The Hundzula Retreat-Based Infrastructure Model for African Natural Language Processing

Johannes Sibeko[1,*] , Seani Rananga[2] , Neo Putini[3] , Hlaudi Daniel Masethe[4]
[1] Nelson Mandela University, [2] University of Pretoria, [3]
University of Kwa-Zulu Natal, [4] Tshwane University of Technology
Port Elizabeth, Pretoria, Durban, South Africa
johanness@mandela.ac.za, seani.rananga@up.ac.za,
nokwandaputini@gmail.com, masethehd@tut.ac.za

Abstract

The development of Natural Language Processing (NLP) resources for African indigenous languages remains constrained by limited data availability, fragmented expertise, and a lack of sustainable, locally grounded infrastructures for enabling language research. While much existing work focuses on producing discrete resources such as corpora or lexicons, less attention has been paid to the social, institutional, and methodological conditions that enable such resources to be created, maintained, and sustained. This paper presents the Hundzula Retreat for NLP and Linguistics as a retreat-based resource infrastructure model that addresses these constraints. We conceptualise Hundzula not as a once-off event, but as a structured, upstream research infrastructure that facilitates human capacity development, interdisciplinary collaboration between linguistics and NLP, ethical data practices, and the early-stage incubation of language resources for African indigenous languages. Drawing on evidence from multiple iterations of the retreat, we describe the design principles, workflows, and governance mechanisms that support resource development, including training pipelines, human-in-the-loop methodologies, and collaborative project formation. Rather than focusing on already formalised outputs, the paper foregrounds the infrastructural conditions that make such outputs possible within under-resourced contexts. In doing so, the paper shifts attention from outputs to the enabling ecosystems required for their production. We argue that retreat-based infrastructures constitute an essential but under-recognised category of language resources and demonstrate how the Hundzula model can be adapted and replicated in other low-resourced language contexts. The paper contributes a transferable framework for sustainable NLP resource development grounded in African linguistic realities.

Keywords: African NLP, Linguistic infrastructure, Human-in-the-loop methods, Community-based research

1. Introduction

Natural Language Processing (NLP) research has made significant advances Khurana et al., 2023; Deekshith, 2024; however, these advances have not been evenly distributed across the world's languages (Adebara, 2024). In fact, African indigenous languages remain severely under-resourced, both in terms of linguistic data and the human capacity required to develop, curate, and maintain such resources (Daniel, 2020; Adebara, 2024). Unfortunately, while recent initiatives have produced corpora, lexicons, and benchmark datasets for selected under-resourced languages (Setaka and Trollip, 2022), the broader ecosystem necessary for sustainable resource development for these languages remains fragile (Ògúnrẹ́mí et al., 2023). This includes limited access to trained researchers, weak interdisciplinary collaboration between linguistics and computer science, and insufficient attention to ethical and contextual considerations in the creation of language data using traditional and Artificial Intelligence-based approaches, including machine learning.

Within this context, we propose that the notion of a "resource" in African NLP requires careful consideration. Typically, understandings of resources in language research tend to privilege tangible artefacts such as annotated spoken and written corpora or computational tools like part of speech taggers (Setaka and Trollip, 2022; Chesire and Kipkebut, 2024; Muhammad et al., 2025). While these tangible language resources are also essential, they are often produced in isolation, without adequate investment in the infrastructures that enable their continued growth, reuse, and contextual relevance. Thus, we posit that for many African languages, the primary bottleneck is not the absence of technical methods, but the lack of sustained, locally grounded mechanisms that bring together linguistic expertise, computational skills, and community knowledge.

This paper argues that structured, community-based research infrastructures should be recognised as a critical category of language resources in their own right (Mayernik et al., 2017). To demonstrate the importance of these infrastructures, we present the *Hundzula*[1] *Retreat for Natural Lan-*

[1]The term *Hundzula* is derived from Xitsonga and is used in this context to mean transformation. The retreat, which has been held annually in February since 2022, is

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 121–127
12 May 2026.

guage Processing and Linguistics as a case study. In this way. our paper contributes to RAIL by foregrounding the upstream infrastructural conditions that enable the development of language resources and systems presented at venues such as this workshop. Beyond raising awareness of the initiative, the model offers guidance on decisions regarding resource allocation, mentoring structures, and the sequencing of learning and research activities, thereby bridging conceptual understanding with actionable practice.

Hundzula is a recurring[2], intensive research retreat designed to support collaboration between linguists, NLP researchers, industry language practitioners, and students working on Southern African indigenous languages. That is, rather than focusing solely on the production of finished datasets, the retreat foregrounds capacity building, shared methodological development, and the initiation of resource pipelines that extend beyond the retreat itself.

The Hundzula retreat-based resource infrastructures model is motivated by three interrelated challenges in African NLP. First, there is a persistent skills gap, particularly at the intersection of linguistics and NLP, which impacts the scale and quality of resource development. Second, existing resources are often developed through different institutions and in isolation (Siminyu et al., 2023; Nahar et al., 2021), leading to duplication of effort and limited sustainability. Third, ethical and cultural considerations specific to African language contexts are frequently treated as secondary concerns, despite their centrality to responsible data practices (Okorie and Omino, 2025; Okorie, 2023).

In this paper, we adopt a design-science case study approach applied selectively, through which the Hundzula initiative operationalises a human-in-the-loop approach to resource development. This combines formal training sessions, collaborative project incubation, and mentored research activities. Participants work on language-specific projects that include corpus design, annotation guideline development, tool evaluation, and exploratory NLP experiments. As indicated by Okorie (2025), these activities are embedded within a governance framework that emphasises credit attribution, collaborative ownership, and sensitivity to linguistic and cultural contexts.

The contribution of this paper is twofold. First, it documents the Hundzula Retreat as an example of an infrastructural resource that enables NLP development for African indigenous languages. Second, it abstracts from this case to propose a transferable model for retreat-based resource infrastructures, outlining design principles, outputs, and evaluative criteria relevant to the broader NLP community. By reframing infrastructure and capacity-building mechanisms as resources, this paper seeks to broaden how resource development is conceptualised and evaluated in African NLP research.

2. Linguistic complexity and resource design

The discussion in this article focuses on the fifth edition of the Hundzula Retreat for NLP and Linguistics[3]. The four-day programme illustrates how different stages of the resource lifecycle—conceptualisation, creation, annotation, modelling, and application—can be integrated within a single infrastructural intervention.

The fifth edition of the retreat deliberately began by foregrounding linguistic complexity, particularly phenomena that pose challenges for computational modelling in African languages. The opening keynote on polysemy in cross-border languages situates NLP resource development within the sociolinguistic realities of language contact, mobility, and variation. For many African languages, lexical meaning is shaped by regional usage, multilingual repertoires, and borrowing, complicating assumptions of stable word–meaning mappings. By centering this discussion early in the programme, it shaped Hundzula's position on linguistic analysis as foundational to resource design, rather than as a post hoc interpretive layer. This emphasis was also reflected in several lightning talks that addressed language varieties and under-documented speech communities. For instance, contributions focusing on Sepitori and everyday spoken Khelobedu highlighted the importance of recognising non-standardised and contact varieties as legitimate targets for resource development. These projects illustrate how the retreat supports the initial stages of corpus construction, including decisions about data selection, representativeness, and orthographic conventions. In doing so, Hundzula enables researchers to move beyond idealised or standard language models and engage with the forms of language that are used by real speakers

therefore explicitly framed around the transformation of research and applications in NLP and Linguistics.

[2]The first and second editions of the retreat, held respectively in 2022 and 2023 were hosted by the University of Pretoria under the leadership of Professor Vukosi Marivate, who initiated the retreat. An inter-institutional organising approach was adopted in 2024 when Dr Johannes Sibeko and Ms Andiswa Bukula hosted the retreat at Nelson Mandela University. Professor Mpho Primus then hosted the event at the University of Johannesburg in the year 2025. The recent edition was hosted by Ms Seani Rananga and Dr Keaobaka Seshoka at the North-West University in 2026.

[3]See https://sites.google.com/view/hundzula-retreat/home for full schedule

and users of the languages. Thus, the retreat is a great resource for NLP and linguistics, especially in the context of (Southern) Africa.

3. Corpus building and lexical resources as collaborative processes

A recurring theme across Day 1 of our fifth Hundzula Retreat was the development of corpora and lexical resources through collaborative and iterative processes. Presentations on spoken language corpora, living dictionaries, and regionally grounded lexical collections demonstrated how the retreat functions as a space for refining methodological choices and aligning them with both linguistic theory and computational requirements. The concept of a "living dictionary", for example, reframes lexicographic resources as evolving datasets that can be incrementally expanded and computationally leveraged, rather than as static reference works.

For this discussion, it is important to note that many of the projects discussed at the retreat are not presented as completed resources, but as initiatives in progress. In fact, our selection processes prioritises ongoing projects. This reflects a broader infrastructural logic. That is, the Hundzula retreat contributes to the initiation and acceleration of resource pipelines rather than the production of finished artefacts within the retreat itself. In this way, the feedback of linguists and NLP researchers during the discussions of ongoing work directly contributes to the shaping of annotation strategies, metadata design, and potential downstream applications. The retreat thus functions as a critical intervention point where resource trajectories can be defined and aligned.

4. Ethics, privacy, and culturally grounded data practices

In reality, even the most well-intentioned NLP can raise concerns about ethics (Field et al., 2021). Resultantly, like previous instalments of the retreat, see Okorie (2025), copyright, community protection, and ethical considerations are integral to the retreat's resource infrastructure. In the fifth edition, this importance was illustrated by the inclusion of work on privacy-preserving word frequency analysis and culturally grounded NLP frameworks. In many African contexts, language data is closely tied to community identity, cultural knowledge, and historical marginalisation (Zhong et al., 2024). Thus, the retreat provides a forum in which ethical design choices—such as the use of differential privacy or community-sensitive data governance—are discussed alongside technical implementation.

This approach contrasts with models of resource development that prioritise scale over contextual sensitivity. By embedding ethical reflection within technical discussions, Hundzula promotes data practices that are both responsible and locally relevant[4]. Such considerations are particularly important for indigenous languages, where the consequences of data misuse or misrepresentation may be more acute.

5. Transitioning from foundational resources to applied NLP systems

The Hundzula retreat is not limited to early-stage resource creation, but also supports experimentation with advanced NLP techniques adapted to low-resourced settings. Thus, the second day of the programme highlighted the retreat's role in supporting the transition from foundational linguistic resources to applied NLP systems. Presentations on data augmentation, multilingual speech recognition, large language model deployment, and domain-specific chatbots illustrated how linguistic insight and resource development feed directly into computational modelling.

Several talks explicitly addressed strategies for overcoming data scarcity, including linguistically informed augmentation and multilingual transfer. These approaches depend on the availability of at least minimal annotated data and linguistic expertise, both of which are fostered through the retreat's earlier focus on corpus and lexicon development. In this way, the programme reveals interdependencies between different resource types and stages, reinforcing the value of an integrated infrastructural model. Throughout these activites, the program circles back to collaboration opportunities, thus, enforcing the removal of silos in our language research.

6. Training and collaboration

6.1. Capacity building as a resource

As indicated earlier, a defining feature of the Hundzula Retreat is its explicit investment in human capacity as a resource outcome. Accordingly, workshops such as data annotation (*presented by Marissa Griesel from the South African Centre for*

[4]See Professor Chijioke Okorie's reflections on the data practices as practised in the Hundzula community and the implications it has for copyright and law-related issues in general, post the retreat. Her reflections can be accessed at https://datasciencelawlab.africa/dr-chijioke-okories-reflections-on-the-3rd-edition-of-hundzula-retreat/

Digital Language Resources), large language models in the social sciences (*presented by Professor Sree Ganesh Thottempudi, whose expertise includes the development of NLP tools for under-resourced languages and the application of digital humanities approaches to ancient heritage and culture*), and automatic tone extraction (*presented by Senekane Makhamisa from the University of Johannesburg*) were deliberately selected to equip participants with practical skills that are directly transferable to their own research contexts. Collectively, these sessions address a critical objective in African NLP: bridging the skills gap between computational experts and linguistics scholars, including those already engaged in digital humanities as well as those newly introduced to computational methods.

By situating training within the same space as active research discussion, the retreat collapses the distinction between learning and production. Similar to processes at the Resources for African Indigenous Language (RAIL) workshops, participants at the Hundzula retreat are not only exposed to tools and methods, but they are encouraged to apply them to their own language-specific projects. Thus, we hope that this human-in-the-loop approach enhances the sustainability of resource development by ensuring that expertise is distributed through training rather than concentrated to specific areas of expertise.

6.2. Collaboration and sustainability

To consolidate collaboration and ensure continuity beyond the retreat, we deliberately scheduled structured general discussion sessions at the end of Days 1 and 2 and invited participants to a shared task project that is planned to run over the course of the year. The end-of-session and end-of-day discussions are not ancillary but central to the infrastructural role of the Hundzula retreat. Specifically, they provide opportunities to identify shared challenges, align research agendas, and explore joint funding or publication opportunities. In doing so, the retreat contributes to the formation of research networks that support the long-term maintenance and expansion of language resources.

One of the major challenges in African NLP is fragmentation (Mbaye et al., 2025, p.7). The emphasis on collaboration at the retreat addresses this common limitation of fragmentation. By bringing together linguists, NLP researchers, industry practitioners, and students working across different languages and methodological traditions, the Hundzula Retreat reduces duplication of effort and encourages the reuse of tools, guidelines, and workflows across projects—an issue that has been widely observed, including within the RAIL workshops (Mabuya et al., 2023, p.133).

7. The Hundzula Infrastructural Model

We formalise the infrastructural components of the Hundzula experience across five interacting layers in order to abstract it into a transferable model. These five strata interact recursively rather than sequentially as illustrated in Figure 1.

The Hundzula Model is composed of five interrelated infrastructural layers that collectively facilitate the sustainable development of NLP for African indigenous languages. First, the Human Capacity Layer is the cornerstone of the system, emphasising skills transfer seminars, mentored annotation pipelines, and intentional cross-disciplinary pairings between linguists and NLP researchers to address expertise gaps.

Second, the Collaborative Network Layer promotes inter-institutional research aggregation, incubates shared-task initiatives, and monitors longitudinal collaborations to guarantee consistency beyond the retreat environment. Third, the human and network capacities are operationalised by the Resource Pipeline Layer through concrete activities such as corpus initiation, co-development of annotation guidelines, prototype modelling, and structured review procedures for ethical conformance.

Fourth, the Governance and Ethical Layer, which incorporates attribution policies, copyright awareness, community-sensitive data practices, and open licensing discussions into the infrastructure, oversees these activities to guarantee responsible and culturally grounded resource development. Lastly, the Sustainability Layer guarantees that initiatives extend beyond the retreat by facilitating follow-up shared tasks, post-retreat publication pipelines, and collaborative grant formulation. This transforms short-term engagements into enduring research ecosystems.

8. Conclusion

Overall, our paper posits that the Hundzula programme supports a broader reconceptualisation of what constitutes a resource in African indigenous language research. While tangible artefacts such as corpora, lexicons, and models remain essential, our paper proposes that infrastructures enabling their creation are equally important. Retreat-based models like Hundzula provide structured environments in which linguistic knowledge, computational methods, ethical practices, and human capacity are brought into sustained interaction.

Recognising such infrastructures as resources has implications for how resource development is evaluated and funded. That is, it suggests the need for assessment criteria that account not only for dataset size or model performance, but also for capacity building, collaboration, and contextual rel-

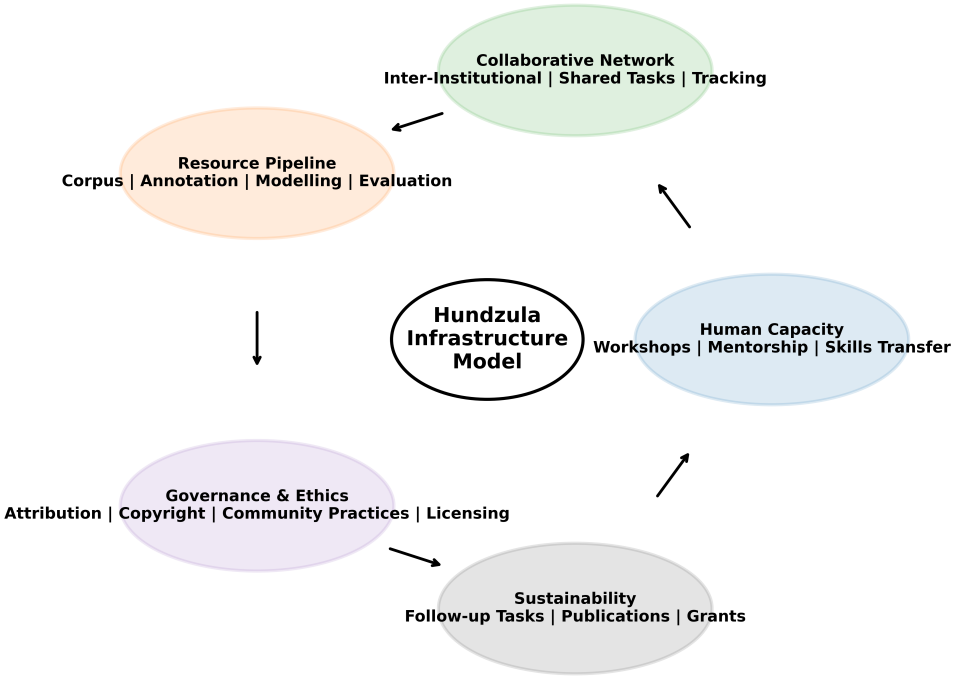

Figure 1: A visual representative of the Hundzula Infrastructural Model.

evance. In low-resourced settings, these factors may ultimately determine whether language technologies are sustainable and socially meaningful.

As indicated in the introduction, the contribution of this article is twofold. First, we document the Hundzula Retreat as an infrastructural resource for scholars working in African Natural Language Processing and across diverse fields of African linguistics and, to some extent, digital humanities. Second, we use the case presented in this article to propose a transferable model for retreat-based resource infrastructure for language research. In doing so, we highlight key affordances of the programme, drawing in particular on the fifth edition of the retreat to demonstrate the intentional design choices aimed at ensuring that the core objectives of the Hundzula Retreat, most notably, the transformation of African NLP and linguistics research, are effectively realised.

9. Limitations

The impact of this article is bounded by the scope and purpose of the Hundzula Retreat itself, which is designed primarily as a resource for scholars and students working in NLP, Linguistics, and Digital Humanities. As such, the retreat prioritises collaboration, skills development, and project incubation rather than the immediate production of finished research outputs. Consequently, many of the projects presented during the fifth and most recent edition of the retreat are still in progress and are therefore not yet ready for public release as stand-alone language resources.

Even so, this limitation should be understood as a defining feature of the Hundzula model rather than a shortcoming. The retreat is explicitly not a conference, and it does not require participants to present completed or novel work. Instead, it provides a protected and generative space in which early-stage ideas, methodological challenges, and partially developed resources can be shared openly, refined collaboratively, and strengthened through interdisciplinary engagement. This design encourages experimentation, honest reflection, and the co-development of tools, workflows, and research directions that may not yet be mature enough for formal dissemination but are essential to sustainable resource development.

In this regard, the outcomes of the retreat are intentionally longitudinal. While the immediate outputs may not align with the typical expectations of venues such as the RAIL Workshop, which we

assume typically foregrounds descriptions of existing resources, the Hundzula Retreat contributes to the conditions under which such resources can emerge. By fostering collaboration, building human capacity, and supporting the early stages of resource pipelines, the retreat functions upstream of the kinds of outputs commonly reported at RAIL.

At the same time, this paper does not provide a systematic evaluation of downstream outputs such as published datasets, shared tasks, or deployed NLP systems, as these remain emergent and not yet consistently documented. Similarly, while the paper draws on multiple iterations of the retreat, it does not offer a formal longitudinal comparison across all editions, instead treating the fifth iteration as a mature instance through which the model is described.

In addition, the study adopts a design-oriented, practice-based perspective but does not implement a fully formalised design-science methodology with explicit artefact evaluation or a detailed case study protocol. The emphasis is on documenting and conceptualising an emerging research infrastructure in context, rather than on methodological formalisation.

Finally, while the paper proposes the Hundzula model as transferable to other low-resourced language contexts, this transferability is argued conceptually rather than empirically demonstrated. Future work could extend this by evaluating the model across different settings and by tracing the progression of retreat-incubated projects into formalised NLP resources.

10. Acknowledgements

Since its inception, the Hundzula Retreat has been supported through funding and institutional contributions from multiple organisations. Institutional support was provided by the Data Science for Social Impact (DSFSI) unit and the African Institute of Data Science and Artificial Intelligence (AfriDSAI), both at the University of Pretoria and the University of the Witwatersrand through the Digital Humanities SARChI Chair; the North-West University through the Language Directorate; and the Nelson Mandela University through the Digital Humanities Hub. We also acknowledge funding from the Department of Science and Technology through the South African Centre for Digital Language Resources (SADiLaR). Furthermore, international support was provided by the UK Government's Foreign, Commonwealth and Development Office, Canada's International Development Research Centre, and Google TensorFlow. We also thank the hosts of previous Hundzula Retreats, including the University of Pretoria (2022 and 2023), the Nelson Mandela University (2024), the University of Johannesburg (2025), and the North-West University (2026). Finally, we acknowledge the contributions of Professors Marivate Vukosi and Mpho Primus, without whom this work would not have been possible.

11. Bibliographical References

Ifeoluwanimi Adebara. 2024. *Towards Afrocentric natural language processing*. Ph.D. thesis, University of British Columbia, Vancouver, BC Canada.

Emmanuel Kigen Chesire and Andrew Kipkebut. 2024. Current state, challenges and opportunities for natural language processing research and development in africa: A systematic review. In *5th Workshop on African Natural Language Processing*.

Jeanne Elizabeth Daniel. 2020. *Applications of natural language processing for low-resource languages in the healthcare domain*. Ph.D. thesis, Stellenbosch: Stellenbosch University, South Africa.

Alladi Deekshith. 2024. Advances in natural language processing: A survey of techniques. *International Journal of Innovations in Engineering Research and Technology*, 8:74–83.

Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett. 2021. Proceedings of the 1st workshop on nlp for positive impact. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, Online. Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Rooweither Mabuya, Don Mthobela, Mmasibidi Setaka, and Menno van Zaanen. 2023. Proceedings of the fourth workshop on resources for african indigenous languages (rail 2023). In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*.

Matthew S Mayernik, David L Hart, Keith E Maull, and Nicholas M Weber. 2017. Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, 68(6):1341–1359.

Derguene Mbaye, Tatiana DP Mbengue, Madoune R Seye, Moussa Diallo, Mamadou L

Ndiaye, Dimitri S Adjanohoun, Cheikh S Wade, Djiby Sow, Jean-Claude B Munyaka, and Jerome Chenal. 2025. Opportunities and challenges of natural language processing for low-resource senegalese languages in social science research. *arXiv preprint arXiv:2601.09716*.

Shamsuddeen Hassan Muhammad, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, et al. 2025. Hausanlp: Current status, challenges and future directions for hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 176–191.

Nadia Nahar, Shurui Zhou, Grace A. Lewis, and Christian Kästner. 2021. More engineering, no silos: Rethinking processes and interfaces in collaboration between interdisciplinary teams for machine learning projects. *ArXiv*, abs/2110.10234.

Tolúlọpẹ́ Ògúnrẹ́mí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing nlp for "low-resource languages": Applying abebe birhane's relational ethics. *GRACE: Global Review of AI Community Ethics*, 1(1):1–13.

Chijioke Okorie. 2023. Copyright, data mining and developing models for south african natural language processing. *Joint PIJIP/TLS Research Paper Series*, 117:1–28.

Chijioke Okorie. 2025. It's the NOODL license–awesome and amazingly geeky! *Available at SSRN https://ssrn.com/abstract=5339254*.

Chijioke Okorie and Melissa Omino. 2025. Addressing inequitable openness in licences for sharing african data and datasets through the nwulite obodo open data licence. *Law, Tech. & Hum.*, 7:94.

Mmasibidi Setaka and Benito Trollip. 2022. Resource repositories and linking resources: An exploratory study. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 4(02).

Kathleen Siminyu, Jade Abbott, Kola Tubọsun, Aremu Anuoluwapo, Blessing K Sibanda, Kofi Yeboah, David Adelani, Masabata Mokgesi-Selinga, Frederick R Apina, Angela Thandizwe Mthembu, et al. 2023. Consultative engagement of stakeholders toward a roadmap for african language technologies. *Patterns*, 4(8):100820.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Weihang You, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.