness rather than attempting to correct for it post hoc. Research that compares mitigation strategies on matched datasets would help establish when each is most effective.

8.3 Community-Centered Annotation and Interdisciplinary Methods

Participatory approaches to annotation, in which community members act as co-designers of labeling schemes rather than as a labor source, offer a promising path toward more culturally appropriate labels for low-resource and underrepresented communities (Nekoto et al., 2020). Practically, this involves community consultation on category design, piloting guidelines with community representatives before deployment, and feedback mechanisms for reviewing externally produced labels. Ethnographic and discourse analytic methods from sociolinguistics can complement quantitative inter-annotator agreement studies by illuminating how

annotators reason about edge cases and what cultural assumptions their reasoning reflects (Hymes, 1974; Duranti and Goodwin, 1992).

8.4 Cross-Cultural Robustness Metrics

A concrete methodological gap is the absence of standardized metrics for cross-cultural robustness in NLP evaluation. Current benchmarks measure accuracy against a single culturally specific gold standard. More informative alternatives include stratified performance reporting by annotator cultural subgroup, agreement metrics computed separately within and across cultural groups, and robustness measures that quantify performance degradation as a function of cultural distance between training and test annotator pools. Developing and standardizing such metrics would give the research community tractable targets for progress and would make cross-cultural generalization a first-class evaluation criterion rather than an afterthought.

8.5 Scope: Where the Argument Applies Most Strongly

We close this section with an explicit note on scope. The case for annotation as cultural interpretation is strongest for tasks high in interpretive load, such as offensive language detection, hate speech classification, sentiment and emotion analysis, sarcasm detection, stance detection, and politeness recognition. It applies with progressively less force to tasks that depend more directly on linguistic form, such as part-of-speech tagging, named entity recognition for non-ambiguous categories, and morphological analysis, although even there genuine ambiguity exists (Plank et al., 2014). The recommendations on annotator diversity and disagreement preservation are most consequential for the high-load tasks; for lower-load tasks, they remain useful but cannot be expected to drive the same magnitude of effect on label distributions or model behavior.

9 Conclusion

Annotation is a central component of NLP dataset construction, yet its interpretive nature is systematically underacknowledged in standard practice. This paper has argued that annotation should be understood as a culturally situated practice shaped by social norms, communicative expectations, and shared cultural knowledge. Drawing on foundational work in sociolinguistics and linguistic anthropology (Hymes, 1974; Duranti and Goodwin, 1992), and on a growing empirical literature on cross-cultural annotation (Davani et al., 2024, 2022), we have shown that the interpretive situatedness of annotation is not an artifact of poor design but a constitutive feature of linguistic meaning. Cultural dimensions of meaning cannot be stripped away through more careful guideline writing; they can only be acknowledged and accounted for in the design of annotation processes.

The practical recommendations we advance, including transparent documentation of annotator context, representation of interpretive disagreement, culturally informed guidelines, diverse annotator recruitment, and cross-cultural evaluation, do not require abandoning the goal of reliable labeled datasets. They require expanding what counts as reliability to include fidelity to the genuine diversity of linguistic meaning across communities. The labels that supervise our models are produced by human beings whose interpretive frameworks are culturally situated. Addressing this fact is a prerequisite for developing NLP technology that genuinely serves the full range of communities it purports to benefit.

Limitations

This paper offers a conceptual reframing of annotation supported by published empirical findings rather than presenting new annotation experiments of our own. Although we draw on work that provides direct evidence for our claims, including the cross-cultural offensiveness study of Davani et al. (2024), the cross-task generality of those findings remains an open empirical question. Future work should test the proposed framework through new studies that systematically vary annotator background and cultural context across a wider range of NLP tasks, languages, and label schemes.

The literature we draw on is weighted toward English-language work and toward tasks involving offensive and pragmatic interpretation. The degree to which our arguments generalize to other language families, annotation tasks, and annotation regimes such as expert linguistic annotation in formal corpora is partly an open question. The structural imbalance in available evidence is itself a manifestation of the problem we identify.

The practical recommendations we advance involve real costs and trade-offs that we have not fully analyzed. Recruiting culturally diverse annotators is expensive; collecting validated metadata on annotator value orientations adds time and

raises privacy considerations; preserving full annotation distributions has implications for storage, evaluation tooling, and downstream model training. Implementation requires careful attention to the specific constraints of each annotation context. Finally, the framework as presented does not provide a quantitative procedure for deciding when sufficient cultural coverage has been achieved for a given task, and developing such criteria is itself an open research problem.

Ethical Considerations

The arguments advanced in this paper have implications for the ethical design of NLP systems. If annotation encodes culturally specific interpretations as universal ground truth, then systems trained on such annotations may perpetuate and amplify biases against communities whose linguistic practices and communicative norms differ from those represented in the annotator pool. This has the potential to cause concrete harm, particularly for communities that are already subject to marginalization.

We encourage NLP practitioners to treat the cultural situatedness of annotation as an ethical consideration alongside technical reliability. Dataset releases should include documentation that enables downstream users to make informed decisions about the scope conditions of the labels. The collection of annotator metadata, including value-orientation data, must be governed by informed consent and appropriate data protection, with aggregate rather than individual-level disclosure as the default (Díaz et al., 2022). Funders and institutions that support NLP research should consider whether the annotation practices used in funded work adequately account for the interpretive diversity of the communities the resulting systems will affect.

We recognize that calling attention to the cultural specificity of annotation creates a risk of being read as undermining the validity of labeled datasets altogether. That is not our intention. Labeled datasets remain essential for NLP research, and the goal of this paper is to promote practices that make annotation more transparent and more equitable, not to abandon the project of labeled data collection.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526.

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

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 2342–2351. ACM.

Alessandro Duranti and Charles Goodwin, editors. 1992. *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge University Press, Cambridge.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Geert Hofstede. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2nd edition. Sage, Thousand Oaks, CA.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Dell Hymes. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage, Thousand Oaks, CA.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press, New York.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142.

Anna Wierzbicka. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press, New York.