SADiLaR

South African Centre for Digital Language Resources

Language resources as enablers

Professor Langa Khumalo
Executive Director: SADiLaR
12 September 2024

science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

The importance of language

- Humanizes us
- Organizes and expresses our thoughts and spurs our innovation
- Allows us to think more complex thoughts
- We speak therefore we are!

The importance of language

- Central to all forms of learning
- Enables knowledge generation and its dissemination

SADiLaR
South African Centre for Digital Language Resources

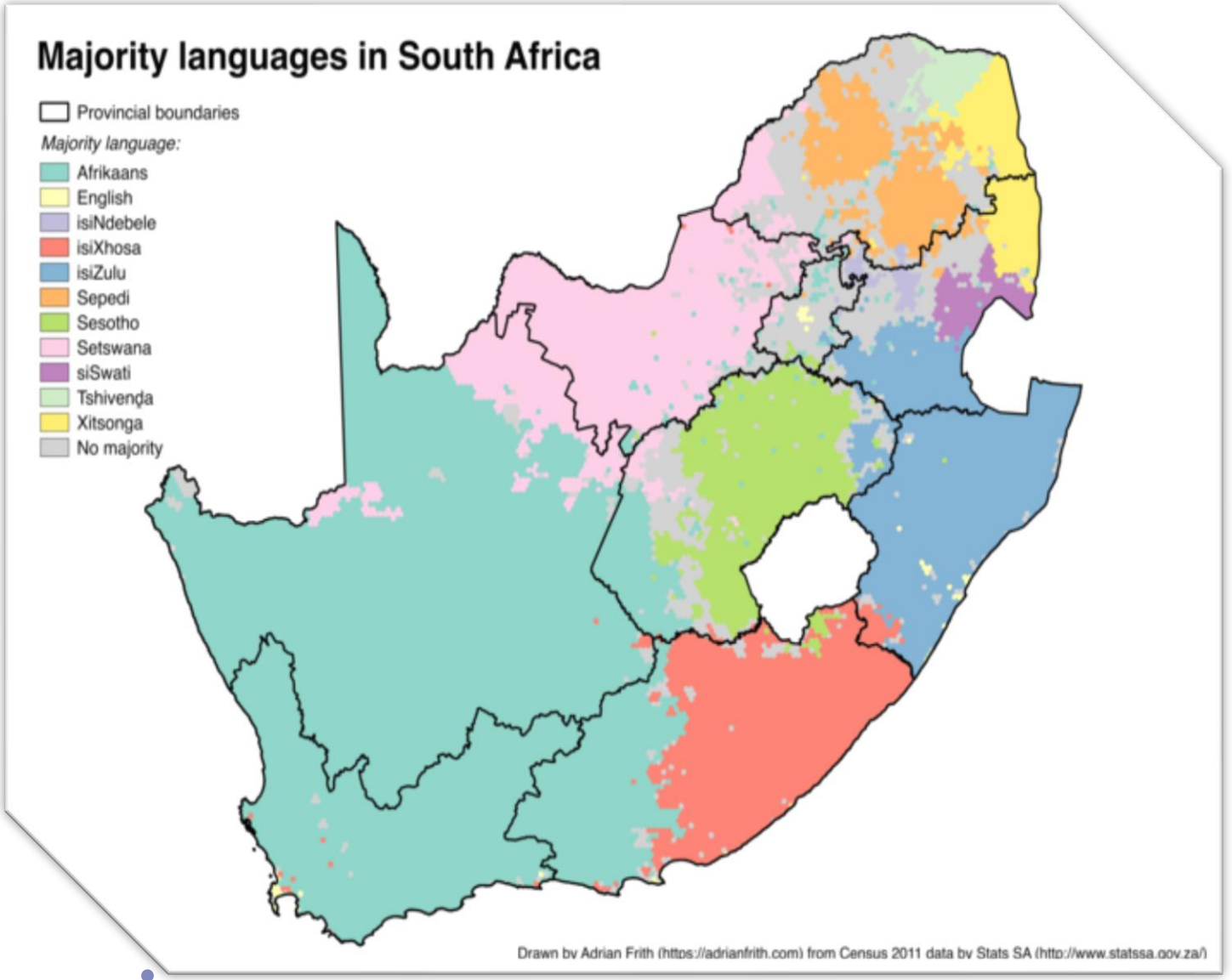
The South Africa Research Infrastructure Roadmap

- SADiLaR was launched in Sept 2016
- As part of SARIR
- 13 RIs established covering 5 scientific domains
- SADiLaR is the only one focusing primarily on the **Humanities**

SADiLaR
South African Centre for Digital Language Resources

Constitutional mandate

- 12 official languages
  - Conjunctively written Nguni languages (isiNdebele, siSwati, isiXhosa, isiZulu)
  - Disjunctively written Sotho languages (Setswana, Sesotho, Sepedi)
  - Disjunctive Xitsonga & Tshivenḓa
  - Afrikaans and English
  - South African Sign Language – official since 2023
- Constitutionally mandated that all languages are treated equally.
- In practice, mostly English.



SADiLaR
South African Centre for Digital Language Resources

Language resource development

- Various HLT/NLP development efforts since turn of the century
  - Universities of Stellenbosch, Pretoria, South Africa, Cape Town (UCT), Johannesburg (WITS)
  - Council for Scientific and Industrial Research (CSIR)
  - Centre for Text Technology (CTexT®)
  - Multilingual Speech Technologies (MUST) @ North-West University
  - African Speech Technology Project (AST) 2000-2004

- Significant progress but
  - Still under-resourced - limited knowledge base

SADiLaR
South African Centre for Digital Language Resources

Low-resource languages

- Notably African languages have low-resources
- This is in terms exhaustive linguistic descriptions, large and specialized corpus resources and machine-readable lexicons
- HLTs are sparse and other computational resources
- Expertise and funding resources also notably low

(Bosch et al. 2007; Pretorius & Bosch 2003; Keet & Khumalo 2014).

SADiLaR: Structure

❖Hub and spoke model, hosted at the North-West University, Potchefstroom

❖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) | Accepted as a full member of CLARIN-ERIC (01.01.2024)
- Linguistic Data Consortium (USA)

CLARIN ERIC
Common Language Resources and Technology Infrastructure

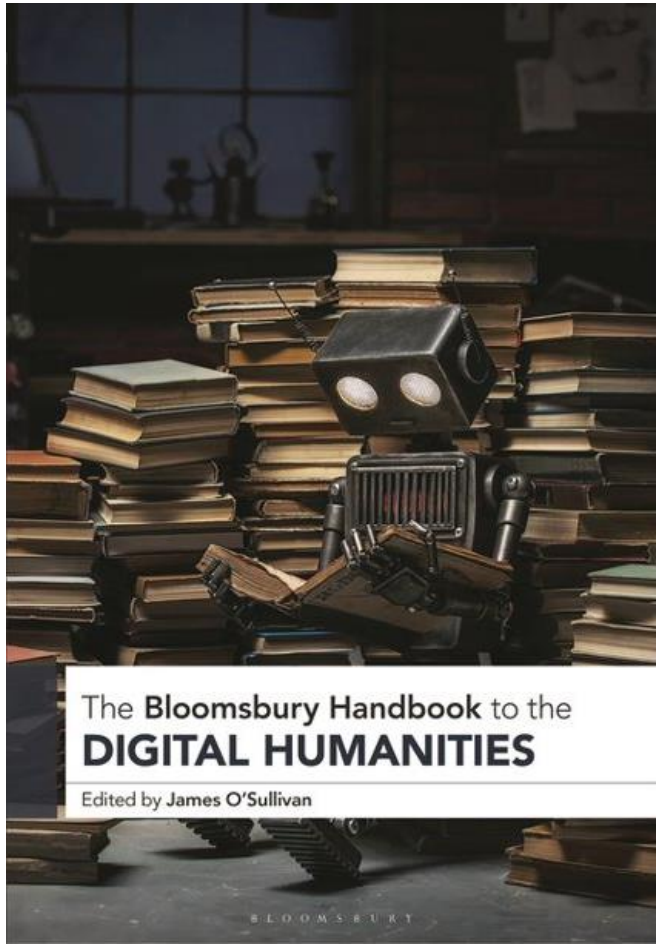
SADiLaR
South African Centre for Digital Language Resources

Digitisation Programme

- 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

- The ESCALATOR Project
  - CoP in DH & CSocSci
  - DH methods in HSS
- The SWiP Project
  - Digital presence
  - Open science



- DH colloquium series
  - Scholarship
  - Training & data sharing
- Data stewardship
  - Capacity building
  - Data curation | management

SADiLaR
South African Centre for Digital Language Resources

Conclusion

- Language resources are key enablers
- Enable languages to be used in all knowledge domains
- Visible in digital | cyberinfrastructure
- NLP and other ML (AI) systems

SADiLaR
South African Centre for Digital Language Resources

science & innovation

Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

SADiLaR
South African Centre for Digital Language Resources

Ngiyabonga
Enkosi
Kea leboga
Dankie
Thank you