Building the wiki-way for low-resource languages

Subhashish Panigrahi
Director, O Foundation
Bhubaneswar
Odisha, India
subhashish@theofdn.org

Sailesh Patnaik
Director, O Foundation
Bhubaneswar
Odisha, India
sailesh@theofdn.org

ABSTRACT

When it comes to internet governance, most indigenous, endangered and other low-resource and marginalized language speakers around the world face a significant challenge both in terms of amplifying their issues through participation and their languages getting benefitted in that process. As Whose Knowledge? underlines, a mere 7% of the 6,500 - 7,000 languages that are spoken around the world, are captured in published material. [1] In this Internet Governance Forum 2021 panel titled "Building the wiki-way for low-resource languages", the representatives engaged primarily around the strategies in the language digital activism and open knowledge platforms, and shared recommendations to grow and sustain the low-resource languages on the digital sphere.

Keywords

Low-resource languages; indigenous; endangered; Internet Governance Forum.

1. INTRODUCTION

To set the context on language digital activism, Eddie Avila, Director of Rising Voices shared how the scarcity of key resources, such as writing systems in the context of oral-only languages or consensus on existing writing systems, access to the internet, consensus on technical terms within a speaker community, and even political implications, continue to remain the major of the barriers behind language digital activism.

Amrit Sufi, native speaker of Angika (endangered language from India) and co-facilitator of the recent Rising Voices organized "Language Digital Activism Workshops for India" [2] series emphasized on the fact that the medium of instruction in her schools were Hindi and English directly impacting the eventual slowdown of her native language Angika to a final cessation. She identified the sense of elitism, associated with not speaking native languages at home environments which are often enforced by parents, to be a primary reason for the disappearance of many languages at homes.

Sardana Ivanova, a speaker of the Sakha language and a doctoral student in Computer Science at the University of Helsinki, demonstrated from her experience how collective volunteer efforts for creating content -- through Wikipedia -- eventually helps Natural Language Processing (NLP) researchers to build tools to better use languages on digital platforms.

Mahir Morshed, a Wikipedia/Wikidata contributor and a doctoral student researching articulatory features and prosodic unit discovery in speech processing, shared inputs to the call to action of this session based on the recent development around Wikidata's lexicography as these tools are making ways to for text generation across languages. [3]

2. KEY QUESTIONS

Some of the key questions that the panelists addressed around existing community strategies, processes and platforms included:

- How language digital activism is helping active engagement of stakeholders in the low-resource language domain?
- How other open and collaborative processes and platforms are helping low-resource languages?
- How is the creation of computational linguistics tools making a long-term shift in access to information in an equitable and decentralized manner, especially in the context of many low-resource languages?
- How Wikipedia and the Wikimedia projects in general, and particularly, the ongoing Wikidata initiatives, are helping low-resourced language speakers reclaim their space on the internet?

Additionally, the questions around specific movements and platforms included:

- How language digital activism is already and is aimed at moving the needle around furthering access to linguistic rights and access to knowledge?
- What is envisioned by many activists for the ongoing initiatives to make a long-term shift as a direct or indirect result of these initiatives?
- What are the learning and recommendations for different stakeholders working on low-resourced languages based on the work around creation of language tools?
- What are the building blocks and the low entry-level barriers in the Wikimedia world for native speaker communities and other stakeholders of low-resourced languages communities?

3. KEY TAKEAWAYS

1. Stakeholders must work collaboratively for supporting low-resource language communities with addressing issues around accessibility and with removing entry-level barriers of platforms.

2. Language technology developers and other stakeholders who are not native speakers must work closely with native speakers to implement the development of language technology based on the advice of the latter.

Some of the panelists directly addressed the questions around the aforementioned takeaways:

- Eddie Avila: We're creating a space for peer learning for language digital activism so that activists can expand their work through such long-term partnerships and share their work with the larger community.

- Amrit Sufi: Many low-resourced language digital activists are currently attempting at "normalizing" the use of their language as their languages are not in active use in public discourse.

Internet Governance Forum 2021, December 8, 2021, Online and Katowice, Poland.

DOI: http://dx.doi.org/10.17613/df9y-nz19

• Sardana Ivanova: Language technology developers who might not be native speakers must work closely with native speakers.

4. CALL TO ACTION

1. Creating spaces for peer learning exchange can be a very powerful tool for many low-resource languages to protect and grow use of languages, and stakeholders must emphasize on creation of such spaces.

2. Stakeholders who must support the creation of Open Educational Resources (OER) for new contributors/potential contributors who are speakers of low-resource languages to remove entry-level barriers to Open and collaborative platforms such as Wikipedia.

Some of the panel inputs/recommendations from the panelists that helped arrive at the aforementioned call to action include:

• Mahir Morshed: There have been many initiatives recently from the Wikimedia Foundation to improve accessibility on Wikipedia and Wikimedia projects. This has helped with the growth of many low-resource languages.

• Mahir Morshed: Creating translation of Wikidata descriptions is one of the easier ways for newbies to contribute in their low-resource language.

• Amrit Sufi: Oral culture documentation as audio and video helps grow visibility for many low-resource languages.

• Eddie Avila: Creating spaces for peer learning exchange can be a very powerful tool for many low-resource languages to protect and grow use of languages.

Avila also emphasized that whether or not language digital activism is moving the needle around furthering access to linguistic rights and access to knowledge is a hard thing to measure. While there are anecdotal evidences to support the impact, there are no clearly defined processes in most cases as each community contributes in their own volunteer spaces and are primarily driven by passion.

5. ACKNOWLEDGMENTS

Our thanks to the four panelists and participants of the panel along with many many collaborators in the open knowledge movement.

6. REFERENCES

[1] Vrana A, Sengupta A, Pozo C and Bouterse S (2020). Decolonizing the Internet's Languages – Summary Report. Whose Knowledge?, 20. (accessed 19 December 2021).

[2] Language Digital Activism Workshops for India · Rising Voices (2021). Rising Voices. Available at https://rising.globalvoices.org/language-digital-activism-workshops-for-india/ (accessed 19 December 2021).

[3] Morshed M (2021). Preparing languages for natural language generation using Wikidata lexicographical data. Septentrio Conference Series. (3):.