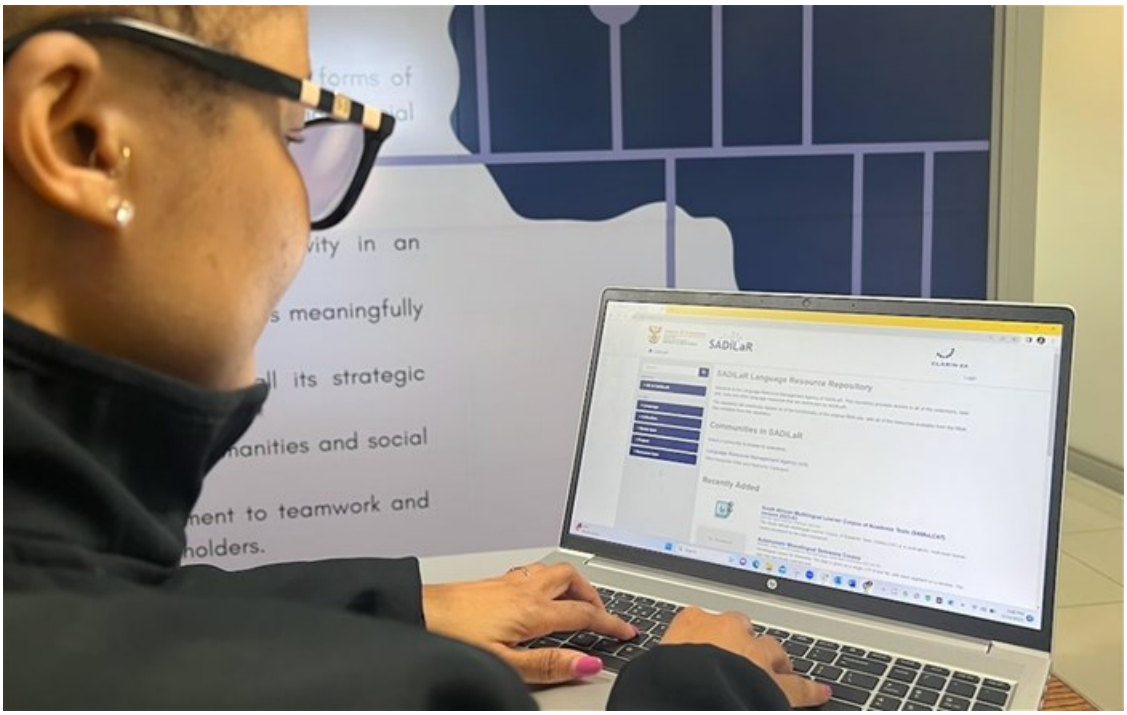
SADiLaR's Language Resource Repository empowers language research

By Birgit Ottermann , issued by North-West University (NWU)

20 Jun 2023

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 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 above. “The SADiLaR Language Resource Repository provides a (in principle) 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 utilise existing linguistic data source if it, or information about it (is of a sensitive nature), 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 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 still are contributing speech data aimed at training automatic speech recognition (ASR) 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 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 in 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.

About SADiLaR

Hosted by the North West University, the South African Centre for Digital Language Resources (SADiLaR) is a national centre supported by the Department of Science and Innovation (DSI) as part of the new South African Research Infrastructure Roadmap (SARIR).

'SARIR is a high-level strategic and systemic intervention to provide research infrastructure across the entire public research system, building on existing capabilities and strengths, and drawing on future needs.' (DST SARIR brochure).

SADiLaR has an enabling function, with a focus on all official languages of South Africa, supporting research and development in the domains of language technologies and language-related studies in the humanities and social sciences. The Centre supports the creation, management, and distribution of digital language resources, as well as applicable software, which are freely available for research purposes through the Language Resource Catalogue.

SADiLaR clients include academic scholars and professionals in all domains of Humanities and Social Sciences, Language Technologies, Natural Language Processing, Computer Science, as well as potential end-users in education, business, and industry.

- **GoAllOut student chapter breaks Guinness World Record** 17 May 2024
- **Seven Eagles to soar at the 2024 Olympic Games in France** 15 May 2024
- **Child speech database research project attracts international attention** 7 May 2024
- **Siya celebrates Rassie** 7 May 2024
- **“Yster” Rassie Erasmus receives honorary doctorate** 7 May 2024

North-West University (NWU)

The North-West University (NWU) is one of South Africa's top five universities; that offers superior academic excellence, cutting-edge research and innovation and teaching and learning. It all starts here.
Profile | News | Contact | Twitter | Facebook | RSS Feed