A Need Finding Study with Low-Resourced Language Content Creators

Hellina Hailu Nigatu
UC Berkeley
USA
hellina_nigatu@berkeley.edu

John Canny
UC Berkeley
USA
canny@berkeley.edu

Sarah Chasins
UC Berkeley
USA
schasins@berkeley.edu

ABSTRACT

Online knowledge repositories like Wikipedia offer a way for communities to share and preserve information about themselves and their ways of living. However, there is a huge gap in the volume and quality of content available for communities that speak high-resourced languages versus communities that speak low-resourced languages—including a majority of African communities.Usually, such online repositories are situated in Western ways of knowledge preservation and sharing, requiring low-resourced language communities to adapt to new forms of interaction and preservation.To understand the challenges faced by low-resourced language content creators on the popular knowledge repository Wikipedia, we collected Wikipedia forum discussions in low-resourced languages and conducted a thematic analysis. For the purpose of this work, we focused on three Ethiopian languages: Amharic, Tigrinya, and Afan Oromo. In this short paper, we report findings from ongoing research aimed at understanding the challenges faced by such content creators.Our analysis reveals several recurrent themes, including (1) how typing in non-Latin scripts on Latin keyboards is challenging and slow, and (2) how content creators' time is consumed by cleaning duplicate and low-quality articles.We hope our study will help inform designers' choices in making such platforms accessible to low-resourced language speakers.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **HCI theory, concepts and models**.

ACM Reference Format:
Hellina Hailu Nigatu, John Canny, and Sarah Chasins. 2023. A Need Finding Study with Low-Resourced Language Content Creators. In *4th African Human Computer Interaction Conference (AfriCHI 2023), November 27–December 01, 2023, East London, South Africa.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3628096.3628738

1 INTRODUCTION

Knowledge repositories are a means to preserve and share information about communities, cultures, and other expertise across the world. *Online* knowledge repositories allow for dissemination of knowledge across borders with a wide set of users. One such platform is Wikipedia, an open-source online knowledge repository available in 332 languages and with over 191 million articles [20]. According to its foundation's vision statement, Wikipedia is a community-rooted effort with the vision to create a world where "...every single person on the planet is given free access to the sum of all human knowledge." [1]. While Wikipedia is available in over 300 languages, not all languages are evenly represented.

In particular, the representation of African communities lags across three axis: quantity, quality, and relevance. Let us take Ethiopia, a country with over 120 million people [21], as an example. There are 6.6 million articles in English Wikipedia compared to 15,190 articles in Amharic, 1258 articles in Afan Oromo, and 257 articles in Tigrinya, three languages spoken in Ethiopia. Additionally, many articles in low-resourced languages are stubs (articles that are too short and incomplete), written in a different language other than the language of the Wikipedia, or have other quality issues such as misspellings (Section 4.1.2). Finally, articles in these languages lack contextual relevance; many articles written in the languages of the communities are about topics unrelated to the community, while articles about the community may not be accessible in their own language. For instance, the article on the front page of the Amharic Wikipedia is about the Big Mac burger while there is no McDonalds in Ethiopia. In contrast, an article about a famous Ethiopian village, Awra Amba, is available in five languages on Wikipedia, none of which are Ethiopian.

When we are focused on building knowledge repositories for a diverse set of users but are not sensitive to how different communities preserve their knowledge, we are implicitly confining the ways as well as their ability to interact with our systems. By making technologies accessible to low-resourced language speakers, we can empower communities to write about their own culture in their own languages. African decolonization scholars' have argued for the importance of African literature to be written in African languages [5]. Building technologies that are inclusive of all languages allow for the voices, indigenous knowledge, and ideologies of local communities to be represented avoiding erasure of history and culture as well as gross misrepresentation[2]. We echo the need for communities to create their own content in their own language and quote from prominent Ugandan law scholar Sylvia Tamel:

> Colonial intellectualism deliberately denigrated Indigenous oral traditions and wisdom as illegitimate methodologies and tools of storing records. Given that Western knowledge systems use the indicator of the written record to separate the human eras of "prehistory" and "history," it is no wonder that traditions that depend on oral wisdom are perceived as lacking history. –Sylvia Tamel [16]

AfriCHI 2023, November 27–December 01, 2023, East London, South Africa

ACM ISBN 979-8-4007-0887-9/23/11.
https://doi.org/10.1145/3628096.3628738

In this short paper, we report on ongoing research to understand the challenges faced by content creators in low-resourced languages. In particular, we collected data from Wikipedia discussion forums of three Ethiopian languages and used inductive thematic analysis [6] to answer the following research question: **What challenges do low-resourced language content creators face when interacting with Wikipedia?**

2 RELATED WORK

Previous HCI literature shows how technology aids in preservation and practice of Indigenous Knowledge (IK) both locally [11, 12, 15] and abroad [4, 17]. But technological advances do not always match IK preservation methods. IK preservation usually involves multi-modal interactions and differs from one community to another. Previous work [8] has criticized the text-only interaction mode of online knowledge repositories for excluding oral tradition based knowledge preservation systems. The shortcoming of current knowledge repository interaction paradigms is not limited to input modalities; previous work [8] criticizes the citation and editing features of Wikipedia, showing how Western-centred editing and participation rules inhibit participation on the platform.

Wikipedia policy requires that "the topic of an article must have already been the subject of publication in reliable, secondary, entirely independent sources" [3]. The policy further states that if one cannot find reliable third party sources, the topic should not have a separate article. While verifiable sources and notability are important, previous studies [13, 14, 19] show how the global research community ignores huge bodies of work by African researchers. Previous work [7, 14] also indicates African researchers are more likely to be cited if they are associated with Western universities. This trend of authorship about African communities coming predominantly from outside of the communities is not limited to research papers; most Wikipedia article contributions about African communities are produced outside the continent [8].

Previous work [9, 10, 18] provides evidence that, despite the hopes of democratization through internet connectivity, there is a geographical divide in information representation. One study [18] finds that only 1% of all Wikipedia articles are in African languages. Another study [9] argues that, aside from digital connectivity, issues such as access to source materials could constrain users from contributing in local languages. Although these works cover important barriers like connectivity and source material availability, they do not touch on whether online knowledge repositories' interaction paradigms work across communities.

3 METHODOLOGY

We collected forum data (discussions among Wikipedia authors) from the Talk Pages on Wikipedia in Amharic, Tigrinya, and Afan Oromo, which we summarize in Table 1. We collected all the forum data available, with no exclusions. We used inductive thematic analysis to analyze all forum posts. The first author did line-by-line open coding for each of the sentences in each post. From those codes, the first author synthesized themes which were discussed in frequent meetings among the authors. Throughout this paper, we will refer to Wikipedia forum participants as "posters." The study was approved by our institution's Institutional Review Board.

Language	Number of Topics	Number of Posters
Amharic	70	29
Tigrinya	10	14
Afan Oromo	6	7

Table 1: Data collected from the Wikipedia forums for the three languages in this study. Topics in the forum have discussion threads where posters deliberate over issues related to the topic.

4 FINDINGS

4.1 Challenges with Wikipedia's Interfaces and Language Supports

4.1.1 Fonts and Allowed Input Languages: Wikipedia interface makes contribution in non-Latin scripts harder. Posters discuss that the Wikipedia interface itself does not support writing in non-Latin (Ge'ez for Amharic and Tigrinya) scripts, especially after updates or new versions are released. Additionally, posters complained the "Wikipedia's Ge'ez font" looked "very distracting" and asked if there was "a thinner version." The font designer, who does not read Amharic, misunderstood the issue and responded that the font does not come with a bold variant. Because the designer themself did not read Amharic, it took several rounds of back and forth before they understood that the issue was that the font looked "bold" for all text (i.e the width of the strokes was too heavy).

4.1.2 Search and Spellcheck: Content creators have to spend time finding and removing duplicate and low-quality articles. Posters discuss how lack of standardized words and phrases lead to duplicate articles about the same concept. We observe threads of discussions on how to avoid repeatedly translating words, especially scientific terms and technological phrases needed for interacting with the Wikipedia interface. Further, Amharic posters raise the issue of homophones and their impact on users' search experiences. One poster gave a detailed explanation stating there are 182 ways of writing "Aste Hailesielase," the name of one of the Ethiopian kings. Some users call on Wikipedia to handle the issue of homophones and ask that their search experience accommodate this characteristics of the Amharic language. Other posters ask for normalization of words and a way to equate all homophones, a position met with opposition from other posters who believe this would lead to loss of the characteristics of the language. [1]

Afan Oromo posters have concerns over widespread misspelling in articles, with concerns that it "may render the Oromo section of Wikipedia unusable." Posters indicate this issue trickles down to their search experience: "you cannot search and find topics b/c they are misspelt when the topic is created." Users also confirm that misspelling has lead to topic duplication: "There are topics I created but later noticed it exist with different/wrong spelling."

We observed posters use number of articles to set goals and motivate adding more articles in a given language. While number

[1] Interestingly, homophone and phonetic differences also present a challenge in Amharic Wikipedia to the way that "Wikipedia" itself is translated. Our analysis revealed multiple treads with discussions on how to properly write "Wikipedia" in Amharic; including a phonetic breakdown of the word and adaptations along the way to make pronouncing it easier for Amharic speakers.

of articles is one indication of the growth, we found that it is not a good measure on its own. Amharic and Tigrinya posters complain of article stubs with little-to-no content. Hence, some posters go through the articles, finding low-quality articles and editing them. Tigrinya posters complain of articles being in English.

4.1.3 (Lack of) Translation Tools: Content creators require translation support in different stages of article writing. Multiple threads of discussion are centered around finding the translation of entity names from a given English, Amharic, or Tigrinya documents. The discussions showed back and forth between posters around what the translations were and which formats made the most sense to native speakers. Posters also discussed issues around how interface words are in English. In Tigrinya forum, there is a post where users provide translation for common interface words. Amharic posters also discussed using *translatewiki* and translation memory for commonly used interface words. Some posters also suggested translation as a way to mitigate the challenges of typing in non-Latin scripts (Section 4.1.1).

4.2 Challenges due to Low-Resourcedness

4.2.1 To the extent that low-resourced languages are connected to low-resourced content creators, there may be financial barriers. Posters discuss the funding source of Wikipedia and ask who has access to the funds. One poster asked for support in terms of mobile data top-up. Another user indicated they needed financial support to access online resources.

4.2.2 Content creators deliberate over how to keep their Wikipedia from being shut down. In the Amharic Wikipedia, we found a discussion thread about creating a Wikipedia page for Ge'ez under the Amharic Wikipedia. This discussion involved a back and forth about creating an independent Wiki for Ge'ez and ended with how the Wikipedia for an Ethiopian language, Afar, was under threat of closure by Wikipedia since it did not have enough contributors.

4.2.3 Lack of media and resources prevent content creators from creating articles about their communities. Posters discussed ways of collecting supporting media such as pictures to support their articles. Some posters requested others who were currently in Ethiopia to take pictures for them while others posted about potentially traveling to Ethiopia to collect media for their articles.

5 DISCUSSION AND CONCLUSION

From our preliminary analysis of Wikipedia Talk Pages in Tigrinya, Amharic, and Afan Oromo, we found that Wikipedia content creators in low-resourced languages face barriers both internally—based on Wikipedia interface design decisions—and externally—based on material conditions. Both categories of barrier hinder local language content creation and contribute to the digital divide. Future work can test and extend our findings via additional data sources such as interviews with content creators and leverage co-design and participatory design methods to prototype alternative platforms, platform modifications, or other interventions. Our results support previous findings (Section 2) that issues like access to source material prevent users from contributing in their own languages and extend this finding by highlighting technological barriers like the Wikipedia web interface's low support for non-Latin scripts. Our results also suggest room for the development and improvement of language technologies. For example, the lack of language-specific technologies like spell-check have negative effects that trickle down to search experience. Our work offers jumping off points for research in making language technologies and interaction with language technologies better for low-resourced language speakers. Overall, we hope our work will offer insights for efforts to build knowledge repositories that are inclusive of low-resourced language speaking communities.

REFERENCES

[1] 2004. Wikipedia Founder Jimmy Wales Responds - Slashdot. https://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds

[2] 2018. Whiteness, the Western Gaze and Africa. In *Imagining Africa: Whiteness and the Western Gaze*, Clive Gabay (Ed.). Cambridge University Press, Cambridge, 1–48. https://doi.org/10.1017/9781108652582.001

[3] 2023. Help:Your first article. https://en.wikipedia.org/w/index.php?title=Help:Your_first_article&oldid=1147040537 Page Version ID: 1147040537.

[4] Kagonya Awori, Frank Vetere, and Wally Smith. 2015. Transnationalism, Indigenous Knowledge and Technology: Insights from the Kenyan Diaspora. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea, 3759–3768. https://doi.org/10.1145/2702123.2702488

[5] H. S. Bhola. 1987. Ngugi wa Thiong'o. Decolonising the Mind: The Politics of Language in African Literature. London: James Currey; Nairobi: Heinemann Kenya; Portsmouth, N. H.: Heinemann; Harare: Zimbabwe Publishing House, 1986. 114 pp. $10.00. Paper. *African Studies Review* 30, 2 (June 1987), 102–103. https://doi.org/10.2307/524049 Publisher: Cambridge University Press.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[7] Hugo Confraria, Jaco Blanckenberg, and Charl Swart. 2018. The characteristics of highly cited researchers in Africa. *Research Evaluation* 27, 3 (July 2018), 222–237. https://doi.org/10.1093/reseval/rvy017

[8] Peter Gallert, Heike Winschiers-Theophilus, Gereon K. Kapuire, Colin Stanley, Daniel G. Cabrero, and Bobby Shabangu. 2016. Indigenous Knowledge for Wikipedia: A Case Study with an OvaHerero Community in Eastern Namibia. In *Proceedings of the First African Conference on Human Computer Interaction*. ACM, Nairobi Kenya, 155–159. https://doi.org/10.1145/2998581.2998600

[9] Mark Graham, Bernie Hogan, Ralph K. Straumann, and Ahmed Medhat. 2014. Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty. *Annals of the Association of American Geographers* 104, 4 (2014), 746–764. https://www.jstor.org/stable/24537592 Publisher: [Association of American Geographers, Taylor & Francis, Ltd.].

[10] Mark Graham and Matthew Zook. 2012. Augmented Realities and Uneven Geographies: Exploring the Geolinguistic Contours of the Web. https://doi.org/10.1068/a44674

[11] Lindah Kotut and D. Scott McCrickard. 2022. Winds of Change: Seeking, Preserving, and Retelling Indigenous Knowledge Through Self-Organized Online Communities. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3502094

[12] Lindah Kotut and Scott D McCrickard. 2021. Trail as Heritage: Safeguarding Location-Specific and Transient Indigenous Knowledge. In *3rd African Human-Computer Interaction Conference*. ACM, Maputo Mozambique, 94–102. https://doi.org/10.1145/3448696.3448702

[13] Richard Maclure. 2006. No Longer Overlooked and Undervalued? The Evolving Dynamics of Endogenous Educational Research in Sub-Saharan Africa - ProQuest. https://www.proquest.com/openview/49d5601ec01cda35c58a012f57ad5064/1?cbl=41677&pq-origsite=gscholar&parentSessionId=tUv%2FywxFchFk9vVd3Pys0hr%2B5lq1Q%2FTd10FcGFH6ezY%3D

[14] Rafael Mitchell, Pauline Rose, and Samuel Asare. 2020. Education Research in Sub-Saharan Africa: Quality, Visibility, and Agendas. *Comparative Education Review* 64, 3 (Aug. 2020), 363–383. https://doi.org/10.1086/709428 Publisher: The University of Chicago Press.

[15] Siang-Ting Siew, Alvin W. Yeo, and Tariq Zaman. 2013. Participatory Action Research in Software Development: Indigenous Knowledge Management Systems Case Study. In *Human-Computer Interaction. Human-Centred Design Approaches, Methods, Tools, and Environments*, Masaaki Kurosu (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 470–479.

[16] S. Tamale. 2020. *Decolonization and Afro-feminism.* Daraja Press. https://books.google.com/books?id=OxnAzQEACAAJ
[17] Jennyfer Lawrence Taylor, Alessandro Soro, Paul Roe, Anita Lee Hong, and Margot Brereton. 2018. From Preserving to Performing Culture in the Digital Era. In *Digitisation of Culture: Namibian and International Perspectives*, Dharm Singh Jat, Jürgen Sieck, Hippolyte N'Sung-Nza Muyingi, Heike Winschiers-Theophilus, Anicia Peters, and Shawulu Nggada (Eds.). Springer Singapore, Singapore, 7–28. https://doi.org/10.1007/978-981-10-7697-8_2
[18] Bert van Pinxteren. 2017. African Languages in Wikipedia – A Glass Half Full or Half Empty? https://doi.org/10.2139/ssrn.2939146
[19] Maina Waruru. 2022. Renowned journal rejects papers that exclude African researchers. https://www.universityworldnews.com/post.php?story=20220603115640789
[20] Wikipedia. 2023. List of Wikipedias - Meta. https://meta.wikimedia.org/wiki/List_of_Wikipedias
[21] Worldmeter. 2023. Ethiopia Population (2023) - Worldometer. https://www.worldometers.info/world-population/ethiopia-population/