CAKE: Cultural Analysis for Knowledge Equity

Jan-Christoph Kalo
University of Amsterdam

Abstract

CAKE (Cultural Analysis for Knowledge Equity) examines how culturally significant concepts are represented across Wikipedia's largest language editions. Despite Wikipedia's global scope, structural biases often marginalize non-dominant cultural perspectives. Using Large Language Models, we extract semantic triples from multilingual articles and align them with Wikidata to compare how shared concepts, such as flatbread or family, are framed across languages in Wikipedia. The outcome is a multilingual, Wikidata-annotated dataset highlighting systemic representation gaps. We plan to validate findings with Wikimedia communities through participatory online workshops. CAKE complements the Abstract Wikipedia project by focusing on analysis rather than generation, contributing to Wikimedia's goals of Knowledge Equity and Knowledge as a Service. All data, tools, and findings will be released openly to support researchers and communities in improving cultural representation in structured knowledge.

Introduction

Wikipedia, as the world's largest collaborative encyclopedia, aims to document "the sum of all human knowledge" across its over 300 language editions. However, despite this multilingual aspiration, Wikipedia's knowledge representation inherently reflects systemic cultural biases. Concepts fundamental to human experience, such as staple foods (e.g., flatbread), familial structures, or rituals, differ significantly across cultures, yet are often disproportionately represented through a Western-centric lens, primarily influenced by larger language editions, such as English. For instance, flatbread, though universally consumed, holds vastly different cultural significance globally, ranging from pita in Middle Eastern cultures to tortilla in Latin American societies. These variations are frequently understated or obscured entirely in dominant language editions, perpetuating cultural marginalization and limiting the diversity and equity of knowledge representation.

This systematic cultural imbalance not only contradicts Wikipedia's foundational goal of neutrality, but it also directly undermines the Wikimedia 2030 strategic objectives—particularly the pillars of **Knowledge Equity** and **Knowledge as a Service**. Knowledge Equity demands the fair and representative inclusion of diverse cultures and communities, countering existing biases and structural gaps. Knowledge as a Service emphasizes the need for structured, reusable, and accessible knowledge formats that support diverse global user communities. Addressing cultural bias is crucial for Wikipedia's ability to serve as a global, inclusive knowledge platform.

Existing initiatives such as Abstract Wikipedia attempt to reduce linguistic and cultural inequities by generating language-independent content from structured symbolic representations [16,17]. However, Abstract

Wikipedia's reliance on centralized content creation can inadvertently amplify dominant narratives unless explicitly guided by culturally inclusive data. Our project **CAKE (Cultural Analysis for Knowledge Equity)** takes a fundamentally different yet complementary approach. Rather than generating new content, CAKE systematically analyses existing multilingual Wikipedia articles using Wikidata's structured properties as a universal formal language. This allows us to rigorously quantify how the representation of culturally relevant concepts diverges across language editions, thus identifying systemic epistemic biases.

To achieve this, we propose using state-of-the-art Large Language Models (LLMs) to extract subject-predicate-object triples from Wikipedia articles across the ten largest language editions [13]. These semantic triples will then be mapped onto Wikidata properties, allowing for precise cross-cultural comparisons. We will evaluate our findings during online workshops with local Wikimedia communities to ensure computational findings resonate with lived cultural experiences. This combined computational and participatory approach will provide both quantifiable metrics of bias and culturally grounded interpretive insights.

Our key research questions include:

1. **How do representations of culturally significant concepts systematically differ across Wikipedia's largest language editions?**
   What types of cultural biases are encoded within Wikipedia's multilingual knowledge base as detectable through LLM-based information extraction?
2. **How can community-driven validation improve the accuracy and cultural sensitivity of computational bias detection methods?**
3. **What insights and tools can effectively inform Wikimedia communities and researchers on addressing representation gaps to advance Knowledge Equity?**

By addressing these questions, CAKE aims to deliver academic and community impact, providing insights, openly available datasets, and tools that researchers and the Wikimedia community can employ to enhance cultural equity in knowledge representation. This project's outcomes will directly contribute to fulfilling Wikimedia's mission, guiding future editorial and strategic decisions towards a more equitable and inclusive global knowledge ecosystem.

Date:
The proposed starting date of the proposal is 1. September 2025 with a total duration of 6 months. The end date is 28. February 2026.

Related work

Bias in natural language processing is a well-documented challenge. A comprehensive survey outlines how varying definitions of bias affect methods and evaluations across the field [1]. Among these, cultural bias is increasingly recognized as a critical concern. Recent work categorizes how cultural norms are embedded in NLP systems, particularly when data disproportionately reflects dominant languages and perspectives [10].

LLMs trained on such data tend to encode and reproduce cultural asymmetries. Evaluations show models generate stereotyped or inappropriate outputs across cultural contexts [11], and benchmarking frameworks have emerged to probe cultural knowledge gaps in LLMs using curated prompts [9]. While effective in revealing internal biases, these approaches rarely apply LLMs to analyze how cultural concepts are represented in external corpora like Wikipedia.

Wikipedia provides a multilingual, community-governed platform in which cultural bias can be studied at scale. NLP-based methods have been developed to detect biased statements in articles [4], while comparative studies of biographies across languages show that framing differences often reflect national or cultural perspectives [2]. Similar disparities are also present in Wikidata, where demographic properties such as nationality or ethnicity are inconsistently modeled across items [13].

Although Abstract Wikipedia introduces a shared, language-neutral representation layer to unify content across editions [16,17], it does not address how concepts are framed differently in existing articles. Nor does it support comparative analysis of culturally significant topics across language versions. Such analysis requires linking textual representations to structured knowledge sources, an area where recent advances in LLM-based information extraction become highly relevant.

Recent work has explored how LLMs can support structured knowledge acquisition through tasks like triple extraction, entity linking, and canonicalization [5,14]. Techniques for aligning entities across languages [3], resolving synonymous relations [6,8,15], and analyzing extraction errors [7] contribute to the creation of consistent, high-quality knowledge graphs.

Despite these technical advances [12, 14], several research gaps remain. No large-scale effort has examined how specific cultural concepts are framed across Wikipedia's language editions, nor systematically aligned those framings with Wikidata. LLMs have not been widely used as tools for comparative cultural analysis in this context. Moreover, bias studies rarely involve participatory validation with the communities affected. This project addresses these limitations by using LLM-based information extraction and canonicalization/alignment to analyze cross-lingual framing in Wikipedia, creating a Wikidata-linked dataset to support cultural equity research.

Methods

The CAKE project systematically investigates how culturally significant concepts are represented differently across multilingual Wikipedia editions, leveraging structured data from Wikidata and advanced information extraction methods.

1. Data Collection

Our study focuses initially on articles covering culturally salient concepts selected systematically based on their universal relevance and substantial differences in cultural framing. For instance, we will begin with broad yet culturally-specific concepts such as "flatbread," which manifests distinctly across various cultures (e.g., pita in the Middle East, naan in South Asia, tortilla in Latin America). Additional culturally central concepts (such as "family," "marriage," and "hospitality") will be systematically identified and included based on initial analysis results and consultation with Wikipedia community members.

We target the ten largest Wikipedia editions by article count and user engagement (including English, German, Polish, French, Japanese, Italian, Dutch, Portuguese, Spanish, and Russian). These editions provide extensive coverage and diverse cultural perspectives.

We will programmatically gather Wikipedia articles corresponding to selected concepts in each of these language editions via the official Wikipedia API.

2. Information Extraction and Canonicalization

We will develop a multilingual pipeline for extracting structured semantic representations from Wikipedia articles, centered on subject–predicate–object (SPO) triples. Our approach combines LLMs for information extraction with knowledge graph alignment techniques for integration into Wikidata [3, 14, 18].

We will employ advanced generative language models, such as LLaMA or equivalent architectures, to perform relation extraction across multiple languages [14]. These models will be fine-tuned or prompted to generate SPO triples from textual content. The generative approach allows for flexible and context-aware extraction of relationships, accommodating the nuances present in different languages and cultural contexts [5, 10].

Recognizing the challenges in Named Entity Disambiguation, especially in a multilingual setting, we will utilize mGENRE (Multilingual Generative Entity REtrieval) for entity linking. mGENRE is a sequence-to-sequence model that generates entity names in an autoregressive manner, effectively linking mentions in text to entries in a multilingual knowledge base [3].

Since each extracted triple will undergo a mapping process to align entities and predicates with corresponding Wikidata identifiers, this ensures a consistent and language-neutral representation of the extracted knowledge. For entities that cannot be confidently mapped to Wikidata entries, we will retain them as noun phrases for subsequent analysis. Similarly, predicates that do not align with existing Wikidata properties will be preserved for further examination.

To address the unmapped entities and predicates, we will implement clustering techniques to group semantically similar elements across different languages. This clustering approach draws inspiration from open information extraction canonicalization [15] and from our work on knowledge graph consolidation [6, 8].

1. We utilize cross-lingual embeddings and similarity measures to cluster noun phrases that likely refer to the same entity across languages.
2. We group predicates based on semantic similarity to identify potential correspondences or novel relations not currently represented in Wikidata.

The outcome of this step will be a multilingual, Wikidata-aligned knowledge graph that captures both universal and culturally specific representations of selected concepts. This enriched knowledge base will serve as the foundation for subsequent cross-lingual semantic analyses and community-driven validation efforts.

3. Cross-lingual Semantic Analysis

We will analyze the semantic triples extracted from Wikipedia articles across different language editions using a combination of quantitative metrics and qualitative comparative methods. The quantitative analysis will focus on measuring both the degree of semantic overlap and the extent of culturally specific representations.

To assess semantic similarity, we will calculate the Jaccard Similarity Index between sets of extracted triples across language pairs. This metric quantifies the overlap between sets, providing insight into how different communities describe the same concept.

In addition to measuring overlap, we will introduce a Cultural Centrality Ratio (CCR), defined as the proportion of triples in a given language edition that include culturally-specific entities or predicates—those that are not shared across multiple editions or cannot be confidently mapped to Wikidata. This metric captures the degree to which a language contextualizes a given concept through its own cultural lens, as opposed to presenting it in a more universal or neutral form. We expect CCR values to vary significantly across both concepts and languages, providing insights into the degree of cultural embedding within Wikipedia narratives. This approach aligns with findings that about a quarter of each Wikipedia language edition is dedicated to representing the corresponding cultural context.

To further enrich our analysis, we will compare the number of unique entities identified in each article, offering a proxy for article size and depth of coverage. Additionally, we will compute betweenness centrality within the constructed knowledge graphs to identify entities that serve as crucial connectors within the graph, highlighting their role in the structure of the article's semantic network.

To complement these quantitative analyses, we will conduct qualitative assessments of concept representations that show substantial cross-lingual variation. This will involve a close reading of triples associated with selected concepts—particularly those with high CCR or low similarity scores—to understand how cultural differences manifest in specific relations or terminologies. These cases will be examined interpretively to explore underlying cultural assumptions, narrative structures, and knowledge framing strategies. Where relevant, we will engage cultural experts and members of the Wikipedia editing community to validate our interpretations and identify potential blind spots. This methodology is informed by studies that analyze cultural borders on Wikipedia through multilingual co-editing activity.

The outcome of this step will be a detailed comparative profile of how culturally salient concepts are framed across Wikipedia's multilingual landscape.

4. Community-driven Validation and Co-creation

We will organize online workshops involving Wikimedia affiliates and local community members corresponding to selected language editions. Communities will review computational findings, assessing their alignment with lived cultural experiences, identifying potential inaccuracies, and ensuring cultural sensitivity.
Furthermore, participants will be able to provide input into refining the tools and help prioritize culturally relevant concepts for further analysis.

Participants will be recruited via presentations at Wikimedia community events (e.g., Wikidata Workshop 2025, Wiki Workshop 2025), direct communication with Wikimedia affiliate groups, and calls for participation disseminated through Wikimedia channels.

Expected output

We aim to produce the following key deliverables:

- **Scientific Publications**:
  Short paper outlining initial ideas at Wikidata Workshop 2025.
  Full paper describing methods and

results at TheWebConf, International Semantic Web Conference, or WebSci.

- **Open-access Multilingual Dataset & Software**:
  Dataset of triples from 10 language editions for a variety of different concepts.
  A publicly tool for cross-lingual semantic analysis available on GitHub.
- **Community Insights**:
  Report recommendations for Wikimedia editors to address cultural biases identified through the analysis.
- **Presentations and Workshops**:
  Presentations at Wiki Workshop 2026, Wikidata Workshop 2026, and Wikimania 2026 to actively engage with Wikimedia communities, researchers, and editors.

Risks

Several potential challenges exist within the CAKE project's scope, primarily related to methodology, community engagement, and data alignment.
(1) Firstly, relying on LLMs for semantic triple extraction carries inherent risks of bias propagation and inaccuracies. To mitigate this, we will carefully evaluate extraction methods on smaller datasets initially, iteratively validating results before scaling the analysis. Additionally, any bias in the underlying models will be transparently documented.
(2) A second risk involves possible gaps or mismatches in aligning extracted semantic triples to Wikidata's structured ontology. To address this, our methodology includes proactive identification of gaps and a clearly documented process to propose new culturally-specific Wikidata properties where necessary, improving Wikidata's expressivity in the process.
(3) Community participation poses another potential challenge, as successful validation depends significantly on active engagement. We will address this proactively by presenting at Wikimedia community events early in the project (Wikidata Workshop 2025) to build relationships, clearly communicate project relevance, and recruit dedicated community participants. Transparent, ongoing communication through Wikimedia channels will further support sustained engagement and ensure community buy-in, thereby reducing the risk of limited participation or misalignment between our computational findings and community insights.

Community impact plan

Beyond academic audiences, the CAKE project is designed to actively involve and positively impact Wikimedia volunteer editor communities, developers, and affiliate groups. Recognizing that genuine cultural representation and knowledge equity require community-driven engagement, we will initiate sustained collaboration with Wikimedia affiliates and local editing communities early in the project. Specifically, presentations at Wikimedia events, such as the Wikipedia Workshop 2025 and subsequent community workshops, will enable direct interactions with editors and community members, fostering trust and reciprocal dialogue.

The project's core deliverables—a multilingual annotated dataset and open-source analytical tools—will be presented to Wikimedia communities clearly and transparently, facilitating practical adoption and use beyond academia. To encourage community ownership and practical applicability, the dataset and software will be openly accessible on GitHub, accompanied by clear documentation and tutorials in Jupyter Notebooks. These resources will empower Wikimedia volunteers and developers to directly apply our findings in assessing and addressing representation gaps within their specific language communities.

Moreover, the structured feedback from community workshops will ensure our outputs remain culturally sensitive and practically useful, enhancing the likelihood of adoption by editors seeking to address knowledge equity actively.

Through this comprehensive approach, the CAKE project aims to foster tangible, community-driven improvements in Wikipedia's global cultural inclusivity.

Evaluation

The success of the CAKE project will be assessed through multiple complementary dimensions.

From a scientific perspective, the primary criterion of success will be the acceptance of our outputs in peer-reviewed venues. Initial validation will come through acceptance and presentation of the preliminary idea at Wikidata Workshop 2025, followed by publication of our comprehensive methodological approach and analytical results at a larger scientific conference, such as TheWebConf or WebSci.

The practical value and impact of our dataset and software will be evaluated by their adoption and use within the research community. Evidence of successful impact will be visible through citations of our dataset and tools in subsequent research by other scholars studying cultural bias, multilingual knowledge representation, or NLP fairness. Additionally, active engagement on GitHub—including repository interactions such as stars, forks, and issue discussions—will provide clear indicators of broader community interest and usage.

A crucial dimension of the project's evaluation is community validation. During workshops with Wikimedia community members and editors, structured qualitative feedback will be gathered to assess how well the analytical outcomes align with participants' lived cultural experiences and editorial practices. Positive evaluations by workshop participants—indicating both cultural sensitivity and practical utility of our insights—will demonstrate the project's value and validate our computational approach.

Adoption by Wikimedia communities will also be signaled through invitations to present our results at prominent Wikimedia community events, including Wiki Workshop 2026 and Wikimania 2026. Such invitations not only confirm the alignment of our work with Wikimedia community interests but also provide opportunities for sustained engagement and ongoing refinement.

Finally, transparency will underpin the entire evaluation process. Regular updates, evaluation results, and feedback summaries will be made publicly accessible via our GitHub repository and Wikimedia research documentation, enabling independent verification by Wikimedia Research Fund chairs and the broader community.

Budget

This project will be carried out by 20h per week over 6 months by a researcher, 20h per week over 6 months by a student assistant, and 5h per week by the project coordinator, for planning and dissemination.
The detailed budget plan is available here.

References

[1] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476.

[2] Callahan, E. S., & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. Journal of the American Society for Information Science and Technology, 62(10), 1899–1915.

[3] De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., ... & Petroni, F. (2022). Multilingual autoregressive entity linking. Transactions of the Association for Computational Linguistics, 10, 274–290.

[4] Hube, C., & Fetahu, B. (2018, April). Detecting biased statements in Wikipedia. In Companion Proceedings of the The Web Conference 2018 (pp. 1779–1786). International World Wide Web Conferences Steering Committee.

[5] Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2022). GenIE: Generative information extraction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3856–3873). Association for Computational Linguistics.

[6] **Kalo, J. C.**, Ehler, P., & Balke, W. T. (2019). Knowledge graph consolidation by unifying synonymous relationships. In The Semantic Web – ISWC 2019: 18th International Semantic Web Conference (pp. 276–292). Springer.

[7] **Kalo, J. C.**, Kruit, B., & Schlobach, S. (2022). Understanding distantly supervised relation extraction through semantic error analysis. In Proceedings of the 3rd Conference on Automated Knowledge Base Construction (AKBC).

[8] **Kalo, J. C.**, Mennicke, S., Ehler, P., & Balke, W. T. (2020). Detecting synonymous properties by shared data-driven definitions. In The Semantic Web – ESWC 2020: 17th International Conference on The Semantic Web (pp. 360–375). Springer.

[9] Keleg, A., & Magdy, W. (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 (pp. 6245–6266). Association for Computational Linguistics.

[10] Liu, C. C., Gurevych, I., & Korhonen, A. (2024). Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. arXiv preprint arXiv:2406.03930.

[11] Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (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) (pp. 16366–16393). Association for Computational Linguistics.

[12] Pan, J. Z., Razniewski, S., **Kalo, J. C.**, Singhania, S., Chen, J., Dietze, S., ... & Graux, D. (2023). Large language models and knowledge graphs: Opportunities and challenges. Transactions on Graph Data and Knowledge (TGDK), 1(1), 2:1–2:38.

[13] Shaik, Z., Ilievski, F., & Morstatter, F. (2021). Analyzing race and country of citizenship bias in Wikidata. arXiv preprint arXiv:2108.05412.

[14] Su, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., ... & Chen, E. (2024). Large language models for generative information extraction: A survey. Frontiers of Computer Science, 18(6), 186357.

[15] Vashishth, S., Jain, P., & Talukdar, P. (2018, April). CESI: Canonicalizing open knowledge bases using embeddings and side information. In Proceedings of the 2018 World Wide Web Conference (pp. 1317–1327). International World Wide Web Conferences Steering Committee.

[16] Vrandečić, D. (2020). Architecture for a multilingual Wikipedia. arXiv preprint arXiv:2004.04733.
[17] Vrandečić, D. (2021). Building a multilingual Wikipedia. Communications of the ACM, 64(4), 38–41
[18] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. Communications of the ACM, 57(10), 78–85.