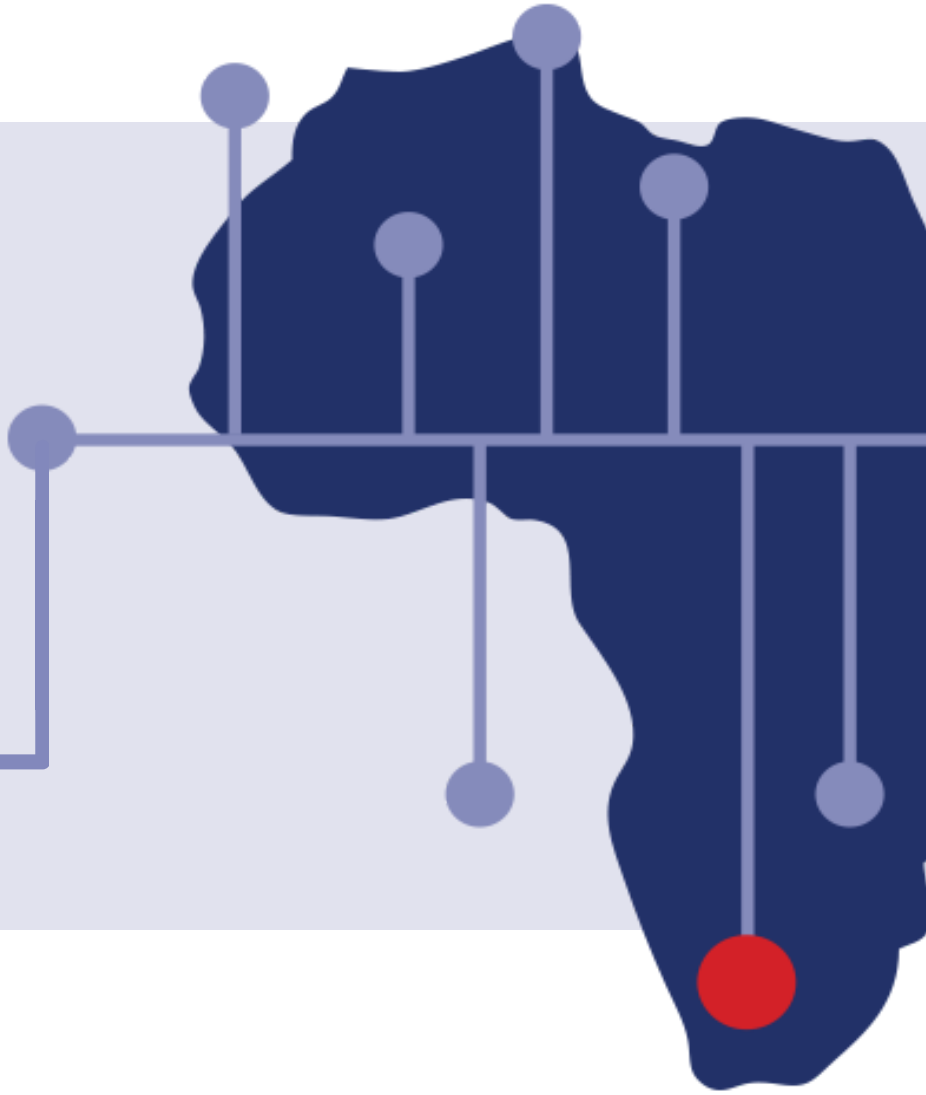
SADiLaR

South African Centre for Digital Language Resources

NEWSLETTER 1 / APRIL 2018

www.sadilar.org



READ ABOUT

Establishing a research infrastructure

English can be regarded as the de facto language of the digital age due to the rapid development of the internet and communication technology over the last three decades. However, speakers of other languages have taken up the challenge and there has been a concerted effort to develop digital tools and other resources in languages other than English to support people's need to use their languages on digital platforms.

The South African Centre for Digital Language Resources (SADiLaR) has been established to foster digital research and development growth in the official languages of South Africa. SADiLaR forms part of the Department of Science and Technology's (DST) new South African Research Infrastructure Roadmap (SARIR), for the large scale development of research capacity in South Africa.

SADiLaR is the only program in SARIR that focuses on the humanities, with the remaining research infrastructure projects focusing on various aspects of health and natural sciences.

Background

South Africa has joined the international playing field with the DST's establishment of SARIR as part of their long-term research and development plan. SARIR is intended to provide a strategic, rational, medium- to long-term framework for planning, implementing, monitoring, and evaluating the provision of research infrastructures necessary for a competitive and sustainable national system of innovation. Thirteen research infrastructures from five scientific domains have emerged as concrete and sufficiently conceptualised proposals for inclusion in SARIR.

In 2008, a ministerial advisory committee addressed recommendations to support human language technology (HLT) development at a national level. SADiLaR resulted from this decision, and from the work of various researchers under the guidance of former Director, Prof. Justus Roux. It provides a vision of developing and supporting a multilingual democracy, through access to digital language resources and language technology development. Ten years after the initial decision, SADiLaR is in its first year of incubation as one of SARIR's research infrastructures, and facilitates an environment for the creation, management, and distribution of digital language resources by offering language data and applicable software that is freely available for research and development purposes for the 11 official South African languages.

The North-West University (NWU) is the host of this multi-partner entity, that has a network of linked nodes consisting of a number of South African universities and agencies (UP, UNISA, CSIR, ICELDA, and CTexT). SADiLaR is the first of its kind in Africa and promotes existing links with similar entities globally, especially with a major counterpart in Europe, the Common Language Resource Infrastructure (CLARIN).

Scope of SADiLaR

Internationally, well-resourced languages have corpora of more than a billion words per language, or thousands of hours of digital speech data, enabling the users of these languages to have access to functional language-based technologies such as automatic speech recognition systems and machine translation applications. Unfortunately, the situation for most of the official languages of South Africa is significantly different, with relatively small data sets available in both the text and speech domains. This in turn limits the opportunities for the development of functional technologies to support the multilingual South African community. The establishment of SADiLaR will enable the future research and development of language technologies in South Africa, as SADiLaR aims to pool computational resources for these purposes.

Beyond the development of technologies that language resources enable, these resources are also a prerequisite for the study of language through digital means, typically encapsulated in the field of digital humanities.

"South Africa is twenty years behind the rest of the world when it comes to the development of Digital Humanities [DH], but SADiLaR ought to enable researchers to develop skills to do DH-related research on par with what is produced internationally," says Prof Attie de Lange, current Director of SADiLaR.

Furthermore, SADiLaR will, through capacity-building initiatives, promote and support the use of digital data and innovative methodological approaches aiding numerous projects in the domains of Humanities and Social Sciences.

As an enabling agency, SADiLaR will provide training for a new generation of researchers through various workshops by both national and international experts. Workshops will also be held on request and will cover a broad spectrum of topics in the DH and HLT domains, such as digitisation, data standards, use of computational mediated approaches through tools such as R and Python, methods to clean, organise and analyse data, as well as thematic workshops relating to the Humanities and Social Sciences.

FUNDED BY:

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

HOSTED BY:

NWU
NORTH-WEST UNIVERSITY
NOORDWES-UNIVERSITEIT
YUNIBESITI YA BOKONE-BOPHIRIMA

PARTNERS:

CSIR
our future through science

UNISA
university of south africa

CTexT
Sentrum vir Tekstegnologie
Centre for Text Technology

ICELDA
Inter-institutional Centre for Language Development and Assessment

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

www.sadilar.org

SADiLaR launches a new website and data repository

FUNDED BY:
science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

SADiLaR is excited to announce the launch of our website, which sets the foundation for us in our expanding role as African leader in the digital humanities (DH) and natural language processing (NLP) domains.

The website offers users an overview of the various activities at SADiLaR, including our vision and mission, our strategic goals, as well as various activities and opportunities at SADiLaR. Users will be able to easily navigate the comprehensive overview of activities at SADiLaR and retrieve the information they need.

The website also includes digitisation and resource procedures that provide guidance on the process of digitisation or obtaining and submitting resources. These features form an integral part of the website, since it offers the opportunity for collaboration between various parties to contribute to the collection and distribution of various language resources.

As part of SADiLaR's expanding role, the new site has also integrated the activities of the Language Resource Management Agency (RMA). Both the existing resource index and catalogue are available from SADiLaR as part of an institutional repository, from which all available SADiLaR resources are distributed.

The SADiLaR site also provides access to various online linguistic research technologies, and it is foreseen that both the repository and technology sections will be extensively updated over the course of the coming year.

Another feature of the website is updates on latest news and events that are happening in the DH and NLP environment in South Africa as well as globally. The SADiLaR newsletter will be published online and visitors can subscribe to receive the newsletter via the website. Users will also be able to get updates on training workshops and request workshops by completing an online form.

"Our hope is that the website will spark national and international collaboration between researchers, academics, scholars and the general public and help us build forward towards a multilingual continent," says Prof Attie de Lange, SADiLaR's director.

During the transition phase with the new website, the existing RMA website will still be available as an alternative for accessing the available resources from SADiLaR.

Please feel free to report any issues or provide suggestions for improving the website via email at: info@sadilar.org

www.sadilar.org

Integrating the RMA into SADiLaR with new technologies

FUNDED BY:
science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

Over the past five years the Language Resource Management Agency (RMA) has been the central repository for the distribution and management of language resources, data and software tools, for the official languages of South Africa. The RMA has provided an excellent foundation for SADiLaR to build on. With the knowledge and skills obtained from the RMA project, we are now ready to advance to a new phase, and integrate the RMA into an international platform that will have not only national, but also global impact.

SADiLaR provides a platform to access linguistic data and reuse this data, while also offering researchers technologies and software to simplify linguistic analysis. The main distribution channel for resources will be a repository that allows interested parties to access any of the language resources distributed by SADiLaR. "The repository will also link to larger international infrastructures and language distribution agencies, such as the European Language Resource Association (ELRA) and CLARIN in Europe, and the Language Data Consortium (LDC) in the USA," says Dr Roald Eiselen, SADiLaR's technical manager.

The move to an institutional repository for the distribution of language resources has primarily been done for the following four reasons:

1. to simplify the access and download procedures for users by moving away from the "shopping cart" experience;
2. to provide all resources with a digital object identifier (DOI), which is integrated into the international digital handle system;
3. to allow easy integration of the data resources into other repositories and data infrastructures, such as CLARIN and LDC; and
4. as a first step in the process of getting the "data seal of approval" for SADiLaR.

SADiLaR will also make available several research enabling technologies such as:

- metadata and data processing infrastructures that are specifically linked to particular projects;
- general language data analytic platforms made available online; and
- automatic language analysis modules that support the development of more complex language technologies.

Although a substantial number of open-source technologies are reused and adapted to the South African context, several of the technologies and services that are being developed will be new technologies that will be distributed for further use by language communities both in Africa and around the world.
Over the coming year, SADiLaR will expand the set of available language resources on an ongoing basis, while also extending the set of automatic analysis tools that are available via web interfaces. It is expected that these technologies will enable end-users to more easily analyse their own linguistic data, or search and analyse the data available from SADiLaR.

A collaborative approach: Nodes @ SADiLaR

FUNDED BY:
science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

One of the benefits of SADiLaR is the collaborative approach to research and development, owing to the fact that various nodes (institutions and universities) work on different projects that contribute to SADiLaR's main vision of developing language resources for the South African languages.

In every edition of the newsletter, we will profile two of the nodes that form part of SADiLaR, and will provide you with an overview of the unit and the projects they are currently working on.

CSIR Meraka Institute (HLT Research Group)

The CSIR Meraka Institute focuses on shaping South Africa's digital future and is known for the research, development and innovation in the information and communication technology sector. Within the Institute, the Human Language Technology (HLT) Research Group focuses on solving communication challenges that South Africa face as a result of the lack of language resources and data. The Research Group delivers text-to-speech, automatic speech recognition and human language analytics to support the government's service delivery and provide access to information. This, in turn, facilitates smarter decision-making.

On 1 July 2017, the CSIR node formally commenced the project entitled "Human Language Technology Audit 2017/2018". The aim of the project is to update the previous HLT audit of 2009 that was conducted by Ms A Sharma-Grover as part of her research theses. The main reason for undertaking the updated HLT audit is the increase in HLT research and development activity in South Africa, specifically an increase in the number of institutions conducting HLT research and development. The HLT audit will provide SADiLaR with updated information on HLT components (software and models) and general language resources (data) for the South African languages, both in South Africa, and internationally.

The project consists of a number of phases, including the audit design, audit instrument development, audit execution, dynamic audit updates and audit results consolidation and reporting.

Two of the phases have been completed, namely:

1. the audit design, including research into audit methodologies, previous audits in the field as well as workshops with HLT experts; and
2. the audit instrument that was developed as an online tool using open source software which is easy to transfer to SADiLaR once the audit execution has been completed.

The HLT audit went live in December 2017 and a number of experts in the HLT community were invited to participate in the project.

The CSIR has been monitoring the progress of the audit responses and is continuing with the dynamic audit updates work. As part of the project, the node tested various accessibility tools in

order to extend the audit to a wider community. These accessibility tools were found to be compatible with most functions of the online audit. Furthermore, the CSIR node has continued to do research on the dynamic audit updates, which is similar to work by CLARIN and the Language Resource and Evaluation map. The HLT audit will deliver valuable data and resources that will all be presented to SADiLaR and to the broader HLT community.

FUNDED BY:

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

University of South Africa (UNISA): Department of African Languages

The Department of African Languages at UNISA is committed to the promotion, development and use of the South African languages. The Department conducts research that contributes to the advancement of knowledge of African languages, promotes scholarship in African languages and serves as a partner by reaching out to the community through its expertise in African languages.

The UNISA node is currently working on two subprojects that will provide valuable language resources for the empowerment of the African Languages within SADiLaR.

African Wordnet (AWN) subproject:

This subproject is a continuation of three previous phases of Wordnet development for five African languages that was done as a collaborative project between the Department of African Languages at UNISA and CTexT at the North-West University.

During the next phase of the project, the development team will focus on adding new synsets, usage examples, and definitions to the current version of the AWN. As a basis for this extension of the Wordnets, this phase of the project will use the SIL Comparative African Wordlist (see https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf.Snider_silewp2006-005.pdf) in order to ensure comparable development in all five languages. Representation of all 1 700 words in this list in the AWN will ensure a broad coverage of concepts shared among most African languages. This will lead to a useful resource not only for comparative purposes but also for language learners, and in particular for further development of other tools in the digital humanities.

Linguistic terminology subproject:

Specialised, multilingual terminology lists play a vital role in the South African environment, by enabling users to find very specific information in a structured, user-friendly manner. This subproject aims to create a term bank to encapsulate linguistic terminology often used in academic texts. The idea was conceptualised from the realisation that there is a wealth of existing linguistic terminology in African languages that is contained in discontinued and out-of-print hard copy publications. Unfortunately, their current format restricts accessibility to the knowledge represented in them. There is a pressing need to preserve and develop electronic versions of these resources to improve their accessibility.

This subproject is based on the incomplete SAMTerm (multilingual linguistics terminology database) which was conceptualised mainly for the purpose of teaching and learning in the Department of African Languages at UNISA, but also for use by researchers and other language practitioners such as translators. The subproject will be linked to the Open Educational Resource Term Bank of the University of Pretoria (UP) and will also benefit from the digitisation and term extraction capacity of UP.

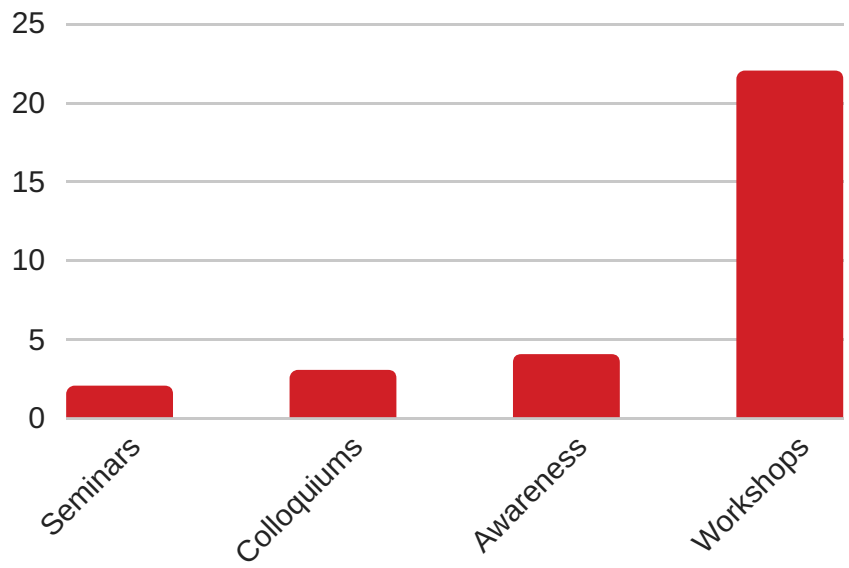
Capacity building initiatives 2017/2018

FUNDED BY:
science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

SADiLaR's Digital Humanities (DH) program facilitates research capacity building by promoting and supporting the use of innovative methodological approaches within the Humanities and Social Sciences. One way of reaching this outcome is by presenting regular workshops on the different approaches, ideas and tools to the broader research community in South Africa.

As a new research infrastructure, SADiLaR is grateful for the various institutions and bodies who have collaborated with us on this venture in support of the DH program.

31 Initiatives



More than 300 participants reached

> 300

www.sadilar.org