# South African Centre for Digital Language Resources

**Justus C Roux**
Research Unit: Languages and Literature in the South African Context
North-West University
Potchefstroom Campus
South Africa
justus.roux@nwu.ac.za

**Abstract**

This presentation introduces the imminent establishment of a new language resource infrastructure particularly focusing on languages spoken in Southern Africa, but with an eventual aim to become a hub for digital language resources within Sub-Saharan Africa. The Constitution of South Africa makes provision for 11 official languages, with equal status although they differ significantly with regard to the number of speakers. The current language Resource Management Agency (RMA) will be merged with the new Centre, which will have a much wider focus than that of data acquisition, management and distribution. The Centre (SADiLaR) will entertain two main programs: Digitisation and Digital Humanities. The digitisation program will focus on the systematic digitisation of relevant text, speech and multi-modal data across the official languages. Relevancy will be determined by a Scientific Advisory Board. This will take place on a continuous basis through specified projects allocated to national members of the Centre, as well as through open-calls aimed at the academic as well as local communities. The digital resources will be enhanced, managed and distributed through a dedicated web-based portal. The development of the Digital Humanities (DH) program will entail extensive academic support for projects implementing digital language based data. SADiLaR will function as an enabling research infrastructure primarily supported by national government.

**Keywords**: infrastructure, digitisation, African languages

## 1. Introduction

The Constitution of South Africa makes provision for 11 official languages, which comprise English, Afrikaans, isiZulu, isiXhosa, Siswati, isiNdebele, Sesotho, Setswana, Sesotho sa Lebowa (Northern Sotho), Tshivenda and Xitsonga. Figure 1 below depicts the balance between the different languages:

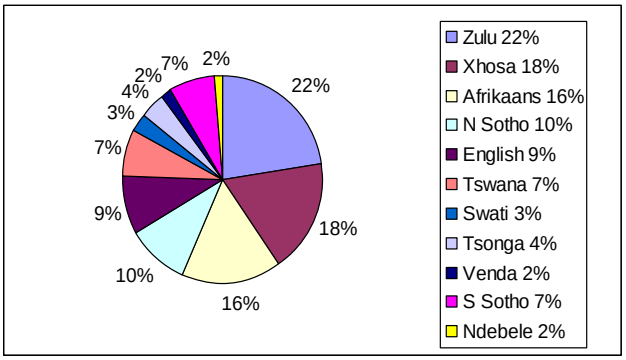


Figure 1 Home languages in South Africa (n=54 mil.)

Apart from English, all of the other languages are regarded as resource scarce languages. This obviously calls for actions to develop resources for these languages, preferably in a coordinated and systematic manner.

## 2. Building a resource infrastructure

In 2008 a proposal to establish a National Centre for Human Technologies (HLT) was approved by Cabinet of National Government with the intent, *inter alia*, to develop applicable language resources. In practice, this became a virtual (administrative) centre under the auspices of the HLT Unit within the Department of Arts and Culture (DAC). A ministerial expert panel on HLT (HLTEP) currently assists the HLT Unit to develop strategic plans and fund development projects in the HLT domain. As funding for HLT projects became available from the DAC and also from the Department of Science and Technology (DST), the necessity for a **central repository** for reusable language resources became crucial. In 2012 the DAC then set up a language *Resource Management Agency* (**RMA**) which is currently being run by the Centre for Text Technology (CTexT©) at the North-West University in Potchefstroom (http://www.rma.nwu.ac.za).

Since its inception four years ago, the RMA has shown significant growth in the number of clients and in the distribution of language resources, including applicable software for various applications. This includes:

- **258 registered users** - 157 from SA, 7 from 5 African countries, and 94 from 16 other countries worldwide
- a total resource **download of 1141 items**, i.e. resources, development platforms, and software.

Given practical considerations and the scale of digital data needed to create a space in the global digital domain for these resource scarce languages, it has become necessary to create a sustainable long term environment for the development of resources and applicable software.

In 2013 the Department of Science and Technology (DST) embarked on a project of redefining a South African Research Infrastructure Roadmap (SARIR), which eventually led to a proposal which was accepted in principle for the establishment of a National Centre for Digital Language Resources. Being on the national research infrastructure roadmap of the country implies long-term funding and support which was not previously possible. The South African Centre for Digital Language Resources (SADiLaR), with its new functions is likely to be formally established in due course of 2016.

## 3. Structure and function of SADiLaR

Structurally this new facility comprises a hub-and-spoke model linking three academic institutions, an inter-university partnership (including two more universities) and an institute of the Council for Scientific and Industrial Research (CSIR) as nodes, with the North-West University (NWU) as coordinator of an independent freestanding hub:

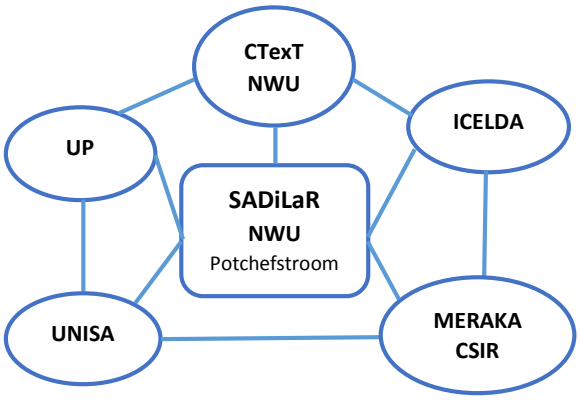


Figure 2 Structure of the Centre

The institutions referred to in more detail below, have taken the initial lead in setting up this national facility with the idea to create more nodes as activities increase and other institutions or groups join up. The SADiLaR hub will formally function as a hosted research entity within the NWU, and will have its own management structure, including a board representing existing and potential role players in the South African context, as well as an international scientific advisory committee (SAC).

The founding members of this initiative all have a record in the field of language resource and language technology developments and comprises:

The **Centre for Text Technology** (CTexT®) linked to the Research Unit for Language and Literatures in the South African Context at the NWU: This entity currently manages the RMA which will be integrated in SADiLaR. The Centre has been involved in major resource and software developments related to the official languages over the last decade as may be observed in the website, http://www.nwu.ac.za/ctext.

The **Department of African Languages at the University of South Africa (UNISA):** This department with its 30 full time staff members represent expertise in all nine official African languages of South Africa. Over and above language content capabilities, the department also has proven expertise in computational applications in African languages. UNISA being a correspondence university, is also extremely well positioned to provide training and upgrading of skills in various domains linked to this initiative.
(http://www.unisa.ac.za/Default.asp?Cmd=ViewContent&ContentID=143).

The **Department of African Languages at the University of Pretoria (UP):** Over at least the last two decades this department has been highly involved in different kinds of data collection particularly related to lexicographic and terminological developments in the African languages. The department has particular skills in statistical assessment of large data corpora which underpins the development dictionaries and domain specific terminologies. The content of its current language archive will easily be accessed through the envisaged national infrastructure, opening up the data to national and international communities on a more formal basis. (http://web.up.ac.za/default.asp?ipkCategoryID=482).

The **Human Language Technology Research Group** at the **Meraka Institute** of the Council of Scientific and Industrial Research (CSIR) is highly specialised in research and development of language technology applications in most of the South African languages. The group has an impressive record of academic publications and is very experienced in the acquisition and processing of digital speech data, as well as the development of applicable software (http://www.csir.co.za/meraka/hlt/).

The **Inter-institutional Centre for Language Development and Assessment (ICELDA) "…** is a partnership of four multilingual South African universities: Pretoria, Stellenbosch, North-West and Free State. ICELDA designs language tests, and supports research in language testing. It is the outcome of collaboration, since 2004, among the partnering universities." (http://icelda.sun.ac.za/). Members of this entity annually create large corpora of digital texts written by university students at different levels as part of academic literacy programs. This data, when made accessible on a national level will not only contribute to an understanding of the nature of language acquisition and performance, but also to the design of software for language testing and remediation.

The above mentioned formal partners (Fig. 2) will each have a senior researcher functioning as a node manager who will have an oversight function regarding the activities assigned to the particular node.

Given that it is foreseen that the Centre will have a life span of at least fifteen years, provision has been made for career paths for 11 researchers (representative of each language), a senior research manager and a technical manager, over and above positions for computational linguists and software developers under the supervision of an executive director. Provision is also made for internships and post-doctoral researchers to contribute to the digital development of the official languages of the country.

## 4. Focus of the Centre

Given that the Centre *inter alia* constitutes an upgrading of the current Resource Management Agency (RMA) with assigned new functions related to the development of Digital Humanities in South Africa, the Centre will focus on the following two programs:

**Digitisation program**: This program pertains to the systematic development of digital text, speech and multimodal resources in all official languages in a pro-active manner, informed by a Scientific Advisory Board that will comprise national and international experts. Resource development will take place at two levels: by

members of the nodes as agreed and overseen by the node managers, as well as by interested and suitably qualified institutions and/or individuals responding to open calls for participation. A percentage of the digitisation development budget will be allocated to approved proposals from the broader humanities and social science sector. Digitisation will commence with a focus on existing academically valuable non-digitised language related resources, followed by the systematic gathering of language data from various domains on a continuous basis. Over and above the digitisation projects run by the nodes, the systematic development of digital language resources will also be linked to post-graduate academic training where data obtained by scholars will be deposited in the Centre. For example, studies on topics related to the development of child language within, and across the official languages, will provide a wide range of digital speech data to be collated at the Centre. The same applies for data collected digitally in dialect, gender and sociolinguistic studies, as further examples. Ample funding and digital equipment will be available to post-graduate students from any university to gather data, which in turn will populate a matrix comprising languages on the X-level and language varieties/modes on the Y-level. The eventual availability of digital data in a populated matrix will enable new students to embark on related projects with at least some digital data as a point of reference. The use and application of these resources will be assessed as a long term metric. This program will be the responsibility of the Manager: Technology.

**Digital Humanities program**: Given the relatively slow uptake of digital methods and digital data in research in the humanities and social sciences (HSS) in South Africa, this Centre intends playing a major facilitating role in promoting DH practice in research. This will include training of researchers and students on methodological issues related to the use of digital tools, large data sets, the interpretation of results and the representation thereof in research projects. Although the main focus will be on language as the carrier of information, the initial scope of the training will include subjects and topics in the humanities and social sciences at large. This entails the implementation of dedicated awareness campaigns and training programs related to language technologies and/or digital methods in HSS research at tertiary institutions. It is necessary to stimulate interest and build capacity in the domain of natural language processing (especially regarding African languages) and hence special workshops will be organised within the tertiary sector at venues across the country. Furthermore, given that academic capacity in the domain of DH in South Africa is still very limited, appropriate national and international experts will be invited to take part in this awareness campaign and training program for at least the first 12 months after establishment of the Centre. Details on these programs, and calls for international participation will be made public soon after the establishment of the Centre.

## 5. Services and products

The Centre will provide the following services:

**Digitisation service**:

Quality controlled digitisation of language resources (text and speech: fully or partially annotated) for use by researchers and/or commercial entities (under licence conditions). This will include

- academically valuable non-digitised language (text and speech) resources as determined by the Scientific Advisory Committee (SAC) to be digitised on a continuous basis;
- new resources developed by the official partners as annual projects approved by the SAC;
- new resources developed by academic communities / individuals responding to open calls for short term projects approved by SAC or delegated committee.

**Information sharing and resource distribution service:**

A multi-purpose on-line **Portal** making provision for

- a catalogue of South African digital language resources available at the Centre with terms and conditions of use specified;
- an inventory of all digital language resources under development at the Centre;
- an inventory of all software tools available at the Centre for use;
- on-line links to national and international resource centres;
- on-line distribution of digital language resources and software tools.

These two services mentioned above will provide researchers and students alike with large volumes of resources that were not previously accessible, and will allow for credible quantitative as well as qualitative interpretations. The availability of software tools to extract information, organise and visualise results in various ways bring new dimensions to research in the domain of HSS.

**Archival language search service:**

Given that in-house data management tools and procedures will be in place, searches for specific types of data within and across databases will be available for users without them having to download large databases. This will be an on-line service to end-users which will initially be conducted manually. However, this service could in the future be fully automated providing users access to current and historic language forms.

This will be a unique service especially as far as it concerns African languages, whilst simultaneously constituting a living archive of the languages at different points in time.

**Academic leadership services:**

In addition to these services the Centre will provide academic research leadership within the domain of Digital Humanities (DH) as previously described. This will initially be available as a service to the wider South African academic community.

The Centre will continue and expand on the provision of the following **products:**

**Annotated digital speech datasets** in all 11 official languages in different modalities, e.g. as

- natural speech in different environments (noisy, clear, studio quality, telephone speech etc. )
- natural speech of different age groups (children to aged subjects)
- natural gender based speech
- natural speech in different dialects
- impaired speech (to follow at a later stage with co-operation of speech therapists)

**Marked up digital text datasets** in all 11 official languages comprising

- running texts of different genres: newsfeed, novels, dramas, poetry, advertisements, websites etc.
- language learner corpora
- sms-texts
- parallel text corpora

These datasets are necessary for, *inter alia*, the development of dictionaries and coining of new terminologies, language use in academic literacy studies, sentiment analysis, manual as well as machine translation systems, interpretation studies, intelligent information retrieval systems, etc.

**Natural language software processing tools** applicable to all official languages, given that some of the African languages, for instance, have conjunctive or disjunctive writing systems, posing a challenge to morphological parsing.

**High level training packages** in principles and methods of research and development within a digital paradigm. This will include

- workshops at the hub
- workshops at the different nodes
- workshops at any other university responding to open invitations.

These workshops will be directed towards academic staff and students, government officials involved in language based services (e.g. staff of departmental language offices), members of the National Lexicographic Units, freelance translators and interpreters etc.

## 6. National and international co-operation

The Centre will be the first of its kind in the African context and is a multipartner language resource centre providing single sign-on access to a web based platform which integrates language-based resources and advanced tools for research and training related to all official languages of South Africa. This shall be implemented by shared responsibilities of partners as drivers of individual nodes.

The Centre will actively market multilingual digital language resources for national and international research, training and commercial purposes. It will interact with similar international centres especially regarding best practices and knowledge transfer. The Centre will also formally link data to reputable centres in order to provide the international research community access to local language data, whilst simultaneously also getting access to language resources and software tools developed internationally.

This research and development centre combines research expertise and essential technical and administrative support to conduct cutting-edge research in text and speech technology, and use that as the basis for the development of innovative and relevant technological applications. The Centre also ensures long-term sustainability for research and development activities, which is ideal for the establishment of valuable partnerships with industrial and business partners.

## 7. Impact

The *South African National Centre for Digital Language Resources* covers a variety of fields within and across languages and associated technologies, and provides resources for language based research and development in the humanities and social sciences, the typical domain of Digital Humanities. Furthermore the Centre has an enabling function to stimulate the development of language based software applications and tools for use in translation, education, social services, health promotion and issues that are relevant in the local environment. The Centre provides, with the support of business partners, appropriate capacity building programs related to methodological aspects of digitisation and the use of digital media in research and development projects.

This Centre with its focus digitisation and the use of digital resources is a realistic response to the numerous challenges of a country dealing with 11 official languages.

***No particular resources to report with this submission***