SADiLaR

ENGLISH NEWSLETTER

SEPTEMBER 2023

READ ABOUT :

Contact us:

BRIDGING THE PAST AND PRESENT: DIGITISATION WORKSHOP AT MAZISI KUNENE MUSEUM

- **Rooweither Mabuya**

In today's rapidly advancing digital age, the preservation and accessibility of historical artifacts and manuscripts has become paramount. The South African Centre for Digital Language Resources (SADiLaR) recently organised a digitisation workshop at the Mazisi Kunene Museum in Durban. Led by digital humanities researchers Rooweither Mabuya and Andiswa Bukula, and SADiLaR's Director of Operations, Juan Steyn, the aim of the workshop was to guide museum staff and stakeholders through the process of digitising their precious manuscripts.

The Mazisi Kunene Museum is a cultural institution dedicated to commemorating the life and work of renowned poet and philosopher, Mazisi Kunene. It is home to a vast collection of invaluable manuscripts and artifacts that offer a glimpse into the rich cultural heritage of South Africa. However, these delicate materials were at risk of deterioration and limited accessibility owing to their physical nature.

Understanding the urgency to preserve these historical treasures for future generations, the workshop's facilitators, Mabuya, Bukula and Steyn, introduced the concept of digitisation. They showcased how modern technology could breathe new life into old manuscripts, making them accessible to a global audience.

The facilitators guided the participants through a step-by-step digitisation process. They demonstrated the proper handling of manuscripts, emphasising the need for utmost care and caution during scanning. The workshop attendees learned about file formats, resolution settings, and metadata management, enabling them to create comprehensive digital archives.

Digitisation promotes collaboration and knowledge sharing by facilitating the exchange of digitised materials with other museums and institutions. It further enhances accessibility by enabling researchers, scholars, and enthusiasts from around the world to study and appreciate these historical works remotely.

Preserving cultural heritage through modern technology

One of the significant highlights of the workshop was the handover of a scanning device, the IRIScan Desk 5 Pro, along with a desktop computer. This equipment will serve as a vital tool in the digitisation process, allowing the museum to efficiently convert its manuscripts into high-quality digital formats. The scanner's advanced features, such as automatic

page-turning detection and image enhancement, will ensure the preservation of the original manuscripts' integrity while capturing every detail with precision.

The Mazisi Kunene Museum's commitment to digitisation signifies a promising future for the preservation and accessibility of cultural heritage. The workshop participants were encouraged to continue the digitisation efforts beyond the event, ensuring a sustainable and comprehensive digital archive. The museum plans to engage in partnerships and collaborations to expand its digital presence and explore innovative ways to engage audiences with the digitised collection.

The digitisation workshop at the Mazisi Kunene Museum served as a milestone in the journey towards preserving and promoting cultural heritage through modern technology. With the guidance of experts and the provision of advanced equipment, the museum staff and stakeholders are well-equipped to embark on a transformative digitisation journey.

Contact us:

SADILAR'S LANGUAGE RESOURCE REPOSITORY EMPOWERS LANGUAGE RESEARCH

- Birgit Ottermann

The curation, distribution and maintenance of reusable digital text and speech resources for South Africa's official languages is of vital concern for research and development in the field of language technology. The data is important not only for the development of tools for facilitation of communication between different language groups but also for empowering local languages for use in modern communication systems. The South African Centre for Digital Language Resources (SADiLaR) has taken on this crucial guardian role through its Language Resource Repository. To date, it contains hundreds of items in multiple languages which are available to the public through an open-access platform.

"SADiLaR's Language Resource Repository has over 400 records of items in multiple languages, even a few languages from outside South Africa," says Dr Friedel Wolff, SADiLaR's Technical Manager. "Some of the items themselves describe a resource that is itself multilingual or, for example, software that supports several languages. Not every resource in your language might interest you, but it might just be what some researcher or software engineer needs to build something exciting for your language."

Giving permanence to resources

The various types of available resources range from electronic text and speech data (such as domain-specific text collections, wordlists, dictionaries, translation memories and aligned multilingual corpora) to multimodal resources and tools, applications and platforms that support the processing of data and development of new technologies.

According to Wolff, the research data stored in SADiLaR's repository is of immeasurable value to researchers. "Much of the research data on the repository was costly and time-consuming to create. Some required expert knowledge or computing power that few of us have access to," he comments. "The repository makes these available to anyone who is interested, and the idea with repositories like these is that the repository should outlive any specific research topic, researcher's interest or industry fad – in other words, it tries to give some permanence to these resources. Providing this permanence is maybe too hard and tedious for many of the creators, and not always easy to justify in their place of employment. This provides a centralised access point, without trying to take away
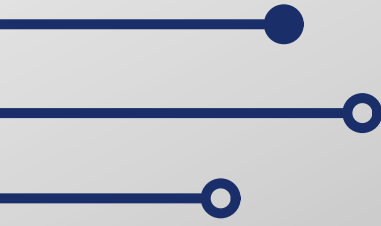
Contact us:

3

Contact us:

any of the credit to the people who put the work into creating them," he explains.

Central point of access

Dr Benito Trollip, a digital humanities researcher at SADiLaR, and enthusiastic user and contributor to the repository, echoes the comments above. "The SADiLaR Language Resource Repository provides in principle a permanent platform for the availing of linguistic data to the broader community that includes not only researchers. It takes one curious person to see what is out there for less well-known languages and they start developing useful technology," says Trollip.

When it comes to the repository being a central point of access, Trollip emphasises how difficult it can be to use the existing linguistic data source if it, or information about it, is of a sensitive nature and is not made available.

"It often took a lot of time and hard work to generate and curate that data. In my humble opinion, we should move away from the mindset of owning, developing and using data solely for our own gain or professional and financial benefit, and rush toward a mindset of sharing data to enable and empower the community at large," he says.

Integral tool

Dr Laurette Marais, manager of SADiLaR's speech node at the Council for Scientific and Industrial Research (CSIR), and her team have experienced the advantages of SADilaR's repository as both contributors and users.

They shared their valuable resources with others, which enabled the development of commercial products, and they also benefited by accessing resources that they did not create themselves.

"For the CSIR Voice Computing research group, also known as the Speech Node of SADiLaR, the Resource Repository has become an integral tool in the planning and execution of our research agenda, both as a reliable venue for sharing the data that we gather and produce, but also as a first port of call when we require language resources for our projects. A notable contribution of ours to the repository was high-quality speech data from our Lwazi 3 project, which we have also used to develop our commercial suite of TTS voices, named Qfrency," says Marais.

"We have in the past and are still contributing speech data aimed at training automatic speech recognition systems. Furthermore, the repository has served as an essential source when we require text data in any of the South African languages. I believe that any student or researcher in language technology in South Africa should be familiar with the repository and what it has to offer, especially given the resource-scarce nature of our languages."

A short history

Interestingly, the repository actually predates SADiLaR. It was launched in 2012 by the North-West University's Centre for Text Technology as the Resource Management Agency (RMA) with funding from the

4

Department of Arts and Culture's National Centre for Human Language Technologies. When SADiLaR was launched in 2019 with the support of the Department of Science and Innovation (following an incubation and development phase since 2016), the RMA was incorporated into SADiLaR's Language Resource Repository. SADiLaR took over full responsibility for the curation and maintenance of the repository thereafter.

Submit a resource

If you have developed a language resource and wish to make it usable and/or discoverable for others, SADiLaR's repository is an excellent option. It is a secure environment with the correct licensing procedures for anyone with research data in the fields of languages, humanities and social sciences. For more information on how to submit a resource, please visit the SADiLaR Resource Guidelines page.

SADILAR–UJ EXTERNSHIP HIGHLIGHTS CAREER POSSIBILITIES IN DIGITAL HUMANITIES

- Birgit Ottermann

A group of linguistics and language practice students from the University of Johannesburg (UJ) recently attended an online workshop with staff from the South African Centre for Digital Language Resources (SADiLaR) to learn more about career possibilities in the field of digital humanities. This one-day externship, which took place on 14 June 2023 via Zoom, offered the students a different perspective on the types of careers they could follow, once qualified.
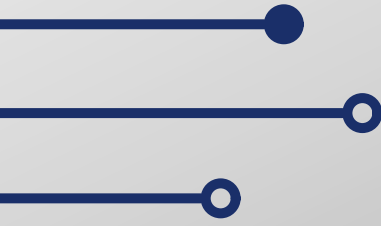
Contact us:

Contact us:

"Our main goal was to provide students with a short-term professional learning experience in digital humanities and language development," says Dr Benito Trollip, digital humanities researcher for Afrikaans at SADiLaR and the main organiser of the externship. "We did this by introducing the students to SADiLaR and its different research nodes; demonstrating some of the tools we use to analyse, annotate, translate or otherwise interact with language in the digital space; and having staff members give personal accounts of their work at SADiLaR."

The idea for an externship originated when Laurinda van Tonder, lecturer in Applied Linguistics at UJ's Department of Languages, Cultural Studies and Applied Linguistics, contacted Trollip to find out whether an externship for the final-year Linguistics and Language Practice students would be possible at SADiLaR.

"Laurinda explained to me that the students normally have one-day internships at institutions such as banks and local government offices (where language practitioners are also working), but she thought that introducing the students to SADiLaR would enhance their view on the available careers within language," Trollip recalls. "She also had to explain to me what an externship is – which is basically an online internship," he adds.

Informative and thought-provoking

According to Trollip, SADiLaR's very first externship was a huge success. "We received very positive feedback from the participants as well as the SADiLaR team members who took part. The students engaged a lot and I feel the interactions between them and the team from SADiLaR were natural, informative and thought-provoking.

"We spent a lot of time going through the website and showing the students how to find the information, resources and tools they could possibly need. Our tutorials on the various tools illustrated to students what is available and how it could assist them with their work and studies in the different linguistic and applied linguistic areas they are busy with or planning to pursue," he says. "Two tools that especially piqued the interest and interaction of the students were the Autshumato machine translation (translating a sentence from English to isiZulu) and searching for words in the Corpus portal."

According to Rooweither Mabuya, digital humanities researcher for isiZulu at SADiLaR, who joined the externship's afternoon session, the students were very keen to learn more about the research happening at SADiLaR. "I shared my research interests with the students and explained how I was introduced to digital humanities and digital scholarship when I joined SADiLaR. I also highlighted how attending various workshops and training opportunities helped me to understand how I could use Digital Humanities in my research." Mabuya who is a PhD candidate, was pleased when a student asked her to share the tools she is using for her research.

Another SADiLaR digital humanities researcher, Mmasibidi Setaka (who specialises in Sesotho), was impressed with how engaged students were. "They asked many questions related to what we do at SADiLaR. Since digital humanities is still young field in South Africa, they found it interesting to learn more about the work we are doing in digitising and creating language resources for the eleven official languages. I also shared how we host different workshops to share the knowledge we have acquired with others. Based on the success of this externship, I would like us to extend it to other institutions too."

The student feedback below speaks for itself, and more externships are sure to follow in the future:

"Thank you so much guys. This was very informative. Continue to host more of these please. Other people would really appreciate it too."

"Thank you to everyone. This was productive, exciting, and very much needed by us as students. "

"I just want to say thank you to the whole SADiLaR team for this experience today. We all learned a lot today, and we are motivated to further in this field and maybe one day join you all. Everyone had a great spirit."

SADILAR-FUNDED RESEARCH GIVES A VOICE TO SEPEDI CHILDREN WITH SEVERE COMMUNICATION DISABILITIES

- **Birgit Ottermann**

The implementation of alternative forms of communication for children who are not yet literate – especially those who have severe communication needs and require picture-based systems to express themselves – has always been a key research interest for Prof Kerstin Tönsing.

"Children who need such systems often have some physical disabilities which make them incapable of using sign language," says Tönsing, who is a speech therapist, audiologist, and professor at the University of Pretoria's Centre for Augmentative & Alternative Communication. "We provide them with picture-based systems that represent words which they can point to or activate on a device in order to convey their message," she explains.

Although this type of intervention has been widely used in the field of augmentative and alternative communication (AAC) to help individuals who cannot rely on their natural speech to communicate – and there are numerous picture-based
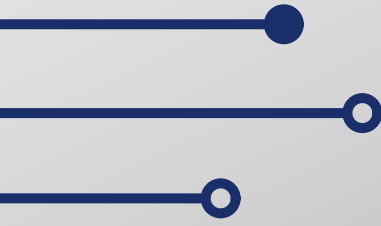
Contact us:

7

Contact us:

vocabulary packages available for various languages – many of the so-called low-resource languages have been left behind. This is also true for many of the South African languages, including Sepedi.

"I started getting increasingly concerned that we are doing a lot of intervention in South Africa in English for children whose home language is not English," Tönsing recalls. "We live in a multilingual country and children should be able to express themselves in their home language – this also applies to children who need a picture-based AAC system to do so. But, what are we doing to foster all of the children's languages?"

Thanks to research funding from the South African Centre for Digital Language Resources (SADiLaR), this reality is about to change for Sepedi children. Tönsing and her team have embarked on a project to develop a research- and stakeholder-informed non-electronic AAC system for Sepedi. The aim is to design a comprehensive system with a range of vocabulary items that allow for language development.

Mother tongue communication

"The communication of preliterate children from a Sepedi (Sesotho sa Leboa) language background is severely restricted as there currently is no AAC system available in their mother tongue – an estimated 80 000 Sepedi children are affected. A picture-based AAC system that allows for expression in Sepedi is therefore an urgent necessity," Tönsing comments.

She is ideally placed to lead this research project in that much of her PhD and postdoc research centred on developing AAC systems that give access to non-English South African languages.

"The project builds on the master's research of Ngwanamashiane Rahab Mothapo, who is now a PHD student and an important collaborator," Tönsing continues. "We established a core vocabulary list based on the analysis of language samples of Sepedi preschoolers. We then identified graphic symbols from existing symbol libraries, which were validated and extended with stakeholders during in-depth discussions. Some additional symbols are currently being designed to represent some of the Sepedi core words for which we could not find suitable symbols."

Human-centred design process

According to Tönsing, they also did a needs analysis and discussed design requirements with the stakeholder group, which comprises speech-language therapists with experience in AAC and providing services in the Sepedi language; parents of children with complex communication needs from a Sepedi background; adult users of AAC systems; and teachers who teach children from a Sepedi background with severe communication disabilities. The participants are based mainly in the Limpopo province where Sepedi is spoken.
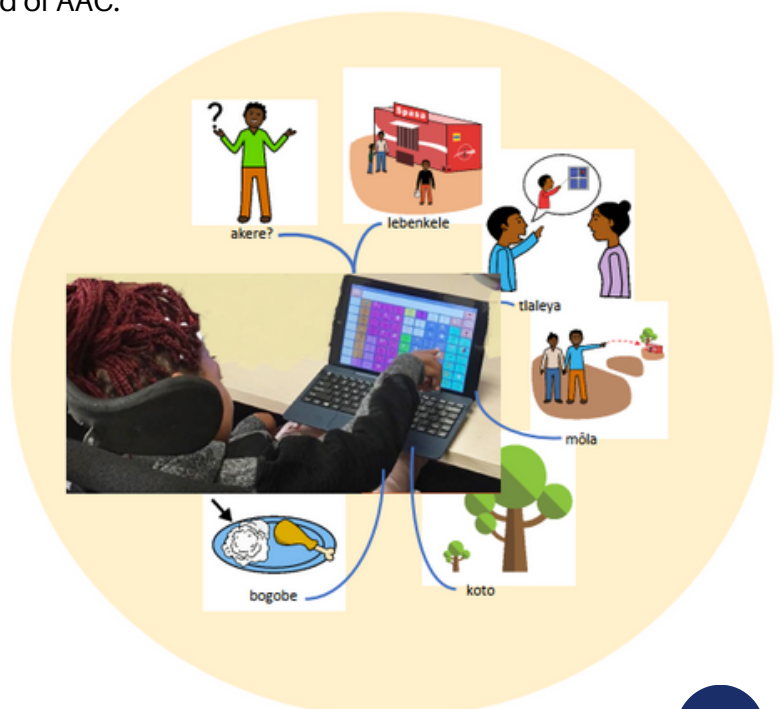All interactions with the focus groups are being transcribed and

Contact us:

thematically analysed. The next step is to start a human-centred design process where the first prototype AAC system is shared with stakeholders for evaluation. "We are on the brink of creating the first prototype, which is very exciting. It is an iterative process with stakeholders where we listen to their feedback and redesign until we reach a stage where the system is ready to be used. Even then, the feedback from users will be ongoing," Tönsing says.

International conference

The results of the first phase of the project will be presented by Rahab Mothapo at the Biennial Conference of the International Society for Augmentative and Alternative Communication in Cancún, Mexico, from 24 to 27 July 2023. The conference, themed 'Communication beyond borders', will be attended by researchers, academics, practitioners, and persons using AAC from across the world, and will showcase the latest research and lived experiences in the field of AAC.

"It is such a privilege to be enabled by SaDiLaR to present research findings of such an important study that cuts through a number of disciplines," Mothapo comments. "I am looking forward to sharing this information with the extremely diverse audience. This for me, marks a small attempt at making communication accessible to the Sepedi-speaking South African population living with communication impairment."

SADILAR CELEBRATES SUCCESSFUL 4TH RAIL WORKSHOP IN CROATIA

- Birgit Ottermann

The South African Centre for Digital Language Resources (SADiLaR) hosted a successful fourth workshop on Resources for African Indigenous Languages (RAIL) in Dubrovnik, Croatia. The annual workshop, which took place on 6 May 2023 as part of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), was organised by SADiLaR's Rooweither Mabuya, Mmasibidi Setaka and Prof Menno van Zaanen, and the CAM Foundation's Don Mthobela.

"The RAIL workshop was a great success. All our attendees were fully engaged throughout the session even though it was a full-day workshop," says Mabuya, who co-chaired the event with Setaka. "The workshop provides an interdisciplinary platform for researchers working on African indigenous languages, particularly those languages that are under-resourced," she explains. "It brings together researchers interested in showcasing their research; provides an overview of the current availability of African indigenous language resources, including data collections as well as tools; and allows for information sharing and discussions on improving the quality and availability of the resources."

Many African indigenous languages currently have no or very limited resources available and are often structurally quite different from the better-resourced languages; therefore, they require the development and use of specialised techniques.

Growing a scientific community

"By bringing together researchers from different fields, such as computational linguistics, sociolinguistics and language technology, to discuss the development of language resources for African indigenous languages, we hope to boost research in these fields," says Setaka. "Ultimately, we aim to create the conditions for the emergence of a scientific community of practice that focuses on data, as well as tools, specifically designed for or applied to indigenous languages found in Africa."

Both Mabuya and Setaka, who are digital humanities researchers at SADiLaR in IsiZulu and Sesotho respectively, have been involved with the RAIL workshop since its inception in 2019. The first two workshops were virtual events co-located at the Language Resources and Evaluation Conference (LREC) in 2020 and Digital Humanities

10

Contact us:

Association of Southern Africa (DHASA) conference in 2021. Post COVID-19, the third RAIL workshop was hosted as an in-person event in 2022 at the 10th Southern African Microlinguistics Workshop in Potchefstroom.

"This year's workshop, which was one of 13 workshops accepted at the EACL 2023 conference, had a total of 14 papers and one findings paper presented," Mabuya remarks. "It was a hybrid format as some participants were not able to travel – seven papers were presented in person and eight were presented virtually."

Excellent feedback

Mabuya and Setaka were fortunate to travel to Croatia and attend the EACL 2023 conference in person, thanks to travel grants that they received to present their papers at the conference. "Mmasibidi Setaka received the Diversity and Inclusion Subsidy aimed at supporting scholars from marginalised regions, whereas I received a grant from the student volunteer programme aimed at supporting early-career scholars to attend the conference and also assist with conference duties," Mabuya says.

Both researchers presented papers on behalf of their co-authors at the RAIL workshop. Mabuya's paper was titled "Unsupervised Cross-lingual Word Embedding Representation for English-isiZulu", with Derwin T Ngomane, Vukosi Marivate, Jade Abbott and Rooweither Mabuya as authors.

"Our paper received such great feedback in comments and questions from the audience," Mabuya recalls.

"One of the participants even asked for a meet-up with my co-authors as he needed some assistance in his own research which was similar to ours."

Setaka was equally pleased with the response she received for the paper she presented, titled "Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words", with Johannes Sibeko and Mmasibidi Setaka as authors. "People were very interested, and so delighted to learn that there's a workshop dedicated to NLP [natural language processing] in Africa."

Making new connections

Reflecting on the conference and overall success of the workshop, Setaka says: "The conference brought together a diversity of people interested in the many aspects of NLP. The workshop itself was a great success with a lot of participation from the audience. The fact that our workshop was accepted at EACL was a great highlight for me, considering the nature of EACL and its standing in the NLP community."

Mabuya adds that she made some great connections. "It was a big conference with numerous interesting talks and presentations, and I got to meet some amazing scholars. Regarding our workshop, it was great that we had authors who have been submitting their research to RAIL each year since it started. This shows the quality of our workshop and the work published in our proceedings. We also have an excellent programme committee who assist with reviewing the submissions."

11

The Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023) are now available, and the SADiLaR team is looking forward to receiving submissions for the fifth instalment of RAIL, once the call for papers is out. "The aim is to grow the workshop to greater heights and also allow other scholars to help organise it going forward," Setaka concludes.

Contact us:

Click here to subscribe to our newsletter

Contact us:

PHD RESEARCH PAPER PUTS SADILAR IN THE GLOBAL SPOTLIGHT
- Birgit Ottermann

The South African Centre for Digital Language Resources (SADiLaR) is enjoying international bragging rights, thanks to a PhD research paper that was included in the Post Conference Proceedings publication of the 2022 CLARIN Annual Conference.

Johannes Sibeko, a PhD student supervised by SADiLaR's Prof Menno van Zaanen, was one of 12 doctoral scholars selected to present an abstract of their research in the PhD Student Session of the 11th annual CLARIN conference, which took place from 10 to 12 October 2022 in Prague, Czech Republic. It was such a success that he was invited to submit a full paper for inclusion in the Post Conference Proceedings, an open-access publication that presents the highlights of the entire conference.

CLARIN, which is short for Common Language Resources and Technology Infrastructure, is a digital infrastructure that offers data, tools, and services to support research that is based on language resources. Sibeko's paper presents his doctoral research project that explores the development of resources for measuring text readability in Sesotho, a Bantu language spoken by more than ten million speakers across southern Africa.

The only student presenter from Africa

"The acceptance letter meant the world to me," says Sibeko, who is a lecturer in Digital Humanities at the Nelson Mandela University in Gqeberha. "I was nervous to present at the conference, but at the same time excited for the opportunity to be on an international stage. Being chosen to represent SADiLaR, as well as being the only student presenter from Africa,

was a great honour for me. I felt very proud."
Unfortunately, Sibeko was unable to attend the conference in person because of visa issues. "I ended up attending online only. It was very disappointing because I couldn't participate in the interesting activities organised by CLARIN and I also missed out on networking opportunities. However, I remain positive that there will be more travel opportunities in the future."

Reflecting on his achievement as an early-career researcher, Sibeko says he should start believing more in himself and should approach international publication outlets with more confidence. "I feel like I am just arriving where I wished to be in my research journey."

Addressing learners' poor reading ability

Asked about the topic for his PhD research, Sibeko explains that South African learners are lacking in reading skills. "In education, teachers are expected to choose and adapt texts to their learners' levels. However, these processes are intuitive and subjective. As a result, there is no objective way of assuring that texts administered for learning, teaching and assessment are of the correct readability levels," he says.

"An objective measure of text readability in Sesotho will help in the selection

13

**For any feedback on the content of the newsletter, please feel free to contact SADiLaR:
info@sadilar.org**

**For any feedback on the translations, please feel free to address your message to the translator when sending an email to SADiLaR:
info@sadilar.org**

Contact us:

and adaptation of texts for different purposes and expected levels. My study therefore aims to develop metrics for measuring text readability that can benefit researchers, authors, teachers, and readers. The aim is to adapt nine existing readability metrics into Sesotho using English as a higher-resourced helper language. All the modules will be published open access on SADiLaR's repository."

Sibeko also hopes to develop a web-based application to provide access to automated text readability analysis that will allow the user to paste texts and receive a readability analysis report.

According to Menno van Zaanen, Professor of Digital Humanities at SADiLaR and North-West University, Sibeko's research illustrates the importance and applicability of digital language resources for South African languages. "Being able to measure the readability of Sesotho texts allows lecturers to select suitable texts for learners, and professional writers to adjust their texts to their relevant audiences. Not only is Johannes' work interesting from an academic perspective, it illustrates how these resources can be used to boost the South African languages, like Sesotho, for the general public," he comments.

Former high school teacher

Interestingly, Sibeko is a former high school teacher in Sesotho and English whose research is driven by his desire to maximise opportunities for language learning. "The challenge of selecting and aligning reading texts with readers goes beyond Sesotho. Nonetheless, Sesotho is selected as the initial approach to tackle this overarching issue. I am hopeful that readability metrics can be developed for other indigenous languages too."

Sibeko was exposed to digital humanities for the first time when he landed a post at Nelson Mandela University and was required to explore this new field. "I had no idea what it was when I first started out," he recalls. "Now I have immersed myself in this field. I am having a really good time doing my PhD and the research is becoming that much more interesting."