Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Disappearing without a Trace: Coverage, Community, Quality, and Temporal Dynamics of Wikipedia Articles on Endangered Brazilian Indigenous Languages

Marisa Vasconcelos[1,2*], **Priscila de Souza Mizukami**[2*], **Claudio Santos Pinhanez**[2]
[1]Universidade Federal de Minas Gerais, Brazil
[2]IBM Research, Brazil
marisa.vasconcelos@gmail.com, priscila.mizukami@gmail.com, csantosp@br.ibm.com

Abstract

Nearly half of Brazil's 180 Indigenous languages face extinction within the next 20 years. What's more concerning is that most of these languages lack a single scientific article describing them, which means they could disappear without leaving any documented evidence of their existence. This work investigates the state of articles about those languages in Wikipedia, both in the English and Portuguese versions, regarded here as indicative of the minimum world-level trace of the previous existence of these languages. Our study shows that over 30% of these languages do not have a single Wikipedia article describing them. It also highlights that the Portuguese and English editing communities are not only distinct, but have different practices, achieving similar levels of quality through different temporal dynamics. These results, although encouraging, suggest that any effort to enhance coverage comprehensiveness in both Wikipedias should consider different strategies for engaging each editing community.

Introduction

About 43% of the 7,000 languages spoken in the world today are in danger of disappearing by the end of this century (UNESCO 2010), with around half of those languages being used only by Indigenous peoples in Africa and the Americas. The UNESCO classifies the danger of disappearance in a scale from 0 to 5, where 0 (*Extinct*) corresponds to the about 1,300 languages which no one speaks or remembers any more; and 1 and 2 (respectively *Critically endangered* and *Severely endangered*) are languages which are in the brink of extinction in the next 20 years, since only the elderly speak them among themselves. Approximately 20% or 1,400 languages are in these categories, corresponding to doubling the already known losses in such unique parts of the humanity's heritage.

Brazil is the country in the world with the 2nd largest group of critically endangered languages: 45 of the 180 languages spoken in Brazil are barely used by elders, sometimes by less than 10 people (UNESCO 2010). Those are likely to compound the cultural loss of about 80% languages since the landing of Europeans in Brazil in 1500 (Franchetto and Balykova 2020), when an estimate of 800 languages were spoken by over 8 million Indigenous people. Half of the currently spoken languages in Brazil has some reasonable scientific description (Moore and Galucio 2016) but the majority faces the prospect of disappearing without a trace.

This work examines the status of Brazilian languages in perhaps the most ubiquitous catalog of human knowledge: *Wikipedia*. In particular, we perform a thorough analysis of the existence and the quality of Wikipedia articles describing each of those languages, both in English and Portuguese versions. To perform an analysis of the existence of descriptive articles, we start by compiling a list of languages spoken by Indigenous peoples in Brazil by aggregating data from different sources and normalizing the different terms used in their nomenclature. We then look at different aspects associated with the quality of the articles (size, number of references, etc.), and the strength of the community which manages and cares for the articles.

We see this work as an assessment of the current state of what can be considered as the bare minimum register of the existence and gathering of information needed by languages facing disappearance. Our results highlight the urgent need for action by government, NGOs, and Indigenous communities to ensure their proper presence and description of their language. In particular, in our analysis, we contrast the Portuguese and English versions of the Wikipedia, those two distinct linguistic communities may require different types of orchestration and motivation to embrace the need of a Wikipedia documentation project. For instance, it is less likely that the Indigenous individuals would be able to produce the articles in English, since the most common non-Indigenous language spoken by Indigenous people in Brazil is Portuguese.

Notice that this work does not address a more important aspect of the documentation and vitalization of Indigenous languages, which would be the creation of a Wikipedia in their own languages. Such effort could act as an encyclopedic registry that would reflect their worldviews, traditions, and cultures. In the case of Brazilian Indigenous languages, we could find only a couple of languages with proto-Wikipedias in the Wikimedia Incubator[1] with more than 20 entries.

[*]The author was affiliated with IBM Research, Brazil at the beginning of this work.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]https://incubator.wikimedia.org/wiki/Incubator:Main_Page

Therefore, we follow here a series of research questions aimed to provide a portrait of the current situation of Wikipedia entries related to Brazilian Indigenous languages:

- R1: How many articles are there in Wikipedia about Brazilian Indigenous languages?
- R2: Are there differences between the Portuguese and English versions according to the level of endangerment?
- R3: How different are the editing communities of the two versions?
- R4: Are there differences between the quality of the content in the two versions of Wikipedia?
- R5: Are the dynamics of the creation and maintenance of Wikipedia articles different in the two versions?

We believe this paper is an important contribution towards achieving a full representation of Brazilian Indigenous languages in Wikipedia. As described in detail in the rest of the paper, we found that about 30% languages are still not represented and, in particular, extinct, critically, and severely endangered ones. We also found evidence that the English editing community is more active and engaged, while the two communities seem to be quite disconnected. However, the study did not find any significant difference in terms of the size and quality of the articles produced by the two communities. Finally, we will publish, as open-source, the list of names and denominations associated with each language, which can be a resource for future research in the area.

Background on Endangered Languages

Languages are the most effective record of human linguistic and cognitive evolution (Hale et al. 1992). Documenting and analyzing languages is as crucial as archaeology for the understanding of humanity's past. Moreover, languages record distinctive and highly informative ways of thinking and comprehending reality and society (Harrison 2008). Maffi (2002) and Loh and Harmon (2014) point out that the amount of human knowledge about nature is intertwined in Indigenous languages, and how the extinction of linguistic diversity is the loss of ancient, important knowledge about biological diversity.

There are essentially two types of work to avoid the loss of a language: *documentation* and *vitalization*. Documentation involves the collection of linguistic data, including utterances, stories, conversations, and written records both in textual form and in media such as recordings, videos, photographs, etc.; and the creation of grammatical, phonetic, phonological, morphosyntactic, and semantic analysis. Thomason (2015, chapter 6) is a good introduction to documentation practices in linguistics.

Vitalization comprises the activities pursued to maintain and grow the number of speakers of a language and, in particular, the efforts to have children learn it in early age. Notice that vitalization efforts may include work where documentation of the language, gathered in the past or from other sources, is used to help to restore knowledge, enlarge the vocabulary, or recover patterns of speech and accents. When such efforts are done in the context of an extinct or critically endangered language, we use the term *revitalization*. Notable examples are the revitalization of *Hebrew*, from used almost exclusively in religious ceremonies in the early 1900s to being the first language of 7 million people worldwide. Pérez et al. (2019) provides a survey of vitalization efforts.

The period from 2022 to 2032 was proclaimed by UNESCO as the Decade of Indigenous Languages to foment the vitalization and sustainability of linguistic diversity. This measure is an incentive for global governments to raise awareness and work together on actions to inhibit extinction, especially of Indigenous languages, and to protect their cultures and rights.

Related Work

Given its popularity and accessibility, many researchers have conducted studies in Wikipedia, exploring various aspects of the platform, such as its content, structure, and social dynamics. Those studies have provided insights into issues such as automatically assessing article quality (Dalip et al. 2009; Shen, Qi, and Baldwin 2017), analyzing citations (Piccardi et al. 2020; Baigutanova et al. 2023), images (He et al. 2018; Rama et al. 2022), and info-boxes (Graells-Garrido, Lalmas, and Menczer 2015; Lewoniewski 2017) or understanding readers' preferences (Lehmann et al. 2014). Other works focused on identifying underrepresented groups (Graells-Garrido, Lalmas, and Menczer 2015; Mandiberg 2023; Gallert et al. 2016; Sethuraman, Grinter, and Zegura 2020; Hoenen and Rahn 2021) and the asymmetry of the coverage across different language versions (Hale 2015; Graham 2011; Lemmerich et al. 2019; Roy, Bhatia, and Jain 2020; Ashrafimoghari 2023). In this paper, we focus on these two areas. Specifically, we examine the representation of Brazilian Indigenous languages in Wikipedia comparing the English and Portuguese language versions.

Underrepresented Communities in Wikipedia

It is well known that Wikipedia has issues with community representation in its articles, including the content written about certain groups and their active participation. Graells-Garrido, Lalmas, and Menczer (2015) found a gender gap in biographical articles with a higher frequency of terms related to marriage events in women's biographies and a disproportional centrality of articles about men. While women are generally well-covered and featured in numerous Wikipedia language versions, Wagner et al. (2015) discovered significant disparities in the way they are depicted compared to men. In another context, Wang, Pappu, and Cramer (2021) found that female musical artists are more represented in Wikipedia due to their visibility but rock artists (predominantly White and male) are more covered than Latin and hip-hop artists.

Editing participation of minorities in Wikipedia was analyzed by Sethuraman, Grinter, and Zegura (2020) by examining articles produced in locations with Indigenous majorities and rural places. They found that content from Indigenous and rural areas tended to be shorter, had more bot contributions, and received less attention from human editors. In their empirical study on collecting oral information from

an Indigenous community in Namibia, Gallert et al. (2016) concluded that oral narratives are just as valuable as written works for creating Wikipedia citations. However, extracting encyclopedic information from a narrator requires prior insights and the ability to ask the right questions. Mandiberg (2023) observed limitations in using Wikipedia data to assess editing diversity and representation of Indigenous and non-dominant ethnic groups in articles.

Asymmetries among Wikipedia Versions

Graham (2011) argued that content discrepancies across Wikipedia's various language versions, produced in different regions of the world, may result in content pertaining to a specific culture not being written in that culture's language. Roy, Bhatia, and Jain (2020) found important details missing on non-English versions of Wikipedia, even in the largest one (English). Hoenen and Rahn (2021) analyzed 45 small Wikipedias and found common article categories like geolocations, animals, plants, and famous people. (Callahan and Herring 2011) compared articles about American and Polish celebrities, finding that English articles are usually longer with more references and links while Polish articles focus more on personal and professional accomplishments.

Despite efforts to enhance inclusion and diversity on the platform, underrepresented communities still face challenges hindering their active participation. For instance, efforts to increase the number of Wikipedia versions is incipient for Brazilian Indigenous languages. The content written in such languages is only available on the Wikipedia Incubator and it is limited to only five languages (*Pemon*, *Tucano*, *Xavante*, *Xerente*, and *Nheengatu*). In this paper we describe a study to understand how information about Brazilian Indigenous communities are present on the English and Portuguese Wikipedias, starting by articles on their languages.

Normalizing the Nomenclature of the Brazilian Indigenous Languages

The main resources used in this study to determine the Indigenous languages spoken in Brazil were the lists from the *UNESCO Atlas of Endangered Languages* (UNESCO 2010), the 2010 census of Indigenous populations in Brazil (IBGE 2010), and the *Ethnologue* list (Ethnologue 2022). In the following sections, we provide a detailed overview of each list and explain how we combined and matched them with Wikipedia articles towards generating a single, normalized, curated list.

Our primary source for this study was the UNESCO Atlas of Endangered Languages (UNESCO 2010). This atlas contains data on over 8,324 languages spoken and recently extinct in over 80 countries. Each language on this list is classified from 0 to 5 as *Extinct*, *Critically Endangered*, *Severely Endangered*, *Definitely Endangered*, *Vulnerable*, *Stable yet Threatened*, and *Safe*. For our purposes, we focused only on the 190 languages spoken in the Brazilian territory.

To augment the UNESCO list, we consulted the latest available Brazilian census, conducted by *IBGE* in 2010 (IBGE 2010), as the 2022 census is still on going. According to the 2010 census, only 37% of Indigenous children up to the age of 5 years old speak an Indigenous language at home. This statistic is alarming considering the vulnerable state of many of these languages according to UNESCO. The census list, however, has its limitations, as some respondents only provided the name of their language family rather than the specific language spoken[2]. Nevertheless, the IBGE list identified 214 languages across 35 linguistic families and two branches, and we extracted the names of languages and their respective speaker counts for our analysis.

		Wikipedia	
	# Languages	**# pt articles**	**# en articles**
IBGE	214	165 (77%)	155 (72%)
Ethnologue	228	171 (75%)	186 (81%)
UNESCO	190	150 (79%)	154 (81%)
Total (unique)	279	191 (68%)	200 (72%)

Table 1: Number of languages, according to the different lists, which have articles in the Portuguese and English Wikipedias. There are 164 articles in both Wikipedias.

The third list used in this study was provided by the *Ethnologue Language of the World* (Ethnologue 2022). Ethnologue is an annual publication which catalogs over 7,000 languages spoken in more than 200 countries and widely used by linguists (Hoenen and Rahn 2021). The UNESCO *Atlas of Endangered Languages* report (UNESCO 2010) also uses data from Ethnologue publications. From the Ethnologue list, we extracted a total of 228 Indigenous languages, either extinct or currently spoken in Brazil.

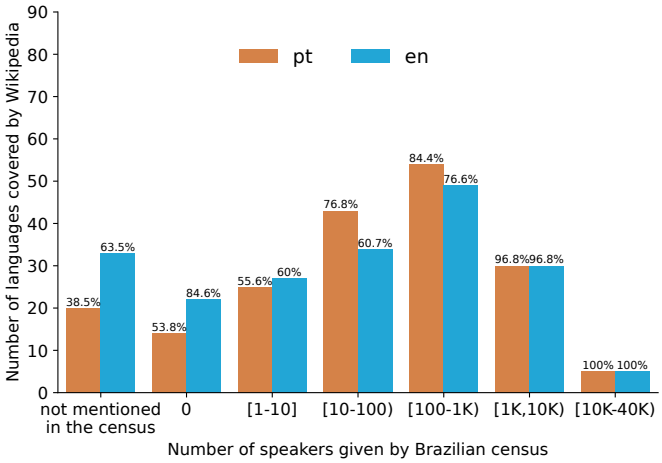
To merge the three lists mentioned above, we selected only the languages which had been or are currently spoken in Brazil. The merging process also considered the Levenshtein distance between the main name and the alternate names (e.g., for Ethnologue) to expand matching. As a result, we were able to match 60% of the languages using string matching similarity, while the remaining ones had to be manually matched. Table 1 presents the number of articles found exclusively in the IBGE census, the Ethnologue list, and the UNESCO Atlas of Endangered Languages, on the Portuguese and English versions of the Wikipedia. We observed an overlap of 52% between the three lists, with the Ethnologue list contributing the largest number of languages (82%) to the merged list. Finally, there may be Brazilian Indigenous languages with articles in Wikipedia which were not listed in our three sources which, unfortunately, we were not able to analyze in this study.

Coverage of Brazilian Indigenous Languages in Portuguese and English Wikipedias

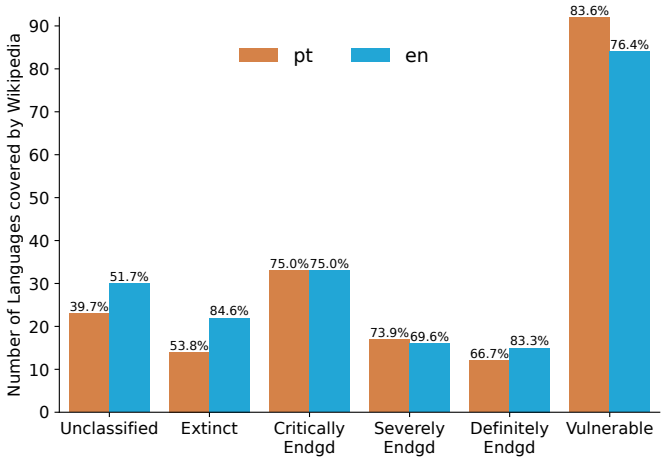
After normalizing the list of Brazilian Indigenous languages, we queried the *Wikipedia API*[3] for each language in the merged list. This resulted in 34,569 articles in Portuguese and 43,679 articles in English. We then applied a filter to select only articles containing the keywords such as *Language*,

[2]In such cases, the census categorized these responses as "unspecified" for each family or subdivision, respectively.

[3]https://www.mediawiki.org/wiki/API:Search



(a) Number of speakers per Indigenous language.



(b) Number of articles per UNESCO degrees of endangerment.

Figure 1: Distribution of Wikipedia Indigenous languages dataset versus offline indicators. The percentages on top of each bar are the number of articles in that bin covered by Wikipedia.

Dialect or *Língua* and *Dialeto* (e.g., Portuguese), resulting in 4,340 and 1,808 articles in Portuguese and English, respectively. Following this filtering, 9% of the languages yield no results in Portuguese and, 6% in English from the API.

From the three language lists, we could extract the ISO 693-3 codes for 86% of the languages in the merged list. The ISO 693-code is a three-letter identifier for all languages, including extinct ones. We use this code to match Wikipedia articles containing that code (21% for Portuguese and 25% for English). After that, we manually inspected these matches to ensure the articles were indeed about the respective language. In this study, our focus is on Wikipedia "Language" articles, which provide information on various facets of a language including classification, phonology, grammar, vocabulary, geographical distribution, historical evolution, and current status.

For that did not matched languages using the ISO code, we adopted an alternative approach based on the similarity between the language's name and its variants, and Wikipedia article titles. For the similarity assessment, we use the Levenshtein distance, considering matches with score greater than 0.5. We then manually inspected these matches, resulting in a total rate of 77% for Portuguese and 81% for English versions. In cases where the API did not yield any results or produced incorrect matches with our approaches, we conducted a manual search on each Wikipedia version to ensure that the article did not exist. This manual search allowed us to found 44 articles in Portuguese and 38 in English.

An overview of the Wikipedia dataset used in this study is described in Table 1. Note that up to 72% of the languages were covered by a Wikipedia article. The number of articles found exclusively in Wikipedia could not be determined, as we cannot estimate that. We explored alternative approaches, such as using the ISO 639-3 provided by the Ethnologue list and querying the SIL website[4], for corresponding Wikipedia articles. However, this approach proved ineffective, as the SIL website only provided links to English Wikipedia articles and some languages share the same ISO 639-3 code. We will release our list of languages and their corresponding Wikipedia article links together with the published paper. Our data shows that when contrasting the number of speakers of each language with its coverage, all the top-25 languages with the highest number of speakers are represented in Wikipedia. Languages with at least one speaker have up to 78% probability of having at least one Wikipedia article, while for languages with no accountable speaker that probability falls to 54%.

Figure 1a shows the relationship between the number of languages covered by the Portuguese and English Wikipedias and the number of speakers for each language, grouped in exponential bins. Languages not mentioned in the Brazilian census but not considered extinct by UNESCO fall into the first bin labeled as *"not mentioned in the census"* due to a lack of available data about the number of speakers. The second bin, labeled with *"0"* comprises languages considered extinct by UNESCO while the following bins contain languages whose number of speakers fall into a certain range of the distribution. We did not find a clear correlation between the number of speakers and the probability to have Wikipedia representation. However, some interesting observations can be made here. For languages not mentioned in the census, extinct languages, and those with fewer than 10 speakers, we notice that the coverage percentage of English articles is higher than in Portuguese, and conversely in the range from 10 to 1,000 speakers. For highly spoken languages (e.g., with at least 1,000 speakers), the probability of the existence of an article about the language is over 96% in both versions.

[4]https://iso639-3.sil.org/code_tables/639/data

		Wikipedia articles	
	# Languages	**# pt**	**# en**
No Glottolog code	19	1	1
Zero references	23	11	9
At least one reference	237	179	190
Total (unique)	279	191	200

Table 2: Number of languages in our dataset found on Glottolog, along with the number of references available there.

	Portuguese	**English**
# of revisions	5,412	15,028
# of unique registered editors	454	1,571
# of unique IPs (anonymous)	254	901
# of unique bots	74	158
# mean revisions/editor (non bot)	7.10	5.41
# of reverts	74	139
# potential vandalisms	8	42

Table 3: Summary of revision activity per Wikipedia version

Next, we examine how Indigenous languages in different levels of endangerment are represented in both Wikipedias. Figure 1b illustrates the distribution of Wikipedia articles across each level of endangerment. The *"Unclassified"* bin refers to languages not reported in the UNESCO Atlas of the World's Languages in Danger (UNESCO 2010). From that plot, we can observe that languages in higher level of endangerment (e.g., from Extinct to Definitely Endangered) are not comprehensive covered by Wikipedia. When comparing the two versions of Wikipedia, we notice that the Portuguese version provides greater coverage for vulnerable languages, while the English version offers better coverage for extinct languages and those not mentioned by UNESCO. The similarity between Figures 1a and 1b is expected, as fewer speakers often indicate a higher risk of language extinction, providing further evidence of more coverage about languages on the brink of extinction by the English Wikipedia, and the opposite for languages severely endangered or above.

Finally, we examined the documentation status of the considered languages using Glottolog's resources and their presence in Wikipedia. Glottolog[5], a bibliographic database for lesser-known languages worldwide, utilizing unique codes assigned to each language, which also integrate into Wikipedia for cross-referencing purposes. For languages absent on Wikipedia, we applied the previously mentioned string similarity approach, leveraging Glottolog's database. These codes enable us to access Glottolog's database to retrieve linguistic references for each language. These references, collected from linguistic libraries worldwide, provide a comprehensive view of available documentation for each language. Table 2 shows the number of languages with Glottolog codes and their presence on Wikipedia. Around 16% of the languages lack representation or any reference in Glottolog. By considering languages with at least one reference as having the potential for new Wikipedia articles, we prospect that 58 and 47 new articles could be created in Portuguese and English, respectively.

Difficulties and Limitations

We encountered significant challenges while compiling the comprehensive list of Wikipedia articles about Brazilian Indigenous languages. One common issue was discrepancies in language names between Portuguese and English, with some languages having significantly different spellings. Additionally, we could not find articles for a few languages, such as the *Brazilian Guarani*, mentioned in the IBGE census. The available article about the Guarani language covered all types of Guarani spoken in Latin America. We also had to expand our list to include derivative languages and dialects explicitly mentioned in the lists. Despite these limitations, we managed to collect separate articles for languages, like *Nhandeva*, *Mbya*, and *Kaiowá*. Finally, we employed manual inspection to match languages with Wikipedia articles. We are aware of the limitations of this approach, primarily due to the scarce resources available for Brazilian Indigenous languages. To address these limitations in future work, we intend to collaborate with linguistic experts to enhance our validation process.

The Editing Practices of the Distinct Portuguese and English Communities

In this section, we analyze the differences between the communities involved in editing the Portuguese and English Wikipedia articles in our datasets. Table 3 presents the main statistics of each Wikipedia version in terms of the revisions[6] made by their respective editors. From the table, we observe that English articles exhibit higher activity compared to the Portuguese version. The number of unique editors engaged in the Portuguese version is much less than in the English one (7.10 to 5.41, respectively), what makes the workload for Portuguese editors larger.

Reverts primarily occurred in articles with the highest revision activity as shown in Table 4, constituting up to 1.4% of the total revisions in both datasets. Regarding potential vandalism, we observed that it was not concentrated in a single article. Using manual inspection of the data, we saw that in the Portuguese articles, reverts manifested through the use of offensive language while in English articles it appeared as nonsensical phrases or the blacking out of entire paragraphs.

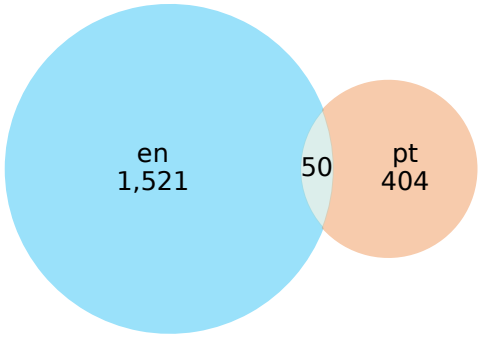
Next, we analyze the composition of the Wikipedia communities in our dataset. Figure 2a presents a Venn diagram illustrating the registered editors (excluding bots and anonymous users) of both Wikipedia versions. From the diagram, we observe a tiny overlap (around 2.6% of editors) who made revisions in both versions. For those editors who made changes in both version, we examine the prevalence of each version in terms of editing. Figure 2b shows all the revisions made by such bi-version editors, categorized by the article version. The editors are ranked based on their total number of revisions (combining Portuguese and English revisions). We note that most of the editors exhibit a preference for editing in one of the languages, with 84% of them

[5]https://glottolog.org/

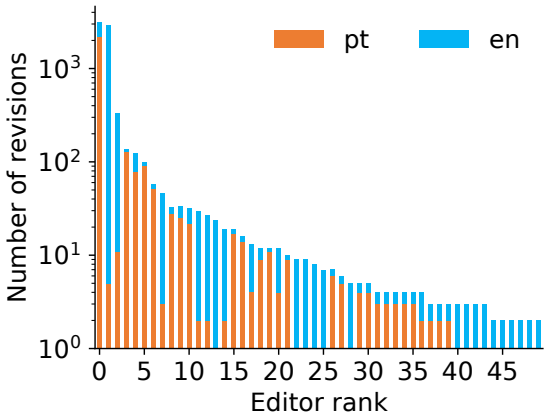
[6]We use revisions and edits interchangeably.

Portuguese (pt)					English (en)				
Article title	**ISO**	**Total**	**Editors**	**Rev/editor**	**Article title**	**ISO**	**Total**	**Editors**	**Rev/editor**
Língua Nheengatu	yrl	441	88	4.2	Guarani language	grn	1,081	308	2.31
Língua Guarani	grn	318	95	1.64	Pirahã language	myp	839	263	2.33
Língua Pirarrã	myp	128	36	2.05	Tupi language	tpk	636	142	3.44
Língua Maxacali	mbl	110	26	3.54	Nheengatu	yrl	367	106	2.96
Língua catuquina-canamari	knm	102	9	11.22	Ticuna language	tca	214	61	2.54

Table 4: Top-5 articles with the highest number of revisions.



(a) Number of registered editors in each Wiki community



(b) Number of revisions made by the same editor in both versions (y-axis in log scale).

Figure 2: Wikipedia Editing Communities

having made significantly more revisions in Portuguese. As observed by Park et al. (2021), editors tend to be more active in their primary language and make smaller revisions in their secondary languages. This suggests the possibility that the majority of revisions to articles on Brazilian Indigenous languages are being made by Portuguese speakers. In sum, the two editing communities seem to be highly focused on just one of the versions: there is a very small number of bi-version editors and even them seem to have clear preferences for just one of the versions.

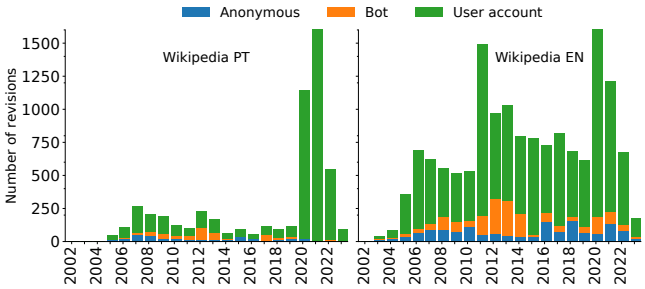


Figure 3: Revision activity 2002-2023.

To better understand how both communities behave over time, we analyze the revision dynamics over time. Figure 3 shows the number of revisions over time for both versions, categorized by the type of editors who made the edition. As shown in Table 3, the number of revisions for English

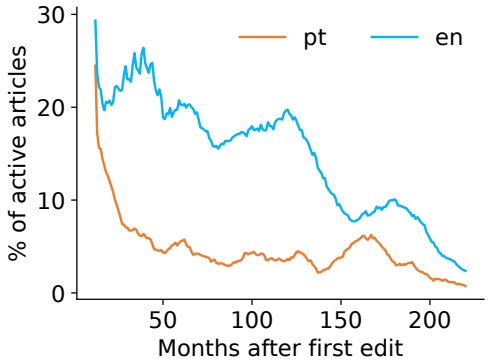


Figure 4: Number of articles active within the last year. An article is active if it has at least one edit during this period.

articles is significantly higher and more intense compared to the Portuguese version. In general, Wikipedia versions with a higher number of active editors tend to be more frequently updated (Roy, Bhatia, and Jain 2020). The revision timelines also reveals that Portuguese articles gained traction post-2020, while English articles experienced activity spikes in 2011 and 2020. The surge in 2020 could be linked to external factors, as the pandemic and political events related to Indigenous people in Brazil impacting Wikipedia content (Watch 2023; Times 2023).

Next, we analyze the frequency of revisions made to the

article after its creation. Figure 4 shows the monthly percentage of active articles since their creation in each Wikipedia version. An article is considered active if it has at least one revision in a given month. We then plot a moving average for each month after the article's creation (window size = 12 months). Both versions show a similar decrease in activity after the initial two months, with up to 37% of the articles remaining active). However, Portuguese articles have a higher dropout rate after this period. Indeed, after one year, only 13% of the articles written in Portuguese were still active, in contrast to 32% of English articles. This result suggests that articles in Portuguese are possibly abandoned at a much higher rate compared to English articles.

Lastly, Table 4 lists the most edited articles in our dataset. For both versions, articles about *Nheengatu*, *Guarani*, and *Pirahã (Pirarrã)* languages are the most active. *Nheengatu* is the 13^{th} most commonly spoken Indigenous language in Brazil, according to the 2010 census. The *Pirahã* language has been a subject of discussion among linguists due to its lack of recursion, related to the ability to form complex sentences (Futrell et al. 2016). However, there is ongoing debate among linguists regarding recursion, generating considerable discussion and activity in that article. Finally, some languages have articles covering multiple language variants across different countries, such as *Guarani*, also known as *Paraguayan Guarani*, a co-official language in Bolivia and Paraguay, with approximately 6 million speakers. This may explain that high activity in Wikipedia.

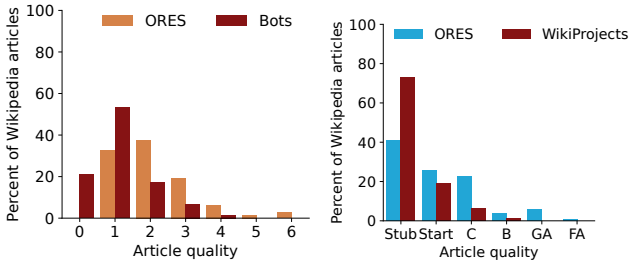
The observed disparities between English and Portuguese Wikipedia can be attributed to several factors. Firstly, English Wikipedia being the largest community in terms of both editors and articles, enjoys a global reach and broader accessibility, naturally attracting a more diverse contributor base and content. Conversely, Portuguese Wikipedia, ranking 18^{th} in articles and 7^{th} in editors[7], though substantial, is limited to Portuguese-speaking regions, potentially restricting its contributor base. Secondly, differences in editing culture, community norms, and incentives affect editors numbers and activity on each Wikipedia.

A Comparison of the Quality of the Articles in Portuguese and English

In this section, we analyze the differences between the Brazilian Indigenous language articles in the Portuguese and English versions of the Wikipedia, considering the content present in the articles. To compare both version we use different metrics such as article quality, article length, number of citations, images and so on.

Article Quality ORES Assessement

The Wikipedia communities for each language version has defined grading systems to assess the quality of their articles. For instance, the Portuguese version employs a six-level grading system, with the highest index corresponding to the most favorable quality assessment. Automatic assessments are also conducted by Wikipedia's bots for levels 1 to 4, as these levels are more objectively determined based on factor such as number of images, and references. User evaluations can replace these bot-generated assessments. In the case of the English version, article quality is classified into six classes (from lowest to the highest): Stub, Start, C, B, Good Article, and Featured article. Wikipedia also includes WikiProjects, groups of editors that specialize in specific areas of knowledge or topics. They collaborate to improve and expand related articles, and often conduct quality assessments to monitor progress.



(a) Wikipedia PT (b) Wikipedia EN

Figure 5: Article quality assessments: In the Portuguese Wikipedia, a zero score indicates an unassessed article.

Another method to gauge article quality is the Objective Revision Evaluation Service (ORES)[8] and adopted by several research studies of Wikipedia article quality (Teblunthuis 2021; Zhang and Terveen 2021). This service analyzes various structural attributes of articles (e.g., number of sections, presence of infoboxes, quantity of references), to predict a quality label for each article.

Figure 5 shows the distribution of article quality assessed by WikiProjects, bots (only for Portuguese version) and the ORES service. If we had multiple quality evaluations, which is the case for pages assessed by more than one project, we considered the highest score for that article. We observe that, for Portuguese articles, all articles were assessed by automated bots. We can see that the proportion of high-quality articles (rated 5 and 6 for Portuguese articles, and GA and FA for English articles) is relatively small in both Wikipedia versions[9]. This observation aligns with the broader Wikipedia statistics, where such articles comprise around 0.3% on average in each language (Lewoniewski, Wecel, and Abramowicz 2017). We observe that 65% and 48% of the Portuguese and English articles, respectively, had ORES scores higher than their respective assessments conducted by projects and bots. This overestimation may be attributed to two main factors: potential prediction errors within the ORES system and irregular, infrequent assessments by volunteer editors, leading to ratings lagging behind actual changes in article quality (Teblunthuis 2021). Given the well-established nature of the ORES score, we will maintain its use for our next analysis.

We also contrast other metrics used previously for qual-

[7]https://meta.wikimedia.org/wiki/List_of_Wikipedias/Table

[8]https://ores.wikimedia.org

[9]We didn't compare the quality scores between the 2 versions, as the quality criteria for each version are significantly different.

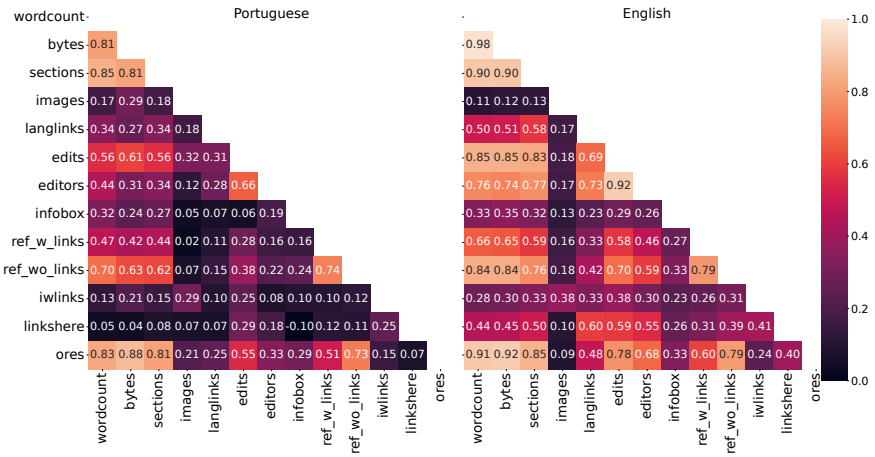
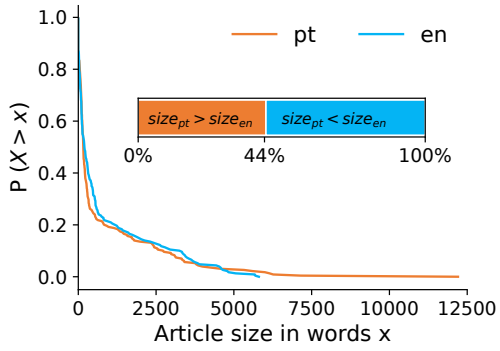
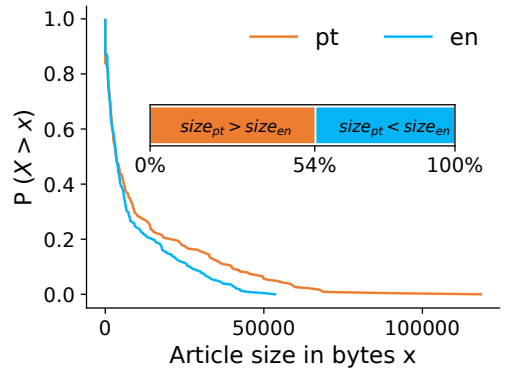


Figure 6: Spearman Correlation matrix of articles metadata for each Wikipedia version.



(a) Article size in words



(b) Article size in bytes

Figure 7: Article length cumulative distributions.

ity assessment of Wikipedia articles, such as structural features (e.g., article length, number of sections, images, citations) and edition activity (e.g., unique editors and revisions) (Dalip et al. 2009; Shen, Qi, and Baldwin 2017). Figure 6 presents a matrix where each element is the Spearman correlation (ρ) between each feature and the ORES score. Notably, most features in both versions exhibit very weak to moderate correlations ($\rho \leq 0.59$). In the English version, we observe a higher number of moderately and strongly correlated features, with 32 correlations compared to 21 in the Portuguese version. When examining the correlations with ORES scores, we find that high-quality articles tend to be more verbose, as indicated by the correlations between word count, number of bytes, and the number of sections. Additionally, for the English version, we observe a strong correlation between the number of editors and article quality. To provide a more in-depth analysis of these metrics, we will explore their variability among the articles in our next step.

Title	Length diff.	ORES
Língua Tucano (Tucano language)	113,993 chars	pt=3, en=Start
Língua Galibi (Carib language)	11,261 words	pt=3, en=C
Lanc-Patuá (Karipúna French Creole)	-41,719 chars	pt=1, en=FA
Língua Caro (Ramarama language)	-5,422 words	pt=2, en=B

Table 5: Largest differences in article length for parallel articles. Negative values indicate that articles in English are larger than their respective counterparts in Portuguese.

Wikipedia Article Statistics

Amount of Content. We initially assessed the size of articles on Brazilian Indigenous languages in both Wikipedia versions. Figure 7a and 7b show the complementary cumulative distribution function (CCDF) of word count and bytes, in both Wikipedia versions. These length distributions did not show statistically significant differences between the two Wikipedia versions (*Kolmogorov-Smirnov* p-value > 0.05).

As for the distribution of the length of the Brazilian Indigenous language articles in both versions, we found articles ranging from 0 to 118,391 bytes (12,204 words) for Portuguese and 0 to 53,519 bytes (5,804 words) for English. Zero characters or words indicate the absence of an article

Portuguese		English	
Domain	**Total**	**Domain**	**Total**
etnolinguistica (pt)	22%	ethnologue (en)	24%
sociomabiental (pt)	9%	ethnolinguistica (pt)	9%
unb (pt)	7%	archive (various)	8%
archive (various)	6%	socioambiental (pt)	6%
worldcat (en)	6%	unicamp (pt)	3%
editoracrv (pt)	4%	semanticscholar (en)	3%
unicamp (pt)	3%	uchicago (en)	3%
sil (en)	3%	jstor (en)	2%
geocities (pt)	3%	berkeley (en)	2%
scielo (pt)	3%	wordcat (en)	2%
Total links	677	**Total links**	828
Zero links	2.0%	**Zero links**	4.2%

Table 6: Top-10 most popular domains for external links in the Reference section.

on Wikipedia. Both word count distributions, as shown in Figure 7a, are very similar up to 5,800 words. Particularly lengthy articles in Portuguese, include those about the languages *Galibi* (*Carib*) and *Tucano* with 12,204 words and 118,391 bytes, respectively.

We also analyze the article length differences between parallel articles – those about the same Indigenous language – in both Portuguese and English versions. In 55% of the articles, the English articles are larger than their Portuguese counterparts, with an average of 1,045 more words. Conversely, in terms of bytes (Figure 7b), 54% of Portuguese articles are larger than their English versions. Notably, extreme differences (Table 5) were observed in articles such as *Tucano* and *Galibi (Carib)*, which stand out in terms of bytes and word count, respectively, when comparing Portuguese and English versions. Conversely, the most significant disparities favoring the English version were observed in articles like *Lanc-patuá (Karipuna French Creole)* for size in bytes and *Caro (Ramarama)* for word count. Discrepancies in article quality, indicated by the ORES score, suggest that translating these articles into their respective versions could enhance the available information.

Citations. We analyze the citations listed in the *References* section of each Wikipedia article, including citations with and without links to external sources. Citations are crucial, as Wikipedia guidelines require editors to support their edits with appropriate references to ensure verifiability (Baigutanova et al. 2023). Indeed, the number of citations (ref_wo_link) and links (ref_w_link) in the Reference sections correlates with the ORES metric, as show in Figure 6. This suggests that high-quality articles are associated with the presence of citations and links. While, the distributions of citations differ significantly between the two versions (*Kolmogorov-Smirnov* p-value < 0.05), the distributions of links are quite similar (*Kolmogorov-Smirnov* p-value > 0.05). Around 20% of articles in both English and Portuguese share at least 7 and 9 citations, respectively. Notably, 7.8% of Portuguese articles and 4.5% of English articles have no links in the *References* section. Additionally, 1.57% of Portuguese articles and 3.5% of English articles have no citations. These findings suggest that most of Wikipedia articles about Brazilian Indigenous languages in both versions are well-referenced.

Table 6 lists the top 10 most referenced domains in the links. We found that up to 41% of the links lead to sites that no longer exist. Among the frequently cited domains are those cataloging languages (e.g., Ethnologue, Etnolinguistica, SIL), digital university repositories (e.g. UNB, Unicamp, Uchicago, and Berkeley), and digital libraries (e.g., WorldCat, JSTOR, Scielo, and Semantic scholar), as well as Portuguese-language books (editoracrv). This suggests that many sources of information in the articles are reputable. We also observed external references to Internet archives like *archive.org*. An official source of information on Indigenous peoples in Brazil, the *Instituto Socio Ambiental*, is often cited in both versions. In Portuguese articles, there are few links to an Indigenous vocabulary website (e.g., Geocities) without official citations. Notably, most links lead to content in the same language as the article; Portuguese articles primarily link to Portuguese-written sites, while English articles do the same in English.

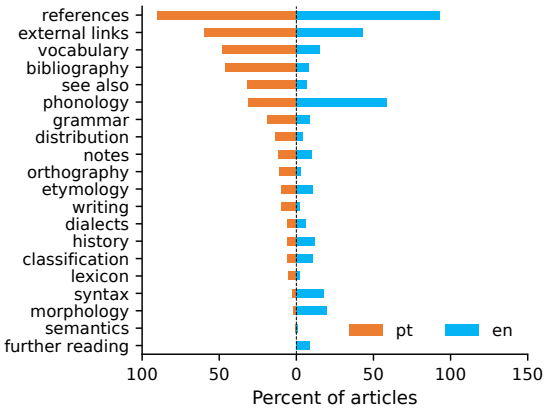
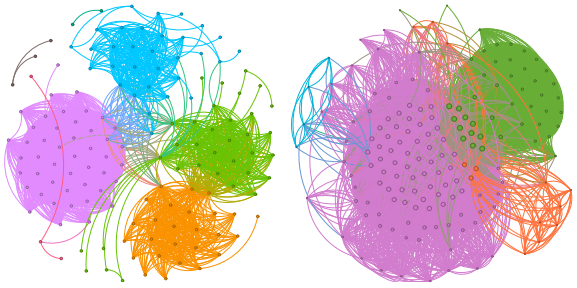


Figure 8: Distribution of top-20 article sections

Articles Network We generate a Wikipedia graph based on the links between our set of articles in each version, excluding links outside this set, including redirects. To compare the two versions, we focused only on languages present in both Wikipedias, resulting into 164 languages. Table 7 summarizes the main characteristics of each graph. When examining the properties of the two networks, we observe that the English articles exhibit a higher connectivity, as indicated by the higher number of edges, average degree, and density when compared to the Portuguese version.

To understand the communities within each graph, we applied the Louvain algorithm to community detection and assessed community membership using the Normalized Mutual Information (NMI) metric (Ferreira et al. 2021). This metric is used to evaluate the similarity between two sets of communities, quantifying the extent to which information is shared between them. A higher NMI value indicates greater similarity between the two sets of communities, with 1 indicating perfect agreement and 0 indicating no agreement. We obtained a value of 0.5579, suggesting an overlap between



(a) PT Wikipedia articles (b) EN Wikipedia articles

Figure 9: Wikipedia Indigenous language article network. Both graphs include the same Indigenous languages.

the communities of English and Portuguese articles. In other words, there are sub-communities common to both versions.

In case of Portuguese, we observed that communities are defined by several factors, including language families (for example, the purple and green communities mainly consist of *Tupi* and *Macro-Jê* family languages, respectively), languages spoken in countries bordering Brazil (orange cluster), and groups of languages without a family affiliation (such as *Aruak* in blue cluster). On the other hand, in the English articles, the clusters are less distinct compared to Portuguese. However, we do observe a cluster containing the majority of languages from *Tupi* family (green cluster).

High connectivity in Wikipedia articles has both advantages and disadvantages, which can be examined from two perspectives. On one hand, it affects the accessibility and discoverability of Brazilian Indigenous language knowledge and culture for a broader audience. Highly connected articles are more likely to be discovered and explored by users, fostering a deeper understanding of Indigenous topics. This accessibility benefits readers to delve deeper into the subject, and it enables editors to contribute to a broader knowledge ecosystem while maintaining consistent cross-referencing. On the other hand, the presence of a higher number of communities and higher modularity in Portuguese articles, as discussed earlier, indicates the existence of more distinct semantic clusters. These clusters can assist readers in locating specific information tailored to a specific language or family of languages, facilitating a better understanding of the relationships between different aspects of a topic. For the editors' community, these clusters can be valuable in identifying missing articles or areas in need of improvement.

Finally, we computed the Spearman correlation between the ORES article quality score and the total number of links to and from each article, taking into account links that extend beyond the set of Indigenous language articles. For the Portuguese version, we observe that these two metrics exhibited no significant correlation, with values up to $\rho = 0.14$. However, in the English version, we identified a moderate correlation (up to $\rho = 0.41$). This suggests that, in the case of English articles, the quality of an article may be related to its connectivity with the broader Wikipedia ecosystem.

Sections Structure. We examined how articles are organized into sections. In terms of the number of sections, we found no statistical difference between Portuguese and English articles (p-value > 0.05). The quality of an article is closely tied to the number of sections, which is also correlated with the article's length (word count and number of bytes). Investigating article structure, Figure 8 shows the distributions of the top-20 article sections in both English and Portuguese Wikipedia versions. Most articles in both versions contain a "References" section. However, a small percentage, 7% in English and 10% in Portuguese, lack a "References" section. These percentages exceed those of articles with no citations nor links, as some references may appear in other sections, like "Source" or "Notes". As previously discussed, maintaining a high article quality relies on including references. Figure 8 also reveals that articles in Portuguese tend to have a higher frequency of sections related to links and bibliography, such as *External Links*, *See Also*, and *Bibliography*. In contrast, English articles often focus on technical aspects of language, including *Phonology*, *Morphology*, and *Syntax*. We searched for Wikipedia guidelines on writing language articles and found a language article template that includes recommended sections and an infobox template[10]. This template currently suggests 10 sections, such as *Classification*, *History*, *Grammar*, *Phonology*, and some of which align with sections found in the collected articles. These findings suggest that the articles may have been written by experts, possibly linguists. Confirming this require individual analysis of the articles in our dataset.

Other features In our analysis of the other features mentioned in the correlation matrix, we found that Portuguese articles tend to include more images. However, in both versions, the number of images is only weak correlated with article quality. Concerning the number of versions (*langlinks*), we found that, on average, articles in the dataset have around 7 versions in Wikipedia. The article with the most versions is "Guarani", with 103 versions in Portuguese and 104 in English. English articles show a moderate correlation between the number of versions and quality. Infoboxes are prevalent in 78% of Portuguese articles and 99% of English articles. In English articles, common infobox attributes include the number of speakers, year of extinction, and ISO 639-3 code while Portuguese articles often contain geographical attributes. Nevertheless, the presence of infobox information has only a weak correlation with article quality.

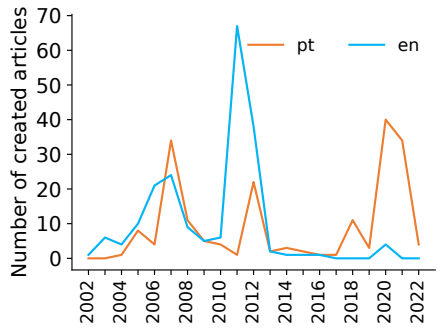
The Temporal Dynamics of Creation and Maintenance of the Articles

In this section, we examine the timeline of article creations over the years. Figure 10a shows the creation dates of the articles based on the timestamp of their first revision. We did not find statistical differences in the distributions (*Kolmogorov-Smirnov* p-value > 0.05). Our dataset includes articles spanning two decades, with some created as recently as a year ago. Note that the peaks of article creation are not synchronized between the English and Portuguese versions. English articles had a pronounced peak around 2011 with 67 articles, while the Portuguese version experienced a peak in

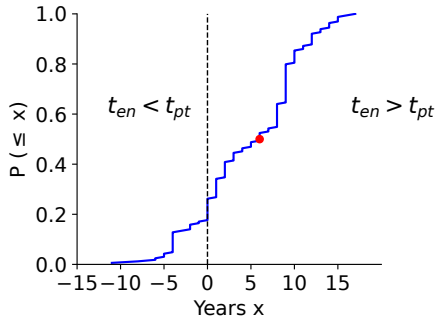
[10] https://en.wikipedia.org/wiki/Wikipedia:WikiProject Languages/Template

	# nodes	#single nodes	# edges	Avg degree (std)	Density	#C.C	#Comm.	Mod.
Wiki PT	145	19	2,668	36.8 (30.12)	0.1278	59	7	0.4976
Wiki EN	164	0	9,799	119.5 (64.54)	0.3666	14	4	0.2115

Table 7: Network metrics for the article graphs.



(a) Creation timeline



(b) Creation intervals

Figure 10: Article creation temporal dynamics.

2020 with 40 articles. These temporal differences in creation peaks are also reflected in Figure 10b, indicating that most articles were initially created in English before Portuguese.

Lastly, we analyzed the creation timeline of articles in both versions by examining the time difference between their creation in the Portuguese and English. Figure 10b shows the cumulative distribution of time differences for articles present in both versions. Positive values indicate that the English version (t_{en}) was created before the respective Portuguese version (t_{pt}). Note that most of the articles (80%) exhibit positive time differences, indicating English version precedence. However, 9% of articles have both versions created in the same year, and 17% were first created in Portuguese. The red dot represents the median number of articles, indicating that 50% of the articles had the English version created before the Portuguese, within a 6-year time span. These findings support Figure 10b, indicating that the English community initiated articles about Brazilian Indigenous languages much earlier than the Portuguese.

Conclusion

We analyzed Brazilian Indigenous Languages articles in Wikipedia and found that most spoken Indigenous languages are present, and even extinct languages attracted the attention of editors. However, more community engagement is needed, as only 68% of Indigenous languages are covered. This signifies a concerning lack of crucial information about these cultures in the digital world.

Our study explored the dynamics of the English and Portuguese communities in creating and editing these articles. We found evidence suggesting that the English community is more engaged and potentially more diverse. An important finding is that the English and Portuguese communities are quite distinct, having a tiny number of editors in common. Even those who edit in both versions tend to prefer one over the other. Comparing content produced by both communities, we found no significant differences in article lengths and small differences for other indicators like the number of references. We also noticed temporal differences, with English versions often preceding Portuguese versions by several years. Understanding these dynamics can shed light on the unique ways these linguistic communities engage with and contribute to Indigenous languages on Wikipedia representation. Further research can guide more inclusive contributions across linguistic communities.

To boost coverage, we suggest reporting these gaps to public authorities, notably considering Brazil's new ministry for Indigenous population and the UNESCO efforts of the Decade of Indigenous Languages. Additionally, the existing articles require more frequent updates, as our results reveal declining edit rates after two years. As future work, we plan to analyze other Wikipedia versions for the same set of languages. We also intend to explore articles about the people and ethnicities associated with these languages and other aspects of the languages, such as subfamily and families. Finally, we were unable to explore a critical aspect of the problem, which involves assessing the engagement level of Indigenous communities themselves in creating and maintaining these articles. This is currently unavailable in the Wikipedia records and, therefore, must be obtained by other means. We acknowledge that Indigenous participation and leadership are fundamental in any effort to document and vitalize Indigenous languages, following the *"Nothing for us without us"* principle established in the *Los Pinos Declaration* of 2020 (UNESCO 2020) and understand that their active involvement in the documentation process is crucial.

Ethical considerations. We do not foresee a negative societal impact coming from this research, on the contrary, it may positively contribute to UNESCO Decade of Indigenous languages initiative and Brazilian public policies for Indigenous languages' preservation.

References

Ashrafimoghari, V. 2023. Detecting Cross-Lingual Information Gaps in Wikipedia. In *Proc. of the WWW*.

Baigutanova, A.; Myung, J.; Saez-Trumper, D.; Chou, A.-J.; Redi, M.; Jung, C.; and Cha, M. 2023. Longitudinal Assessment of Reference Quality on Wikipedia. In *Proc. of WWW*.

Callahan, E. S.; and Herring, S. C. 2011. Cultural bias in Wikipedia content on famous persons. *JASIST*, 62(10): 1899–1915.

Dalip, D.; André Gonçalves, M.; Cristo, M.; and Calado, P. 2009. Automatic Quality Assessment of Content Created Collaboratively by Web Communities: A Case Study of Wikipedia. In *Proc. of the JCDL*.

Ethnologue. 2022. O Brasil Indígena: Estudos especiais. https://www.ethnologue.com/country/BR/.

Ferreira, C.; Murai, F.; Silva, A.; Almeida, J.; Trevisan, M.; Vassio, L.; Mellia, M.; and Drago, I. 2021. On the dynamics of political discussions on Instagram: A network perspective. *OSNEM*, 25: 100155.

Franchetto, B.; and Balykova, K. 2020. Introdução. In *Índio não fala só tupi: Uma viagem pelas línguas dos povos originários no Brasil*. 7Letras.

Futrell, R.; Stearns, L.; Everett, D. L.; Piantadosi, S. T.; and Gibson, E. 2016. A Corpus Investigation of Syntactic Embedding in Pirahã. *PLOS ONE*, 11(3): 1–20.

Gallert, P.; Winschiers-Theophilus, H.; Kapuire, G.; Stanley, C.; Cabrero, D.; and Shabangu, B. 2016. Indigenous Knowledge for Wikipedia: A Case Study with an OvaHerero Community in Eastern Namibia. In *Proc. of AfriCHI*.

Graells-Garrido, E.; Lalmas, M.; and Menczer, F. 2015. First Women, Second Sex: Gender Bias in Wikipedia. In *Proc. HT*.

Graham, M. 2011. Wiki Space: Palimpsests and the Politics of Exclusion in Critical Point of View: A Wikipedia Reader. *Critical Point of View: A Wikipedia Reader*, 269–282.

Hale, K.; Krauss, M.; Watahomigie, L.; Yamamoto, A.; Craig, C.; Jeanne, M.; and England, N. 1992. Endangered languages. *Language*, 68(1): 1–42.

Hale, S. A. 2015. Cross-language Wikipedia Editing of Okinawa, Japan. In *Proc. of CHI*.

Harrison, K. 2008. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

He, S.; Lin, A.; Adar, E.; and Hecht, B. 2018. The_Tower_of_Babel.jpg: Diversity of Visual Encyclopedic Knowledge Across Wikipedia Language Editions. In *Proc. ICWSM*.

Hoenen, A.; and Rahn, M. 2021. Migration of Small and Endangered Languages into the Wikipedia. *ComputEL*, 2.

IBGE. 2010. O Brasil Indígena: Estudos especiais. https://indigenas.ibge.gov.br/estudos-especiais-3/o-brasil-indigena/download. Accessed: 2023-03-14.

Lehmann, J.; Müller-Birn, C.; Laniado, D.; Lalmas, M.; and Kaltenbrunner, A. 2014. Reader preferences and behavior on Wikipedia. In *In Proc. of HT*.

Lemmerich, F.; Sáez-Trumper, D.; West, R.; and Zia, L. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proc. of WSDM*.

Lewoniewski, W. 2017. Completeness and Reliability of Wikipedia Infoboxes in Various Languages. *Business Information Systems Workshops*, 303.

Lewoniewski, W.; Wecel, K.; and Abramowicz, W. 2017. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics*, 4: 43.

Loh, J.; and Harmon, D. 2014. *Biocultural diversity: threatened species, endangered languages*. WWF Netherlands.

Maffi, L. 2002. Endangered languages, endangered knowledge. *ISSJ*, 54(173): 385–393.

Mandiberg, M. 2023. Wikipedia's Race and Ethnicity Gap and the Unverifiability of Whiteness. *Social Text*, 41(1 (154)): 21–46.

Moore, D.; and Galucio, A. V. 2016. Perspectives for the documentation of indigenous languages in Brazil. In *Language Documentation and Revitalization in Latin American Contexts*, 29–58. De Gruyter Mouton.

Park, S.; Kim, S.; Hale, S.; Kim, S.; Byun, J.; and Oh, A. 2021. MultilingualWikipedia: Editors of Primary Language Contribute to More Complex Articles. In *Proc. of ICWSM*.

Pérez Báez, G.; Vogel, R.; and Patolo, U. 2019. Global Survey of Revitalization Efforts: A mixed methods approach to understanding language revitalization practices. *Language Documentation & Conservation*, 13: 446–513.

Piccardi, T.; Redi, M.; Colavizza, G.; and West, R. 2020. Quantifying Engagement with Citations on Wikipedia. In *Proc. of WWW*.

Rama, D.; Piccardi, T.; Redi, M.; and Schifanella, R. 2022. A Large Scale Study of Reader Interactions with Images on Wikipedia. *EPJ Data Science*, 11(1).

Roy, D.; Bhatia, S.; and Jain, P. 2020. A Topic-Aligned Multilingual Corpus of Wikipedia Articles for Studying Information Asymmetry in migration Resource Languages. In *Proc. of LREC*.

Sethuraman, M.; Grinter, R.; and Zegura, E. 2020. Approaches to Understanding Indigenous Content Production on Wikipedia. In *COMPASS '20*.

Shen, A.; Qi, J.; and Baldwin, T. 2017. A Hybrid Model for Quality Assessment of Wikipedia Articles. In *Proc. of ALTA*.

Teblunthuis, N. 2021. Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression. In *Proc. of OpenSym*.

Thomason, S. 2015. *Endangered languages*. Cambridge University Press.

Times, N. Y. 2023. As Bolsonaro Keeps Amazon Vows, Brazil's Indigenous Fear 'Ethnocide'. https://www.nytimes.com/2020/04/19/world/americas/bolsonaro-brazil-amazon-indigenous.html. Accessed: 2024-04-06.

UNESCO. 2010. Atlas of the World's Languages in Danger. https://unesdoc.unesco.org/ark:/48223/pf0000187026. Accessed: 2024-04-06.

UNESCO. 2020. Los Pinos Declaration (Chapoltepek) – Making a Decade of Action for Indigenous Languages. https://unesdoc.unesco.org/ark:/48223/pf0000374030. Accessed: 2024-04-06.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proc. of ICWSM*.

Wang, A.; Pappu, A.; and Cramer, H. 2021. Representation of Music Creators on Wikipedia, Differences in Gender and Genre. In *Proc. of ICWSM*.

Watch, H. R. 2023. Bolsonaro's Plan to Legalize Crimes Against Indigenous Peoples. https://www.hrw.org/news/2020/03/01/bolsonaros-plan-legalize-crimes-against-indigenous-peoples. Accessed: 2024-04-06.

Zhang, C.; and Terveen, L. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *Proc. of OpenSym*.

AAAI ICWSM Paper Checklist

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes. Every result explanation follows and explanation about the limitations.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. We do not foresee a negative societal impact coming from this research, on the contrary, it may positively contribute to UNESCO Decade of Indigenous languages initiative and Brazilian public policies for Indigenous languages preservation.
   (g) Did you discuss any potential misuse of your work? Yes. See the answer of the previous item.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA. This paper exclusively employs exploratory analysis.
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA. We are the first to study Wikipedia articles for Brazilian Indigenous languages..
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes. We better describe them in the Section .
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes. We have used the ORES service to assess the quality of each Wikipedia article.
   (b) Did you mention the license of the assets? Yes. ORES is open source, and more details are available on the website.
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. Our dataset is from Wikipedia, a publicly accessible platform where content is contributed voluntarily. Wikipedia's content is covered by Creative Commons licenses, allowing for free use and sharing without requiring individual consent.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. Wikipedia has policies and community guidelines in place to moderate and remove offensive or inappropriate content, and we have taken precautions to work with sanitized and anonymized data to avoid any inclusion of PII.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see fair)? NA. We are not releasing the dataset since anyone can collect it using the Wikipedia API without any difficulty. Wikipedia already implements FAIR principles for its data.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see gebru2021datasheets)? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA