BILINGUAL PUBLISHING GROUP

**Forum for Linguistic Studies**

https://journals.bilpubgroup.com/index.php/fls

**COMMUNICATION**

# A Report for the SADiLaR-Wikipedia-PanSALB Project for South African Languages

***Muzi Matfunjwa*** [*] **, *Nomsa Skosana* , *Lebogang Boemo***

*South African Centre for Digital Language, North-West University, Potchefstroom 2531, South Africa*

## ABSTRACT

South African languages are underrepresented in online encyclopaedias, which limits access to free and open information in these languages. This report provides an overview of the SADiLaR-Wikipedia-PanSALB (SWiP) project and how it was utilised to create content in Wikipedia for South African Languages. The creation of Wikipedia content was facilitated through training participants across 11 public universities in South Africa wherein they were taught how to contribute to Wikipedia. The participants mainly consisted of students, lecturers, language practitioners, and community members. They learned how to activate Wikipedia translation tools on their Wikipedia accounts, search for their preferred language and translate selected articles. These participants were also trained to edit the translated and published articles, create new ones in their respective languages, and structure them according to Wikipedia's guidelines for article creation. The training also included explanations on how to add references and images, as well as how to link articles to other Wikipedia pages. The SWiP project resulted in the creation of 737 articles, 160 images and 1960 references. The new corpora developed on Wikipedia are currently used by human language technology developers to create and improve tools for South African languages. Therefore, the SWiP project has significantly enhanced the visibility of South African languages on Wikipedia and improved access to information in these languages.

***Keywords:*** SWiP; Corpora; Wikipedia; South African Languages

*CORRESPONDING AUTHOR:
Muzi Matfunjwa, South African Centre for Digital Language, North-West University, Potchefstroom 2531, South Africa; Email: muzi.matfunjwa@nwu.ac.za

ARTICLE INFO
Received: 12 March 2025 | Revised: 20 April 2025 | Accepted: 24 April 2025 | Published Online: 7 May 2025
DOI: https://doi.org/10.30564/fls.v7i5.9068

CITATION
Matfunjwa, M., Skosana, N., Boemo, L., 2025. A Report for the SADiLaR-Wikipedia-PanSALB Project for South African Languages. Forum for Linguistic Studies. 7(5): 598–603. DOI: https://doi.org/10.30564/fls.v7i5.9068

# 1. Introduction

In this digital era, how we access and interact with online information is rapidly evolving. In South Africa, online information is dominated by English, leaving a significant underrepresentation of indigenous languages. Although the South African Constitution recognises twelve official languages, there is a noticeable disparity between these languages, especially on Wikipedia. To alleviate the gap between the South African languages on Wikipedia, the SADiLaR-Wikipedia-PanSALB (SWiP) project was initiated. This project focused on creating digital corpora for South African languages, underlining those that are underrepresented on Wikipedia. The SWiP project is a collaboration between the South African Centre for Digital Language Resources (SADiLaR), Wikipedia, and the Pan South African Language Board (PanSALB). The project aimed to preserve and advance South African languages in Wikipedia [1]. Therefore, this report provides an overview of the SWiP project and how it was utilised to create content in Wikipedia. Human language technology (HLT) developers are presently using these newly created Wikipedia corpora to develop HLT tools for South African languages, such as machine translation systems. Some of these corpora are also used to evaluate and improve current HLT and natural language processing (NLP) systems that were underperforming due to insufficient corpora utilised in their development.

# 2. Organisations Involved in the SWiP Project

SADiLaR is a national centre supported by the Department of Science, Technology and Innovation. This centre has an enabling function, with a focus on official languages of South Africa, supporting research and development in the domains of language technologies, language-related studies, and social sciences [2]. Wikipedia is a free online encyclopaedia created and edited by volunteers worldwide. It is hosted by the Wikipedia Foundation, and it facilitates knowledge creation and dissemination, making it an invaluable resource for language documentation and research [3]. PanSALB is an organisation established to promote multilingualism, develop the 12 official languages, and protect language rights in South Africa. This Board was established in accordance with Act 59 of 1995 by the Parliament of South Africa. As a statutory body, PanSALB plays a pivotal role in language policy formulation, language promotion, and the development of language resources to ensure equitable access to information and services for all language communities in South Africa [4].

# 3. Materials and Methods

This section describes how the SWiP project was conducted, leading to the creation of corpora in Wikipedia. This was a joint project and language specialists, community contributors, lecturers, students, and language enthusiasts participated voluntarily. It adopted crowdsourcing as a method of soliciting content for Wikipedia in South African languages from these people. The SWiP project followed systematic stages as outlined:

## 3.1. Launch

The SWiP project was launched at the University of South Africa (UNISA) on the 22nd of September 2023 in which SADiLaR, Wikipedia and PanSALB participated as organisations that run this project. External stakeholders such as His Majesty King Makhosonke Mabena II, South African Broadcasting Corporation and isiNdebele language community were invited and partook in this launch. This event aimed at raising awareness about the project in all South African universities in which workshops were to be conducted. It was during this launch that it was highlighted that isiNdebele needed more attention as it was the only official South African language that was not yet available on Wikipedia but still in Wikimedia's incubation stage. Wikimedia incubation is a platform for developing, testing, and evaluating language projects to see if they are appropriate for hosting by the Wikimedia Foundation [5]. An invitation was made to all South African language speakers to contribute in great measure to Wikipedia.

## 3.2. Workshops and Training

The second step was conducting two-day workshops at 10 South African universities wherein lecturers, students, and language enthusiasts were trained on how to contribute to Wikipedia. The universities were Walter Sisulu University, University of Mpumalanga, Cape Peninsula University of Technology, Vaal University of Technology, UNISA, Tshwane University of Technology, Central University of Technology, Sol Plaatje University, University of Venda and University of Zululand. These universities were selected from each province of South Africa according to their status, that is, historically disadvantaged institutions except for UNISA.

The workshops were divided into two parts: theory and practice. In the theory session, PanSALB promoted the use of standardised language practices such as orthography and spelling rules for each language, helping to correct common linguistic errors and prevent their inclusion in Wikipedia. Meanwhile, Wikipedia trainers described five pillars of Wikipedia which are best practices of contributing to the platform [6]. They also discussed the history of

Wikipedia, as well as the support and projects offered by the Wikimedia Foundation.

In the practical session, participants began by creating usernames to log into their Wikipedia pages which enabled them to contribute on this platform. The usernames were subsequently added to the SWiP project dashboard to keep track of participants' activities and statistics on the entire project [7]. After they logged into their Wikipedia accounts, each participant was guided on how to activate the Wikipedia translation tool, which was essential for contributors working on multiple languages. This was followed by finding different languages on Wikipedia. Participants then learned to use the visual editor which is a user-friendly tool that enhances the appearance of a Wikipedia article when editing. The session included exercises where participants translated content from English into various targeted languages. They also had to review and edit the translated content to correct grammar and spelling errors before it could be published on Wikipedia, as required by the translation tool. This approach integrated the use of the translation tool and visual editor which reinforced the skills that are needed for contributions on Wikipedia. The sessions continued with a guide on how to create new articles and providing exercises; these practical exercises help to increase content in the different languages.

Moreover, participants were taught how to link articles and add references within Wikipedia. The addition of references to articles was emphasised to increase the articles' credibility. They also learned to categorise articles to increase content organisation on Wikipedia. The workshops concluded by adding pictures to Wikipedia's Commons project, which is important for the visual appeal of articles. Wikipedia's Commons is where these pictures are added, and these can be used across various Wikipedia articles.

Overall, these workshops imparted valuable skills to participants that they can use at a time convenient to them. The skills also help to increase content in the different South African languages and ensure that participants' contributions are impactful and meaningful. To facilitate the continuation of contributions beyond these workshops, a SWiP resource page was developed [8]. This page serves as a guide for participants who may need assistance while working independently as it contains video recordings of the content provided in the workshops.

### 3.3. Content Creation Initiative

A content creation initiative that aimed at motivating potential contributors to write and edit articles as well as add pictures to Wikipedia was rolled out. This was to increase the number of contributors on Wikipedia in the South African languages, expand the corpus and improve the quality of language resources available online. This initiative was carried out through a SWiP writing competition, which was open to everyone including those who had participated in the SWiP training workshops. The competition ran from 14 July–19 August 2024. This writing competition was advertised on various social media platforms, the SADiLaR website, radio as well as television stations. Winners were awarded incentives at the end of the competition.

## 4. Results and Discussions

The results of the SWiP project are presented in **Tables 1**, **2** and **3** and can be accessed online.

**Table 1.** SWiP project dashboard.

| Items | Quantity |
|---|---|
| Number of Editors | 318 |
| Number of articles created | 737 |
| Number of articles edited | 1,220 |
| References added | 1,960 |
| Number of pictures uploaded | 160 |

**Table 2.** SWiP writing competition dashboard.

| Items | Quantity |
|---|---|
| Number of Editors | 56 |
| Number of articles created | 145 |
| Number of articles edited | 223 |
| References added | 619 |
| Common uploads (Number of pictures uploaded) | 57 |

**Table 3.** Activities undertaken in each language on the SWiP dashboard as of 4 March 2025.

| Language | Number of Edits | Number of Articles edited | Number of articles created |
|---|---|---|---|
| English | 515 | 228 | 0 |
| Afrikaans | 87 | 40 | 17 |
| isiZulu | 927 | 381 | 239 |
| isiXhosa | 98 | 59 | 45 |
| Xitsonga | 103 | 47 | 36 |
| Siswati | 270 | 148 | 126 |
| Setswana | 509 | 113 | 101 |
| Sesotho sa Leboa | 80 | 45 | 39 |
| Sesotho | 221 | 97 | 86 |
| Tshivenda | 153 | 63 | 48 |
| IsiNdebele | | | 133 |

In **Table 1**, the work done on the SWiP project, including the SWiP writing competition is summarised. These figures were taken from the SWiP project dashboard [7], which constantly changes due to ongoing editor activities on Wikipedia. The project has recruited 318 editors who are actively participating with their contributions; this was seen in the number of new articles that were created, 737, and the editing of existing articles, 1,220, which demonstrates an expansion in content development and improvement of existing articles. One thousand nine hundred sixty references were added, and this enhanced the credibility and reliability of information on Wikipedia. One hundred sixty pictures were uploaded, and these uploads helped boost the visual appeal of the articles on Wikipedia.

In **Table 2**, only work done on the SWiP writing competition is presented. These figures were taken from the SWiP writing competition dashboard which started on 14 July 2024 and closed on 19 August 2024 [9]. This table indicates that the writing competition also played an important role in increasing the articles, pictures and editors for Wikipedia in the SWiP project.

In **Table 3**, the activities on Wikipedia in the South African official languages are presented. The data shows that isiZulu has 927 edits compared to other languages with 381 articles edited and 239 new articles created. It is important to note that isiNdebele has 133 articles. Before the SWiP project began, isiNdebele was the only spoken South African official language missing on Wikipedia. It was through this project that this language moved from the incubator stage to the main Wikipedia page.

The results of the SWiP project reveal that there have been strides in increasing content across the languages. However, the data also shows disparities in the number of articles created, especially in indigenous languages. For example, in the SWiP dashboard, isiZulu is leading compared to Sesotho sa Leboa, which has the least number of created articles. This means that Sesotho sa Leboa, among other languages, still needs more attention to achieve better representation on Wikipedia.

# 5. Challenges Encountered During the SWiP Project

The SWiP project encountered several challenges that impacted its effectiveness. Most challenges relate to the workshops conducted during this project. These included logistical issues and technical limitations which delayed training processes and affected the overall experience for participants and trainers. For instance, there was a water crisis that forced the shortening of some workshops to a few hours instead of the full day as originally planned. This not only disrupted the workshop schedule but also limited the content that could be covered. This compromised the depth and breadth of the workshop, impacting the participants' learning experience. Technical challenges included participants arriving without pre-registered Wikipedia usernames. The creation of usernames on-site proved difficult as Wikipedia only allows six usernames to be created from a single IP address. This challenge led to frustration among participants who were unable to follow the workshop without having usernames and passwords to log into their Wikipedia accounts. The inability to create Wikipedia usernames before the commencement of the workshop not only delayed the workshop process but also caused significant dissatisfaction among attendees. Moreover, participants were blocked in some workshops from accessing the Wikipedia platform in their language, making it impossible for them to contribute and fully participate in the workshop.

Load shedding and power outages were also disruptive during the workshops which not only affected internet connectivity but also delayed the workshop flow as hands-on activities were supposed to be done on the Wikipedia website. These mishaps created a stressful environment for both participants and trainers. To alleviate these problems, portable generators were utilised to provide laptops and modems for internet connectivity with power.

The challenges encountered during the SWiP workshops emphasised the need for improved planning and contingency measures to ensure smooth execution and achievement of desired outcomes. By reflecting on these

experiences and implementing effective solutions, future workshops can be prepared better to meet their objectives. Addressing these issues is crucial for ensuring that upcoming workshops are conducted more systematically and efficiently.

# 6. Lessons Learned from the SWiP Project

The SWiP project has provided valuable insight into the preservation and promotion of South African languages on Wikipedia. It is important to reflect on the lessons learned from the project to ensure that similar projects are successfully implemented. The lesson learned from the SWiP project is the importance of community engagement; involving language communities in the project developed a sense of ownership and pride among participants. The sense of ownership provided by this project was crucial in encouraging active contributions to Wikipedia, which helped to expand the presence of South African official languages online. The project also highlighted the value of collaboration between different institutions and universities. The partnership between SADiLaR, Wikipedia and PanSALB shows how collaborative partnerships can amplify the impact of language preservation projects.

# 7. Conclusion

The SWiP project was set out to enhance the presence of South African languages on Wikipedia. There were 318 trained editors through this project, and they contributed collaboratively to the creation of 737 new articles, 1,220 edited articles, 1,960 added references and 160 uploaded pictures. This shows that the SWiP project was successful in creating corpora for South African official languages, which contained content from a variety of fields on Wikipedia. This project does not only improve the representation of South African languages on Wikipedia, but it enables the development of NLP and HLT tools for these languages by providing developers with the corpora needed to develop and train these tools. The project has also equipped participants with skills to contribute continuously and independently to Wikipedia which in turn enhances their computational skills. It has elevated the participants from being passive Wikipedia content consumers to useful contributors who patriotically contribute significant information in their language from their viewpoint and alleviate misrepresentation of their social, economic and cultural life. The involvement of these participants has made the project sustainable as people continue to voluntarily add content on Wikipedia and pass these skills to other language users.

The project has also demonstrated the positive effects of collaboration between SADiLaR, Wikipedia and PanSALB. PanSALB's engagement ensured that no South African official language was marginalised on Wikipedia and standardised spelling and orthography rules for these languages were adhered to. SADiLaR and Wikipedia provided the computational skills needed for the success of the project. This working relationship proved that if resources are directed towards completing a project and organisations are willing to learn from each other's abilities, it benefits not only them but everyone in the community. This collaborative approach involving the stakeholders has laid a strong foundation for future projects aimed at enriching the digital presence of the languages. Therefore, the SWiP project has bridged the existing gap in the digital corpora for South African official languages on Wikipedia and safeguarded their presence in the digital era for future generations.

## Author Contributions

Conceptualization, M.M., N.S. and L.B.; methodology, M.M., N.S. and L.B.; software, N.S. and L. B.; validation, M.M., N.S.,and L.B.; formal analysis, M.M., N.S. and L.B.; investigation, M. M., N.S. and L.B.; resources, N.S. and L.B.; data curation, N.S. and L.B.; writing—original draft preparation, M. M., N.S. and L.B.; writing—review and editing, M.M., N.S. and L.B.; visualization, M.M., N.S. and L.B.; supervision, M.M., N.S. and L.B.; project administration, N.S. and L.B.; funding acquisition, M.M.. All authors have read and agreed to the published version of the manuscript.

## Funding

This work was supported by the South African Centre for Digital Language Resources, a research infrastructure established by the Department of Science, Technology and Innovation of the South African government as part of the South African Research Infrastructure Roadmap.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

The data supporting reported results can be found at https://outreachdashboard.wmflabs.org/courses/SADiLaR,_Wikipedia,_PanSALB/SWiP_Workshops and https://outreachdashboard.wmflabs.org/courses/SADiLaR,_Wikipedia,_PanSALB/SWiP_Writing_Competition_(2024)/.

## Acknowledgments

The authors would like to acknowledge all stakeholders that were involved in the SWiP project.

## Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

[1] SWiP Project, n.d. SWiP Project. Available from: https://sadilar.org/en/swip/ (cited 12 November 2024).
[2] SADiLaR , n.d. SADiLaR. Available from: https://sadilar.org/en/ (cited 20 August 2024).
[3] Wikipedia Foundation, n.d. Wikipedia. Available from: https://www.wikipedia.org/ (cited 6 October 2024).
[4] PanSALB, n.d. PanSALB. Available from: https://www.pansalb.org/ (cited 12 September 2024).
[5] McDonough, D.J., 2017. Expanding the sum of all human knowledge: Wikipedia, translation and linguistic justice. The Translator. 23(2), 143–157. DOI: https://doi.org/10.1080/13556509.2017.1321519
[6] Five pillars of Wikipedia, n.d. Five pillars of Wikipedia. Available from: https://en.wikipedia.org/wiki/Wikipedia:Five_pillars (cited 15 December 2024).
[7] SWiP project dashboard, n.d. SWiP project dashboard. Available from: https://outreachdashboard.wmflabs.org/courses/SADiLaR,_Wikipedia,_PanSALB/SWiP_Workshops (cited 4 February 2025).
[8] SWiP Resource page, n.d. SWiP Resource page. Available from: https://meta.wikimedia.org/wiki/SWiP_Resource_Page (cited 4 February 2025).
[9] SWiP writing competition dashboard, n.d. SWiP writing competition dashboard. Available from: https://outreachdashboard.wmflabs.org/courses/SADiLaR,_Wikipedia,_PanSALB/SWiP_Writing_Competition_(2024)/ (cited 5 September 2024).