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

Marisa Vasconcelos & Priscila Mizukami & Claudio Pinhanez
marisavas@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. Most of these languages lack a single scientific article, risking disappearance without trace. This work examines Wikipedia articles about these languages revealing over 30% lacking representation. Additionally, Portuguese and English editing communities, while distinct, achieve similar quality levels with different practices and temporal dynamics. Efforts to enhance coverage in both Wikipedias should consider community-specific strategies.

Indigenous Languages Status

- About 43% of the world's 7,000 languages are endangered, many spoken exclusively by Indigenous peoples in Africa and the Americas.
- UNESCO proclaimed the period from 2022 to 2032 as the Decade of Indigenous Languages to promote the vitalization and sustainability of linguistic diversity.
- Brazil has the 2^{nd} largest group of critically endangered languages in the world. Of its 180 languages, 45 are barely used by elders, sometimes by fewer than 10 people.
- Despite half of Brazil's currently spoken languages having some scientific description, most face the prospect of disappearing without a trace.
- This work investigates the representation and quality of Brazilian Indigenous languages in Wikipedia in both English and Portuguese versions.

Degree of endangement	Speaker population
Safe	Language is spoken by all generations
Vunerable	Most children speak the language, but it may be restricted to certain domains (e.g., home)
Definitely endangered	Children no longer learn the language as a mother tongue in the home.
Severely endangered	The language is spoken by grandparents and older generations. While the parent generation may understand it, they do not speak it to children or among themselves.
Critically endangered	The youngest speakers are grandparents and older, and they speak the language partially and infrequently.
Extinct	There is no one who can speak or remember the language.

Figure 1: UNESCO language endangerment classification

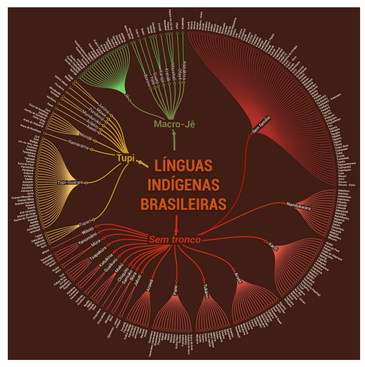

Figure 2: UNESCO language endangerment classification

Research Questions

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

Methodology

- Compiled a list of Indigenous languages spoken in Brazil using UNESCO Atlas of Endangered Languages, the 2010 Brazilian census, and the Ethnologue catalogue [1, 2, 3].
- Used string similarity methods to merge and match language names and their variants across datasets.
- Queried the Wikipedia API for articles related to each identified language, applying a filter to select articles containing relevant keywords such as "Language", "Dialect", "Língua", and "Dialeto".
- Extracted ISO 693-3 codes from each article for language matching with Wikipedia articles, and utilized name similarity when ISO codes were unavailable.

Figure 3 illustrates the entire process of collecting Wikipedia articles.

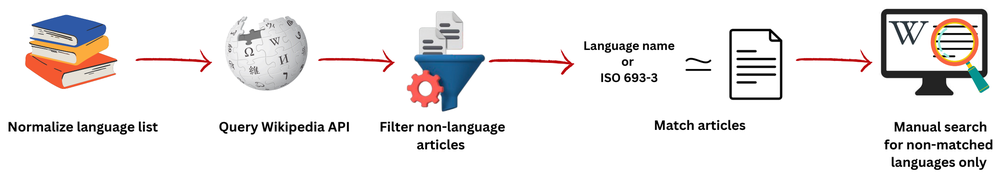

Figure 3: Workflow for Collecting and Validating Wikipedia Data on Brazilian Indigenous Languages.

Coverage of Brazilian Indigenous Languages in Wikipedia

		Wikipedia	
	# Languages	# pt articles	# en articles
Brazilian Census	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: Brazilian Indigenous languages.

- Top-25 languages with the highest number of speakers are all represented on Wikipedia.
- Languages at higher endangerment levels (e.g., Extinct to Definitely Endangered) often lack representation.
- English and Portuguese Wikipedia prioritize languages differently based on speaker count.
- Collaborative efforts and using resources like Glottolog can enhance coverage.

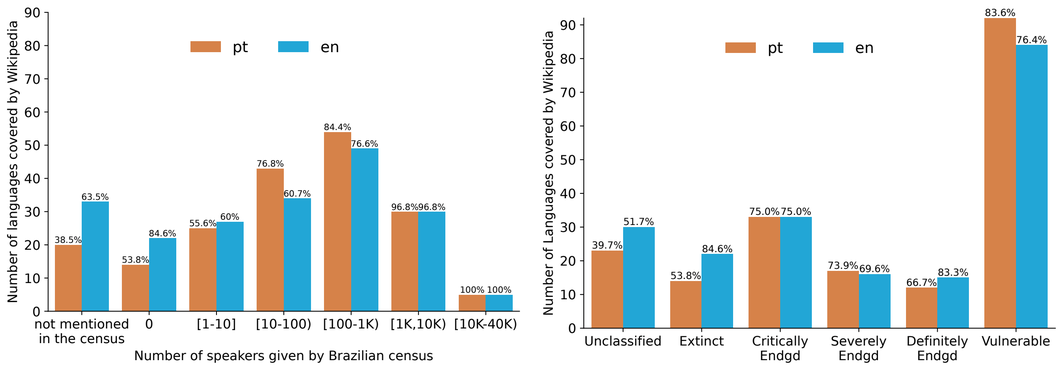

Figure 4: Wikipedia Indigenous languages dataset versus offline indicators. Percentages on bar represent covered articles.

The Editing Practices of the Portuguese and English Communities

The editing practices in the Portuguese and English Wikipedia communities exhibit significant differences.

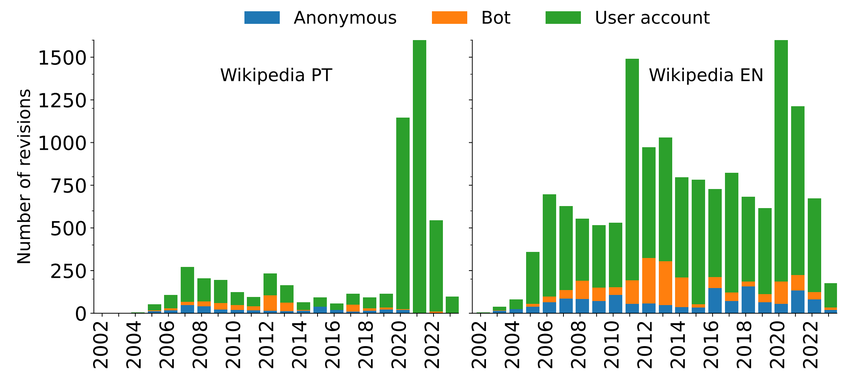

Figure 5: Revision activity 2002-2023.

- English Wikipedia shows higher editing activity compared to Portuguese.
- Minimal overlap exists between editors of Portuguese and English Wikipedia articles, with most editors showing a clear preference for one version.
- Temporal editing patterns vary, with Portuguese Wikipedia gaining traction post-2020.
- While both versions initially decline in activity post-creation, Portuguese articles show a higher dropout rate over time, potentially indicating abandonment, with only a small proportion remaining consistently active.

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

- Both versions have a small proportion of high-quality articles: 65% of Portuguese and 48% of English articles had higher ORES scores than assessments by projects and bots.
- Portuguese and English Wikipedia versions differ in article length distributions, suggesting content depth and detail discrepancies.
- About 41% of the top referenced domains in Wikipedia articles lead to unavailable sites, with Portuguese articles primarily linking to Portuguese-written sites, akin to English articles.
- English articles show stronger interconnections, while Portuguese articles emphasize language families and neighboring countries' languages.
- Portuguese articles often include more link and bibliography sections, while English articles emphasize technical language aspects, suggesting possible expert authorship in linguistics.

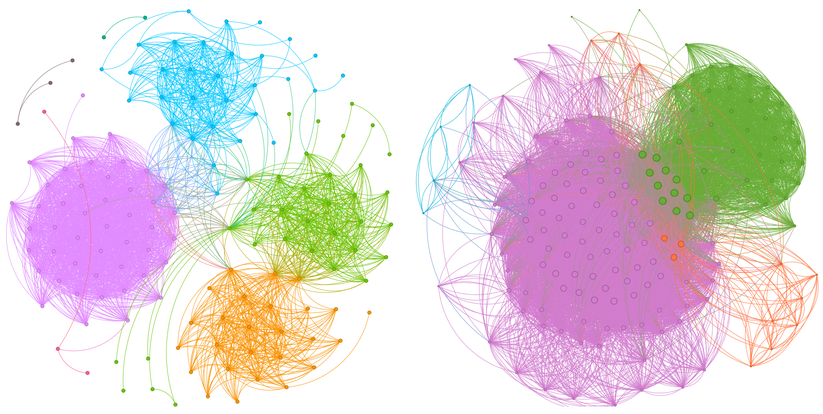

Figure 6: Wikipedia Indigenous language article network. Both graphs include the same Indigenous languages.

The Temporal Dynamics of Creation and Maintenance of the Articles

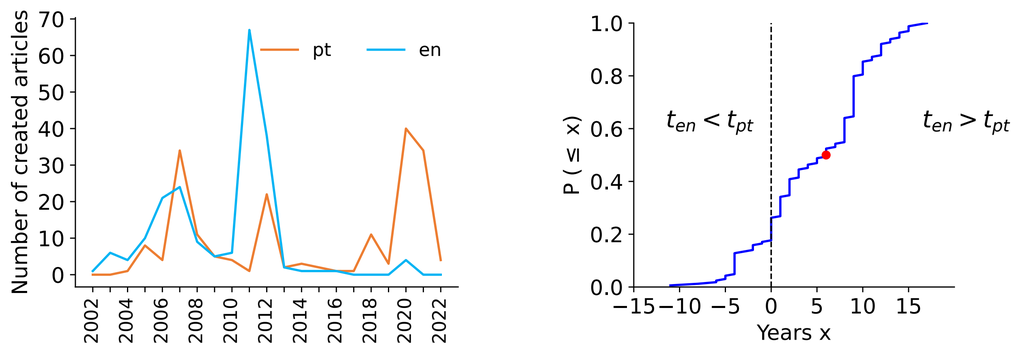

Figure 7: Article creation temporal dynamics.

- The dataset spans two decades, with recent articles created up to a year ago, and peaks in article creation do not align between English and Portuguese versions.
- Most articles were initially created in English with a median time span of six years.
- English dominance in the early stages is followed by increased Portuguese article creation.

Conclusion and Future Directions

- Indigenous languages on Wikipedia are underrepresented with only 68% coverage.
- English Wikipedia shows higher engagement and diversity compared to Portuguese, despite similar content quality.
- Suggestions include addressing reporting gaps, updating articles, and engaging Indigenous communities.
- Future research will explore other Wikipedia versions and engage further with Indigenous communities, aligning with the "Nothing for us without us" principle.

References

[1] UNESCO. Atlas of the world's languages in danger. https://shorturl.at/ZEcVF, 2010.
[2] IBGE. O Brasil Indígena: Estudos especiais. https://shorturl.at/bZbPt, 2010.
[3] Ethnologue. Languages spoken in Brazil. https://www.ethnologue.com/country/BR/, 2022.