Culture is Not Trivia: Sociocultural Theory for Cultural NLP

Anonymous ACL submission

Abstract

The field of cultural NLP has recently experienced rapid growth, driven by a pressing need to ensure that language technologies are effective and safe across a pluralistic user base. This work has largely progressed without a shared conception of culture, instead choosing to rely on a wide array of cultural proxies. However, this leads to a number of recurring limitations: coarse national boundaries fail to capture nuanced differences that lay within them, limited coverage restricts datasets to only a subset of usually highly-represented cultures, and a lack of dynamicity results in static cultural benchmarks that do not change as culture evolves. In this position paper, we argue that these methodological limitations are symptomatic of a theoretical gap. We draw on a well-developed theory of culture from sociocultural linguistics to fill this gap by 1) demonstrating in a case study how it can clarify methodological constraints and affordances, 2) offering theoretically-motivated paths forward to achieving cultural competence, and 3) arguing that localization is a more useful framing for the goals of much current work in cultural NLP.

1 Introduction

Language and culture are closely linked: language can be conceptualized simultaneously as an artifact of culture as well as a process through which culture is created (Ochs, 2009). As language technologies become increasingly integrated into the everyday lives of a diverse set of users, it is imperative that they are robust to cultural differences between user bases (Hershcovich et al., 2022).

Cultural NLP, sometimes also known as cultural alignment, is a subfield within the NLP and ML communities that has experienced drastic growth in recent years to meet this challenge. Work in cultural NLP usually involves building or evaluating systems that 1) have knowledge of cultural facts and 2) apply this knowledge appropriately in specific situations where cultural knowledge is relevant (Adilazuarda et al., 2024; Liu et al., 2024b). This can include building new evaluation benchmarks or fine-tuning datasets that contain cultural knowledge of some kind, either manually (Lee et al., 2024; Koto et al., 2024) or automatically from a large corpus (Shi et al., 2024; Wang et al., 2024a), or creating systems that generate culturally-relevant output (Khanuja et al., 2024).

Most work relies on various proxies for defining both cultural boundaries and cultural objects. Proxies of cultural boundaries commonly include nationality, religion, ethnicity, or other demographic features. Proxies for cultural objects might include culture-specific knowlege of foods, values, or norms (Zhou et al., 2024a; Sorensen et al., 2024; Dwivedi et al., 2023). These works constitute an important step forward in understanding how to build fairer, more inclusive language technologies. However, the disparate array of cultural proxies being evaluated is symptomatic of a theoretical gap: to achieve culturally-competent NLP systems, we must make progress towards a clearer, more unified conception of culture, and what it means for the systems we build to be responsive to that.

Fortunately, cultural NLP is not alone in the search for a useful notion of culture, and its theoretical challenges are not new. Dissatisfaction with the coarseness of demographic cultural boundaries led to the second wave of sociolinguistics, which refocused efforts on identifying local cultural meaning within communities of practice (Eckert, 2012). Larger questions, like the utility of the culture concept, have been debated in fields like sociocultural anthropology, where some researchers have abandoned culture altogether as being essentializing and othering (Vann, 2013). Indeed, epistemological and empirical tensions as they relate to the study of culture have been grappled with across such fields as anthropology, sociolinguistics, sociology, cultural studies, among many others.

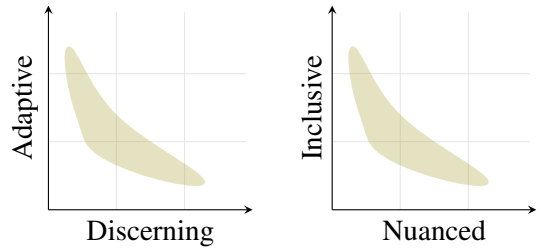

Figure 1: Two separate spaces of desiderata; the **left** represents aspects of *competence*, while the **right** represents aspects *coverage*. In each space, there often exists a trade-off between the axes, so that most cultural NLP work falls into the conceptual area that is shaded.

Contributions In this paper, we draw on theoretical developments in *sociocultural linguistics* (Bucholtz and Hall, 2005) — itself a collection of several adjacent disciplines — to clarify the status of cultural knowledge in building culturally competent NLP systems.

We first review the goals of cultural NLP, and enumerate specific desiderata for culturally-aware language technologies that current works pursue (§2). Then, we highlight recurring difficulties in cultural NLP (§3) by providing a survey of common self-stated limitations in existing papers. We introduce sociocultural linguistics as a field with a useful theoretical framework which we can apply to better understand culture as an object of study (§4), and provide a case study to illustrate how the theory of *indexicality* can be applied to clarify the distinction between learning cultural knowledge and learning stereotypes (§5). Finally, we discuss the broader implications granted by this understanding of culture, offering two main claims. First, we highlight existing methodological and theoretical gaps in achieving the ambitious goal of cultural competence, and provide theoretically-motivated suggestions for making progress on the task (§6.1). Then, we argue that the goal in cultural NLP might be reasonably understood as localization instead of cultural competence or understanding, providing a more tractable and situated framing with which to build useful NLP systems (§6.2).

2 The goals of cultural NLP

Though the field of cultural NLP does not necessarily agree on a definition of culture, there is general agreement on the goal: to build culturally-*competent* NLP systems (Bhatt and Diaz, 2024). Here, we break down this high-level goal into several more specific desiderata that are frequently mentioned in cultural NLP papers. We want our language technologies to be:

Adaptive. A foundational premise of cultural NLP is that language technologies should be culturally *sensitive*. In other words, culturally competent language technologies should be responsive to specific cultural contexts when designing their outputs. It would be insufficient for an NLP system to produce the same output for all cultural contexts; many works on bias in NLP have shown and problematized the tendency of language technologies to represent, exaggerate, and perpetuate a hegemonic set of values and structures (Voigt et al., 2018; Sheng et al., 2019; Bender et al., 2021).

There has been relatively limited exploration in building systems that are explicitly reflexive (Sorensen et al., 2024). Instead, a wide body of work focuses on assessing whether NLP systems can generate different, appropriate outputs in response to different cultural contexts — by answering value-oriented survey questions in a manner consistent with the target culture, for example (Cao et al., 2024a; Huang and Yang, 2023). Some works that center adaptation as a value focus on *extrinsic* evaluation: instead of probing whether language models *know* specific cultural facts, they test whether NLP systems *respond* in a way that demonstrates this knowledge (Bhatt and Diaz, 2024).

Discerning. At the same time, there is also a desire that NLP systems not perpetuate reductive stereotypes. Past work has demonstrated that users have differing expectations for cultural adaptation (Lucy et al., 2024), and that cultural adaptation is not equally desired in all settings. For example, users may want technologies to understand their regional or ethnic dialects, but not generate them (Blaschke et al., 2024). This has motivated work on stereotype mitigation (Jha et al., 2023; Ma et al., 2023), in which datasets of harmful stereotypes are collected in order to evaluate or engineer systems to avoid generating them.

Inclusive. Cultural NLP values *breadth*: language technologies should perform well across a large number of cultures. This value is represented by many works in the genre which build benchmarks for a large number of different cultures (Bhutani et al., 2024); these papers usually use nationality and surveys as tractable ways of achieving large scale (Zhao et al., 2024; Ramezani and Xu, 2023). Some text mining methods for accumulat-

ing cultural knowledge also reflect the emphasis on large-scale, broad coverage (Fung et al., 2024; Nguyen et al., 2023).

Nuanced. In addition to breadth, there is a desire for depth in the form of more granular and extensive cultural understanding. The value of nuance motivates works which build resources for specific languages or resources (Koto et al., 2024; Son et al., 2024; Li et al., 2024b) and explore locally meaningful cultural categories (Dev et al., 2023). These works might rely on local informants to provide cultural knowledge (Koto et al., 2024), trading breadth of coverage for richer cultural knowledge that large-scale survey-based methods cannot capture.

Though these four desiderata are not mutually exclusive, they coalesce into two sets of values that are largely unrelated (visualized in fig. 1). The first two desiderata reflect two kinds of cultural *competence*: the knowledge of how to respond differentially, and the knowledge of when it is appropriate to do so. The second two desiderata reflect an orthogonal value of cultural *coverage*: we want systems that cover many cultures, as well as systems that cover many aspects of each culture.

3 Recurring troubles

Explicitly stating these desiderata can shed light on the motivations of current work, but they do not themselves offer any answers about what "culture" is. This becomes apparent when we look into the limitations sections of many cultural NLP papers, where we find recurring themes that point to challenges posed by overly narrow definitions of culture. We survey the self-stated limitations of 57 papers from 2022-2024 which explicitly mention culture, as well as the cultural proxies they use.[1] This is not meant to be an exhaustive survey, but rather illustrative of the general state of the field.

The most commonly cited limitation was one of coverage (40% of papers): the dataset or evaluation being presented was only collected with respect to a small subset of cultures. Partly, this can be explained by the proxies being used for setting cultural boundaries. Of the papers we surveyed, 36% of them used nationality as a demographic proxy (Adilazuarda et al., 2024). However, many papers problematize this choice in the limitations section, since nations are politically defined and not culturally homogeneous (Bickham et al.), and language labels usually reflect a hegemonic notion of a standard variety (Lippi-Green, 2011).

Another common limitation was a lack of dynamicity (12% of papers): culture is constantly constructed through social negotiation (Ochs, 2009), but benchmarks are largely static collections of examples or facts (Son et al., 2024; Keleg and Magdy, 2023; Jin et al., 2024; Li et al., 2024b). In most works, there is no granularity in the temporal dimension, failing to achieve an aspect of the desired *nuance*. Uncertainty around the definition of culture also limits nuance in cultural technologies, since most papers focus only on a small subset of culturally-relevant objects through proxies like food, etiquette, or values, without a framework to unify them. Roughly 37% of papers directly problematize their choice of a particular cultural proxy as being limited in its ability to represent culture as a whole, or too coarse to capture intragroup variation (28% of papers mention this specifically).

Finally, the tension between adaptation and discernment results in uncertainty about how to address stereotypes in data. Some papers (14%), which are largely intended for use in aligning models, view the potential of collecting stereotypes as cultural knowledge to be a limitation (Shi et al., 2024). Other papers explicitly collect stereotypes in order to build systems which can avoid generating them (Bhutani et al., 2024).

Other limitations mentioned in various works include an overemphasis on English-language data and methods, the lack of extrinsic evaluation in favor of multiple-choice knowledge tests, the use of pretrained models to construct datasets, and various concerns with crowdsourced or human-annotated data, including the possibility that individual preferences are being construed as cultural ones.

Little progress has been made on rigorously addressing these limitations. In many of these instances, there are questions in clear need of theoretical answers: how do we move past static, global categories when defining culture; how do we conceptualize culture in a way that respects its dynamic and constructed quality; how can we unify different facets of culture; how do we appropriately model and study stereotypes to build fairer systems?

4 A sociocultural solution

The notion of cultural competence is most commonly referenced in social and health services research, where culturally competent care has been

[1]The full list of papers can be found in the appendix.

encouraged as a way to reduce disparities in care quality and outcomes (Alizadeh and Chavan, 2016). Similar to cultural NLP, these works face operational challenges in identifying what cultural competence should look like (Kirmayer, 2012). Often, this literature draws on sociological and anthropological work to resolve these challenges. We will do the same here, focusing our attention on the fields of linguistic anthropology and sociolinguistics, which study language and culture in tandem.

For computational linguists, cultural competence might evoke the notion of linguistic competence, or Gumperz's more general idea of *communicative competence* (Gumperz, 1997): the "knowledge of linguistic and related communicative conventions that speakers must have to initiate and sustain conversational involvement." Gumperz argues that communication must be understood not only in the context of linguistic systems of grammar, but within a semiotically rich social space. This idea has been widely accepted and refined in both linguistic anthropology and sociolinguistics,[2] and we take *social context* as a point of departure for understanding *culture*.

4.1 Text and context

Insofar as culture can be construed as a structured social phenomenon, it makes sense to understand it as the aspects of (extralinguistic) social context which make themselves interactionally relevant. In Gumperz's terms, this context contributes to a more general level of sensemaking in an interaction. Edwards (1991) points out that even our linguistic categories are not subject only to cognitive processes, but also to social ones: the semantic category of "bird" might evoke an image of a robin or sparrow in a test-taking setting, but certainly indexes a different one at Thanksgiving dinner.[3]

Addressing culture as social context shifts the ambiguity from one term to the other. The question becomes, what do we take to be social context? It is useful to look at the evolution of sociolinguistics as another quantitative discipline in which this question is at the fore. Early sociolinguistic studies focused frequently on sociological categories like socioeconomic class (Labov, 1985; Guy, 2011) or gender (Lakoff, 1973), placing the speaker as a passive member of an externally-imposed category (Eckert, 2012). This led to concerns about the limits of coarse macrosociological categories, much like the critique of nationality in cultural NLP today. In response, the second wave of sociolinguistic research incorporated ethnographic methods to better understand *local* dynamics of language variation, relying on social networks and locally-relevant social categories. While second wave studies focused on local meaning, they still treated social categories as static, an essentializing assumption that equates identity with group affiliation. Third wave studies focus on *identity* as a performance constructed from a diversity of semiotic resources including, but not limited to, language style. Social context, then, becomes the space within which identity is constructed and performed.

This evolution represents not only theoretical developments in response to empirical challenges in sociolinguistics, but also a steady convergence of ideas with other disciplines. Bucholtz and Hall (2005) provide a well-integrated framework for analyzing language and sociocultural identity in the form of *sociocultural linguistics*, which synthesizes a convergent set of ideas from across disciplines to analyze language as well as other semiotic practices. Thus, sociocultural linguistics should not be thought of as a single theory of culture, but rather a concordant collection of theories that have seen relative convergence across fields that study language, culture, and society.

We take this framework as the point of departure for the rest of this paper. We describe the foundational concepts (§4.2) and provide a case study for how they can clarify the objects and goals of cultural NLP (§5). Then, we take a broader look at how sociocultural linguistic theory can inform computational work on cultural competence (§6.1) and advance the goals of cultural NLP (§6.2).

4.2 A primer on sociocultural linguistics

Bucholtz and Hall (2005) lay out five core principles of sociocultural linguistics. The terminology they use centers on the idea of *identity* as the "social positioning of self and other." As we explained above, this is a useful way of conceptualizing the cultural system more generally.

Emergence. Identity emerges through interaction. This is the view that language does not *come from* culture, but rather that culture is constituted *through* linguistic (and other forms of) interaction. This draws on, among others, the ideas of identity

[2] Indeed, Gumperz was greatly influential in the establishment and progress of both these fields.

[3] In the United States, turkey is often a centerpiece in the Thanksgiving meal.

performance (Butler, 1988) and audience design (Bell, 1997). Emergence supports a more *nuanced* representation of culture as one that is dynamic, and a more complex notion of *adaptation* that supports the idea that even an individual can inhabit multiple cultural roles. This offers one solution to the challenge of defining cultural categories: it may make sense to instead induce cultural categories latent in the data.

Positionality. Identity includes multiple levels of categories, including macro-level demographics, locally-specific meaning, and contextually-specific stances and styles. This is a more *inclusive* and *nuanced* notion of culture that speaks to one of the core limitations of current work. National identity is only one level at which identity occurs, and the idea of positionality insists that we understand more granular categories of identity as well, including ones that are local to a specific community or even an interaction.

Indexicality. Identity is constructed through an indexical process involving signs and their conceptual referents (Silverstein, 2003; Eckert, 2008). Indexicality offers a *mechanism* through which culture is constructed; it is the process of drawing links between linguistic (and other) forms and social meaning. These can play into cultural ideologies about language, construct stances local to specific interactions, consist of overt references to identity, and more; this provides a unified mechanism through which we can conceptualize culture. In section 5, we explore an example of how indexicality provides a useful theoretical account of stereotype in cultural NLP.

Relationality. Identity takes on social meaning in relation to other identities. This provides a useful way of conceptualizing culture that aligns with findings in machine learning: contrastive learning of feature spaces often result in stronger representations than supervised learning among predefined categories. Similarity and difference are not the only relations available within this framework, which also includes authentication-denaturalization and authorization-illegitimization, among others. If the cultural space is structured through these relations, computational methods might benefit from considering how to encode them.

Partialness. Finally, any account of culture is necessarily incomplete, since it is itself situated contextually in relation to the subject it describes. Indeed, a person's identity at a given point in time may be partially deliberate, partially habitual (and subconscious), partially attributable to perception, partially conditioned by the interactional context, and partially subject to the ideologies that surround the interaction. That there is no single ground truth is a troubling statement for those who want to build robust, generalizable systems. However, it also provides a certain freedom from a positivist mirage. Instead, researchers and system designers are encouraged to think more critically about their position, and the assumptions encoded in the technology they build, with respect to the users whom these systems impact.

5 Case study: culture from text

In §4.1, we distinguish between language (the text) as the traditional object of study in linguistics and various ways of assessing culture as the surrounding social context. It has become essentially paradigmatic within NLP that we should expect to derive extratextual information (such as world models, for example) from training on text alone. It is reasonable, then, that there is a vein of cultural NLP work that mines for facts about specific identities from culturally-centered discourse on social media or the Internet (Shi et al., 2024; Fung et al., 2024; Nguyen et al., 2023). In some cases, large language models are prompted to *generate* specific culturally-relevant scenarios (Qiu et al., 2024b). In many of these papers, authors note the dangerous potential for extracting biased or stereotyping information. How should we understand the epistemic status of these surfaced facts as cultural knowledge? In this section, we apply indexical theory to demonstrate that these works, in fact, can *only* learn stereotypes.

These papers aim to construct an indexical field. Cultural facts aggregated by these systems generally resemble the form:

> In **cultural group**, **belief** is widely accepted.

This is effectively a mapping between the space of beliefs and the cultural groups that they index. Indexicality can occur at different levels of social awareness; a *first-order* index evidences membership in a group. For example, the use of "pop" over "soda" might index membership in the population of Midwestern U.S. English speakers.[4] However,

[4] In **the Midwest**, it is widely accepted that **fizzy, sugary drinks are called "pop."**

higher-order indices occur as these associations themselves become embedded in cultural ideology. As such, the understanding that Midwesterners say "pop" is *in and of itself* a piece of cultural knowledge (Eckert, 2008).[5]

5.1 Stereotypes all the way down

The implication of this idea is that works which study cultural discourse are **primarily studying the stereotypes** embedded in the ideologies of the groups that generated this data, and only incidentally studying the cultures that are the objects of discourse. In fact, Labov (1973) defines stereotype exactly as the linguistic forms which are subject to metapragmatic discussion.

By applying the theory of indexical order, we gain clarity on the aspects of culture being studied. In the case of papers that mine cultural knowledge from cultural discourse, we find the subjects of study to be different from what we initially assumed. We are not learning about a diverse set of international cultures, but rather the world-view of the text authors, situated in a specific interactional context (perhaps posting about culture shock).

This is not just a matter of naming, nor a dismissal of the utility of these datasets. Instead, indexical theory clarifies the extent to which they are useful. It shows that these datasets exclude, by construction, cultural knowledge that is *not* subject to metapragmatic discussion. It shows that higher-order indices can still be useful because their meanings are tied to the lower-order ones from which they arise. But it also illustrates complications that we must contend with: higher-order indices might persist even when lower-order ones are no longer as salient. "Authentic" Pittsburghers, for example, might be described as unpretentious, hospitable, sports-loving, etc. But this style originally indexed the immigrant-heritage, working-class history of the formerly industrial city, an identity that may not necessarily apply to its current residents, many of whom work in the health-care or higher-education sectors (Johnstone, 2014).

5.2 Indexical values are contextual

It is also a mistake to assume that a given style from a given speaker always indexes the same thing, because the indexical value is also dependent on context, and interactionally interpreted.

Chun (2007), for example, provides an account of a "foreign speaker" language style as deployed by Asian American high schoolers. She notes how this style can be employed both as accommodation to foreign speakers (e.g., a child speaking to her immigrant parents) and as mockery (e.g., between two peers at school). Sometimes, quotatively, an utterance can even fulfill both roles depending on the interactional frame through which it is interpreted. The social meaning of an utterance is determined situationally within a specific interactional context.

6 Paths forward

Sociocultural linguistics paints a picture of culture as a complex, dynamic system through which sense-making occurs. It is one that has proven useful in accounting for and describing how semiotic systems are constructed and deployed for social action in everyday interactions.

6.1 Culturally competent NLP

But there exists a gap between this model of culture and our current computational methods for approaching culture. There is opportunity for NLP work to fill in these gaps.

Sociocultural linguistic theory tells us that culture is *emergent*, and cultural NLP acknowledges that culture is a dynamic process, but currently our datasets are limited to static snapshots of cultural artifacts. It may be fruitful to instead analyze discursive sequences in which cultural knowledge is suggested or contested. When and how are norms enforced in interaction? How is cultural knowledge shared, and how is it taken up by the rest of the community? As an example, consider this interaction between two Latina high school students from Mendoza-Denton (2008):

Lupe:	¿Qué me ves? *(What are you looking at?)*
Patricia:	Tschhh, don't EVEN talk to me in Spanish, 'cause your Spanish ain't all that.

Through contextual information like the participants' posture, make-up, and social networks (including the fact that they are rival gang members), we can understand the setting of this interaction: Patricia has interpreted Lupe's question to be a claim to authenticity. But through the interaction itself we can see the cultural process in action: as Mendoza-Denton (2008) notes, Lupe asserts

[5] In **the U.S.**, it is widely accepted that **Midwestern U.S. English speakers use "pop" over "soda."**

her Mexican-ness symbolically through her use of Spanish. Through both her assertion and Patricia's contestation, the social importance of Spanish is *reinforced* as indexing their Mexican identities. Analogous computational work might study comment threads for these kinds of interactions, and additionally incorporate contextual mechanisms like flairs or voting that users can employ to express affiliation or pass judgment on platforms like Reddit (Gaudette et al., 2021).

Sociocultural theory tells us that culture is *positional*, operating at multiple levels of identity and often composed of features from many different styles, but current methods impose coarse, usually unidimensional, boundaries like nationality on cultural categories. Relationality and indexicality offer mechanisms through which cultural sensemaking occurs — how can we better model positionality as a contextually legible field of identities by identifying instances of cultural categories being constructed in relation to other categories, or identities being assembled by combining different indexical signs? Castelle (2022) suggests that modern language models can be usefully conceptualized more generally as effective learners of semiotic systems; how might we build systems that learn representation spaces for other kinds of meaning beyond semantics, like social or discursive meaning?

Indexicality also motivates the need for datasets that are contextually rich: culture is the combination and construction of different semiotic resources that make reference to social meaning, yet our methods are deployed on datasets that largely consist of decontextualized text. Data that contains social context in other forms (e.g., metadata or other kinds of world state) could be one way of addressing this limitation (Nguyen, 2025). For example, the STAC corpus (Asher et al., 2016) consists of dialogue situated in a game scenario, and includes information about the game state and actions. This places linguistic interaction within a broader context; future works might extend this paradigm to other, more socially relevant metadata.

Indexical fields also exist beyond text, reaching into other modalities in the form of gesture, prosody, and even extralinguistic semiotic systems like fashion (Chun, 2007) and make-up (Mendoza-Denton, 2008). Not only is it important to represent non-text modalities to capture culture, but combining modalities can also be a promising direction to learning social meaning (Zhou et al., 2024b).

There is also much theoretical work at hand to account for how a software system might differ from a human in how it is taken up as an interlocutor in interaction. Creating systems that perfectly replicate human behavior is neither desirable nor felicitous. Consider this podcast transcript introducing the findings of a scientific paper:

> **Host A:** Think about those old Hollywood films, the ones your grandma might watch.
>
> **Host A:** Do those performances feel different than what you might see in movies today?
>
> **Host B:** Hm, yeah I guess they do. It's, like, more dramatic. The emotions are way more, out there?

This serves the discursive purpose of simultaneously motivating a finding and establishing rapport with the listener by drawing on the presenter's personal experience. However, if this same script is generated with an LLM,[6] the social action becomes infelicitous. The LLM has no grandmother, whose past movie-going experiences are being imagined and described. Instead, the audience must reinterpet this sequence as a post-hoc rationalization of the source material that is about to be presented, failing to motivate the finding or establish rapport. Not all semiotic resources available to humans are available to the language technologies we build.

6.2 Localized NLP

All told, we are far from building culturally competent systems, given these clear and pressing theoretical and methodological gaps. But cultural NLP also faces more immediate goals, which are perhaps more central to the field as it is currently configured. We want to create, e.g., web agents that will not make food purchases that violate religious dietary laws (Qiu et al., 2024a) or image generation models that show the local currency when displaying money (Khanuja et al., 2024). Do we need to achieve cultural competence in the general sense for these more immediate applications?

Many would argue that machine translation systems have not yet achieved linguistic competence (and this is perhaps an easier case to make in the multilingual setting). Yet, individual software applications have been internationalized long before MT achieved even its most recent success. When building systems that accommodate more users, a

[6] As, indeed, it was, by NotebookLM (Google, 2024).

more useful, immediate framing might be one of *localization* rather than cultural competence. Understanding the task at hand as building localized NLP applications helps us locate ourselves in the space of desiderata (fig. 1).

Localization is tractable because it forces us to specify the application domain, constraining the relevant depth of knowledge. Localized translations are generated only for the necessary text within an application; culturally localized systems can focus on the domain-specific *nuances* of cultural knowledge. Sociocultural approaches to culture are contextual and situated, and localization forces us to evaluate cultural performance in a situated application setting.

Localization also forces us to enumerate our audience, constraining and making explicit the *coverage* of our systems. Localized translations are not provided for an arbitrary, unconstrained set of languages or an arbitrary set of text. Furthermore, a website that offers its interface in, e.g., "Spanish" rarely allows users to choose a specific regional dialect, even though different varieties of Spanish often show lexical and syntactic variation. This is a pragmatic choice, but also an ideological one about which language varieties to support, and it is better that the ideological choices be made explicitly and transparently. In the cultural setting, this can also make the choice of cultural boundaries less arbitrary. If the goal is to build a culturally localized healthcare chatbot, for example, differing levels of medical literacy may be a more salient cultural boundary with more actionable interventions than something like nationality.

Finally, localization forces us to consider the NLP system as an interlocutor in the human-computer interaction. While many existing cultural knowledge benchmarks probe large language models removed from the specific context of how they will be used, approaching the task as localization forces us to define the expected *behavior* within a given application context. Developers of a recipe application might improve user experience by offering culture-specific ingredient substitutions (She et al., 2024), but a healthcare application might benefit from adopting a stance of cultural *humility* instead of potentially stereotyping or stigmatizing *adaptation* (Lekas et al., 2020). Defining the bounds of expected cultural performance specifies where we want a particular application to lie in the *discerning / adaptive* space.

Thus, localization allows us to focus on particulars that may be more tractably implemented and evaluated in real-world systems today.

7 Conclusion

It is important to build language technologies that are responsive to cultural values. However, the current field of cultural NLP has not found agreement on what it means to model culture, settling instead for a wide array of cultural proxies for both categories of identity and categories of indexical features. In this paper, we deconstruct the goals of cultural NLP and highlight how recurring discomforts in current work are illustrative of a lack of theoretical alignment. We propose drawing on convergent theoretical insights from a variety of social-scientific disciplines which have centered the study of culture in the context of language and other semiotic systems.

When studying such a complex, multifaceted, and dynamic object as culture, it is equally challenging and imperative that the object of study be well-defined. We demonstrate how sociocultural linguistics provides a useful theoretical framework that treats culture as an enacted process, not a static artifact. We explore the implications of this: we show that learning cultural facts through metapragmatic discourse is limited to learning about stereotypes; we make the case that building culturally competent computational systems requires a dynamic model of culture as a process, not a collection of trivia, and that sociocultural linguistics provides a powerful model of this process, but methodological and theoretical gaps still loom large; finally, we argue that many current works in cultural NLP can be usefully reframed as localization, which encourages situated, participatory design and evaluation of systems.

The growth of cultural NLP reflects the more general state of natural language processing. Until recently, NLP has been largely preoccupied with learning the semantic meaning of textual symbols. But other fields of linguistics have long established that the world around us cannot be extricated from the words we produce and interpret. Gumperz argued fifty years ago that communicative competence reaches beyond grammatical knowledge. As computational methods have become more powerful in representing textual semantic meaning, it becomes both tractable and necessary to consider other kinds of meaning, like sociocultural meaning, as equally important objects of study.

8 Limitations

Though we highlight certain representative works in making the case for a theory-led approach to cultural NLP, this paper is not meant to be a survey of the field. Adilazuarda et al. (2024) and Liu et al. (2024b) offer good overviews of recent work.

We also introduce the specific theoretical framework of sociocultural linguistics. Though other theories of culture certainly exist, we focus on sociocultural linguistics for its linguistically-oriented approach to culture (see §4 for detailed discussion about this).

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Somayeh Alizadeh and Meena Chavan. 2016. Cultural competence dimensions and outcomes: A systematic review of the literature. *Health & Social Care in the Community*, 24(6):e117–e130.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Allan Bell. 1997. Language Style as Audience Design. In Nikolas Coupland and Adam Jaworski, editors, *Sociolinguistics*, pages 240–250. Macmillan Education UK, London.

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*, pages 610–623, Virtual Event Canada. ACM.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16055–16074, Miami, Florida, USA. Association for Computational Linguistics.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Jennifer Bickham, Nancy A Naples, and NYU Press. 1. Border Politics: Contests over Territory, Nation, Identity, and Belonging.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614.

Judith Butler. 1988. Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal*, 40(4):519.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024a. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024b. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of*

the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Castelle. 2022. Sapir's Thought-Grooves and Whorf's Tensors: Reconciling Transformer Architectures with Cultural Anthropology. In *Workshop on Cultures in AI/AI in Culture at NeurIPS 2022*.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Elaine Wonhee Chun. 2007. *The Meaning of Mocking: Stylizations of Asians and Preps at a U.S. High School*. Ph.D. thesis, University of Texas, Austin, Austin, TX.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socioculturally inclusive stereotype resources with community engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 4365–4381. Curran Associates, Inc.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for Analyzing LLMs for Etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, 41(1):87–100.

Derek Edwards. 1991. Categories Are for Talking: On the Cognitive and Discursive Bases of Categorization. *Theory & Psychology*, 1(4):515–542.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. *Preprint*, arXiv:2402.09369.

Tiana Gaudette, Ryan Scrivens, Garth Davies, and Richard Frank. 2021. Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, 23(12):3491–3508.

Google. 2024. NotebookLM. Google.

John J. Gumperz. 1997. Communicative Competence. In Nikolas Coupland and Adam Jaworski, editors, *Sociolinguistics*, pages 39–48. Macmillan Education UK, London.

Gregory R. Guy. 2011. Language, social class, and status. In Rajend Mesthrie, editor, *The Cambridge Handbook of Sociolinguistics*, 1 edition, pages 159–185. Cambridge University Press.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, Bangkok, Thailand. Association for Computational Linguistics.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Barbara Johnstone. 2014. "100 % Authentic Pittsburgh": Sociolinguistic authenticity and the linguistics of particularity. In Véronique Lacoste, Jakob Leimgruber, and Thiemo Breyer, editors, *Indexing Authenticity*, pages 97–112. DE GRUYTER.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

Laurence J. Kirmayer. 2012. Rethinking cultural competence. *Transcultural Psychiatry*, 49(2):149–164.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

W. Labov. 1973. *Sociolinguistic Patterns*.

William Labov. 1985. Hypercorrection by the Lowe Middle Class as a Factor in Linguistic Change. In William Bright, editor, *Sociolinguistics*, pages 84–113. DE GRUYTER.

Robin Lakoff. 1973. Language and woman's place. *Language in Society*, 2(1):45–79.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand. Association for Computational Linguistics.

Helen-Maria Lekas, Kerstin Pahl, and Crystal Fuller Lewis. 2020. Rethinking Cultural Competence: Shifting to Cultural Humility. *Health Services Insights*, 13:1178632920970580.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024b. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024c. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. In *First Conference on Language Modeling*.

Rosina Lippi-Green. 2011. The standard language myth. In *English with an Accent*, 2 edition. Routledge.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024b. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Preprint*, arXiv:2406.03930.

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. "One-Size-Fits-All"? Examining Expectations around What Constitute "Fair" or "Good" NLG System Behaviors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional Stereotypes in Large Language Models: Dataset and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Norma Mendoza-Denton. 2008. *Homegirls*, 1 edition. John Wiley & Sons, Ltd.

Farhad Moghimifar, Shilin Qu, Tongtong Wu, Yuan-Fang Li, and Gholamreza Haffari. 2023. NormMark: A weakly supervised Markov model for socio-cultural norm discovery. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5081–5089, Toronto, Canada. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Dong Nguyen. 2025. Collaborative Growth: When Large Language Models Meet Sociolinguistics. *Language and Linguistics Compass*, 19(2):e70010.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting Cultural Commonsense Knowledge at Scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pages 1907–1917, New York, NY, USA. Association for Computing Machinery.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural Commonsense Knowledge for Intercultural Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1774–1784.

Elinor Ochs. 2009. Linguistic Resources for Socializing Humanity. In Nikolas Coupland and Adam Jaworski, editors, *The New Sociolinguistics Reader*, pages 405–423. Macmillan Education UK, London.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Haoyi Qiu, Alexander R. Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2024a. Evaluating Cultural and Social Awareness of LLM Web Agents. *Preprint*, arXiv:2410.23252.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2024b. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 189–198, Bangkok, Thailand. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. *Preprint*, arXiv:2404.12464.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cedric Archambeau, and Danish Pruthi. 2023. Geographical erasure in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12310–12324, Singapore. Association for Computational Linguistics.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Michael Silverstein. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4):193–229.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. KMMLU: Measuring Massive Multitask Language Understanding in Korean. *Preprint*, arXiv:2402.11548.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. 2024. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.

Elizabeth F. Vann. 2013. Culture. In James G. Carrier and Deborah B. Gewertz, editors, *The Handbook of Sociocultural Anthropology*, chapter 1, pages 30–48. Routledge, Abingdon.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2024. Navigating Cultural Chasms: Exploring and Unlocking the Cultural POV of Text-To-Image Models. *Preprint*, arXiv:2310.01929.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy Chen. 2024a. CRAFT: Extracting and tuning cultural instructions from the wild. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47, Bangkok, Thailand. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024b. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024c. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Winston Wu, Lu Wang, and Rada Mihalcea. 2023. Cross-cultural analysis of human values, morals, and biases in folk tales. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5113–5125, Singapore. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj

Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023b. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2024a. Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge. *Preprint*, arXiv:2404.06833.

Naitian Zhou, David Jurgens, and David Bamman. 2024b. Social meme-ing: Measuring linguistic variation in memes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3005–3024, Mexico City, Mexico. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023a. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023b. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.

A List of surveyed works

L[1]>p1

Paper	Limitations
Good Night at 4 pm?! Time Expressions in Different Cultures (Shwartz, 2022)	proxy, individual, multilingual
Probing Pre-Trained Language Models for Cross-Cultural Differences in Values (Arora et al., 2023)	proxy, survey
GD-COMET: A Geo-Diverse Commonsense Inference Model (Bhatia and Shwartz, 2023)	intrinsic, stereotype
Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study (Cao et al., 2023)	language as culture, proxy
Sociocultural Norm Similarities and Differences via Situational Alignment and Explainable Textual Entailment (CH-Wang et al., 2023)	coarseness, individual, coverage
Toward Cultural Bias Evaluation Datasets: The Case of Bengali Gender, Religious, and National Identity (Das et al., 2023)	coverage
Building Socio-culturally Inclusive Stereotype Resources with Community Engagement (Dev et al., 2023)	context, coverage, multilingual
NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly (Fung et al., 2023)	stereotype, culture is dynamic
Multilingual Language Models are not Multicultural: A Case Study in Emotion (Havaldar et al., 2023)	coverage, coarseness
Culturally Aware Natural Language Inference (Huang and Yang, 2023)	coarseness, stereotype
SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models (Jha et al., 2023)	context, coarseness, subjectivity
Multi-lingual and Multi-cultural Figurative Language Understanding (Kabra et al., 2023)	coarseness, unmarked culture
DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models (Keleg and Magdy, 2023)	proxy, coarseness, crowdsourced
Evaluating the Diversity, Equity, and Inclusion of NLP Technology: A Case Study for Indian Languages (Khanuja et al., 2023)	proxy, language as culture, coarseness
NormMark: A Weakly Supervised Markov Model for Socio-cultural Norm Discovery (Moghimifar et al., 2023)	coverage
FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models (Palta and Rudinger, 2023)	proxy, coverage, annotator, reductive

Table continues

Paper	Limitations
Knowledge of cultural moral norms in large language models (Ramezani and Xu, 2023)	coverage, culture is dynamic
NLPositionality: Characterizing Design Biases of Datasets and Models (Santy et al., 2023)	proxy
Geographical Erasure in Language Generation (Schwöbel et al., 2023)	multilingual
Modeling Cross-Cultural Pragmatic Inference with Codenames Duet (Shaikh et al., 2023)	stereotype
Cross-Cultural Analysis of Human Values, Morals, and Biases in Folk Tales (Wu et al., 2023)	coverage, multilingual
Cross-Cultural Transfer Learning for Chinese Offensive Language Detection (Zhou et al., 2023a)	multilingual, coarseness
Cultural Compass: Predicting Transfer Learning Success in Offensive Language Detection with Cultural Features (Zhou et al., 2023b)	proxy, coverage, coarseness
NormBank: A Knowledge Bank of Situational Social Norms (Ziems et al., 2023a)	coverage, stereotype
Multi-VALUE: A Framework for Cross-Dialectal English NLP (Ziems et al., 2023b)	proxy
Investigating Cultural Alignment of Large Language Models (AlKhamissi et al., 2024)	coverage, proxy, reductive, coarseness
Extrinsic Evaluation of Cultural Competence in Large Language Models (Bhatt and Diaz, 2024)	proxy, multilingual
Bridging Cultural Nuances in Dialogue Agents through Cultural Value Surveys (Cao et al., 2024a)	language as culture, pluralism
Cultural Adaptation of Recipes (Cao et al., 2024b)	proxy, coarseness, coverage
The Echoes of Multilinguality: Tracing Cultural Value Shifts during LM Fine-tuning (Choenni et al., 2024)	language as culture, proxy, coarseness
Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking (Fung et al., 2024)	LLM-derived
ViSAGe: A Global-Scale Analysis of Visual Stereotypes in Text-to-Image Generation (Jha et al., 2024)	subjectivity, annotator, reductive, proxy, discerning
KoBBQ: Korean Bias Benchmark for Question Answering (Jin et al., 2024)	subjectivity, proxy
CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean (Kim et al., 2024)	proxy, coarseness, coverage, stereotype, discerning
The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models (Kirk et al., 2024)	discerning, preference v morality, exploited annotators

Table continues

Paper	Limitations
IndoCulture: Exploring Geographically-Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces (Koto et al., 2024)	culture is dynamic, coverage
KorNAT: LLM Alignment Benchmark for Korean Social Values and Common Knowledge (Lee et al., 2024)	culture is dynamic, unmarked culture, annotator, intrinsic
CultureLLM: Incorporating Cultural Differences into Large Language Models (Li et al., 2024a)	intrinsic
CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting (Li et al., 2024c)	intrinsic, coarseness
Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings (Liu et al., 2024a)	proxy, coverage
Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions (Masoud et al., 2025)	culture is dynamic, multilingual, proxy
Having Beer after Prayer? Measuring Cultural Bias in Large Language Models (Naous et al., 2024)	coarseness, coverage
Cultural Commonsense Knowledge for Intercultural Dialogues (Nguyen et al., 2024)	crowdsourced, LLM-derived, stereotype
Evaluating Cultural and Social Awareness of LLM Web Agents (Qiu et al., 2024a)	multilingual, coverage, context
NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models (Rao et al., 2024)	proxy, coarseness, culture is dynamic, multilingual, coverage
DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures (Seth et al., 2024)	coverage, intersectionality, multilingual
CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies (Shi et al., 2024)	multilingual, unmarked culture, stereotype
KMMLU: Measuring Massive Multitask Language Understanding in Korean (Son et al., 2024)	copyright, intrinsic
Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties (Sorensen et al., 2024)	LLM-derived
Navigating Cultural Chasms: Exploring and Unlocking the Cultural POV of Text-To-Image Models (Ventura et al., 2024)	coverage, based on existing model
CRAFT: Extracting and Tuning Cultural Instructions from the Wild (Wang et al., 2024a)	multilingual

Table continues

Paper	Limitations
SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning (Wang et al., 2024b)	coverage, intrinsic
Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models (Wang et al., 2024c)	survey, coverage
Benchmarking Machine Translation with Cultural Awareness (Yao et al., 2024)	proxy
RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations (Zhan et al., 2024)	multilingual
WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models (Zhao et al., 2024)	culture is dynamic, proxy, individual
Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge (Zhou et al., 2024a)	coverage, multilingual, crowdsourced

Table complete

B Category key

Category	Description
coverage	Makes note of insufficient or limited coverage.
proxy	Notes the limitations of the chosen proxy.
coarseness	Mentions that cultural boundaries are too coarse.
multilingual	Mentions that only one (or a limited number) of languages are studied.
stereotype	Mentions concerns about learning or perpetuating stereotypes.
culture is dynamic	Mentions limitations in capturing the dynamicity of culture.
intrinsic	Mentions the limitations of only performing intrinsic evaluation.
language as culture	Mentions the limitations of treating language as cultural boundaries.
crowdsourced	Mentions concerns about crowdsourced data.
discerning	Mentions that not all cultural attributes should be treated the same way.
LLM-derived	Mentions concerns that some part of the dataset was generated by LLMs.
reductive	Mentions concerns that culture is reduced to the proxies chosen.
annotator	Mentions concerns about human-annotated data.
unmarked culture	Mentions concerns with "universal" cultural attributes or values.
individual	Mentions that individual values in the data may not be reflective of cultural ones.
subjectivity	Mentions that cultural annotations are subjective.
context	Notes that the system or datasets lacks social or situational context.
survey	Notes the limitations of relying on survey data for cultural knowledge.
pluralism	Mentions concerns with choosing which value to align to in pluralistic situations.
preference v morality	Notes that aligning to preferences may not be desirable behavior.
exploited annotators	Notes that annotators providing preference data do not generally share in the benefits.
intersectionality	Mentions limitations in covering intersectional identities.
copyright	Notes that some data points were removed due to copyright.
based on existing model	Notes that automated evaluation is based on existing model.

Table continues

Table complete