International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 4, Issue 12, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i12.1074

Designing Culturally Inclusive NLP Models for Low-Resource Languages and Communities

Nneoma Udeze[1]

Publication Date: 2025/12/30

Abstract

Over the last few years, Natural Language Processing has developed at an extremely fast pace, but the majority of these improvements focus on popular, well-resourced languages, which means that billions of speakers of under-resourced languages are not able to access modern language technologies. More to the point, the contemporary approaches to the development of Natural Language Processing do not tend to consider the social contexts, viewpoints, and requirements of the communities, the language of which is being modeled. The result of this is insufficient technologies that can even be detrimental to linguistic and cultural diversity. This paper examines the complex issues of planning culturally based Natural language Processing (NLP) for limited-resource languages and communities, aiming to transcend specific concerns to address fundamental questions about belief systems, power, representation, and local linguistic rights. This paper uncovers important features of equity in NLP through the analysis of linguistic diversity, technical constraints, and cultural dynamics. These include culturally based representations that reflect the language characteristics of marginalized languages, participatory design, which focuses on community knowledge and goals, cultural protocols, evaluation frameworks aligned with community values rather than Western academic standards, and sustainable methodologies that mitigate the effects of environmental and data extraction. Technical strategies, including transfer learning, multilingual language models, active learning, and few-shot learning, are also reviewed in the paper, though limitations of these methods are evaluated critically with respect to culturally different languages.

Keywords: *Natural Language Processing, Cultural Inclusion, Limited Resource Languages, Collaborative Research, Local Data Freedom, Linguistic Diversity, Language Technology, Critical AI, Ethical NLP.*

I. INTRODUCTION

The Natural Language Processing (NLP) research community worldwide has made breakthroughs in the field over the years, as language generation models develop more human-like language understanding and generation capabilities. But these developments are not evenly spread. Although English, Chinese, and a limited number of other popular languages have access to advanced voice assistants, real-time translators, and AI-based writing tools, the largest of the 7,000 or so languages in the world do not even have a grammar checker (Joshi et al., 2020). This technological gap is indicative of and strengthens more fundamental cultural marginalization. The models of NLP that are primarily trained on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies encode particular worldviews, values, and knowledge structures that can be incompatible and even harmful in other cultural settings (Henrich et al., 2010). When such models are implemented across linguistic and cultural borders without being adapted, they may impose foreign cognitive patterns, kill culturally specific meanings, and speed up the disappearance of traditional systems of knowledge.

The languages with limited online presence, limited computational resources, and research interest are limited resource languages that pose technical and ethical challenges to NLP development. Although the lack of data and the performance of models have become the subject of increasing research, the cultural aspects of NLP participation have not been addressed. What does the cultural appropriateness of an NLP model entail? What can technology contribute to cultural identity and make it stronger instead of weaker? And who is to be master of language material and the course of technological evolution?

The paper will explore these questions by discussing in detail what culturally sensitive NLP design can be like in the case of limited-resource languages and communities. It combines the input of language technologies, human-computer interaction, indigenous studies, science and technology studies, and collaborative

Udeze, N. (2025). Designing Culturally Inclusive NLP Models for Low-Resource Languages an Communities. *International Journal of Scientific Research and Modern Technology*, *4*(12), 115–119. https://doi.org/10.38124/ijsrmt.v4i12.1074

research to suggest structures of Natural Language Processing that reinforce instead of erode marginalized languages and cultural diversity. The discussion is presented by exploring the issues of linguistic and cultural diversity, technical methods and their weaknesses, inclusive design, and case studies in various community settings.

II. UNDERSTANDING LOW-RESOURCE LANGUAGES AND CULTURAL CONTEXT

- *Defining and Characterizing Limited Resource Language*

Limited resource languages are typified by a lack of digital language data, low-quality software tools, and low financial resources devoted to the development of language technology. Blasi et al. (2022) approximate that only around 100 languages in the world, out of the estimated 7,000, have enough technological support to support the modern NLP systems. This lack is not only the mirror of the populations of speakers but a complicated combination of historical, political, and economic processes.

The categorization of "low-resource" is rather shocking in terms of inequity. Small languages, including Icelandic or Estonian, have a good institutional base, governmental support, and good digital infrastructure, which allow advanced language technologies. On the other hand, millions of languages, such as Yoruba in West Africa, Quechua in the Andes, or any other Indigenous language in any part of the world, are technologically marginalized even though they are very active and culturally important. The cause of this difference lies in historical injustices such as colonialism, continuous economic discrimination, and exclusion as a system in the development of technologies globally.

Language resource distribution is essentially a reflection of colonial histories, economic inequality, and trends of systemic marginalization instead of the language properties and constraints themselves. The linguistic complexity, cultural richness, and in many cases large numbers of speakers of the indigenous languages, minority languages, and languages of the Global South are disproportionately categorized as low-resource languages. This scarcity of resources is a social construct, which is created by decades of underfunding, non-inclusion in the technology development pipeline, and preference for economically dominant languages in research and commercial use. This social aspect is crucial to comprehend to come up with ethical and effective methods of inclusive NLP.

- *Language Diversity and Typological Differences*

There is remarkable structural diversity in the languages of the world, and NLP models based mostly on English and other Indo-European languages are challenged by it. This linguistic diversity covers several dimensions such as morphology, syntax, phonology, and semantic organization, each of which has far-reaching consequences in computational modeling.

The morphological complexity is vastly different among languages. Whereas languages such as Mandarin Chinese or Vietnamese have fairly straightforward morphological systems with little inflectional marking, polysyntactic languages such as Inuktitut, Mohawk, or Greenlandic permit single words to encode what would otherwise represent whole sentences in English, including multiple arguments, tense markings, and grammatical relationships within their word structures. Agglutinative languages such as Turkish, Finnish, or Swahili construct words by means of systematic suffixation, producing large morphological paradigms that put a strain on tokenization and segmentation methods used to isolate languages.

Syntactic structure shows no less varied patterns. Word order varies between the comparatively fixed Subject-Verb-Object arrangement of English to free word order arrangements of languages such as Latin, Russian, or Warlpiri, in which grammatical relationships are indicated by elaborate case systems but not positional restrictions. Some languages use verb-initial orders, such as most Austronesian languages and Mayan languages, and there are also those that use object-verb-subject order as well. These structural differences essentially influence parsing plans, dependency structures, and syntactic annotation plans.

These typological differences directly and problematically apply to NLP systems. Space-delimiting word segmentation algorithms that are based on comparatively short, space-delimited words often do not work with languages with productive compounding (German, Dutch), polysynthesis (Mohawk, Inuktitut), or scriptio continua systems of writing (traditional Chinese, Thai). Tagging systems based on the English or European grammatical categories of nouns, verbs, adjectives, and adverbs may essentially misrepresent languages with different structural categories, such as no distinct adjective-verb category or use of a noun category. Dependency parsing and constituent structure analysis are based on syntactic structures that are not necessarily in agreement with the grammatical properties of most languages in the world.

Moreover, trained language models capture distributional regularities and semantic relationships that are sensitive to the cultural backgrounds of the training data. Such models not only acquire linguistic structure but also cultural concepts, social relationships, and value systems that are incorporated in texts. They may be exploited to distort or misrepresent local ideas, introduce alien categorical structures, and propagate cultural bigotry when used in other linguistic and cultural contexts (Bender and Koller, 2020). A model that is mostly trained on English Wikipedia, e.g. will encode Western cultural beliefs about kinship, social hierarchies, time, and a myriad of other aspects that might not translate well to other cultural situations.

III. CULTURAL EXCLUSION IN CURRENT NLP PARADIGMS

- *The Hegemony of English and WEIRD Languages*

Modern research and development in NLP is still too dominated by English, which poses a systematic obstacle to linguistic diversity and cultural inclusion. According to Joshi et al. (2020), 64 percent of papers at the ACL, the largest conference in the field, were based on English, and 28 percent included both English and other languages, so less than ten percent of publications covered non-English languages. This is not just limited to academic research to commercial applications, where English-language models are the recipients of the overwhelming majority of investment, computational resources, and optimization.

This trend is demonstrated by the development path of powerful pretrained models. The original versions of BERT and GPT were only trained in English, and only later, as extensions, not as design considerations, were multilingual or language-specific versions introduced. This model considers non-English languages as secondary appendices and not as equal priorities, and in most cases, languages that are structurally or culturally different from the English language perform very poorly. The amount of computational resources, human labor, and fine-tuning that goes into English models is many times greater than the amount of computational resources, human expertise, and fine-tuning that goes into other languages, which only serves to perpetuate and amplify existing inequalities.

This English-centric bias goes beyond the selection of language, but it imparts profound cultural suppositions into models, data sets, measures of evaluation, and research agendas. The prevalence of WEIRD (Western, Educated, Industrialized, Rich, Democratic) views on NLP can be seen on several levels. The annotation schemes often capture the Western emotional display rules, psychological categories, and social concepts that might not be applicable in the cross-cultural context. As an illustration, emotional recognition systems that have been trained on Western samples can be incorrectly classified on emotional displays of other cultures where the display conventions are different. Named entity recognition systems follow Western naming rules, which often follow given name, family name, and do not cope with a wide range of naming systems such as patronymies, matronymies, teknonyms, or mononyms that are common in most cultures.

Sentiment analysis tools encode Western valence associations, potentially misinterpreting culturally specific expressions of politeness, indirectness, or collective orientation. The co-reference resolution systems presuppose the Western kinship systems and social relations. Even so-called language-neutral activities contain some cultural assumptions: question-answering systems assume that there is some particular knowledge structure, information-seeking behavior, and discourse conventions that differ greatly across cultures.

Perhaps the most problematic, it is a common thing in multilingual language models to merely copy English-based semantic representations into other languages by aligning them in embedding spaces, instead of being able to learn culturally grounded representations in each language. This strategy poses a threat of systematic erosion of local linguistic subtleties, cultural ideas, and world-views, and the imposition of Western conceptual frameworks on the various linguistic and cultural settings in the name of universal language comprehension.

IV. TECHNICAL APPROACHES FOR LIMITED RESOURCE NLP

Transfer learning has become a leading paradigm of dealing with data scarcity in low-resource languages by utilizing knowledge gained in high-resource languages. The principle behind this method is that linguistic representations acquired through large amounts of data in richly endowed languages can be used to give useful inductive biases in the acquisition of related languages with sparse training data. Multilingual pretrained models (like multilingual BERT, Cross-lingual Language Model - RoBERTa, and multilingual T5) are trained to share representations across dozens or even hundreds of languages at once, allowing zero-shot or few-shot transfer to languages not present during training of the model, and sometimes even to languages that have never been seen by the model (Conneau et al., 2020).

These models have shown remarkable performance in specific tasks, especially when the target languages have typological similarities, genetic relations or contact with better-resourced languages in training data. Cross-lingual transfer has made it possible to rapidly develop simple NLP functionality on many languages that would otherwise have insufficient annotated data to support supervised learning. Language adaptation methods like adapter layers, which include small language-specific components to frozen multilingual models, offer computationally efficient language adaptation, without necessarily retraining the model.

Nevertheless, transfer learning has serious limitations and dangers to culturally sensitive NLP, especially when used blindly in different linguistic and cultural settings. The models that are primarily trained on languages with high resources might spread unsuitable linguistic assumptions, e.g., the expectations of the word order, the complexity of morphology, or the syntactic patterns that are not relevant to the target languages. Languages whose structure is significantly different than that of high-resource languages in the training mix are always disadvantaged by transfer, which only recreates, but never decreases inequalities. An example is that polysynthetic languages, non-concatenative systems of morphology, or languages whose pragmatic conventions are radically different hardly benefit from models designed to be optimised in Indo-European languages.

Furthermore, the multilingual models tend to focus on the major world languages and consider the low-resource languages as peripheral extensions to the models, not as equal partners. The capacity of the model, training processes, and optimization activities are usually biased towards languages with larger training corpora, leading to the dilution of capacity in which low-resource languages are competing over scarce model parameters. It may lead to superficial or distorted representations of marginalized languages.

There is also the risk of transfer learning continuing to propagate or increase cultural biases within the source language data that are present. High-resource training data that has been encoded in semantic associations, stereotypes, and value judgments can be transferred to target languages, bringing with it foreign cultural structures and potentially leading to cultural damage. As an example, a model trained on gender associations or occupational stereotypes using English example data can bring the biases to a language whose gender system or social structure is different.

Finally, culturally sensitive transfer learning must take into account negative transfer, where transferred representations actively damage performance or cultural appropriateness in the target languages. This requires the integration of transfer learning with language-specific training data that is reviewed by the community and can override the unwanted transferred assumptions. The assessment should go beyond the traditional performance measures, such as accuracy or F1 scores, to determine that transferred representations are linguistically correct and culturally relevant to the target communities. This involves looking at how models are able to maintain significant linguistic differences, take into consideration cultural ideas and categories, and not impose foreign structures on local knowledge systems. The participatory evaluation that includes native speakers and cultural professionals is necessary to detect the forms of cultural distortion that cannot be evaluated by quantitative measures only.

V. CASE STUDIES IN CULTURALLY INCLUSIVE NLP

- *Masakhane: Grassroots African NLP*

Masakhane, which means “we build together” in isiZulu, is an African language NLP grassroots movement by African researchers and practitioners (Orife et al., 2020). The project clearly disregards extractive research models in favor of community ownership and benefit. Masakhane has an approach that incorporates participatory workshops, which involve African researchers learning about NLP and producing resources in their own languages, collaborative dataset building that is sensitive to cultural appropriateness and regional diversity, open-source publication of models and resources, and a strong pledge to decolonial research practices.

The initiative has already created NLP materials in more than 40 African languages, and has also created institutional capacity and provided other models of NLP research. The main success factors are the focus on African leadership and priorities, integration of technical innovation with social infrastructure development, the focus on learning and mentorship, as well as research outputs, and community ownership of data and models.

- *Inuktitut Language Technology*

The local language, Inuktitut, of the local communities in the Canadian Arctic presents a great technical challenge to Natural Learning Processing because of its intricate structure and lack of resources. However, the attempts to create NLP tools in Inuktitut have yielded culturally relevant approaches Kodner et al., 2017). Notable aspects of such endeavors are working with integrated community units and elders in the early stages, combining traditional knowledge with modern language usage, developing structural analyzers that do not impose artificial divisions, developing the educational tools in line with the objectives of language revitalization, and ensuring that the community has control over the language data with stringent rules on commercialization. Such projects as the Inuktut Digital Transformation Project in Nunavut show that NLP can help local languages to be revived in case the project is created in the context of true cooperation, taking into account cultural practices and local concerns.

VI. CONCLUSION AND RECOMMENDATIONS

The ability to create culturally inclusive Natural Language Processing (NLP) models of limited resource languages and communities is a technical challenge and a moral imperative. Although technical issues, including data scarcity, limited category coverage, and model accuracy, are being tackled using developments in transfer learning, multilingual language models, and few-shot learning, ethical and cultural issues are also critical. These are respecting community autonomy, avoiding cultural harm, equitable distribution of benefits, sustainable relationships, and a fundamental reconsideration of the way NLP research and development are undertaken.

NLP researchers and developers need to embrace participatory research paradigms, which reorganize the research processes with emphasis on community collaboration. This will involve starting projects through working with communities to know the issues of concern and building fair partnerships where the community members are research partners with a substantial decision-making role, rather than as informants or sources of information. Community needs should be used to formulate the research questions instead of foisting their external agendas. Moreover, ethical data governance should be realized in the framework of such principles as CARE and licenses like Kaitiakitanga, which acknowledge the authority of the community. The informed consent must be continuous and retractable, with

the communities being in control of the access to the data, the limitation of its use, and the benefit-sharing terms.

Funding agencies and institutions are also essential since they focus on grant proposals coordinated by or in true collaboration with language communities. The sources of funding should be opened to local organizations and community groups as well as traditional academic institutions, and easy application procedures and assistance in the development of the grassroots projects. To ensure that data sovereignty is observed by technology companies, it is necessary to avoid language data collection without due community consent and fair sharing of benefits, design clear data management policies, and explain the data collection, use, and retention policies to the involved communities.

In conclusion, true inclusion requires the reorganization of research practices based on community collaboration, maintaining local data sovereignty, building assessment models to gauge cultural as well as technical sufficiency, building local knowledge instead of dependence, and fairly sharing the benefits of technological advancement with communities whose languages and knowledge base technological progress. As the case studies above have shown, culturally sensitive NLP can be realized when the researchers focus on the needs of the community and leadership, adopt responsible data governance, develop linguistically and culturally sensitive approaches, invest in long-term capacity building, and position their work as a service to linguistic diversity instead of building only technical capabilities.

REFERENCES

[1]. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

[2]. Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5486–5505). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

[3]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

[4]. Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

[5]. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (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* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

[6]. Kodner, J., Caplan, S., Xu, H., Marcus, M. P., & Yang, C. (2017). Case studies in the automatic characterization of grammars from small wordlists. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 76–84). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0112

[7]. Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., Degila, K., . . . Bashir, A. (2020). *Masakhane -- Machine translation for Africa* [Conference paper]. AfricaNLP Workshop, ICLR 2020. https://doi.org/10.48550/arXiv.2003.11529