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

Wikipedia offers global access to knowledge across more than 300 languages, but substantial evidence shows that articles differ in tone, emphasis, and detail across language editions. This study investigates how historical topics are represented differently across English, Japanese, and Spanish Wikipedia by analyzing short summaries retrieved through the Wikipedia API. We operationalize multilingual variation through three measurable proxies: word count, word choice, and topical emphasis. Our findings show that English summaries provide the most contextual detail, Spanish summaries often emphasize political or institutional actors, and Japanese summaries favor concise and neutral phrasing. These consistent patterns demonstrate that linguistic and cultural factors shape how historical information is framed across languages. The study contributes a reproducible, proxy-based methodology for detecting multilingual variation and highlights the need for cross-lingual coordination to promote equitable global knowledge representation.

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 focuses on multilingual differences within Wikipedia, particularly in the domain of historical topics. While Wikipedia strives for a neutral point of view, articles about historical events and periods often vary in tone, emphasis, and content depth across different languages. These inconsistencies may influence how readers understand major historical topics, shaping perceptions in ways that reflect broader cultural or editorial priorities.

Understanding these linguistic differences is important because Wikipedia serves as a foundational information source for users around the world. When articles differ across language editions, readers may receive culturally filtered or uneven perspectives on shared global history. A systematic examination of these differences helps reveal how language influences access to balanced and accurate information.

This study distinguishes between the broader motivating research question and the specific operationalized question that can be directly measured with our dataset:

Motivating Research Question: *How do different language editions of Wikipedia shape the way historical topics are presented across cultures?*

Operationalized Research Question: *How do word count, word choice, and content emphasis differ in History-related Wikipedia summaries across English, Japanese, and Spanish, as retrieved through the Python Wikipedia API?*

The main contributions of this work are as follows:

- A multilingual comparison of History-related Wikipedia summaries in English, Japanese, and Spanish.
- Empirical evidence showing how linguistic and cultural factors influence tone, detail, and emphasis across language editions.
- A proxy-based framework for future computational measurement of multilingual Wikipedia bias.

The rest of this paper is organized as follows. Section **??** reviews related work on multilingual Wikipedia bias and governance. Section **??** describes our data collection process and proxy-based comparison approach. Section **??** presents our empirical findings with respect to our operationalized research question. Section **??** discusses broader implications, limitations, and interpretive considerations. Finally, Section **??** summarizes our conclusions and notes directions for future work.

2 Related Work

Research on Wikipedia has examined its community structure, governance mechanisms, and the systemic biases that emerge across topics and languages. Scholars widely acknowledge that while Wikipedia aspires to neutrality, the platform reflects the cultural, linguistic, and ideological perspectives of its contributors. Recent work in particular highlights how multilingual editions of Wikipedia diverge due to translation asymmetries, local editorial norms, and culturally specific framing choices.

2.1 Wikipedia Governance and Community Culture

Reagle's foundational study, *Good Faith Collaboration* [**?**], examines how Wikipedia's norms, policies, and consensus-based editing processes shape participation. While these structures promote openness, they also produce uneven influence, often privileging dominant language communities and long-term editors. Prior work on governance demonstrates that cultural preferences embedded in

community norms can subtly affect how content is written, updated, and moderated.

2.2 Linguistic and Ideological Bias on Wikipedia

Beyond governance, recent analyses directly examine how linguistic and ideological biases manifest in Wikipedia articles. The Wikipedia article on ideological bias [3] outlines patterns of political framing across language editions, while empirical studies such as the Johns Hopkins University Hub report [1] show that multilingual articles differ not only in length but also in emphasis and tone. New Scientist [2] similarly reports that language editions selectively add or omit facts, creating culturally inflected versions of the same topic. Collectively, this work shows that multilingual Wikipedia cannot be treated as a uniform body of information.

2.3 Our Work in Context

While existing research identifies systemic and linguistic bias broadly, few studies have operationalized multilingual comparison through short-form summaries obtained directly through the Wikipedia API. Our project addresses this gap by examining English, Japanese, and Spanish summaries of historical topics using measurable proxies—word count, word choice, and topical emphasis. By focusing on concise API-retrieved summaries, our work offers a reproducible, fine-grained perspective on how historical information diverges across languages. This connects high-level scholarship on Wikipedia bias with a concrete, computational methodolog

3 Methodology

Our project investigates how multilingual Wikipedia articles differ in their presentation of historical topics. While our motivating research question asks how cultural or linguistic factors influence content on Wikipedia, this is too broad to measure directly. Therefore, we operationalize the question by examining three measurable proxies: (1) word count, (2) word choice patterns, and (3) topical emphasis within article summaries. These proxies enable us to observe differences in language usage that may indicate broader cultural framing or editorial tendencies.

To bridge the gap between the motivating and operationalized questions, we treat each proxy as an observable indicator of deeper concepts such as cultural perspective or linguistic bias. For example, shorter summaries may reflect editorial priorities or content availability, while variations in word choice may reveal differences in how societies frame historical events. Although proxies cannot fully capture the complexity of cultural interpretation, they provide a reproducible foundation for comparative analysis.

3.1 Data Collection

Data were collected using the Wikipedia API, which provides structured programmatic access to article content. Our Python script (linked in our GitHub repository) queries the API by specifying a topic name and a corresponding language code. For each historical topic, we retrieved a 500-character summary from the English, Japanese, and Spanish Wikipedia editions. The script stored each summary, along with metadata such as the language and timestamp, into local text files for comparison.

Topics included broad historical themes such as "World War II," "Industrial Revolution," and "History of Japan." All data were collected between October and November 2025 under consistent retrieval conditions to ensure comparability across languages.

3.2 Data Processing

Because the Wikipedia API returns clean text, preprocessing requirements were minimal. Each summary was saved as a separate file and grouped by topic. We verified that the retrieved summaries accurately matched the intended page in each language. For each topic, we aligned the English, Japanese, and Spanish summaries as a triplet to support side-by-side comparison.

To prepare for analysis, we calculated summary word counts and conducted a manual inspection of key vocabulary (e.g., references to dates, political actors, cultural terminology). Future work may incorporate automated preprocessing methods such as translation normalization, tokenization, or part-of-speech tagging to support quantitative linguistic comparison.

3.3 Analysis Methods

Our analysis combines proxy-based measurement with qualitative review. For each multilingual summary set, we evaluated:

- **Word Count (Proxy 1):** Used as a measurable indicator of content depth or editorial emphasis.
- **Word Choice (Proxy 2):** Identified through recurring nouns, adjectives, and historically loaded terms.
- **Topical Emphasis (Proxy 3):** Observed via which facts, figures, or contextual details each language highlighted or omitted.

These proxies allowed us to track concrete differences without requiring full article translations. We conducted a manual qualitative comparison to identify patterns such as differing emphasis on causes vs. effects, regional framing, or political neutrality. While this stage did not rely on computational NLP tools, our approach sets the foundation for scalable quantitative analysis such as keyword frequency counts, sentiment analysis, or semantic similarity scoring in future iterations.

4 Results

This section presents findings addressing our operationalized research question: *How do word count, word choice, and topical emphasis differ in History-related Wikipedia summaries across English, Japanese, and Spanish?* All results derive from the API-collected summaries described in Section **??**.

4.1 Word Count Differences

English summaries were consistently the longest across all topics. For example, the English summary for "World War II" contained 498 characters, compared to 452 in Spanish and 311 in Japanese. A similar pattern appeared for all historical topics tested. This suggests that English Wikipedia tends to provide more context within the summary, while Japanese summaries are more concise and Spanish summaries fall in between.

4.2 Word Choice Differences

Distinct linguistic patterns emerged across languages. English summaries frequently used globally standardized historical vocabulary such as "allies," "conflict," and "industrialization." Spanish summaries used terms emphasizing political actors and institutions (e.g., "potencias," "gobiernos"). Japanese summaries employed neutral, descriptive phrasing and avoided evaluative vocabulary.

These differences reflect each language community's stylistic norms and conventions in presenting historical information.

4.3 Differences in Emphasis and Focus

English summaries tended to foreground causes, consequences, and chronological structure. Spanish summaries sometimes emphasized political interpretation or regional context. Japanese summaries often emphasized definitions, periods, or structural overviews rather than political framing.

These patterns suggest that different language editions prioritize different narrative elements, even for widely shared historical subjects.

4.4 Summary of Findings

Across all three proxies—word count, word choice, and emphasis—multilingual summaries showed systematic differences. English provided the most detail, Spanish offered interpretive nuance, and Japanese prioritized concise and neutral description. These findings demonstrate that even short summaries exhibit culturally and linguistically shaped variation.

5 Discussion

5.1 Interpretation of Results

Our results confirm that multilingual Wikipedia summaries reflect culturally and linguistically distinct framing rather than direct translation. English summaries tended to provide broader contextual detail, Spanish summaries often highlighted institutional or political elements, and Japanese summaries favored concise structural information. These differences show that historical narratives are constructed differently across language editions, shaped by editorial norms and cultural communication styles.

5.2 Implications

These findings have implications for how global audiences consume historical information. Readers relying on different language editions may receive substantively different interpretations of major historical events. For Wikipedia, this underscores the need for improved cross-language alignment tools and awareness among editors. For researchers, the results demonstrate the value of proxy-based multilingual comparison for detecting subtle forms of linguistic bias.

5.3 Limitations

This study is limited by its use of short 500-character summaries, which provide only a narrow window into full article content. The number of topics compared was also small, and no translation normalization or automated linguistic analysis was used. Future work could expand the dataset, incorporate NLP methods, or compare full article structures to achieve a deeper understanding of cross-lingual historical framing.

6 Conclusion

This study examined how Wikipedia's multilingual editions present historical topics differently by analyzing 500-character summaries in English, Japanese, and Spanish. Using word count, word choice, and topical emphasis as measurable proxies, we identified systematic differences across languages. English summaries provided the greatest level of detail, Spanish summaries offered more interpretive framing, and Japanese summaries emphasized concise and neutral description.

These findings show that even short summaries reflect culturally shaped editorial choices. Wikipedia, despite its neutrality policy, presents distinct linguistic versions of history depending on the language edition. Our work provides a reproducible framework for detecting such differences using simple API-based methods.

Future research should expand the scale of analysis, incorporate automated translation alignment or NLP-based metrics, and explore full-article structural differences to better understand multilingual knowledge representation.

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.