SADiLaR
South African Centre for Digital Language Resources

Global Virtual Forum Summer School

Overview of SADiLaR's resources and services supporting African languages and digital humanities research.

Andiswa Bukula
South African Centre for Digital Language Resources
01 July 2024

Funded by:

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Overview

- South African Research Infrastructure Roadmap (SARIR)
- SADiLaR structure
- SADiLaR Programmes
- Research and development projects
- Support and capacity building
- Impact
- Strategic medium-term planning
- Demonstrations

SA Research Infrastructure Roadmap -SARIR

- National Department of Science and Innovation funded initiative
- Support the establishment and running of research infrastructures
  - Facilities, resources, and services
  - For the scientific community
  - Generate, exchange, and preserve knowledge
  - Strong support for notions of open data and open access
- 13 Ris established covering 5 scientific domains
- SADiLaR is the only one focusing primarily on the **Humanities**

SADiLaR: Structure

- Hosted at the North-West University, Potchefstroom, Main Hub.
- Currently 6 participating nodes in different areas of specialisation
  - University of Pretoria – Digitisation
  - University of South Africa – Semantics and terminology
  - ICELDA – Language development and teaching
  - CSIR – Speech resources and technologies
  - CTexT, NWU – Text resources and technologies
  - University of Stellenbosch – Child Language Development
- International representatives in the Management structure
  - CLARIN & ELRA (Europe)
  - Linguistic Data Consortium (USA)

SADiLaR
South African Centre for Digital Language Resources

SADiLaR: Programmes

- SADiLaR: Currently in the middle of the first cycle of 15 years

- The next phase will further build on the established infrastructure, digital language resource management and advancing academic scholarship activities through applied research through three programmes:

    - Digitisation programme
    - Digital Humanities programme
    - Emerging programme – Playing an enabling role towards language policy capacity building, development and implementation

- New strategy approved:

Ensuring a digital future for all official languages in South Africa

SADiLaR
South African Centre for Digital Language Resources

Digitisation programme

- Pro-active effort to extend language resources in official SA languages
- Node projects to develop language resources
  - Digital text and speech corpora from existing, non-digital resources (UP)
  - Wordnets and terminology development (UNISA)
  - Multilingual L2 learner corpus of academic writing and language specific academic literacy testing (ICELDA)
  - Transcribed speech corpora, automatic speech corpus collection and computational grammars (CSIR)
  - Annotated text corpus creation for conjunctive languages (CTexT)
  - Communicative inventories for South African Languages (SUN)
- Digitisation support to external institutions
- Provide assistance in best practices and digitisation efforts and software development

Digital Humanities programme (1)

- The Digital Humanities programme focuses on facilitating the building of research capacity by promoting and supporting the use of digital data and innovative methodological approaches within the Humanities and Social Sciences.

- The ESCALATOR programme stands central to all human capacity development taking place under the DH programme

- More integrated programme – less working / training in silos
  - Work towards an inclusive and active South African community of practice in Digital Humanities and Computational Social Science.
  - Positive effects through broader engagement: Establishment of new DH Centres at

    Rhodes University; Nelson Mandela University and North-West University

Digital Humanities programme (2)

- ESCALATOR programme has three main activities
  - Multiple training/community tracks part of the **DIGITAL TRAINING & CHAMPIONS** initiative
  - **DH-IGNITEs** events continue to function as an important vehicle to bring together regions
  - **STAKEHOLDER MAP** and dedicated slack channel – support and facilitate access to expert practitioners in the field.

- Further highlights include
  - Multiple collaborations that saw mentorship training taking place; general computation training through CODATA-RDA School of Research Data Science and four conferences of the Digital Humanities of Southern Africa Association.
  - The establishment of a Journal for DHASA
  - Some activities saw also broader recognition such as the the N|uu dictionary project that recently won an award.

Support and capacity building

- Linking to the ESACALATOR programme lies a contribution towards general Open Science awareness and capacity building. This is an important point for the South African government to foster a science aware and science literate society.

- Core to this stands SADiLaR's SWiP project - "Preserving Languages: Open, Free and Accessible Knowledge for All"

- SADiLaR been involved in Open Science and Data awareness as part of its contractual milestones in the past. We aim to transition this to a more focused initiative through existing networks and platform by starting a regular, at least yearly Data Stewardship Summer School to strengthen these national efforts.

Emerging programme: Language policy; implementation

- SADiLaR was instrumental to conduct a Language Resources Audit towards the implementation of the New Language Policy Framework for Higher Education Institutions in South Africa.
- SADiLaR approved a post audit secondment programme to continue with activities emerging from the audit. Main findings are available on our repository.
- Stemming from these activities SADiLaR has also been named as a **key partner** in the draft of the national terminology policy.
- The non-competing nature of the RI with a national focus in the national interest of the research but also public sector creates opportunities.

Research & development projects

SADiLaR runs two main types of projects.

- The first is **larger specialisation projects** such as the creation of parallel datasets for machine translation systems run by respective SADiLaR nodes.
- The second is **open call projects**. These allow general researchers to apply for funding towards specific activities that contribute to SADiLaRs broad mandate. It is expected that a more focussed collaboration driven round will be announced towards the end of 2024.

- Some projects are “hub” driven focusing on core infrastructure developments or capacity building initiatives
  - Currently this includes the creation of a **terminology aggregation platform** as well as the **ESCALATOR** programme.

Impact areas within SADiLaR and its Nodes

CSIR – speech resources and technologies node

- The continuous growth of speech resources – through (semi-)automatic harvesting of existing sources of speech data to create resources – can be used to develop new and improve existing speech technologies.

- Two recently approved projects will in turn see the development of a wide coverage resource grammar that can be used to develop downstream applications in isiZulu.

- A second project, with the government communication services as a key partner, will see the first building blocks to be put in place to effectively create captions for official governmental speeches.

CTexT – text processing technologies and text resources node

Using specialisation project data in conjunction with the existing NCHLT datasets:

- **12 improved core technologies** were developed (lemmatisers, part of speech [POS] taggers and morphological decomposers). These core technologies can be used to improve existing and create new end-user software, such as the existing machine translation systems, web services and spelling checkers.

- As part of a recently approved project, CTexT will complete the development of a **dictionary mobile application** that would place dictionary and terminology-related language resources in the hands of all South Africans.

- A **nationwide distribution of spelling and grammar checkers** is currently underway. The development of open-source tools, such as an open-source Apple and Android Dictionary app, allows for reuse by different domains

ICELDA – language development and testing node

The ICELDA node brings together expertise and resources from higher education institution partners across the country.

- It produced an **academic writing tool** through core collaborative assistance by KU Leuven to guide student writing, and in turn, to improve academic literacy in South Africa.

- **Node-generated learner corpora** and annotated learner corpora for future research and product development. These are already used by PhD and MA students, and by academics for research and product development.

- **A multilingual generic academic wordlist** has also been completed, including POS, definitions and examples in all official languages. These wordlists will be distributed through an online platform with the assistance of the hub.

- **Tests of academic language ability** for use in secondary and higher education have been developed and translated into indigenous languages.

- **Academic writing assistance to university students** including educational videos in all official languages as well as worksheets and exercises in English (following the principles of translanguaging where English is the source while being supported by another language as resource).

- Full description

SADiLaR
South African Centre for Digital Language Resources

UNISA – wordnet and terminology development node

- African Wordnet development
  - 10 000 Synsets and definitions for five languages: Setswana, isiZulu, isiXhosa, Sepedi, Tshivenda
  - New Wordnets for isiNdebele, Xitsonga, Siswati, Sesotho – first 1 000 synsets now included
  - AfWN data for localisation of Mozilla Common Voice for SA languages
  - Co-host of the Global Wordnet Conference in 2021

- Terminology development
  - 500 linguistic terms frequently used in university classrooms for nine languages added to the Lexonomy platform
  - Literary terms in progress

- Full description

University of Pretoria (UP) – digitisation node

- The primary impact of the datasets compiled is to serve as a digital resource for research in African languages, contributing to the intellectualisation of these languages.

- Furthermore, the datasets underpin the development of a variety of HLTs which potentially could enable all South Africans to participate in various spheres of civil society in their strongest language. Examples of these are language applications (apps), machine translation, speech recognition and speech-to-text transcription.

- Full description

Stellenbosch University – child language development node

- The main function of the child language development node is to **promote research on child language development in all South African languages**

- **Digitisation of child language development data** so that it is freely available on the SADiLaR platform for all scientists working on language, cognition, child development, language learning and language disorders.

- The scientific and applied aims of the node are to **advance knowledge about children's language development in African languages.**

- Data on African languages to inform the **development of valid diagnostic tools and interventions** to promote the language and cognitive development of South Africa's children in health and educational settings.

- Full description

SADiLaR
South African Centre for Digital Language Resources

Impact

- **Long-term digital preservation and legal distribution of resources** still lies at the central to SADiLaR longer term goals.

- In terms of **capacity building and awareness** workshop and events SADiLaR has been involved in nearly **200 such activities reaching more than 5300 participants** across activities.

- Since the inception of SADiLaR - the addition of more than **150 new language resources as part of its repository.**
- SADiLaR's role as a non-competing entity has the **potential to enable long term impact in South Africa** to help ensure that isolated initiatives are linked to broader activities.

Strategic medium-term planning

- Focusing on systematically executing on the objectives of the new strategy
- Support for individual research communities – open science awareness and open call projects
- Expanding offering of language resources through specialization projects
- Integration into international and regional infrastructures
- Updates and maintenance to current technologies and Integration of technologies into toolchains to allow users to process and search their own data in the infrastructure
- Create dissemination pathways for project outputs and broader research community

Demonstrations

- SADiLaR.org
- Repository
- Corpus portal
- Technologies
  - NCHLT Web Apps
  - Autshumato
  - Qfrency
- Language portals
  - The African Language Grammar Portal

Contact details

Web: sadilar.org

Email: info@sadilar.org

X/Twitter: SADiLaR_ZA

Facebook: South African Centre for Digital Language Resources

SADiLaR
South African Centre for Digital Language Resources