Multilingualism: A case of the South African Centre for Digital Language Resources in developing language resources

Rooweither Mabuya

10 April 2024

Embracing Multilingualism

- South Africa's approach to embracing and promoting multilingualism stands in contrast to the policies of many other African nations.

- My passion for multilingualism stems from a desire to preserve and promote our rich cultural heritage.

"If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart."
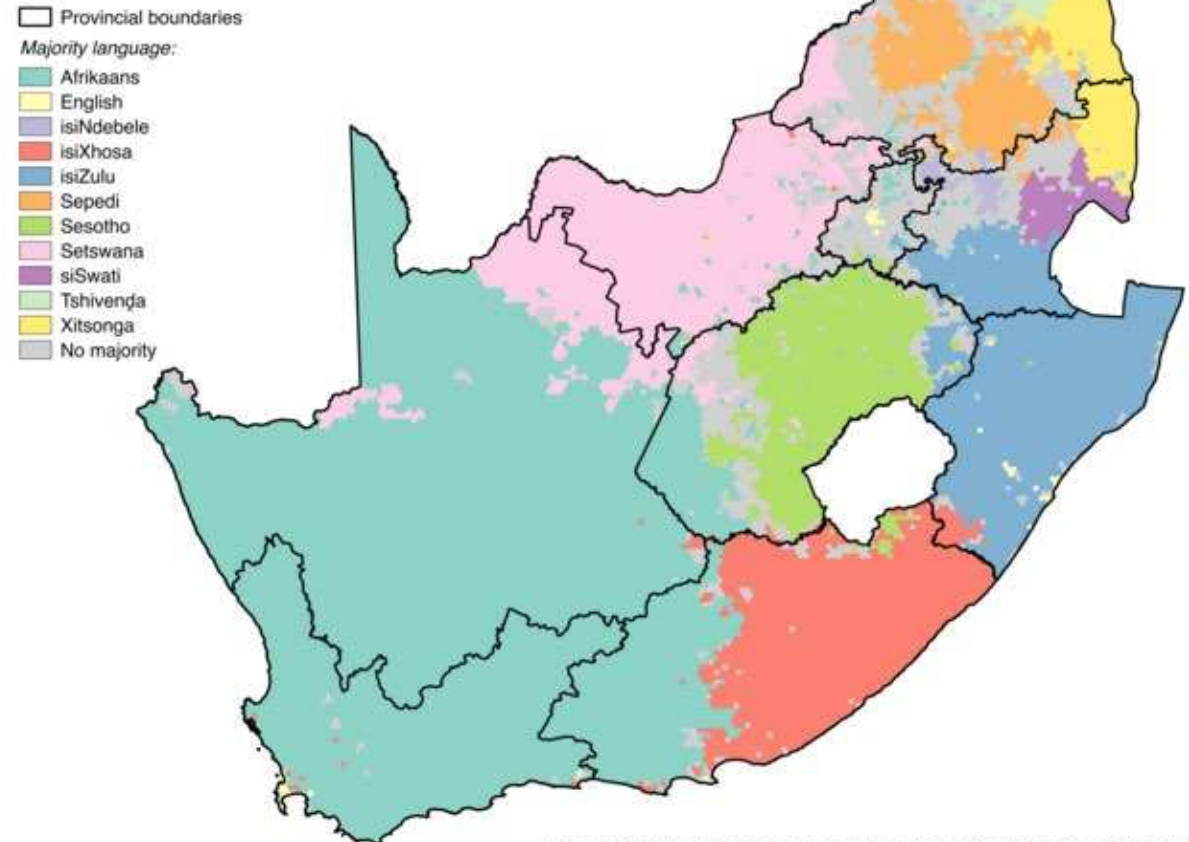
Nelson Mandela

The Significance of Multilingualism in South Africa

- 12 official languages
  - **Nguni languages**
    (IsiNdebele, Siswati, isiXhosa, isiZulu)
  - **Sotho languages**
    (Setswana, Sesotho, Sesotho sa Leboa)
  - Xitsonga, Tshivenda, Afrikaans, English and Sign Language
- Constitutionally mandated that all languages be treated equally
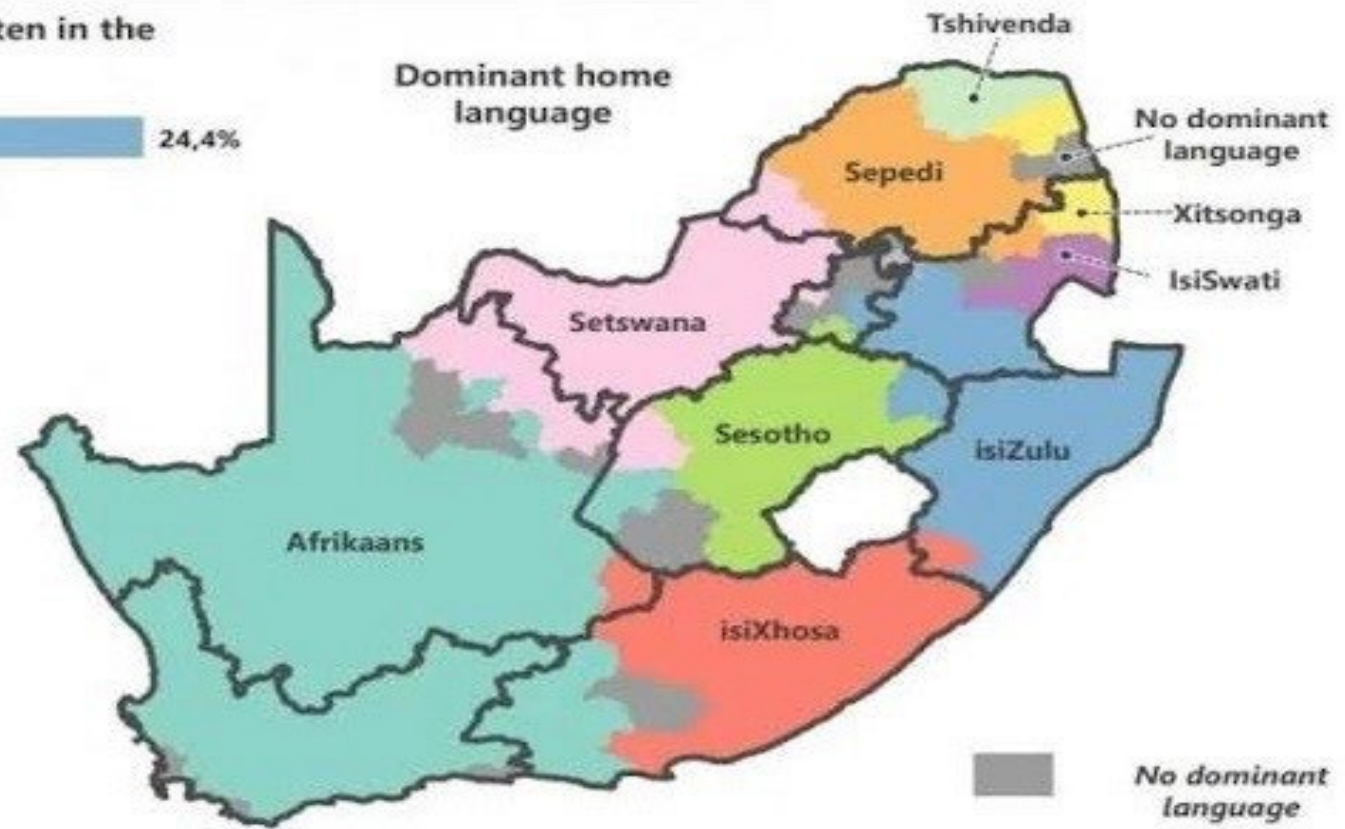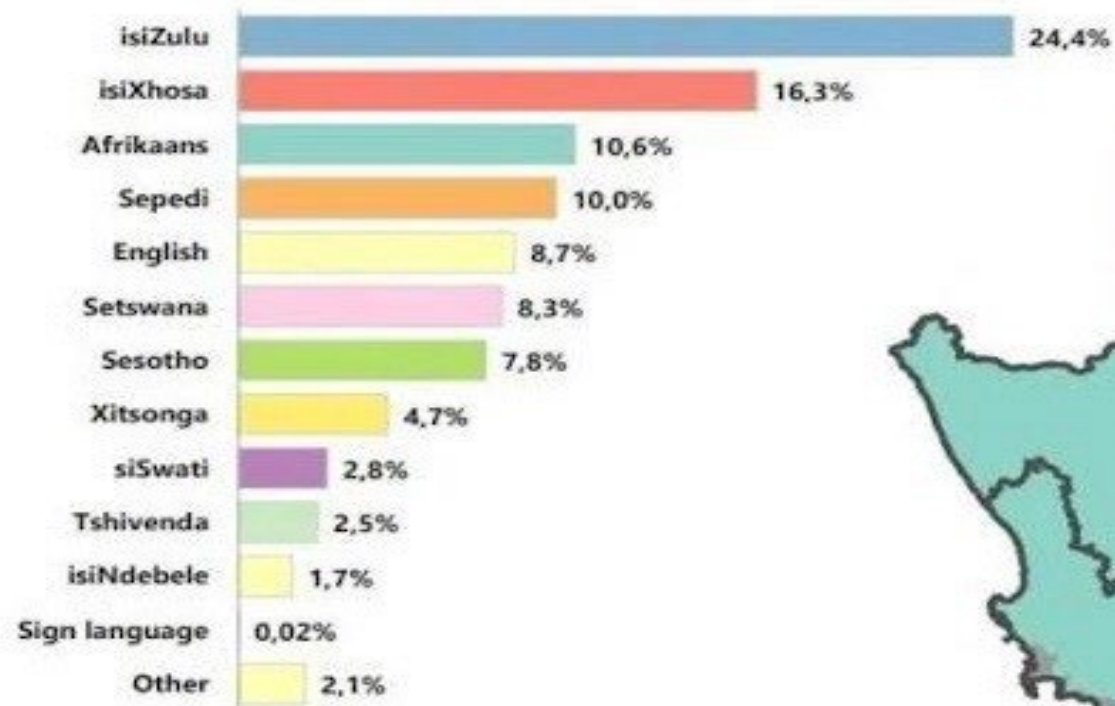
Majority languages in South Africa

☐ Provincial boundaries

Majority language:
- Afrikaans
- English
- isiNdebele
- isiXhosa
- isiZulu
- Sepedi
- Sesotho
- Setswana
- siSwati
- Tshivenda
- Xitsonga
- No majority

Drawn by Adrian Frith (https://adrianfrith.com) from Census 2011 data by Stats SA (http://www.statssa.gov.za/).

The Significance of Multilingualism in South Africa

**Population distribution by language spoken most often in the household, 2022**

- isiZulu — 24,4%
- isiXhosa — 16,3%
- Afrikaans — 10,6%
- Sepedi — 10,0%
- English — 8,7%
- Setswana — 8,3%
- Sesotho — 7,8%
- Xitsonga — 4,7%
- siSwati — 2,8%
- Tshivenda — 2,5%
- isiNdebele — 1,7%
- Sign language — 0,02%
- Other — 2,1%

Dominant home language

Tshivenda
No dominant language
Sepedi
Xitsonga
IsiSwati
Setswana
Sesotho
isiZulu
Afrikaans
isiXhosa

No dominant language

Source: Census 2022 Stats SA (http://www.statssa.gov.za)

Challenges in Language Resource Development

- Scarcity of computational resources for languages other than English, hindering research and technological advancements.

- Limited availability and quality of resources for South African languages, impacting natural language processing tasks.
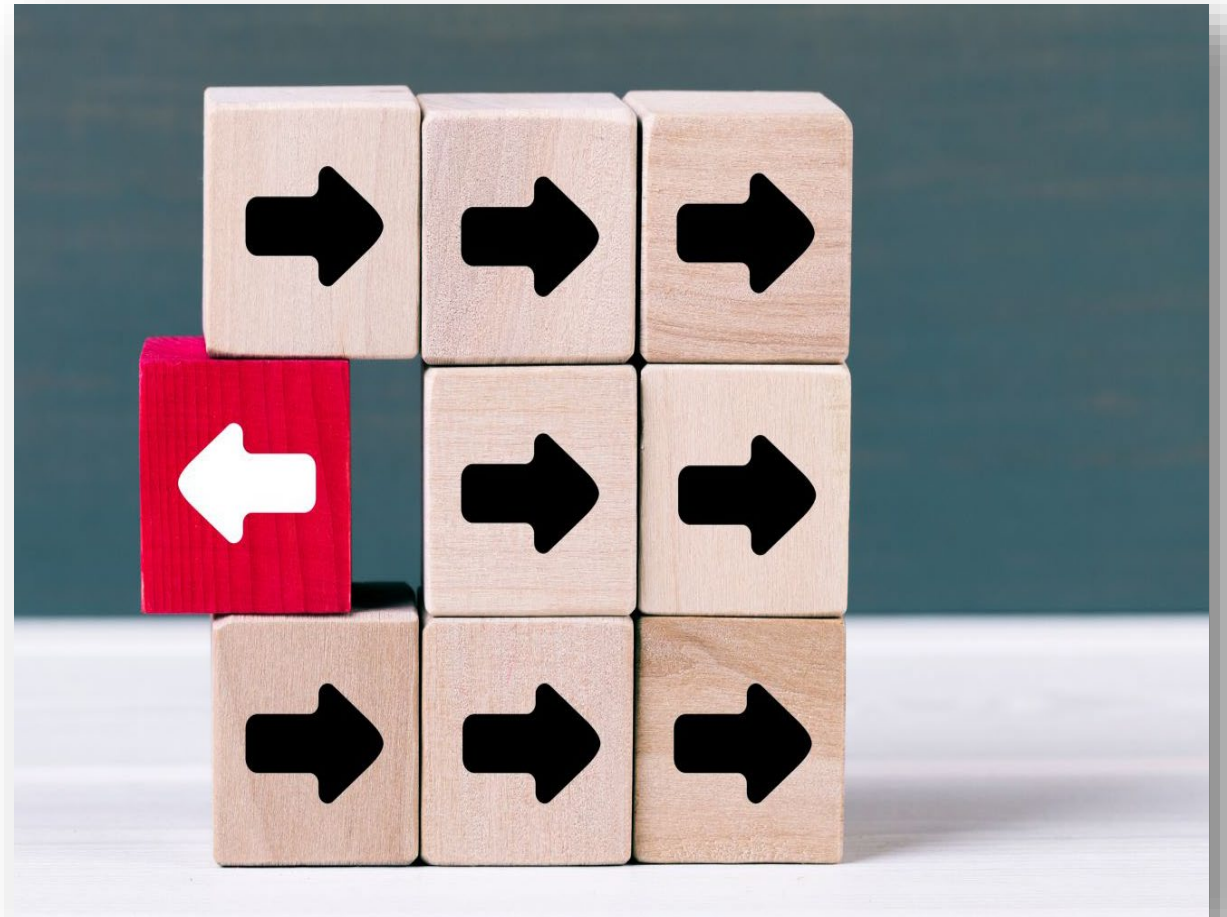
Challenges in Language Resource Development

English: above
Afrikaans: bo
Synonym: hierbo
Synonym: bokant
IsiZulu: ngaphezulu
IsiXhosa: ngaphezulu
Siswati: etulu
IsiNdebele: ngehla
Setswana: - godimo
Sepedi: ka godimo
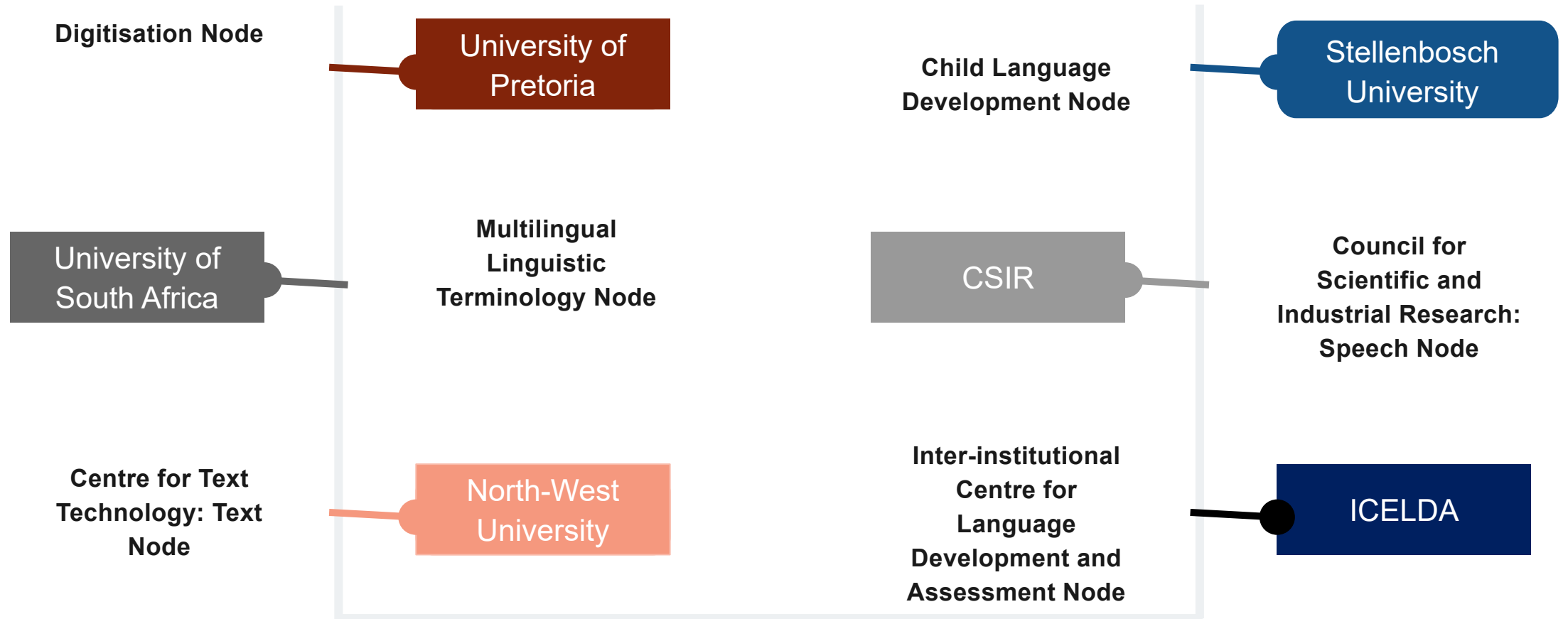Sesotho: ka hodima
Tshivenda: nṱha ha
Xitsonga: henhla ka

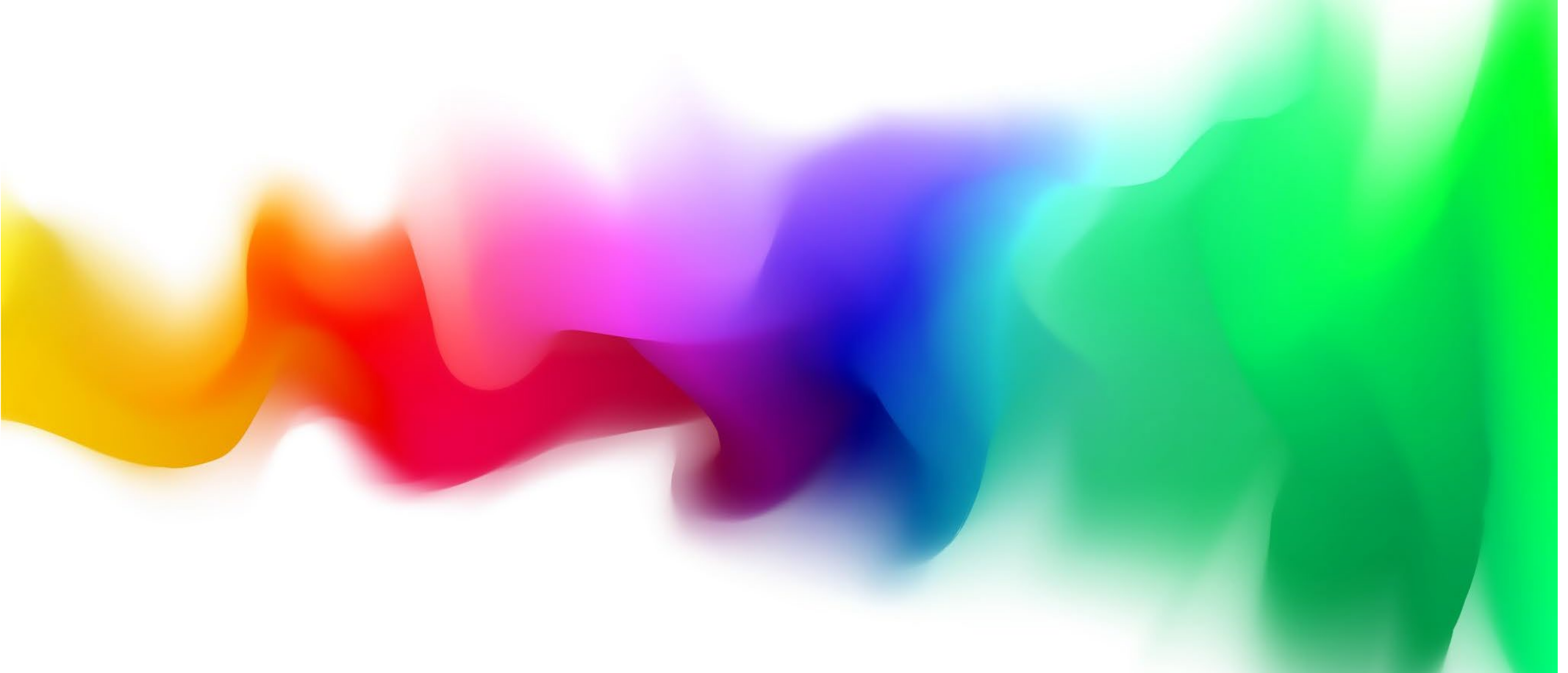# Introduction to South African Centre for Digital Language Resources

Introduction to SADiLaR

Digitisation Node

University of Pretoria

Child Language Development Node

Stellenbosch University

Multilingual Linguistic Terminology Node

University of South Africa

CSIR

Council for Scientific and Industrial Research: Speech Node

Centre for Text Technology: Text Node

North-West University

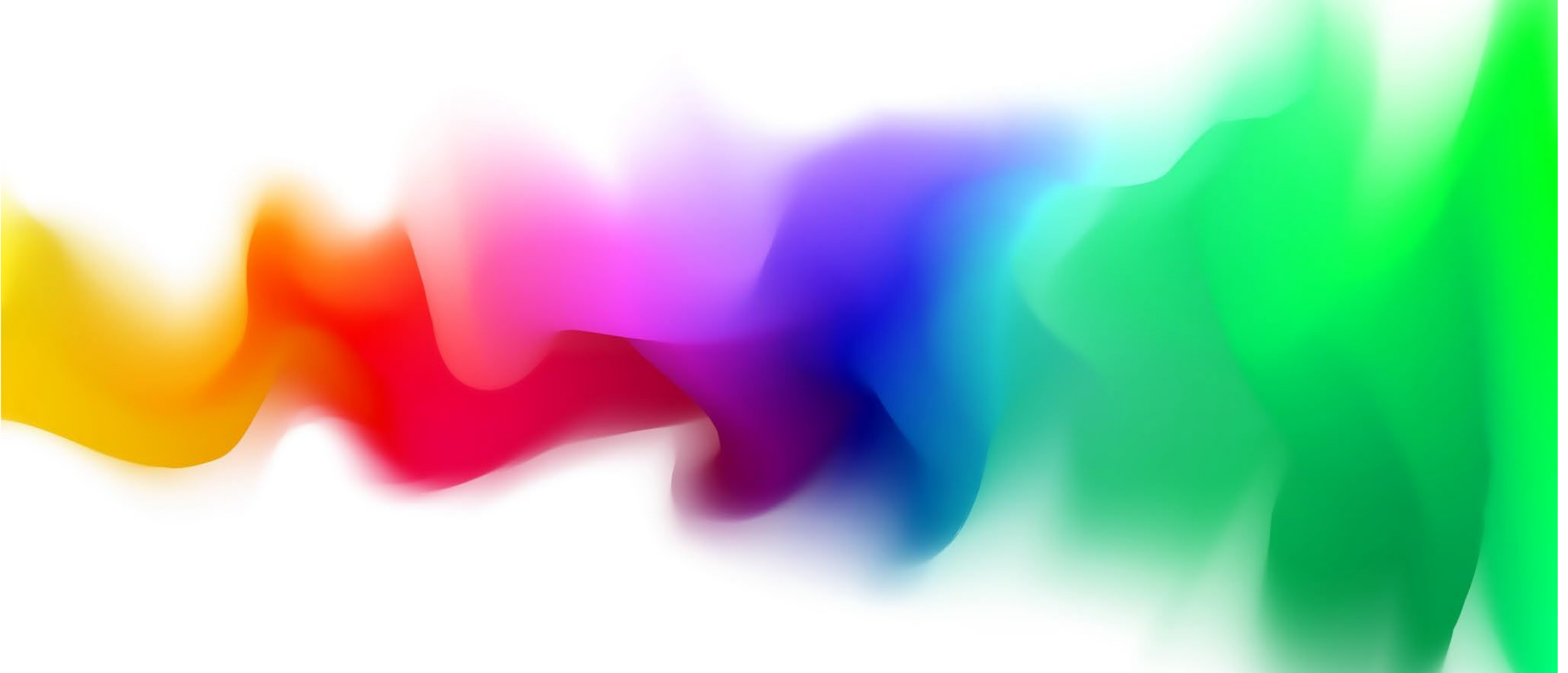**Inter-institutional Centre for Language Development and Assessment Node**

ICELDA

Our Background

What do we do?

SADiLaR's Approach to Language Resource Development

- **Establishment:** This marked a significant step towards addressing the scarcity of resources for indigenous languages.

- **Funding:** The Department of Science and Innovation (DSI) provides crucial funding for SADiLaR's operations.

- **Programs:** SADiLaR's digitisation and digital humanities programs are at the core of its efforts. These programs focus on converting linguistic data collections into digital format.

Case Studies: SADiLaR's Impact

Curation, development, distribution, and maintenance of language resources

Language Resource Index

- Various public and private institutions, both nationally and internationally
- All resources are available for download
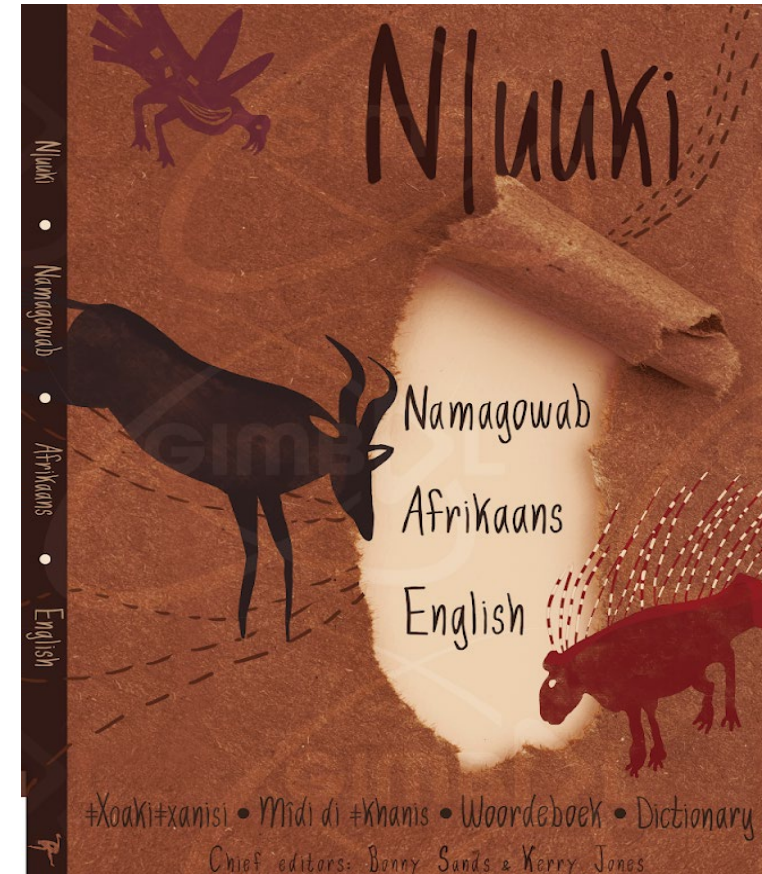
Language Resource Catalogue

- A digital collection of language resources, in various modalities, that are available for download
- Upload metadata and resources via the repository

General technologies

- A group of developed and hosted services - corpus searches, automatic linguistic processing and machine translation

Case Studies: SADiLaR's Impact

Collaboration and Partnerships

Language Resources Audit:

Implementation of the Department of Higher Education and Training Language Policy Framework 2022.

Conducted Audit in 26 public universities in 2023.

SADiLaR-Wikipedia-PanSALB:

The SWiP project is aimed at promoting all of South Africa's indigenous languages online.

To encourage language communities to actively contribute content to the free encyclopedia (Wikipedia).

Conclusion

Development of digital language resources is crucial for indigenous languages.

These resources enable full functionality and intellectualization of indigenous languages.

They help to preserve cultural heritage and to promote inclusivity and diversity.