Language Bias in Wikipedia

CPSC-298 Wikipedia Governance Research Project

Cole Tamashiro, Michael Tamura, Kody Wong
Chapman University
Orange, California, USA
ctamashiro@chapman.edu,mtamura@chapman.edu,kowong@chapman.edu

Abstract

Our research investigates how language bias shapes the information presented across different language editions of Wikipedia. While Wikipedia strives to be a neutral and global knowledge platform, articles often vary significantly in tone, emphasis, and detail depending on the language in which they are written. Using the Wikipedia API, we collected 500-character summaries from various topics—including politics, history, and science—in English, Japanese, and Spanish. By comparing the outputs, we observed that certain details and perspectives were emphasized differently between languages, revealing cultural and linguistic biases embedded within the content. Although our analysis was primarily qualitative, the findings show that Wikipedia is not a direct translation across languages but rather a set of culturally shaped narratives. Understanding these differences is important because language bias can influence global access to reliable information and reinforce unequal knowledge representation online. Our project highlights the need for continued awareness and technical methods to detect and mitigate language bias in multilingual digital platforms.

Keywords

Wikipedia, governance, [add your keywords here]

1 Introduction

Wikipedia is one of the most widely used information platforms in the world, available in over 300 languages and read by millions of users daily. Despite its mission of neutrality and open collaboration, numerous studies and examples have shown that the way information is written and presented can differ depending on the language edition. These differences raise questions about how cultural and linguistic contexts influence the perceived objectivity of shared knowledge.

This paper investigates the presence of language bias within Wikipedia articles across different languages. Although Wikipedia strives to maintain a neutral point of view, the content, tone, and emphasis of articles can vary significantly depending on the language of publication. Such inconsistencies may alter how users understand global topics, particularly in fields like politics, history, and science.

Understanding language bias in multilingual platforms is critical because it affects access to accurate and balanced information worldwide. When information differs across language editions, readers may receive culturally filtered or incomplete perspectives. Addressing these discrepancies can promote greater transparency, equity, and inclusivity in digital knowledge sharing.

This paper investigates the following research questions:

(1) What Wikipedia topics show the most variation in language (e.g., politics, history, science)?
(2) How does language bias affect access to reliable information across underrepresented languages?
(3) How do editors' demographics and linguistic backgrounds shape the content and tone of Wikipedia articles?

The main contributions of this work are:

- A comparative analysis of Wikipedia articles across three language editions (English, Japanese, and Spanish) using the Wikipedia API.
- Evidence demonstrating how linguistic and cultural differences influence tone and emphasis in multilingual content.
- A framework for future computational analysis of bias using natural language processing tools.

The rest of this paper is organized as follows. Section 2 reviews related research on Wikipedia bias and multilingual governance. Section 3 describes our data collection and comparison approach using the Wikipedia API. Section 4 presents our main findings. Section 5 discusses the broader implications and limitations of our study. Finally, Section 6 concludes and outlines directions for future work.

2 Related Work

Research on Wikipedia has examined its community structure, editorial governance, and systemic bias across topics and languages. Scholars have long noted that while Wikipedia aspires to neutrality, it reflects the perspectives of its contributors and the linguistic communities that sustain it. Recent work has expanded beyond governance to explore how multilingual differences and translation asymmetries introduce subtle forms of cultural and ideological bias.

2.1 Wikipedia Governance and Community Culture

Reagle's foundational study, *Good Faith Collaboration* [**?**], explores how Wikipedia's open editing model and community norms support collaboration and self-governance. These practices, while promoting transparency and participation, also create conditions where biases can emerge through consensus-driven editing. Studies of Wikipedia governance have shown that editorial authority and content decisions are unevenly distributed, often reflecting dominant language groups or cultural perspectives. This line of research establishes the foundation for understanding how community-driven systems can both sustain openness and reproduce bias.

2.2 Linguistic and Ideological Bias on Wikipedia

Beyond governance, recent analyses have focused on how bias manifests linguistically. The Wikipedia article on ideological bias [3] summarizes political and cultural asymmetries across the platform, highlighting ongoing debates about neutrality. The Johns Hopkins University Hub report [1] provides empirical evidence that Wikipedia's multilingual content contains translation and framing differences shaped by cultural contexts. Similarly, New Scientist [2] demonstrates that facts and emphasis vary across language editions, influencing how topics are understood by readers in different regions. Together, these works show that language is a key factor in how bias operates within Wikipedia's ecosystem.

2.3 Our Work in Context

While prior research identifies systemic and linguistic bias in Wikipedia, few studies have directly compared multilingual article content using live data from the Wikipedia API. Our project extends this discussion by developing a simple computational framework to observe real-time differences between English, Japanese, and Spanish articles. By combining existing scholarship on governance and ideological bias with direct programmatic analysis, we provide a practical demonstration of how language shapes the representation of information. This work contributes to a growing effort to make Wikipedia's multilingual landscape more transparent, equitable, and linguistically aware.

3 Methodology

Our goal was to examine how the same topics on Wikipedia are represented differently across multiple languages. To achieve this, we developed a Python-based program that uses the Wikipedia API to extract short summaries of topics in English, Japanese, and Spanish. The analysis focused on identifying variations in language use, tone, and emphasis to explore how linguistic and cultural factors may influence content presentation. This approach allowed us to directly compare multilingual outputs for the same subject under consistent retrieval conditions.

3.1 Data Collection

We collected data using the Wikipedia API, which allows access to structured article content through simple queries. For each test case, a topic name (such as "Politics," "Climate Change," or "History of Japan") was entered under the `page` variable, and the program retrieved a 500-character summary. The same topic was then queried in different language editions—English, Japanese, and Spanish—by specifying the appropriate language code. This dataset provided side-by-side textual samples for qualitative comparison. All data were gathered between October and November 2025.

3.2 Data Processing

Because the Wikipedia API returns clean text snippets, minimal preprocessing was required. We stored the outputs in local text files for comparison. The main processing step involved aligning the English, Japanese, and Spanish summaries for each topic and reviewing them manually to note differences in phrasing, terminology, and emphasis. We also checked that each summary corresponded to the correct topic and language version. Future iterations of this project may include additional preprocessing, such as translation normalization or tokenization, to support automated linguistic analysis.

3.3 Analysis Methods

Our analysis was primarily qualitative. We compared summaries across the three language editions to identify differences in word choice, tone, and focus. Each team member independently reviewed the results, highlighting variations in emphasis (e.g., historical framing, political orientation, or omission of details). These findings were then discussed collectively to determine recurring patterns of linguistic bias. While no computational models were used at this stage, this manual approach provided insight into the types of bias that might later be quantified using natural language processing (NLP) tools such as sentiment or keyword frequency analysis.

3.4 Ethical Considerations

All data used in this study were publicly available through the Wikipedia API and complied with Wikimedia's Terms of Use. The project did not involve any private user information or identifiable data. We ensured that all references and excerpts from Wikipedia were properly cited, and that our analysis was limited to observing language patterns for academic and educational purposes.

4 Results

The analysis revealed consistent differences between Wikipedia articles across English, Japanese, and Spanish editions. While all three language versions covered similar topics, the emphasis, tone, and level of detail varied significantly. These findings demonstrate that linguistic and cultural factors influence how information is framed, even when describing the same subject.

4.1 Variation in Emphasis and Content Across Languages

The English editions of articles tended to include more comprehensive summaries with neutral phrasing and balanced structure. Japanese summaries, in contrast, often emphasized historical or cultural context, using more formal tone and indirect sentence structures. Spanish versions sometimes included more interpretive phrasing or additional regional examples, suggesting that editors localized content for their audiences. For example, when comparing political topics, the English summaries focused on factual chronology, while Spanish and Japanese summaries included contextual explanations or omitted controversial points.

4.2 Differences in Word Choice and Tone

Across multiple topics, language-specific nuances reflected distinct cultural communication styles. English summaries frequently relied on concise, technical vocabulary; Japanese entries employed respectful or honorific phrasing; and Spanish texts used more descriptive adjectives. These stylistic differences indicate that "neutrality" is interpreted differently across language communities. Even for scientific topics, such as "climate change" or "evolution," certain keywords were prioritized differently, showing how language affects framing and emphasis.

4.3 Observed Patterns of Omission and Addition

Another recurring pattern was selective omission or addition of information. Certain facts present in one language version were missing in others—most notably in politically sensitive or historically contested topics. For instance, while the English version of an article might include data or external references, the Japanese or Spanish version sometimes provided more narrative or interpretive explanations instead. These differences were subtle but consistent, supporting the hypothesis that editorial and cultural contexts shape the presentation of "neutral" information.

4.4 Summary of Findings

Overall, our results reveal that multilingual versions of Wikipedia do not simply replicate the same content in different languages. Instead, they represent localized perspectives shaped by linguistic structure, editorial norms, and cultural context. This outcome highlights the persistence of language bias even in platforms designed around collective neutrality, and underscores the need for more cross-lingual coordination among editors to ensure balanced global representation.

5 Discussion

5.1 Interpretation of Results

The results of our analysis show that Wikipedia's multilingual articles are not simple translations of one another but reflect distinct linguistic and cultural framing. Across English, Japanese, and Spanish, the same topics differed in tone, detail, and emphasis—particularly in politically or historically sensitive subjects. These differences suggest that editorial decisions, community norms, and cultural context shape how knowledge is represented. Our findings answer the main research question by confirming that language bias significantly influences how information is structured and presented across languages, even within a supposedly neutral platform like Wikipedia.

5.2 Implications

These findings have important implications for how readers interpret global information sources. For Wikipedia, understanding the extent of language bias highlights the need for greater cross-lingual collaboration among editors and tools that flag inconsistencies across versions. For researchers, this study underscores the value of examining multilingual content to uncover systemic bias. In practice, our results demonstrate the importance of critical digital literacy—users should recognize that the "neutral" presentation of facts may differ based on the language they read.

5.3 Limitations

Our analysis was limited in both scale and automation. The Wikipedia API program retrieved only short 500-character summaries, which capture limited context. The study also compared a small number of languages, meaning our findings cannot be generalized across all of Wikipedia.

6 Conclusion

This research investigated how language bias influences the content and interpretation of Wikipedia articles across different language editions. Although Wikipedia is built on the principle of neutrality, our analysis revealed noticeable differences in emphasis, tone, and detail between English, Japanese, and Spanish versions of the same topics. Using the Wikipedia API, we collected and compared summaries to illustrate how cultural and linguistic contexts shape what is presented as "neutral" knowledge.

Our findings demonstrate that Wikipedia is not simply translated content but a reflection of localized editorial norms and community perspectives. This highlights the importance of considering linguistic diversity when evaluating information reliability and accessibility. By showcasing how even factual entries can differ by language, this project contributes to the broader understanding of systemic and cultural bias in online information platforms.

Future work could expand this approach by integrating natural language processing tools to quantify bias across a larger dataset and by exploring strategies for promoting greater consistency and inclusivity in multilingual digital knowledge systems.

Acknowledgments

We thank … for ….

References

[1] Johns Hopkins University. 2025. *Finding Hidden Biases in Wikipedia's Multilingual Content.*
[2] New Scientist. 2016. *Wikipedia Facts Depend on Which Language You Read Them In.*
[3] Wikipedia contributors. 2024. *Ideological bias on Wikipedia.*

A AI Usage Documentation

A.1 Literature Review

We used an AI-assisted literature review workflow (see the file literature-review.prompt.md) to summarize and analyze key sources about bias and multilingual representation on Wikipedia. The agent processed three core articles from Wikipedia, Johns Hopkins University, and New Scientist. It extracted summaries, methodologies, and key findings, which were then manually reviewed and incorporated into our related work section. The AI also helped generate BibTeX entries for our references file.

A.2 Data Analysis

AI assistance was not directly used to analyze the data. Instead, we manually compared outputs from the Wikipedia API in English, Japanese, and Spanish to identify qualitative differences in phrasing, tone, and content. However, we plan to extend this work by integrating a natural language processing (NLP) model in the future to automatically detect patterns of bias.

A.3 Writing Assistance

We used ChatGPT to support various parts of the writing process, including improving the clarity and structure of the abstract, generating section outlines for consistency with the provided ACM LaTeX template, and editing sentences for readability. All content

was reviewed and revised by group members before inclusion in the final paper.

A.4 Code Development

ChatGPT was used for limited code guidance when setting up and debugging the Wikipedia API script (week7.py). The AI provided suggestions on how to retrieve summaries, handle different languages, and format results in a readable output. The final implementation was reviewed and tested by our group.

A.5 Verification

All AI-generated material was reviewed and verified by the authors. Summaries were cross-checked with original sources, and all code suggestions were tested for accuracy. No AI-generated text or code was used without human editing or confirmation. Factual claims were verified against primary data and research papers to ensure accuracy and integrity.