Machine-assisted Content Creation on Peer Production Platforms: Insights from Machine Translation on Multilingual Wikipedia

Abstract

We examine how AI-powered technology shapes content creation in peer production platforms by studying Wikipedia's integration of Google Translate in January 2019. Using difference-in-differences analysis across over 100 Wikipedia language editions, we find that this technological change increased translation volume by 149% while maintaining content quality and reader engagement. However, these benefits are not equally distributed - popular and high-resource languages experience significantly larger gains compared to low-resource languages, suggesting technology may amplify existing resource advantages. Analysis of biographical and geographical content reveals progress in addressing representation gaps, though substantial disparities persist. Our findings on the interplay between technological capability and structural inequalities carries important implications for platform design as well as leveraging AI technology for democratizing knowledge access.

Keywords: machine translation, artificial intelligence, digital language divide, multilingual environment, peer production platform

1 Introduction

Language barriers create substantial costs in digital content creation and knowledge sharing, leading to unequal access to online information (Ginsburgh and Weber, 2020). This "digital language divide" is particularly acute within collaborative platforms that aim to democratize information access. Despite its role as a major source of open information, Wikipedia exhibits substantial content disparities across languages as well (Miquel-Ribé and Laniado, 2020; Johnson and Lescak, 2022). For example, English Wikipedia contains over 6 million articles, while Bengali Wikipedia, representing the sixth most spoken native language globally, has only about 100,000 articles. These disparities limit both information access and exposure to diverse perspectives, creating persistent gaps in global knowledge sharing (Nagaraj, 2022; Pereira et al., 2024).

The importance of addressing these language barriers extends beyond Wikipedia to fundamental questions of information equity in the digital age. Research demonstrates that differential access to online information affects a wide range of social and economic outcomes, from consumer behavior (Hinnosaar et al., 2023) and financial decision-making (Xu and Zhang, 2013) to scientific knowledge production (Thompson and Hanley, 2018) and judicial decision-making (Thompson et al., 2024). As digital platforms become central channels for information dissemination, linguistic constraints can reinforce existing socioeconomic disparities and limit the potential benefits of digital technologies.

Recent advances in artificial intelligence, particularly machine translation technologies, have the potential to substantially reduce the costs of cross-language communication (Brynjolfsson et al., 2019). Unlike earlier rule-based systems, modern neural machine translation can process semantic context and linguistic nuance, enabling more natural translations at scale. This technological capability to efficiently translate content across languages could theoretically reduce information barriers and expand multilingual digital infrastructure (Goldfarb and Tucker, 2019; Agrawal et al., 2022). However, the realized benefits of such tech-

nologies may depend on existing platform structures, community resources, and systematic inequalities in digital ecosystems.

In this study, we examine **whether machine translation technology can effectively address the shortage of multilingual content on peer production platforms**. Specifically, we investigate how reducing translation barriers through improved technology affects both the volume and distribution of content creation across Wikipedia's diverse language communities. Our research design leverage Wikipedia's integration of Google Translate into its Content Translation system in January 2019 as a natural experiment, exploiting variation in translation technology access across languages and time.

While machine translation has proven effective for content consumption, its application to content production presents distinct challenges in maintaining accuracy and quality. Wikipedia's integration of Google Translate provides an ideal setting to study these challenges, as it combines volunteer-driven content creation with structured human-machine collaboration in translation workflows. The Google Translate integration enables us to analyze how combining editorial expertise with machine translation technology affects both the efficiency and quality of content creation in a decentralized platform.

Using this natural experiment and a difference-in-differences strategy, we examine both the aggregate and distributional impacts of machine translation on Wikipedia's content ecosystem. On the supply side, we first analyze: (1) changes in content production by measuring article translation volume, and (2) content quality maintenance through deletion rates and structural metrics. On the demand side, we investigate: (3) reader engagement patterns to assess the value of translated content. Finally, we examine distributional effects through: (4) changes in information flow patterns across language editions, and (5) whether reduced translation costs help address systematic content coverage gaps.

Our empirical analysis reveals substantial but uneven impacts of machine translation technology. Integration of Google Translate increased translation volume by 149%, representing 71 additional articles per month per language edition. This growth coincides with im-

proved content quality, as measured by decreased article deletion rates and sustained reader engagement. However, these benefits are not uniformly distributed. High and medium-resource languages experience significantly larger gains compared to low-resource languages, suggesting that technological advancement may amplify existing resource advantages.

Moreover, we find that machine translation reshapes patterns of knowledge flow across languages. While English increasingly serves as a central source language, we observe greater diversity in target languages receiving translations. Analysis of specific content domains reveals varying impacts on representation gaps. Biographical content shows progress toward gender balance relative to simulated benchmarks, though substantial disparities persist. Geographic content demonstrates particularly strong growth in regions with high linguistic diversity, suggesting technology particularly benefits areas where translation barriers were previously most acute.

These findings advance our understanding of how artificial intelligence affects decentralized knowledge production. While machine translation reduces content creation costs, its benefits depend on existing platform structures and resource distributions. This interaction between technological capability and structural inequalities carries implications for platform design and knowledge democratization efforts. Our results suggest that addressing digital divides requires combining technological solutions with targeted support for under-resourced communities.

The paper proceeds as follows: Section 2 reviews background and literature on digital language barriers and machine translation. Section 3 describes our empirical setting. Section 4 details data sources and empirical strategy. Section 5 presents findings on machine translation's impact on content creation and knowledge accessibility. Section 6 concludes with implications for bridging digital language divides.

2 Background and Literature

Our study examines how artificial intelligence affects content creation costs in digital platforms. This focus on machine-assisted content production is particularly relevant given recent developments in generative capabilities powered by AI technologies. While existing research has primarily examined machine translation for content consumption (such as reading foreign websites or understanding cross-border communication), we analyze its role in content creation - a distinct challenge that requires maintaining quality standards and integrating with human workflows.

Research has extensively documented various knowledge gaps in digital content, with disparities across gender, geographic regions, and knowledge domains (Halavais and Lackaff, 2008; Graham et al., 2014; Wagner et al., 2015, 2016). Studies show that although organized editing campaigns have enhanced content, these efforts often struggle to achieve sustained visibility (Zhu et al., 2020; Langrock and González-Bailón, 2022). The linguistic dimension of digital knowledge gaps remains relatively understudied (Miquel-Ribé and Laniado, 2020; Johnson and Lescak, 2022), despite language barriers presenting fundamental constraints on content creation and distribution.

Research on socio-technical systems reveals how platform architecture shapes information distribution and accessibility. Recent work by Parker (2016) and Constantinides et al. (2018) demonstrates that digital platforms are not merely technical infrastructure but complex systems where technology interacts with social dynamics. The concept of 'platform affordance' explains how technological features can either democratize knowledge or perpetuate existing gaps, often amplified by technical constraints and social structures (Forman et al., 2008; Aaltonen and Seiler, 2016). This framework helps understand how technological changes, such as improved translation tools, might have heterogeneous effects across user communities.

The evolution of machine translation technology, particularly through neural networks, represents a significant advancement in addressing linguistic barriers. Modern approaches

using transformer architectures, exemplified by Google Translate, achieve substantially lower error rates by processing entire sentences and incorporating global context (Bahdanau et al., 2014; Vaswani et al., 2017). These improvements in neural machine translation enable more accurate and nuanced translations (Wu et al., 2016), potentially reducing the costs of cross-language content creation. The technical progress in machine translation extends beyond accuracy improvements to enhance the accessibility and inclusiveness of digital information systems (Turovsky, 2016).

However, empirical evidence on the real-world impacts of machine translation in digital platforms remains limited. While the technology has been rapidly adopted across sectors including business, legal, and healthcare, comprehensive analyses of its effects on platform operations and user behavior are scarce. A notable exception is Brynjolfsson et al. (2019), who examine how eBay's machine translation system affects cross-border e-commerce. Their research shows that reducing language barriers through machine translation significantly increases international trade, demonstrating how technological reduction of communication costs can expand market activity.

Our study bridges these literature streams by examining how machine translation technology affects content creation and knowledge distribution in a collaborative platform. We analyze both aggregate productivity effects and distributional consequences, providing evidence on how technological change interacts with existing platform structures to shape information accessibility. This investigation contributes to understanding the evolving role of artificial intelligence in democratizing access to multilingual information, while highlighting potential limitations when technological solutions encounter systemic inequalities.

3 Empirical Setting

The Wikipedia Content Translation system serves as our empirical setting for examining how machine translation technology affects multilingual content production. Traditionally, Wikipedia editors create articles by writing content from scratch, which requires substantial time and language expertise. Wikipedia introduced the Content Translation system in 2015 to reduce these barriers, providing a structured environment for translating existing articles across language editions. The system displays source and target content in a side-by-side interface, enabling editors to work systematically while preserving article structure and metadata. This design supports both manual translation and machine-assisted content creation, offering a more efficient alternative to writing articles from scratch.

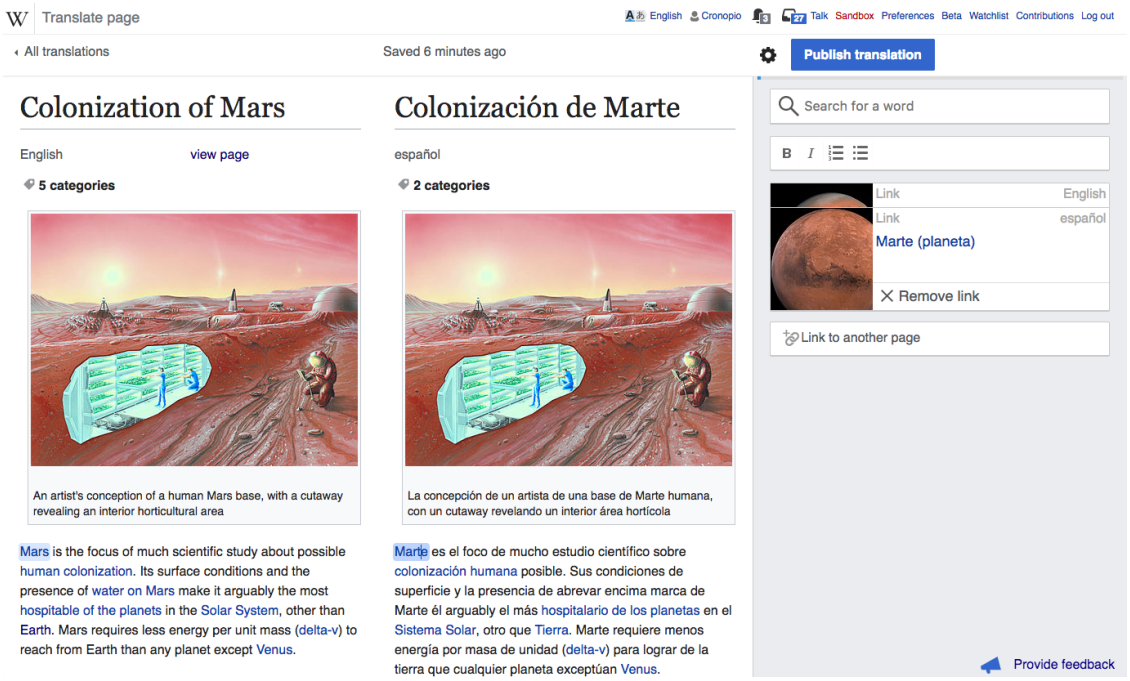
The translation workflow consists of several key components. First, editors select an article from one Wikipedia language edition to translate into another language. The system then presents a dual-pane interface (Figure 1) that displays the source article on the left and the translation workspace on the right. This layout enables direct comparison between source and target content throughout the translation process. The system also automates technical aspects such as preserving internal links, maintaining category tags, and transferring media assets, allowing editors to focus on content accuracy and quality.

Editors can choose between fully manual translation or machine-assisted drafts that they subsequently review and refine. This flexibility accommodates different translation preferences while maintaining Wikipedia's quality standards. The system's adoption has been substantial - by late 2024, it had facilitated the creation of over 2 million articles across Wikipedia's language editions[1].

Prior to 2019, the Content Translation system integrated machine translation engines including Apertium and Yandex. While these tools provided basic translation assistance, their capabilities were limited in terms of accuracy and language coverage. The integration of

[1]See this page: https://en.wikipedia.org/wiki/Special:ContentTranslationStats

Figure 1: Wikipedia's Content Translation Tool Interface



Note: The interface displays a side-by-side translation environment where editors can view the source article (left panel) and create the translated version (right panel). The example shows an article being translated from English to Spanish, demonstrating the tool's core functionality for cross-language content creation.

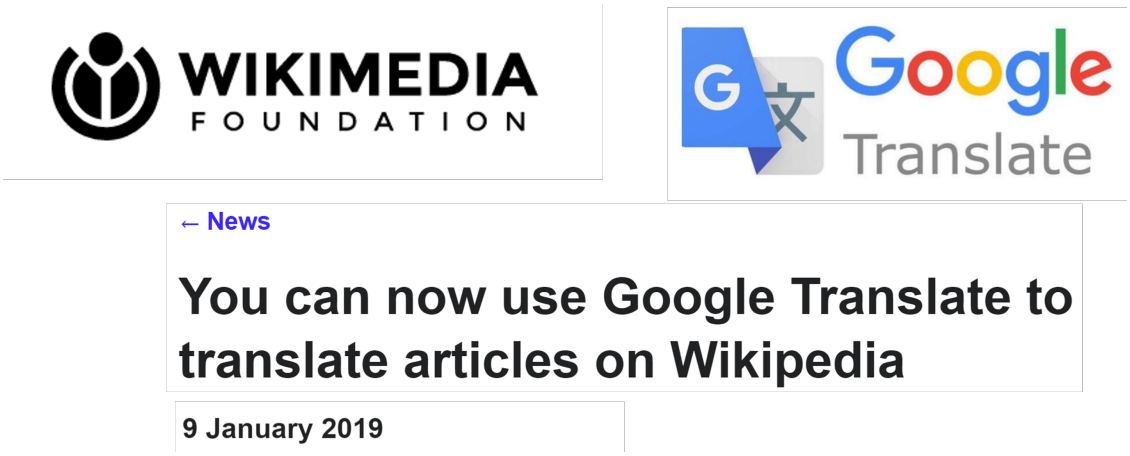
Google Translate in January 2019 marked a significant technological enhancement (Figure 2), particularly for languages previously served by less sophisticated translation tools[2].

Google Translate's neural machine translation architecture represents a substantial advancement over previous translation technologies. The system employs transformer-based models that process entire sentences as coherent units, enabling more nuanced understanding of language context and meaning. Published evaluations indicate that Google Neural Machine Translation (GNMT) reduces translation errors by 55-85% across major language pairs compared to previous approaches.

Importantly, Wikipedia implements Google Translate within a human-in-the-loop framework. Editors use machine-generated translations as initial drafts, applying their expertise

[2]See the announcement from Wikimedia Foundation: https://wikimediafoundation.org/news/2019/01/09/you-can-now-use-google-translate-to-translate-articles-on-wikipedia/

Figure 2: Wikimedia Foundation's Announcement of Google Translate Integration



Note: Official announcement from January 9, 2019, marking the integration of Google Translate into Wikipedia's Content Translation tool. This update provided editors with neural machine translation capabilities while maintaining Wikipedia's open knowledge principles.

to review, revise, and refine the content. This approach acknowledges that while machine translation can efficiently generate draft translations, human oversight remains essential for ensuring accuracy and preserving cultural nuances across languages.

The integration of Google Translate into the Content Translation system represents a significant step in Wikipedia's efforts to enhance information the availability across languages. Additionally, all machine translated content fall under Wikipedia's free license framework, maintaining consistency with the platform's open knowledge principles. The implementation also preserves user privacy by avoiding transmission of personal data to external servers.

Finally, Content Translation system tracks machine translation usage patterns, generating data about technology adoption and impact. This empirical setting provides an ideal context for examining how artificial intelligence affects content production in peer production platforms. The clear pre/post technology integration timeline, variation in language coverage, and detailed usage data enable rich analysis of both aggregate and distributional effects of machine translation technology.

4 Data and Method

4.1 Data

Our empirical analysis draws on data from several sources. First, we obtain comprehensive translation records from Wikipedia's Content Translation API, which documents article translations between language pairs. Each record contains metadata including source and target article titles, URLs, publication dates, and the specific machine translation engine used (if any)[3]. The sample spans from January 2017 to December 2020, providing approximately two years of observations both before and after Google Translate's integration.

Second, we leverage fine-grained article-level pageview data to analyze content consumption patterns. These data enable us to measure reader engagement with both source and translated articles and evaluate whether machine-assisted content provides value from the demand perspective.

Finally, we combine several complementary datasets to characterize article content and quality. These include: (1) article topic classifications from a language-agnostic model for consistent categorization across languages, (2) structured metadata from Wikidata[4] for article attributes such as biographical and geographical information, (3) complete article editing and deletion histories to measure content maintenance, (4) structural quality metrics developed by Wikimedia Foundation to assess article completeness, and (5) the Knowledge Gap Index to establish baseline measures of content availability across languages and categories[5]..

Our final sample comprises 540,488 translated articles spanning 132 target languages. We aggregate this article-level data into a monthly panel from January 2017 to December 2020[6]. The average monthly translation volume per language is 85 articles. Table 1 provides

[3]For more information about data on Content Translation on Wikipedia, please see Appendix A

[4]Please refer to Wikidata's official web page for more information: https://www.wikidata.org/

[5]For more details and data access, please see Appendix B

[6]The sample was constructed with the following specific criteria: 1) Target languages were required to have a minimum of 20 translated articles throughout the sample period. 2) English Wikipedia was excluded as a target language due to its unique restriction on machine translations. 3) Language editions with abnormal spikes in translations were excluded. The abnormal spike is defined as those with a single month exceeding

summary statistics for the sample.

Table 1: Summary Statistics of Wikipedia Content Translation Activities, 2017-2020

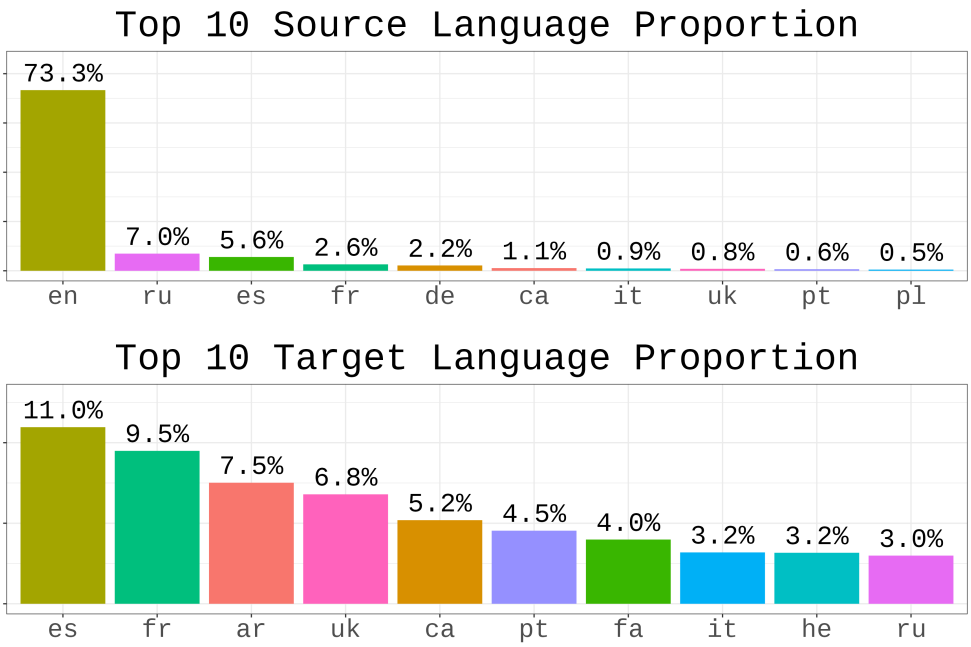
Metric	**Value**
Number of Translations	540,488
Sample Start Time	2017-01-01
Sample End Time	2020-12-31
Number of Unique Target Languages	132
Average Translations Volume per Language/Month	85

Analysis of source and target language distributions reveals clear patterns (Figure 3). English serves as the dominant source language, accounting for 73.3% of translations, with Russian, Spanish, and French as secondary sources. Target languages show greater diversity, with Spanish receiving 11% of translations, followed by French, Arabic, and Ukrainian. This pattern indicates concentrated sourcing from English while achieving broader distribution across target languages.

The topical distribution of translated articles reveals concentrated activity in certain domains. As shown in Table 2, Culture (45.12%) and Geography (30.0%) account for the majority of translations, substantially exceeding STEM (15.3%) and History and Society (9.5%). Further analysis at the subcategory level (Figure 4) identifies Geography.Regions and Culture.Biography as the most frequently translated topics. This distribution informs our subsequent analysis of how machine translation affects knowledge gaps across content domains.

20 times the monthly average, potentially indicating organized campaigns

Figure 3: Distribution of Top Source and Target Languages in Wikipedia Article Translations, 2017-2020

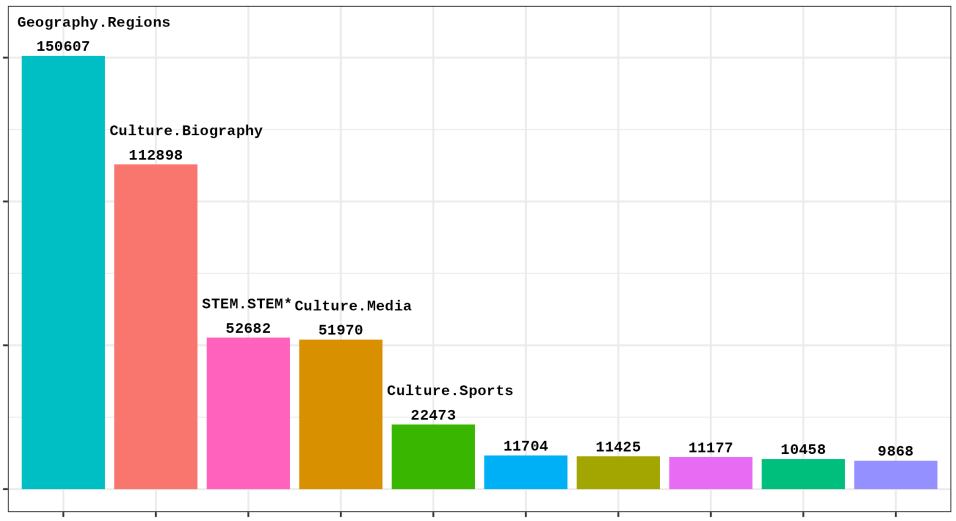


Note: Bar charts showing language distributions in Wikipedia translations. English dominates as source language (73.3%), while target languages show greater diversity with Spanish (11%), French, Arabic, and Ukrainian as primary recipients.

Table 2: Topical Distribution of Translated Wikipedia Articles, 2017-2020

Article Topic	**Article Count**	**Article Percentage**
Culture	231830	45.12%
Geography	154260	30.0%
STEM	78726	15.3%
History and Society	48680	9.5%

Figure 4: Distribution of Translated Wikipedia Articles by Level 2 Topic Categories



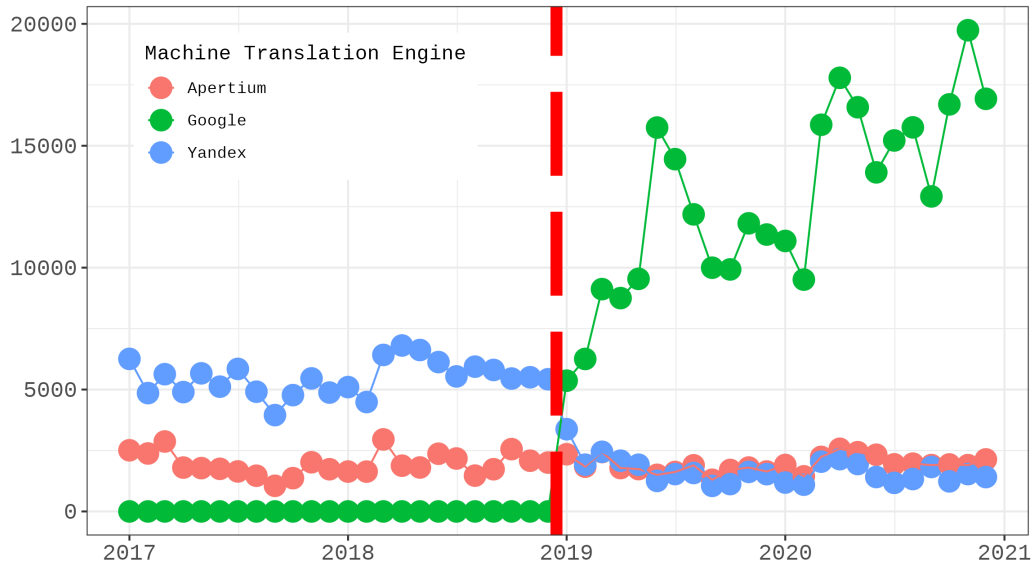
Note: Bar chart displaying translation frequency across topic subcategories. Geography.Regions and Culture.Biography are the most translated topics, together comprising close to 50% of all translations.

4.2 Model-free Evidence

Before examining our main empirical results, we analyze descriptive patterns in Wikipedia's translation data. This analysis focuses on three key metrics: the adoption of different machine translation engines, overall translation volume, and the balance between machine-assisted and manual translation.

Figure 5 tracks the monthly usage of machine translation engines from 2017 to 2020. Prior to Google Translate's integration in January 2019 (marked by the vertical red line), editors primarily used Yandex and Apertium, with relatively stable monthly volumes. Following integration, Google Translate shows rapid adoption, accompanied by a corresponding decline in Yandex usage. This pattern suggests both technological substitution and increased overall reliance on machine translation.

Figure 5: Monthly Usage Trends of Machine Translation Engines in Wikipedia, 2017-2020

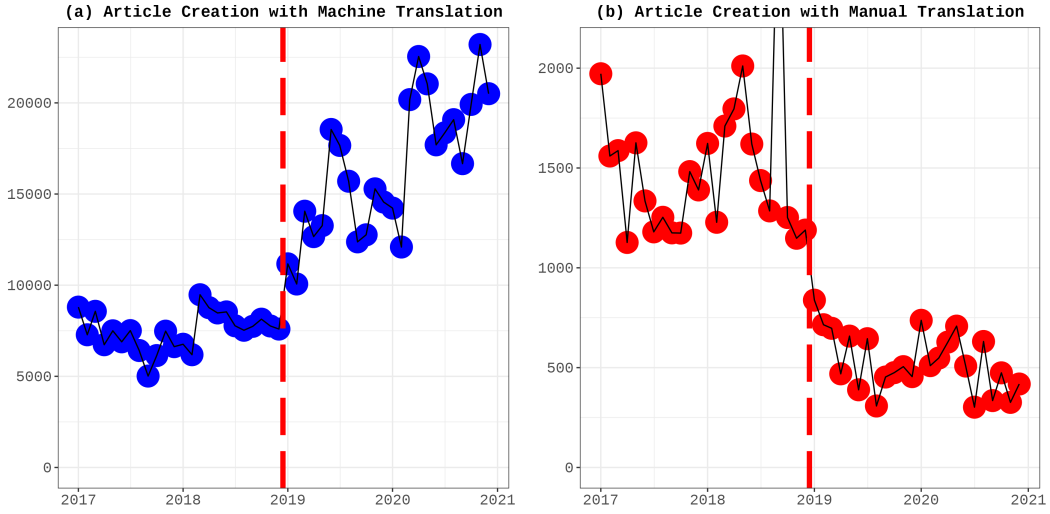


Note: Time series showing article translations by machine translation engine. The vertical red line marks Google Translate's January 2019 integration, after which it becomes the dominant engine while Yandex usage declines.

The adoption of Google Translate coincides with broader changes in translation patterns.

Panel (a) of Figure 6 shows the total volume of machine-assisted translations. While pre-integration volume fluctuated around 8,000 articles monthly, the post-integration period shows an immediate increase followed by sustained higher volumes. This aggregate increase in machine-assisted translation is mirrored by a decline in manual translation, as shown in Panel (b). The reduction in manual translation volume suggests editors are shifting toward machine-assisted methods, likely due to efficiency gains.

Figure 6: Article Translation Volume With and Without Machine Assistance, 2017-2020



Note: Panel (a) shows monthly volume of machine-assisted translations, while Panel (b) shows manual translations. The vertical red line marks Google Translate's January 2019 integration, revealing an increase in machine-assisted translations and corresponding decline in manual translations.

These descriptive patterns suggest three preliminary findings. First, Google Translate achieved rapid adoption within Wikipedia's translation ecosystem, becoming the dominant machine translation engine. Second, this technological change appears to have increased overall translation activity rather than merely substituting between translation methods. Third, the shift from manual to machine-assisted translation indicates changes in editor behavior following the technological intervention.

While these patterns provide suggestive evidence of Google Translate's impact, they do not establish causality. The following sections employ formal empirical strategies to identify

causal effects and examine heterogeneity across languages and content types.

4.3 Empirical Strategy

Our empirical strategy exploits Wikipedia's integration of Google Translate as a natural experiment to examine how machine translation technology affects content production in peer production platforms. We implement a difference-in-differences framework that leverages three sources of variation in the rollout of Google Translate across Wikipedia's language editions.

First, the timing of Google Translate integration stemmed from a top-down partnership between Wikimedia Foundation and Google, independent of individual language communities' needs. This organizational decision provides exogenous variation in technology adoption. Second, Google Translate's language support capabilities created natural treatment and control groups, as the system could initially translate content for approximately one-third of Wikipedia's language editions. We focus our analysis on languages ranking in the top 150 by Wikipedia size to ensure comparability between treated and control groups. Third, while most supported languages integrated Google Translate in January 2019, some languages experienced later integration through 2020, providing additional temporal variation (Table 3).

Table 3: Timeline of Google Translate Integration across Wikipedia Language Editions

Language Grouping	Integration Date	Number of Languages
Control Languages	No Integration	42
Treated Languages	2019-01-01	78
	2019-06-01	1
	2019-10-01	1
	2020-04-01	5
	2020-06-01	2
	2020-08-01	3

Base on the empirical strategy described above, our baseline model specification is as follows:

$$D_{it} = \beta \times Google\ Translate\ Integration_{it} + \gamma_i + \delta_t + \epsilon_{it}$$

where D_{it} represents outcomes (e.g., monthly translated articles) for Wikipedia language edition i at month t. $Google\ Translate\ Integration_{it}$ indicates technology availability, while γ_i and δ_t represent language and time fixed effects respectively. The identifying assumption is that treated and control languages would have exhibited parallel trends in translation activity absent Google Translate integration.

For model estimation, we employ multiple contemporary difference-in-differences (DiD) estimators to ensure robust inference. These include traditional two-way fixed effects models, Sun and Abraham interactions, Callaway and Sant'Anna estimators, and counterfactual imputation framework. Results remain highly consistent across these methodologies, with detailed comparisons presented in Appendix C. Our main analysis focuses on the counterfactual imputation method using matrix completion, which offers particular advantages for our empirical setting.

The matrix completion method (Athey et al., 2021) constructs a panel data matrix where rows represent language editions, columns represent time periods, and entries capture outcome variables. The method exploits stable patterns across both dimensions to impute counterfactual outcomes for treatment effect estimation. This approach offers advantages in our setting: it accommodates heterogeneous treatment effects across languages and time periods, and utilizes both control units and pre-treatment patterns for counterfactual construction.

To further verify robustness, we also implement a simplified analysis focusing only on languages that integrated Google Translate in January 2019. This approach avoids econometric complications from staggered adoption while capturing the majority of treated languages. Results from this simplified specification closely match our primary findings. We employ this framework particularly for our heterogeneity analyses, where its clearer identification supports examination of effect variation across language and topic types.

5 Results

5.1 Primary Effects on Content Production

We begin our empirical analysis by examining the causal impact of Google Translate integration on Wikipedia's content translation volume. This establishes the foundation for our subsequent investigations into quality, engagement, information flow, and representation effects. Our analysis employs a matrix completion approach in a difference-in-differences framework to estimate changes in translation activity.

Figure 7 presents the estimated dynamic effects on translation volume. The event study plot shows translation patterns over 24 months before and after integration, with the black line representing point estimates and the shaded area indicating 95% confidence intervals. The pre-integration period exhibits stable translation volumes with minimal variation around zero, establishing a clear baseline. Following integration, we observe an immediate and sustained increase in translation activity. While the widening confidence intervals post-integration suggest heterogeneous adoption across language editions, the lower bound remains consistently above zero, indicating a robust positive effect.

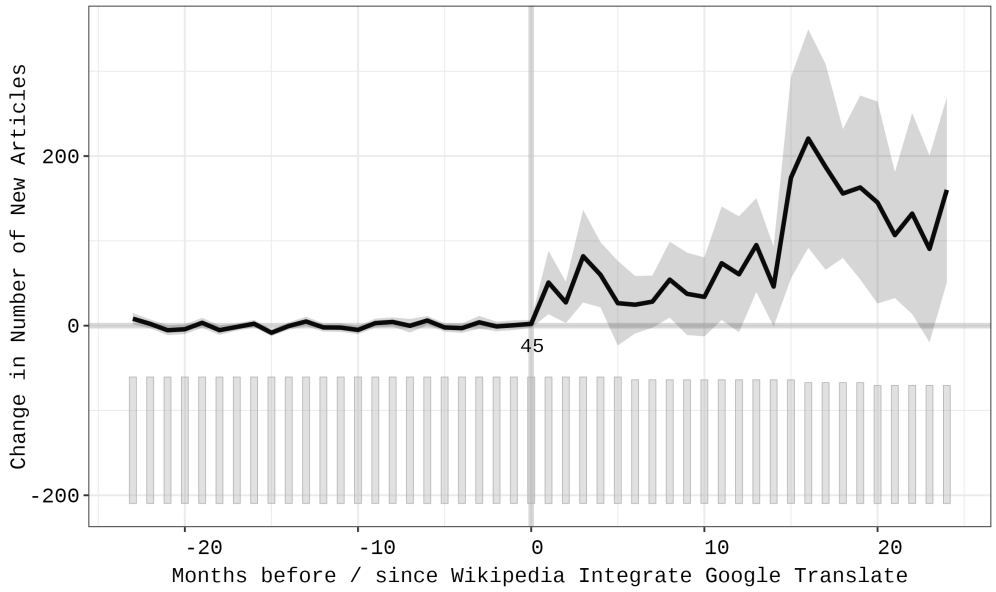
We quantify these effects in Table 4. The Average Treatment Effect on the Treated (ATT) indicates an increase of 71.2 additional translated articles per language edition monthly (SE = 16.8, $p < 0.01$). This represents a 149% increase relative to the pre-integration baseline of 47.7 articles, demonstrating substantial scaling of content production through machine translation.

Table 4: Estimated Impact of Google Translate Integration on Monthly Translation Volume

ATT	S.E.	CI.lower	CI.upper	P.value
71.2	16.8	38.4	104.1	0.0

To validate the causal interpretation of these estimates, we implement two diagnostic

Figure 7: Impact of Google Translate Integration on Monthly Translation Volume

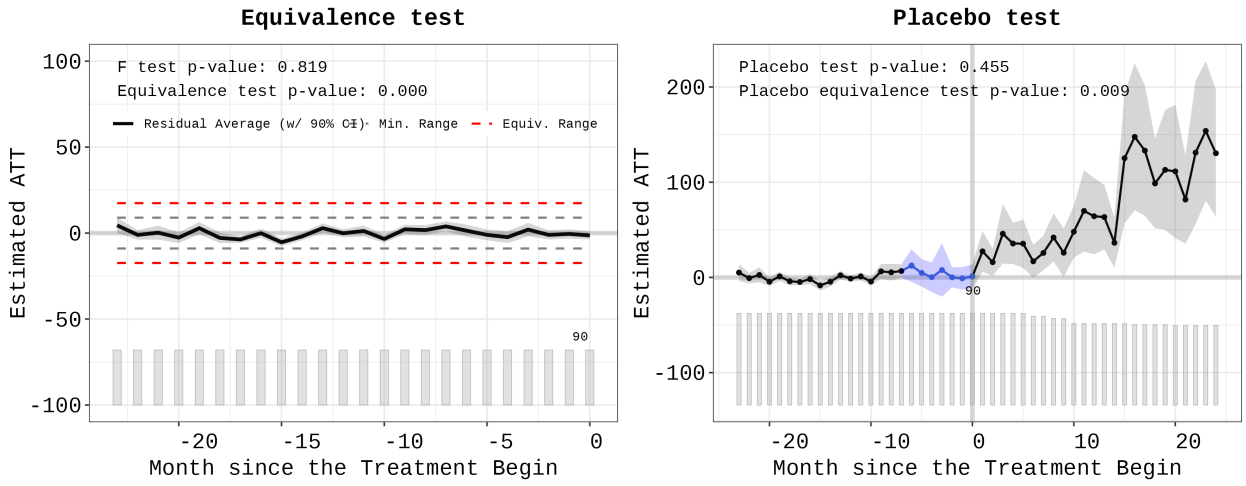


Note: Event study plot showing changes in translation volume relative to Google Translate integration (t=0). Black line represents point estimates with 95% confidence intervals shaded. Pre-integration period shows stable trends while post-integration period demonstrates sustained increase in translation activity.

tests (Figure 13). The equivalence test confirms the absence of pre-trends (F-test p-value = 0.819) while establishing significant treatment effects ($p < 0.01$). A placebo test examining artificial treatment timing six months before integration shows no significant effects ($p = 0.455$), further supporting our identification strategy.

These results establish that Google Translate integration substantially increased content production through translation. This finding motivates our subsequent analyses of how this expanded content creation affects article quality, reader engagement, information flow, and knowledge representation across Wikipedia's diverse language editions.

Figure 8: Diagnostic Tests: Equivalence Test and Placebo Test for Translation Volume Analysis



Note: Left panel shows equivalence test confirming absence of pre-trends (F-test p=0.819). Right panel displays placebo test with artificial treatment timing (blue region), showing no significant effects during placebo period (p=0.455).

5.2 Quality Assessment of Machine-Assisted Content

The significant increase in translation volume following Google Translate integration raises questions about content quality. To assess whether increased translation efficiency compromises quality, we examine three complementary quality signals: community-based quality control through article deletion patterns, structural quality metrics, and reader engagement patterns. This multi-dimensional approach helps address the challenge of measuring quality across Wikipedia's diverse language editions.

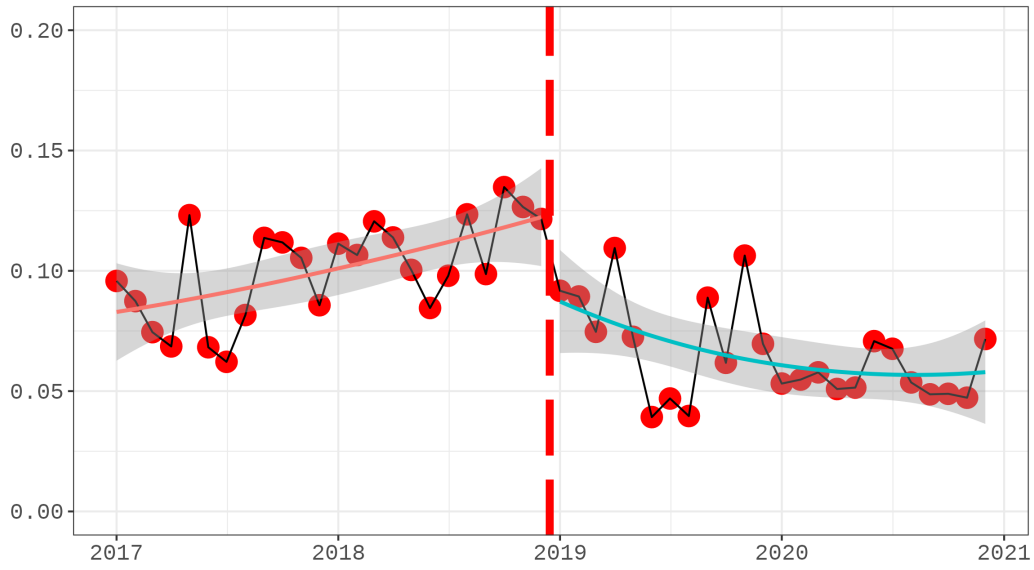
5.2.1 Quality Signal 1: Community Acceptance

We first examine article deletion ratios as an indicator of quality. On Wikipedia, articles that fail to meet community standards for quality or relevance are removed through a consensus-based deletion process. The deletion ratio - the proportion of translated articles that are subsequently deleted - serves as a community-validated quality metric. This measure is

particularly valuable because it reflects the judgment of experienced Wikipedia editors and is a consistent metric across language editions. The Wikimedia Foundation also monitors this metric as part of its quality control for content translation.

To analyze deletion patterns reliably, we must account for the varying scale of translation activities across languages. Small Wikipedia editions may have few monthly translations, making their individual deletion ratios volatile. We address this by calculating aggregate deletion ratios for all languages that adopted Google Translate in January 2019. Figure 9 presents these monthly deletion ratios, with the trend line estimated using piecewise LOESS regression. The data shows a clear downward trend in deletion ratios after Google Translate integration, suggesting improved compliance with Wikipedia's quality standards.

Figure 9: Monthly Article Deletion Ratios Before and After Google Translate Integration



Note: Time series of deletion ratios for translated articles, with LOESS regression trends. The vertical red line marks Google Translate integration in January 2019. Declining deletion ratios post-integration suggest improved content quality maintenance.

To quantify this change precisely, we implement an Interrupted Time Series (ITS) analysis:

$$Y_t = \beta_0 + \beta_1 \times t + \beta_2 \times A_t + \beta_3 \times (t - T_I) \times A_t + \epsilon_t$$

where Y_t represents the deletion ratio at time t, and T_I marks the integration of Google Translate. The indicator variable A_t equals 1 for the post-integration period and 0 otherwise. The coefficients capture the baseline level (β_0), pre-integration trend (β_1), immediate change at integration (β_2), and change in trend post-integration (β_3).

The analysis reveals a 43% reduction in deletion ratio immediately following integration ($\beta_2 = 4.3\%$ compared to a 10.0% baseline) and a continued improvement in trend ($\beta_3 = -0.3\%$). These results counter initial concerns about quality degradation from automated translation. Instead, they indicate that Google Translate integration has enhanced rather than compromised content quality, as measured by community acceptance of translated articles.

5.2.2 Quality Signal 2: Structural Measurement

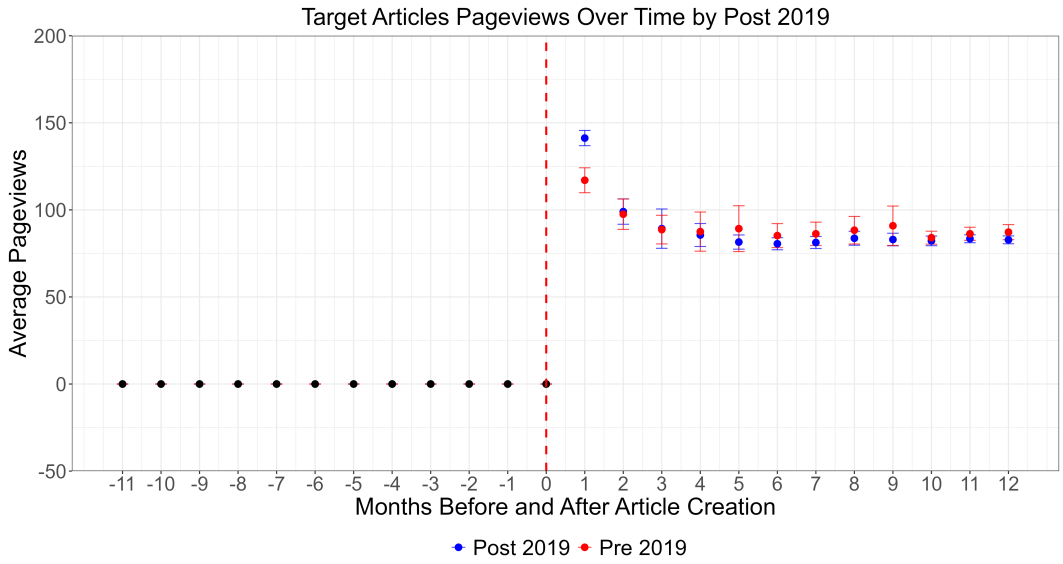
To complement our analysis of deletion ratios, we examine a language-agnostic quality metric developed by the Wikimedia Foundation. This metric evaluates article quality based on structural characteristics including reference count, section organization, article length, and multimedia content. This official quality assessment provides a standardized measure across Wikipedia's diverse language editions, offering an objective evaluation of article structure and completeness. Our analysis of this metric reveals no significant differences in the structural quality of articles translated before and after the integration of Google Translate, suggesting that machine translation maintains consistent article structure standards.

5.2.3 Quality Signal 3: Reader Responses

Reader engagement provides our third signal of article quality, operating under the premise that consistently engaged readership indicates content value. We compare pageview patterns of articles translated before and after Google Translate integration, as shown in Figure 13.

The analysis reveals equivalent pageview levels between pre- and post-integration translated articles, with both groups maintaining stable readership over time. This consistency in reader engagement suggests that the increased translation volume enabled by machine translation does not compromise content utility from the reader perspective, indicating successful scaling of content production while maintaining quality.

Figure 10: Monthly Pageviews of Translated Articles Pre- and Post-Google Translate Integration



Note: Comparison of monthly pageviews between articles translated before and after Google Translate integration. Both groups show equivalent engagement levels, suggesting that increased translation volume through machine assistance maintains consistent reader value.

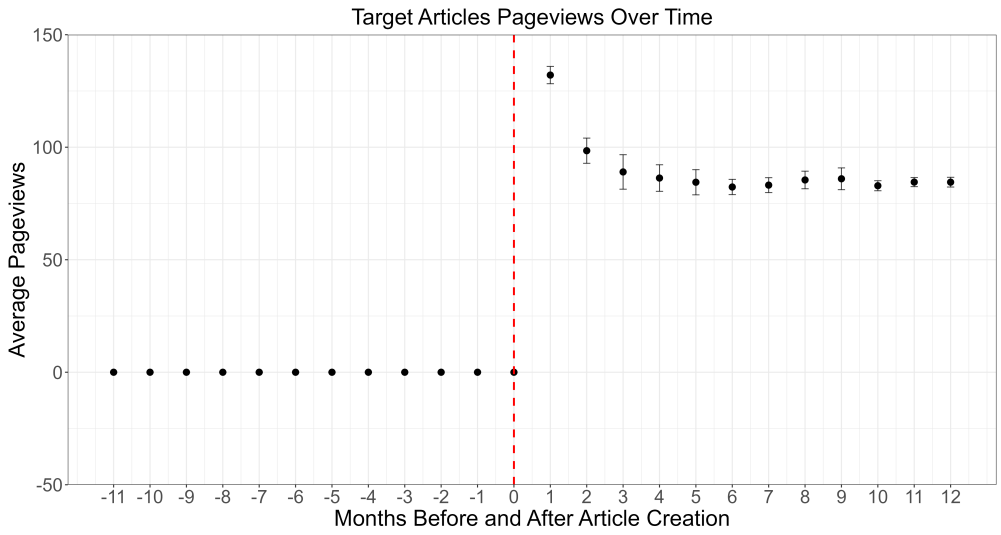
These three quality signals - declining deletion ratios, maintained structural quality, and sustained reader engagement - collectively indicate that the integration of Google Translate has not compromised article quality despite enabling increased translation volume. Rather, the evidence suggests that machine translation, when combined with human oversight in Wikipedia's translation workflow, can maintain or even enhance content quality while expanding multilingual knowledge accessibility.

5.3 Reader Engagement and Content Distribution

Having established the impact on content creation and article quality, we now further examine whether translated articles effectively reach readers and serve their intended purpose of increasing information access. Our analysis of reader engagement focuses on three key aspects: the general readership of translated articles, potential cannibalization effects on source articles, and variations across Wikipedia editions of different sizes.

First, we investigate whether translated articles attract and maintain readership. Our analysis reveals that translated articles receive an average of 80 pageviews per month, demonstrating consistent reader engagement over time. As shown in Figure 11, the pageview patterns remain stable, suggesting sustained reader interest in translated content. This level of engagement provides evidence for the value of translation efforts and indicates substantial demand for multilingual content on Wikipedia.

Figure 11: Monthly Pageviews of Translated Articles

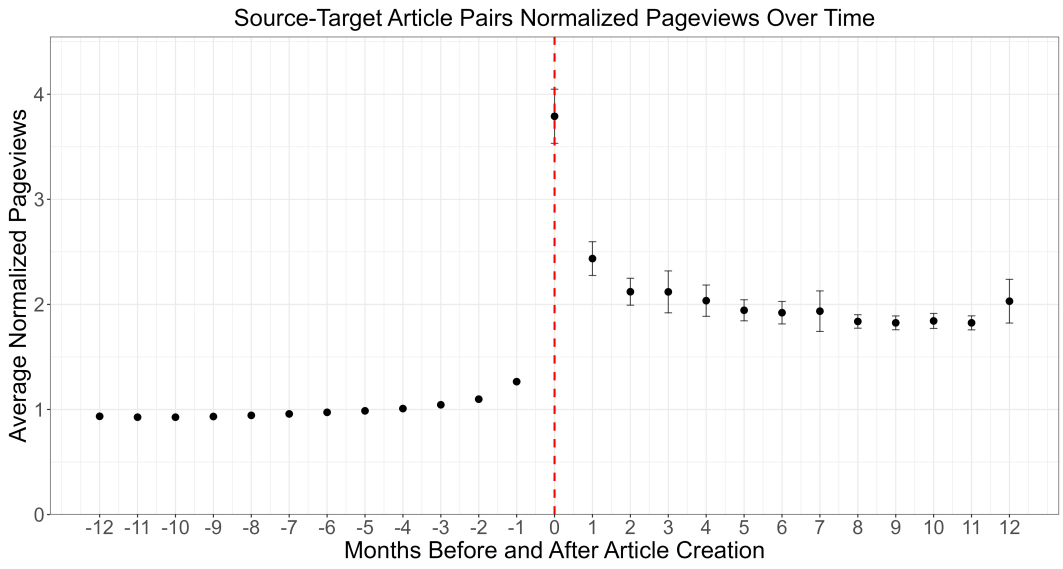


Note: Time series showing the average monthly pageviews per translated article. The stable pattern around 80 pageviews per month demonstrates sustained reader engagement with translated content across the sample period, indicating consistent demand for multilingual Wikipedia content.

A potential concern is whether translations might cannibalize readership from source articles, merely redistributing rather than expanding overall readership. To investigate this,

we normalize pageviews by the average pageviews of source articles in the 12 months before translation. As shown in Figure 13, source articles maintain stable normalized pageviews around 1 in the pre-translation period, by construction. After translation, the combined normalized pageviews of source and target articles increase substantially to approximately 2, indicating translations expand the overall reach of Wikipedia content, allowing readers to access information in their preferred languages. This doubling effect demonstrates that translations expand rather than redistribute readership, with the combined audience reaching about than twice the original readership. The sustained elevated level of normalized pageviews (around 2.0) in the months following translation suggests these gains are durable rather than temporary.

Figure 12: Combined Source and Target Article Pageviews After Translation

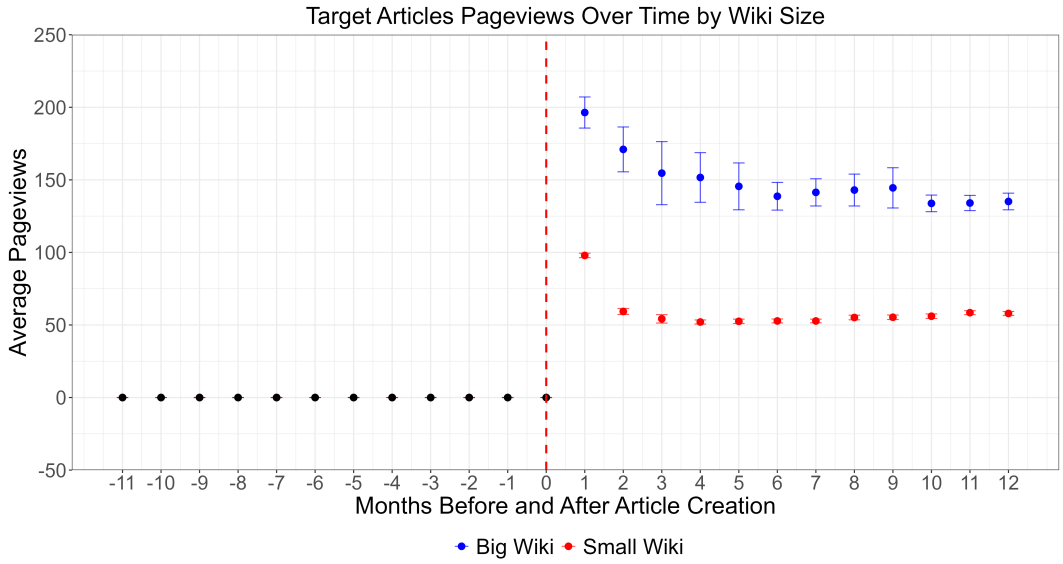


Note: Time series showing normalized pageviews for source articles (baseline=1.0) and combined source-target article pairs. The increase to approximately 2.0 in combined views indicates translations expand rather than redistribute readership.

Lastly, our analysis of reader engagement patterns across different Wikipedia editions reveals variations in how translations serve different language communities. Large Wikipedia editions (those ranking in the top 50 by size) demonstrate higher levels of reader engagement, with translated articles receiving approximately 100-120 monthly pageviews. In contrast,

smaller Wikipedia editions show more moderate engagement levels, typically garnering 60-70 monthly pageviews per translated article. This pattern suggests that translations in larger Wikipedia editions benefit from existing network effects and established readership bases. These findings have implications for platform design and resource allocation, suggesting a need for additional support mechanisms to boost visibility and engagement for translations in smaller Wikipedia editions.

Figure 13: Monthly Pageview Patterns by Wikipedia Edition Size



Note: Comparison of pageview trends between large (top 50) and small Wikipedia editions. Large editions show higher engagement (100-120 monthly views) compared to smaller editions (60-70 views), suggesting network effects in content visibility.

These findings collectively demonstrate that machine translation effectively serves its intended purpose of expanding information access. The consistent reader engagement, net growth in readership, and varied patterns across Wikipedia editions suggest that translations bridge language barriers and meet the information needs of diverse language communities.

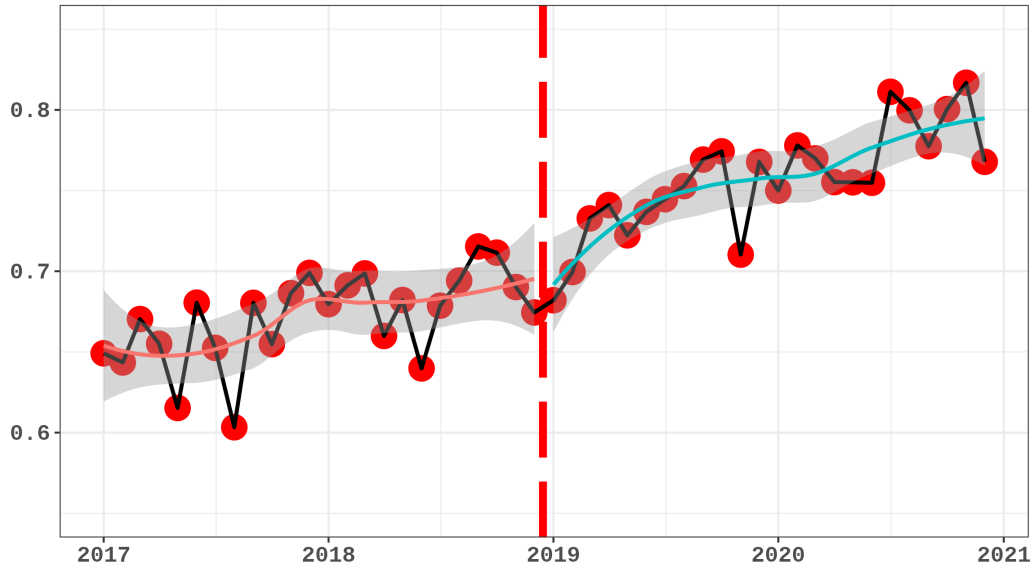
5.4 Patterns of Cross-Language Information Flow

Building on our findings of increased translation volume and maintained quality standards and reader interests, we now examine how machine translation shapes the flow of information

across Wikipedia's language editions. This analysis reveals the distributional implications of translation technology by examining the source and target language patterns, as well as heterogeneous effects across language editions of different sizes.

We begin by analyzing trends in English-sourced translations, as English Wikipedia represents the largest potential source of content. Following our methodological approach, we aggregate data across languages that integrated Google Translate in January 2019. Figure 14 documents an increasing concentration of translations from English following Google Translate integration, with the slope increasing by 68% ($\beta_3/\beta_1 = 0.15/0.22$) in the post-integration period.

Figure 14: Proportion of Articles Translated from English Wikipedia, 2017-2020



Note: Time series with LOESS regression showing the increasing share of English-sourced translations. The vertical red line marks Google Translate integration, after which the slope increases by 68%, indicating growing centralization of English as source language.

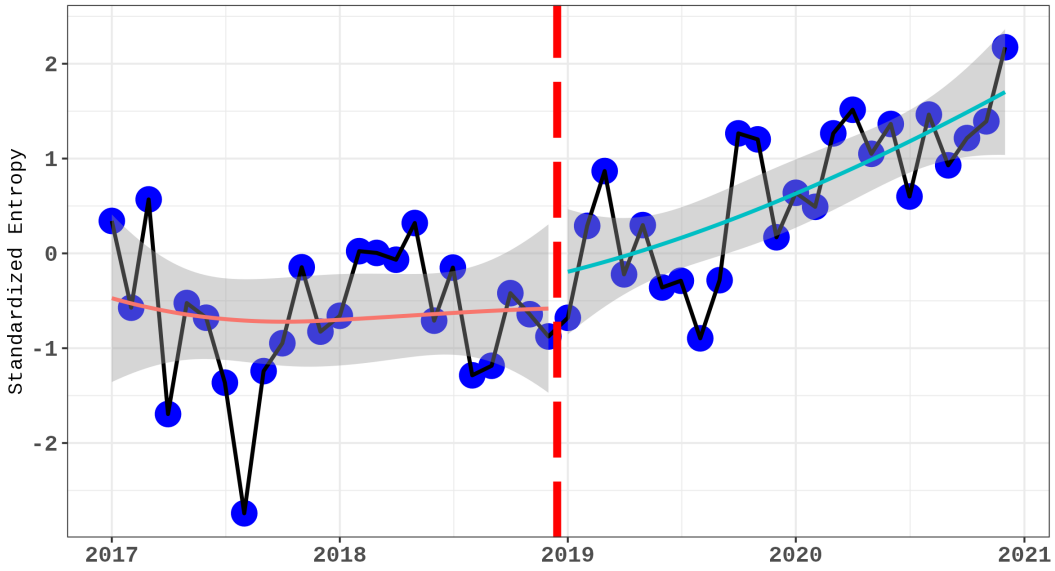
To complement this analysis, we evaluate target language diversity using an entropy-based metric

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where H represents the entropy of target language distribution X, and $P(x_i)$ denotes the proportion of translations into language i.

Figure 15 shows increased target language diversity post-integration, with a significant change in trend ($\beta_3 = 0.008$ compared to pre-integration $\beta_1 = 0.0002$). While source material increasingly comes from English, the technology enables wider dissemination across language communities.

Figure 15: Diversity of Target Languages in Wikipedia Translations, 2017-2020



Note: Entropy-based measure of target language diversity over time, with LOESS regression. The vertical red line marks Google Translate integration, showing increased diversification in target languages despite source language concentration.

Building on these aggregate patterns, We further examine how translation benefits vary with language resources by classifying Wikipedia editions into three tiers based on article count: high-resource (top 15), medium-resource (ranks 16-50), and low-resource (below rank 50). Table 5 reveals substantial heterogeneity: high-resource languages show the largest gains

(140.4 additional articles monthly, 294% increase), while low-resource languages experience smaller, insignificant increases (13.1 articles, 27% increase).

Table 5: Differential Effects of Google Translate Integration by Wikipedia Language Resource Level

Dependent Variable:	Number of Translated Articles
Model:	(1)
GT Integration	13.1
	(9.2)
GT Integration × Medium-resource	78.8**
	(34.6)
GT Integration × High-resource	140.4**
	(60.3)
Fixed-effects	
Language	Yes
Year-Month	Yes
Fit statistics	
Observations	6,048
R^2	0.77540
Pre-integration Mean	47.7

Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 5 reveals substantial heterogeneity across these categories. High-resource languages show the largest gains (140.4 additional articles monthly, 294% increase), followed by medium-resource languages (78.8 articles, 165% increase). Low-resource languages experience smaller, statistically insignificant increases (13.1 articles, 27% increase).

These disparities suggest that machine translation technology may amplify existing resource advantages. This amplification could occur through several mechanisms: (1) larger Wikipedia editions have more active editor communities for quality control, (2) high-resource languages benefit from better machine translation model performance due to larger training corpora, and (3) network effects in readership and internal linking create additional advantages for larger editions.

These findings have implications for platform design and policy interventions. They suggest that merely providing equal access to translation technology may not be sufficient to achieve equitable outcomes across language communities. Instead, platforms may need to implement targeted support mechanisms for low-resource languages, such as additional editorial assistance, specialized quality control processes, or incentive systems that prioritize translations into under-resourced languages. This interaction between technological capability and existing resource inequalities provides context for our subsequent analysis of specific knowledge domains, where we examine how these disparities manifest in particular types of content.

5.5 Representation and Content Gap

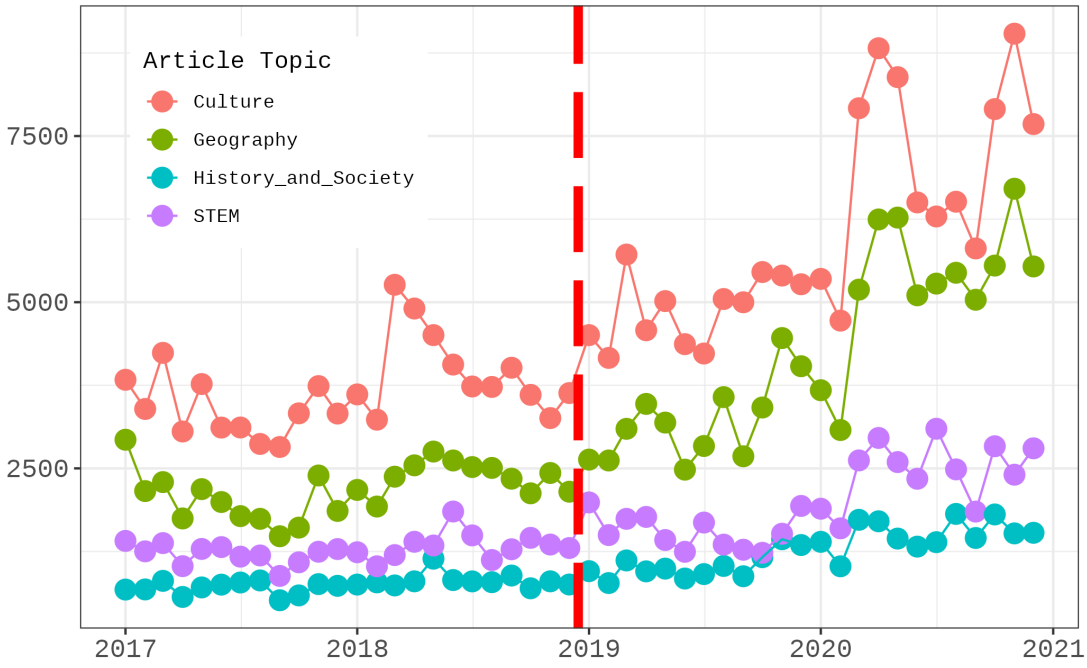
Our previous analyses documented how machine translation affects content production volume, quality, reader engagement, and cross-language information flows. We now examine whether these technological changes contribute to addressing systemic gaps in Wikipedia's content coverage. This investigation focuses on two dimensions: the distribution of knowledge across content domains and representation within key article categories.

Building on our finding that machine translation disproportionately benefits resource-rich languages, we first analyze whether similar patterns emerge across knowledge domains. Topic classification in a multilingual setting presents methodological challenges due to linguistic and cultural variations. We address these challenges by leveraging Wikidata as a language-agnostic classification system, following Johnson et al. (2021). This approach maps articles to a standardized 64-category taxonomy based on WikiProject, enabling consistent topic identification across languages. For methodological details, see Appendix E.

Figure 16 reveals heterogeneous growth across content domains following Google Translate integration. Culture and Geography articles show particularly pronounced increases, while History and Society and STEM topics exhibit more modest growth.

We quantify these patterns using difference-in-differences estimation. Table 6 shows that

Figure 16: Evolution of Translation Volume across Article Topics Following Google Translate Integration



Note: Time series showing monthly translation volume by major content categories from 2017 to 2020. The vertical red line marks Google Translate integration in January 2019. Culture and Geography articles show particularly pronounced increases in translation volume post-integration, while History and Society and STEM topics exhibit more moderate growth.

Google Translate integration led to the largest increases in Culture (25.2 articles per month, $p<0.01$) and Geography (21.4 articles, $p<0.01$). The effects on STEM (7.9 articles, $p<0.01$) and History and Society (5.6 articles, $p<0.05$) are significant but smaller in magnitude.

These topical patterns motivate deeper examination of two critical content areas: biographical and geographical articles. Together comprising nearly 50% of translated content, these categories serve as important channels for representing human diversity (through biographies) and spatial knowledge (through geography).

Table 6: Differential Effects of Google Translate Integration by Content Domain

Dependent Variables:	Culture	Geography	STEM	History and Society
Model:	(1)	(2)	(3)	(4)
GT Integration	25.2***	21.4***	7.9***	5.6**
	(5.9)	(5.4)	(2.1)	(2.3)
Fixed-effects				
Language	Yes	Yes	Yes	Yes
Year-Month	Yes	Yes	Yes	Yes
Fit statistics				
Observations	6,336	6,336	6,336	6,336
R^2	0.76	0.64	0.62	0.58
Pre-integration Mean	20.8	12.5	7.2	4.3

Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

5.5.1 Gender Content Gap

We first analyze how machine translation affects gender representation in biographical articles. Prior research documents substantial gender imbalances in Wikipedia's coverage of notable individuals. We examine whether reducing translation barriers affects these historical disparities.

Our empirical approach links translated articles to Wikidata entries, enabling consistent gender identification across languages. This methodology provides a systematic framework for analyzing representation across Wikipedia's multilingual ecosystem. For detailed documentation of our attribute extraction process, see Appendix F.

Table 7 presents difference-in-differences estimates of Google Translate's impact across gender categories. The results show statistically significant increases in articles about females (3.8 articles monthly, $p<0.01$), males (7.2 articles, $p<0.01$), and non-binary individuals (0.06 articles, $p<0.01$). While all categories show growth, the larger absolute increase in male biographies suggests persistent representation challenges.

Table 7: Effects of Google Translate Integration on Monthly Biography Translations by Gender

Dependent Variables:	Male Biography	Female Biography	Non-binary Biography
Model:	(1)	(2)	(3)
GT Integration	7.2***	3.8***	0.06***
	(2.5)	(1.3)	(0.02)
Fixed-effects			
Language	Yes	Yes	Yes
Year-Month	Yes	Yes	Yes
Fit statistics			
Observations	6,288	6,288	6,288
R^2	0.72317	0.54092	0.26982
Pre-integration Mean	12.5	7.8	0.06

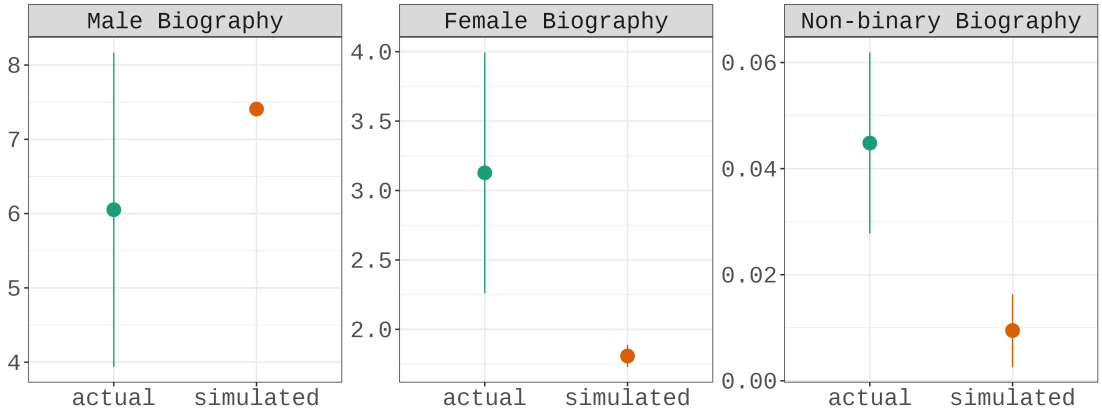
Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

To interpret these changes meaningfully, we must consider the baseline distribution of biographical content. As of January 2019, English Wikipedia's biographies comprised 82.2% males (1.29 million articles), 17.7% females (0.28 million articles), and 0.1% non-binary individuals (1,390 articles). Against this backdrop, translation patterns from 2019-2021 show movement toward greater balance: 62.4% male (30,202 articles), 37.2% female (17,976 articles), and 0.4% non-binary (184 articles).

To formally assess progress in representation, we implement a simulation analysis that establishes an expected baseline given existing content distribution. We simulate translation patterns that would emerge if editors selected articles randomly from the existing stock of English Wikipedia biographies. This creates a null model against which we can compare actual translation patterns.

Figure 17 compares actual translation patterns to these simulated benchmarks. Female and non-binary biography translations exceed simulated expectations, while male biography translations fall below the simulated benchmark. This divergence from the null model suggests that translation patterns are actively shifting gender representation beyond what

Figure 17: Gender Representation in Biography Articles: Actual vs. Simulated Translation Patterns



Note: Comparison between actual translation coefficients (with 95% confidence intervals) and simulated benchmarks (averages from 100 replications). Female and non-binary biography translations exceed simulated expectations, while male biography translations fall below simulated benchmarks.

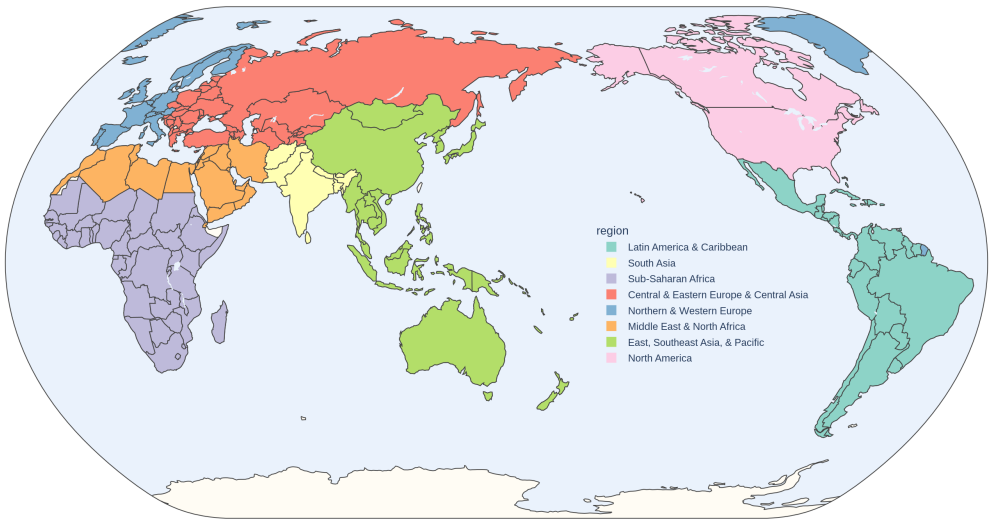
mechanical reproduction of existing content would predict.

5.5.2 Geography Content Gap

We extend our representational analysis to geographic content, examining how machine translation affects the spatial distribution of knowledge. Geographic articles comprise 28% of translations and serve as key vectors for disseminating local knowledge globally.

Following Wikimedia Foundation's regional classification system, we analyze translation patterns across eight global regions (Figure 18). This framework enables systematic comparison of content flows across geographic areas while accounting for regional linguistic diversity.

Figure 18: Global Distribution of Wikimedia Foundation's Regional Classifications



Note: World map showing Wikimedia Foundation's eight regional classifications: Latin America & Caribbean, South Asia, Sub-Saharan Africa, Central & Eastern Europe & Central Asia, Northern & Western Europe, Middle East & North Africa, East/Southeast Asia & Pacific, and North America.

Our difference-in-differences estimates reveal substantial regional heterogeneity in translation growth. Central and Eastern Europe & Central Asia shows the largest gains (8.6 articles monthly, $p<0.01$), followed by East/Southeast Asia & Pacific (7.3 articles, $p<0.01$) and Northern/Western Europe (5.4 articles, $p<0.01$). Regions like Sub-Saharan Africa (0.4

articles, $p<0.05$) and Middle East/North Africa (0.3 articles, $p<0.05$) show more modest increases.

Table 8: Effects of Google Translate Integration on Monthly Geography Translations by Regions

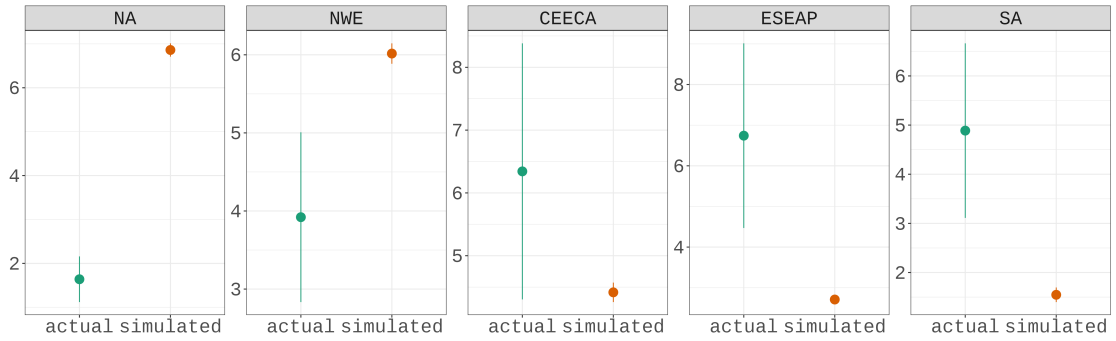
DVs:	ESEAP	SA	MENA	SSA	NA	SCA	CEECA	NWE
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
GT Integration	7.3***	5.3***	0.3**	0.4**	1.7***	1.3**	8.6***	5.4***
	(2.4)	(1.8)	(0.1)	(0.2)	(0.5)	(0.5)	(2.6)	(1.4)
Fixed-effects								
Language	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	6,336	6,336	6,336	6,336	6,336	6,336	6,336	6,336
R^2	0.49	0.44	0.54	0.26	0.47	0.24	0.55	0.60
Pre-integration Mean	4.7	2.8	0.2	0.2	1.0	0.6	5.0	3.4

Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

As with our gender analysis, we implement simulation benchmarks based on existing geographic coverage in English Wikipedia. Figure 19 reveals that translations about Asia and Eastern Europe exceed simulated expectations, while North American and Western European content falls below simulated benchmarks. This pattern suggests that machine translation particularly benefits regions with high linguistic diversity, where translation barriers may have previously limited content dissemination.

These findings reveal important patterns in how machine translation technology shapes content representation. Our analyses indicate that the technology's impact varies systematically across representation dimensions. These patterns suggest that reducing technical barriers alone may not fully address content inequality - while machine translation success-

Figure 19: Regional Representation in Geography Articles: Actual vs. Simulated Translation Patterns



Note: Comparison between actual translation coefficients (point estimates with 95% confidence intervals) and simulated benchmarks (averages from 100 replications). Regions like Asia and Eastern Europe exceed simulated expectations, while North American and Western European content falls below benchmarks.

fully reduces language barriers, achieving balanced representation likely requires complementary interventions to address other systemic constraints such as source content availability or community capacity. These results have implications for platform design, suggesting the need for targeted policies that account for how technological tools interact with existing platform structures to shape information accessibility and representation in digital knowledge systems.

6 Conclusion

This study provides empirical evidence on how artificial intelligence affects content creation and knowledge distribution in peer production platforms. Our analysis of Wikipedia's integration of Google Translate reveals both the promise and challenges of using machine translation to bridge digital language divides. The findings demonstrate substantial efficiency gains, with translation volume increasing by 71 articles per language per month (149% increase) following integration. This growth occurs while maintaining or improving content

quality, as evidenced by reduced deletion rates and sustained reader engagement patterns.

Our investigation reveals complex patterns in how machine translation shapes information flows across languages. While English increasingly serves as a central source language, we observe greater diversity in target languages receiving translations. However, the benefits are not uniformly distributed - high and medium-resource languages experience significantly larger gains (140.4 and 78.8 additional articles monthly respectively) compared to low-resource languages (13.1 articles). This pattern suggests that technological advancement may amplify existing resource advantages rather than automatically democratizing content creation.

The analysis of specific content domains reveals varying impacts on representation gaps. Translation patterns in biographical content show progress toward gender balance relative to simulated benchmarks, though substantial disparities persist. Geographic content demonstrates particularly strong growth in regions with high linguistic diversity, suggesting the technology particularly benefits areas where translation barriers were previously most acute. These findings indicate that while machine translation can help address content gaps, its effectiveness depends on existing platform structures and resource distributions.

Our findings have implications for platform design and digital inclusion efforts. They suggest that merely providing access to better translation technology may not be sufficient to achieve equitable outcomes across language communities. Instead, platforms may need to implement targeted support mechanisms for low-resource languages, such as additional editorial assistance or incentive systems that prioritize translations into under-resourced languages. The interaction between technological capability and existing resource inequalities highlights the need for comprehensive approaches that combine technological solutions with structural support for under-resourced communities.

These results contribute to our understanding of how artificial intelligence shapes knowledge production and dissemination in digital platforms. While machine translation substantially reduces content creation costs, its benefits depend heavily on existing platform

structures and resource distributions. This interplay between technological capability and structural inequalities carries important implications for platform design and knowledge democratization efforts. Future research could further explore the dynamics of human-AI collaboration in content creation, particularly focusing on how different communities adapt and utilize these technologies within their specific linguistic and cultural contexts.

References

Aaltonen, A. and Seiler, S. (2016). Cumulative growth in user-generated content production: evidence from wikipedia. *Management Science*, 62(7):2054–2069.

Agrawal, A., Gans, J., and Goldfarb, A. (2022). *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Brynjolfsson, E., Hui, X., and Liu, M. (2019). Does machine translation affect international trade? evidence from a large digital platform. *Management Science*, 65(12):5449–5460.

Constantinides, P., Henfridsson, O., and Parker, G. G. (2018). Platforms and infrastructures in the digital age.

Forman, C., Ghose, A., and Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information systems research*, 19(3):291–313.

Ginsburgh, V. and Weber, S. (2020). The economics of language. *Journal of Economic Literature*, 58(2):348–404.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Eonomic Literature*, 57(1):3–43.

Graham, M., Hogan, B., Straumann, R. K., and Medhat, A. (2014). Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764.

Halavais, A. and Lackaff, D. (2008). An analysis of topical coverage of wikipedia. *Journal of computer-mediated communication*, 13(2):429–440.

Hinnosaar, M., Hinnosaar, T., Kummer, M., and Slivko, O. (2023). Wikipedia matters. *Journal of Economics & Management Strategy*, 32(3):657–669.

Johnson, I., Gerlach, M., and Sáez-Trumper, D. (2021). Language-agnostic topic classification for wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 594–601.

Johnson, I. and Lescak, E. (2022). Considerations for multilingual wikipedia research. *arXiv preprint arXiv:2204.02483*.

Langrock, I. and González-Bailón, S. (2022). The gender divide in wikipedia: Quantifying and assessing the impact of two feminist interventions. *Journal of Communication*, 72(3):297–321.

Miquel-Ribé, M. and Laniado, D. (2020). The wikipedia diversity observatory: A project to identify and bridge content gaps in wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–4.

Nagaraj, A. (2022). The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry. *Management Science*, 68(1):564–582.

Parker, G. G. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you*. WW Norton & Company.

Pereira, M., Greenstein, S., Sadun, R., Tambe, P., Darre, L. R., Glazer, T., Kim, A., Dodhia, R., and Ferres, J. L. (2024). The new digital divide. Technical report, National Bureau of Economic Research.

Thompson, N. and Hanley, D. (2018). Science is shaped by wikipedia: evidence from a randomized control trial.

Thompson, N., Luo, X., McKenzie, B., Richardson, E., and Flanagan, B. (2024). User-generated content shapes judicial reasoning: Evidence from a randomized control trial on wikipedia. *Information Systems Research*.

Turovsky, B. (2016). Found in translation: More accurate, fluent sentences in google translate. Google Blog.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xu, S. X. and Zhang, X. (2013). Impact of wikipedia on market information environment: Evidence on management disclosure and investor reaction. *MIS Quarterly*, pages 1043–1068.

Zhu, K., Walker, D., and Muchnik, L. (2020). Content growth and attention contagion in information networks: Addressing information poverty on wikipedia. *Information Systems Research*, 31(2):491–509.

Part I

Web Appendix

A Data Source: Published Translations on Wikipedia

The primary data for our analysis comprises published translations from Wikipedia's Content Translation system. Wikimedia Foundation provides data access through Application Programming Interface (API) that lists published translations, offering comprehensive details essential for our study:

- **API Access**: The API provides lists of all published translations across languages, including details like translation IDs, translation date, source and target titles, source and target languages, URLs of the articles, and translation statistics indicate the extent of human versus machine translation.
- **Parallel Text Dump**: The tool also allows access to parallel text data, encompassing the source content and the final user-translated content.

This data is provided under an open license and represents an ever-improving resource for cross-linguistic research and machine translation development. For more information, refer to the MediaWiki page: Content Translation - Published Translations

B Data Source: Wikipedia Knowledge Gap Index

The Knowledge Gap Index developed by the Wikimedia Foundation is instrumental in quantifying knowledge disparities across various dimensions on Wikipedia. The dataset is structured to provide a clear mapping of content gaps across languages

- **Content Mapping:** The Knowledge Gap Index employs a content mapping system, categorizing Wikipedia articles into specific domains and themes. This system allows us to assess disparities in content availability and representation across diverse topics and regions.
- **Gender Gap Measurement:** the Knowledge Gap Index addresses the Gender Gap. This measure assesses the balance in representation of different genders across Wikipedia articles. It aids in quantifying and understanding the skewness in gender portrayal,
- **Geographic Gap Measurement:** Another useful aspect of the Knowledge Gap Index is its focus on the Geographic Gap. This measurement evaluates the extent of content coverage pertaining to different geographic regions.
- **Open Access and Continuous Updates:** The Knowledge Gap Index is publicly accessible and is regularly updated, reflecting the dynamic nature of content creation on Wikipedia.

The Knowledge Gap Index offers a detailed perspective on the existing content distribution and the potential areas where machine translation can have a significant impact. For further details about the dataset and its applications, please refer to the Wikimedia Meta-Wiki page on the Knowledge Gap Index and its Geographic Gap Measurement.

C Robustness for Difference-in-Differences Modeling

C.1 Two-way Fixed Effect

This model estimates the Average Treatment Effect (ATE) of Google Translate's integration on the volume of translated articles. The statistically significant treated coefficient of 61.4 ($p<0.01$) reinforces the substantial impact of Google Translate on content translation activities in Wikipedia.

C.2 Sun and Abraham Weighted Interactions

The Sun and Abraham Weighted Interactions method highlights the temporal dynamics of the treatment effect, indicating a sustained positive impact of Google Translate over time on the translation volume.

C.3 Callaway and Sant'Anna Estimators

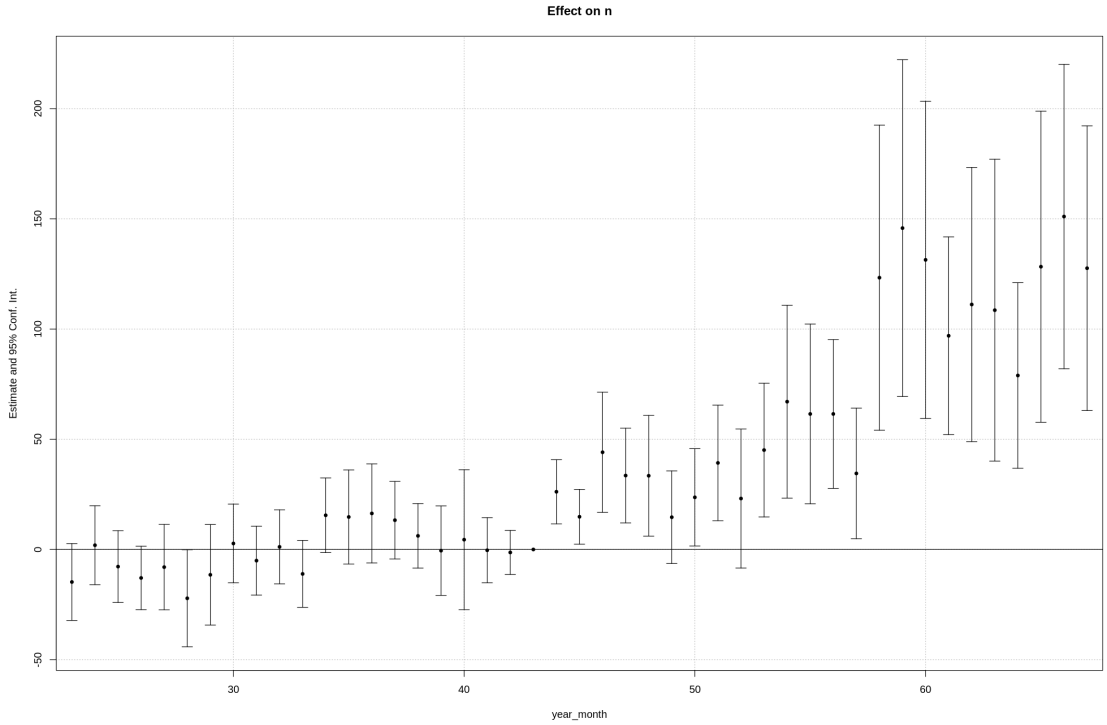
We further assess the robustness of our findings using the Callaway and Sant'Anna Estimators. This approach is designed to capture more nuanced effects in staggered adoption settings, like the Integration of Google Translate.. The Callaway and Sant'Anna Estimators confirm the robustness of our main findings, showing consistent positive effects of Google Translate's integration on translation volume across various Wikipedia language editions.

Table 9: Two-way Fixed Effect Model Results

Dependent Variable:	Number of Translated Articles
Model:	(1)
Google Translate Integration	61.4***
	(14.9)
Fixed-effects	
Language	Yes
Year-Month	Yes
Fit statistics	
Observations	6,336
R^2	0.77

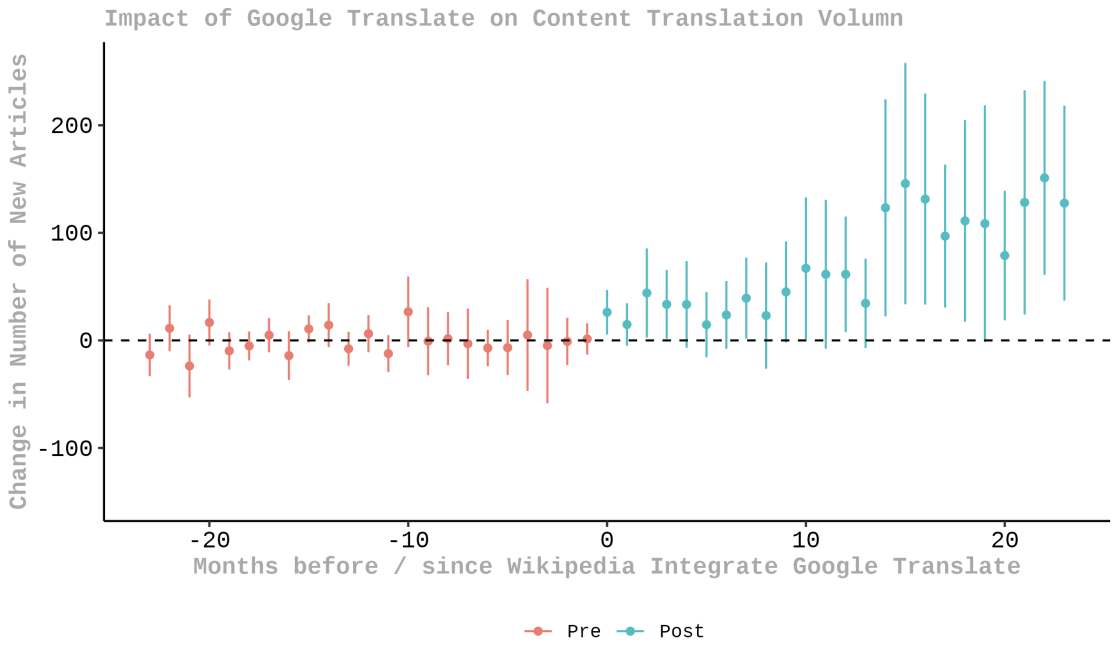
Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Figure 20: Sun and Abraham Weighted Interactions: Translation Volume Over Time



Note: The graph depicts the results from the Sun and Abraham weighted interactions model, emphasizing the enduring positive effects of Google Translate integration across different time periods.

Figure 21: Callaway and Sant'Anna Estimators: Impact on Translation Volume



Note: This figure illustrates the estimated effects using the Callaway and Sant'Anna approach. It highlights the positive influence of Google Translate integration on the translation volume, consistent with our main results.

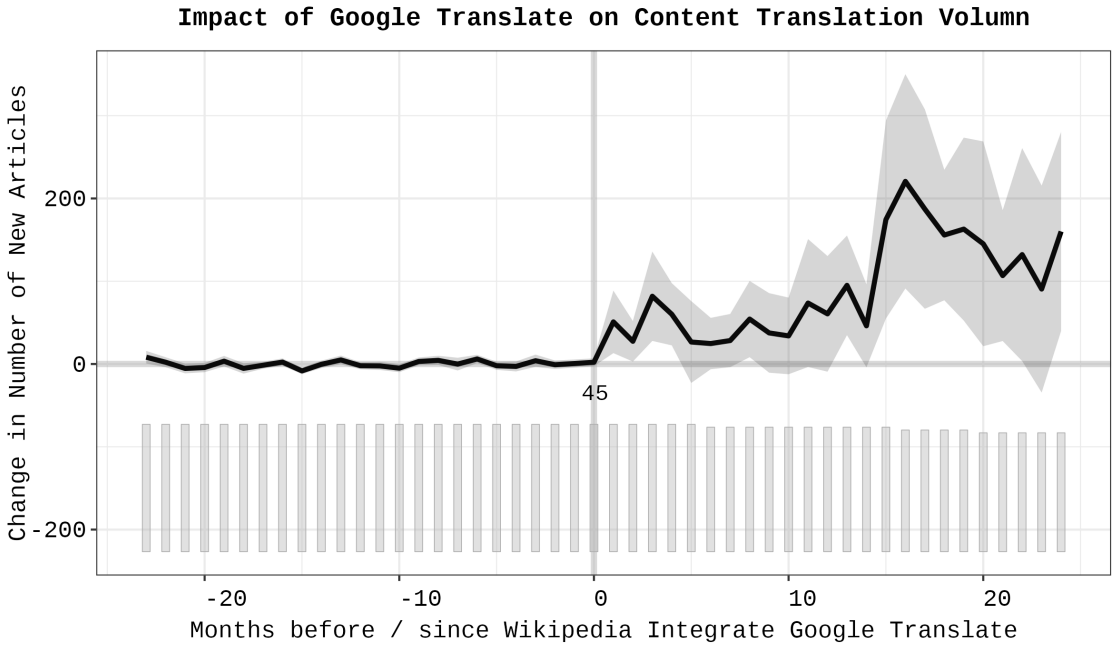
D Alternative Model Specification for High vs. Low Resource Languages

This section presents an alternative analysis focusing on the differential impact of Google Translate integration between high and low resource Wikipedia languages. To isolate resource-driven effects, we limit our sample to adopted editions, enabling a controlled contrast strictly across resource tiers sharing integration timing. Our treatment group comprises languages with size ranks below 60 while controls exceed 60. Using this approach, we can present dynamic temporal effects with counterfactual imputation.

D.1 Temporal Trends and Diagnostic Tests

Figures 22 and 23 provide visual representation and diagnostic analysis of the impact across language resource levels. Figure 22 shows the temporal trend in translation volume, highlighting the pronounced influence of Google Translate on high-resource languages. Figure 23 presents diagnostic tests that substantiate the robustness of our findings, ensuring the reliability of the observed impact across language resource levels.

Figure 22: Temporal Trend of Translation Volume Across Language Resource Levels

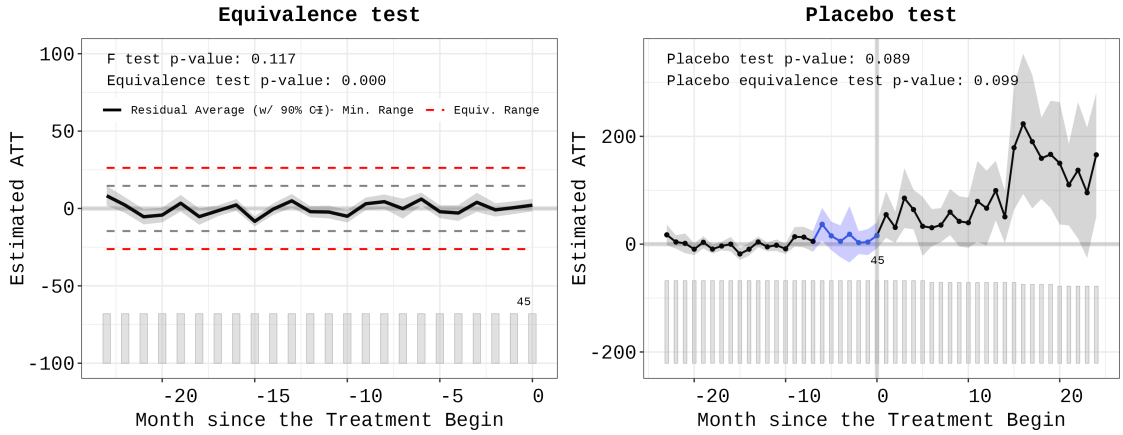


Note: This figure illustrates the temporal trend in translation volume across high vs. low resource languages, highlighting the pronounced impact of Google Translate on high resource languages.

D.2 Average Treatment Effect

Table 10 presents the average treatment effect (ATT) of Google Translate on high vs. low resource languages. The high average treatment effect value (92.34) with a significant p-value indicates a substantial impact on the translation volume, particularly for high-resource languages.

Figure 23: Diagnostic Tests for Heterogeneous Impact across Language Resource Levels



Note: The figure presents diagnostic tests assessing the heterogeneous impact of Google Translate on translation volume for different language resource levels, substantiating the robustness of our findings.

Table 10: Average Treatment Effect on Translated Article Volume by Language Resource Level

	ATT.avg	S.E.	CI.lower	CI.upper	p.value
1	92.34	31.01	31.56	153.11	0.00

E Methodology for Classifying Topics in Wikipedia Articles

Our approach to classifying article topics across different languages in Wikipedia builds on the language-agnostic methodology developed by Johnson et al. (2021). This method involves several steps:

1. **Topic Taxonomy**: Utilizing a taxonomy of 64 topics derived from WikiProject tags, encompassing a broad range of subjects from science and technology to arts and culture.

2. **Link-Based Representation**: Articles are represented by Wikidata items of contained links, facilitating a language-independent content representation.

3. **Mapping Articles to Topics**: Machine learning algorithms map articles to the pre-

defined topics based on link similarities.

4. **Ensuring Cross-Linguistic Consistency**: The focus on link-based features rather than textual content assures consistent classification across languages.

5. **Validation**: Classification accuracy is validated and refined through comparison with a manually classified article sample.

This methodology enables a comprehensive and consistent analysis of article topics across multiple language editions of Wikipedia, overcoming the limitations of language-specific models and ensuring uniform topic classification.

F Methodology for Extracting Attributes from Wikidata

We utilized Wikidata to extract demographic and geographical data for our study. We used the Wikidata Query Service, which allows for sophisticated and precise queries, to retrieve the required information efficiently and accurately

- **Biographical Articles**: Information such as gender, sexual orientation, and ethnicity of individuals was systematically extracted from their respective Wikidata entries. This process involved querying Wikidata's structured data to obtain these specific attributes for each biographical article.
- **Geographical Articles**: For articles pertaining to geographical locations, we extracted location data, including coordinates and administrative hierarchies. This was achieved by querying the corresponding Wikidata items linked to each geographical article.

This methodology ensured a standardized and accurate extraction of demographic and geographical information, useful for our analysis.

G Other Attributes of Biographical Wikipedia Articles

The results from our model reveal modest increases in the translation of biographical articles about LGBTQ individuals and those pertaining to specific ethnicities, with coefficients of 0.08 and 0.2 respectively (see Table 11). The positive coefficients suggest that post-integration of Google Translate, there is a slight but statistically significant rise in the representation of these groups. While these results point towards an incremental step in diversifying the biographical coverage on Wikipedia, they also indicate that much work remains to be done to close the representation gap.

Table 11: Impact of Google Translate Adoption on Article Volume by Other Attributes

Dependent Variables:	LGBTQ Biography	Ethnic-Minority Biography
Model:	(1)	(2)
Google Translate Integration	0.08**	0.2**
	(0.04)	(0.1)
Fixed-effects		
Language	Yes	Yes
Year-Month	Yes	Yes
Fit statistics		
Observations	6,288	6,288
R^2	0.42	0.44

Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

H Geographical Wikipedia Articles by Continents

The table 12 presents the difference-in-differences analysis results, capturing the impact of Google Translate's roll-out on the volume of articles across different continents. The results show that regions like Asia and Europe experienced a significant increase in article volume, with coefficients of 9.0 and 7.8, respectively, suggesting a more pronounced adoption of machine translation in these areas. America and Africa also saw increases, albeit at a lower magnitude, while Oceania demonstrated a modest but significant rise. These findings reflect the variable impact of machine translation across diverse geographical contexts.

Table 12: Impact of Google Translate Roll-out on Article Volume by Geography Regions

Dependent Variables:	America	Europe	Asia	Oceania	Africa
Model:	(1)	(2)	(3)	(4)	(5)
Google Translate Integration	3.0***	7.8***	9.0***	0.5***	1.5**
	(1.0)	(2.0)	(3.3)	(0.2)	(0.6)
Fixed-effects					
Language	Yes	Yes	Yes	Yes	Yes
Year-Month	Yes	Yes	Yes	Yes	Yes
Fit statistics					
Observations	6,336	6,336	6,336	6,336	6,336
R^2	0.42505	0.58168	0.42917	0.29379	0.45671
Within R^2	0.00453	0.00887	0.00744	0.00226	0.00466

Clustered (Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*