SADiLaR

South African Centre for Digital Language Resources

NEWSLETTER

FUNDED BY:

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

HOSTED BY:

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

PARTNERS:

CSIR
Touching lives through innovation

UNISA
university of south africa

NWU® | CTexT®

ICELDA
Inter-institutional Centre for Language Development and Assessment

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

ENGLISH: JUNE 2022

FUNDED BY:

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

READ ABOUT :

HISTORICAL CORPORA: RESEARCH POSSIBILITIES FOR THE TRACING HISTORY TRUST'S VOC DAY REGISTERS

- Roné Wierenga

The Tracing History Trust's corpus (hereafter the THT corpus) is a digital corpus that is compiled from handwritten diary entries dating back to the 17th and 18th centuries (1687-1714). These diary entries were written by members of the Dutch East India Company and tell the stories of their daily lives in the Cape Colony. The diaries provide interesting insights into the lives of the Dutch colonists on the Cape coast and include descriptions of market days, governmental and political events, and interpersonal relationships between families - both South African and Dutch. These texts are written in a form of Early Afrikaans that is neither Dutch nor standard Afrikaans.

A digital corpus, like the THT corpus, is a big collection of texts (i.e. corpus) that is machine readable (i.e. digital) and therefore usable as a research resource. The benefits of a digital corpus include that it is not tied to a specific geographical space, it is analysable through the use of different software and - in most cases - it is freely available to researchers. The corpus includes, amongst other things, family registers, a list of birthdates and death dates, shipping registers and slave registers. This information enables new and innovative research in fields like genealogy, anthropology and even sociology. For historians, the corpus is a treasure trove of information on the history of South Africa's geographical and political landscape. It provides researchers a unique perspective on South African heritage and culture.

The value of this corpus is not limited to historical research. Due to the fact that the diary entries were made before Afrikaans was standardised, the corpus is a valuable resource for research on spelling variation and the development of spelling, lexicographical research about the vocabulary of Early Afrikaans, syntactic research about the use of sentence structure and many other linguistic phenomena. The diary entries could even be used for author identification research, discourse analysis and comparative studies between Early Afrikaans and 17th century Dutch.

The corpus sheds new light on the community and society of the Cape of Good Hope in the 17th and 18th centuries. It tells the story of Portuguese sailors and the Khoi-San that already inhabited the Cape and the origin of the colony that would ultimately become the catalyst for the establishment of the Republic of South Africa.

FUNDED BY:

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

For more information on the extent of the corpus, read Liebenberg (2018) or refer to Wierenga and Breed (2021) as an example of the types of research that the THT corpus could be used for.

BIBLIOGRAPHY:

Liebenberg, H. 2018. Die Wes-Kaapse Argief en die begin van Afrikaans. Tydskrif vir Geesteswetenskappe, 58(2):204-236.

Wierenga, R. & Breed, C.A. 2021. Diachroniese benadering tot die ontwikkeling van die progressiewe perifrastiese konstruksies in Afrikaans en Nederlands: 'n Korpusondersoek. Tydskrif vir Geesteswetenskappe, 61(2):588-619.

LAUNCHING THE DIGITAL HUMANITIES OPEN EDUCATIONAL RESOURCES CHAMPIONS INITIATIVE

- Natalie Simon

The South African Centre for Digital Language Resources (SADiLaR) and the North-West University's (NWU) UNESCO Chair on Multimodal Learning and Open Educational Resources (OERs) are proud to announce the first intake of our Digital Humanities OER Champions Initiative. This programme, offered through SADiLaR's ESCALATOR Digital Champions Initiative, seeks to stimulate activism and research around the use and/or creation of OERs for the digital humanities at universities in South Africa.

To this end, we are pleased to announce that 26 projects have been accepted into this programme to support and fund the creation and adaptation of OERs in the digital humanities.

"As the name implies, OERs are any teaching, learning and research materials which are available in the public domain and permit no-cost access, use and adaptation and redistribution," says Professor Jako Olivier, NWU UNESCO Chair on Multimodal Learning and OER. "OERs are an integral tool in the building of digital humanities skills and expertise in South Africa as this important field needs to develop in response to the specific needs of academics in the humanities in South Africa."

"Digital humanities is a relatively new research field in South Africa," says Professor Menno van Zaanen, digital humanities professor at SADiLaR, "It is the practice of using computational tools in the broad area of the humanities. Digital technologies allow humanities and social sciences researchers to analyse larger amounts of data

Contact us:

FUNDED BY:

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

(such as text), allowing them to answer research questions more objectively or even to answer completely novel research questions."

The 26 projects chosen are distributed across institutions in South Africa. They have objectives ranging from using OERs to introduce a multilingual, decolonial and interdisciplinary perspective in journalism education, supporting indigenous language robotics education in South Africa, to introducing digital humanities and computational thinking into legal research.

In the coming months, SADiLaR will showcase the work of the projects involved in the programme on its various platforms.

In addition to a research grant, the programme has a strong capacity-building focus. Programme participants will do an online short course on OERs, which includes webinars and workshops and creating a space to share best practices. The programme will also include support for champions from the humanities to research the process, participate in a colloquium on their research and work towards a publication on their work. An essential outcome of the programme is to build a network of digital humanities researchers and practitioners in South Africa to develop the fledgling field.

HUNDZULA RETREAT: BRIDGING THE DIVIDE BETWEEN COMPUTER SCIENCE AND LINGUISTICS

- Natalie Simon

Africa is home to a great many riches, not least of which are its languages. According to the African Language Program at Harvard University, approximately one third of the world's languages are from the African continent. The challenge for Africa's linguists and language scholars now is to keep those languages active in an increasingly digital world.

The Hundzula Retreat, which took place in February 2022 at the University of Pretoria, brought together African linguists and natural language processing practitioners in an attempt to break down the silos between the fields and encourage greater collaboration to build the digital footprint of South Africa's indigenous languages. Four members of the South African Centre for Digital Language Resources (SADiLaR) attended and presented at the retreat. The members were: Rooweither Mabuya, digital humanities (DH)

Contact us:

FUNDED BY:

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

researcher with a special interest in isiZulu; Andiswa Bukula, DH researcher with an interest in isiXhosa; Mmasibidi Setaka, DH researcher with a focus in Sesotho; and Respect Mlambo, DH researcher focusing in Xitsonga. Mabuya was part of the organising committee and both she and Setaka gave presentations.

Building understanding, breaking down barriers

Natural language processing (NLP) involves linguistics, computer science and artificial intelligence in order to get computers to perform useful tasks involving human language. Linguistics is the study of language and its structure, which includes elements like grammar, syntax and phonetics. While on paper these two seem a match made in heaven, in reality their paths do not cross as much as one would think.

'Even though NLP practitioners make the tools that linguistics researchers use, we as the linguists do not know what process they follow, while the NLP practitioners often have gaps in their understanding of, say, the syntax or morphology of a language,' says Bukula.

The tools developed by the NLP practitioners range from software as ubiquitous as spell checkers to tools more specialised for research, such as text analysis software.

A big focus of the retreat was for the two specialised disciplines to provide insight into their respective fields. Associate Professor Vukosi Marivate, Chair of Data Science at the University of Pretoria, gave a presentation titled: 'Machine Learning for Everyone - what the 4IR?'. In his presentation, he not only explained core concepts in machine learning but also how those external to the field can engage with machine learning and, more importantly, provided a pathway for linguists to learn more about machine learning. Other presentations around NLP included 'Introduction to Natural Language Processing' by Jade Abbot, machine learning engineer at Retro Rabbit and co-ordinator of Masakhane, a grassroots NLP organisation.

Mabuya gave an introduction to linguistic studies, in which she broke down what linguistics is, and also explained the various disciplines and branches of the field, while Setaka presented on lexicography for under-resourced languages, with a particular focus on the importance of NLPs and digital technologies for dictionary compilation in Africa's indigenous languages.

For the SADiLaR researchers the retreat was insightful and enriching.

'What was useful for me was to understand how these computational tools are developed, and how important the data that they are trained on are,' says Bukula. 'Now instead of just dismissing a tool because it produces low accuracy levels, I recognise that it is just a question of feeding that tool more isiXhosa data or working with the NLP practitioner to give feedback on the language syntax or morphology.'

Contact us:

FUNDED BY:

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

Mabuya says it was a highlight for her to be part of the bringing together of young minds to discuss how to bridge the gap between linguistics and natural language processing.

Laying the groundwork for collaboration

The retreat also included lightning talk presentations from all the participants. These were short presentations about each individual's research and research interests. This allowed those present to identify potential collaborators on future projects. Bukula says she is already writing up a proposal to collaborate with colleagues from Stellenbosch University and University of Limpopo. Mabuya says she was able to share ideas with NLP practitioners also working on isiZulu.

Opening the door to computational possibilities

The retreat helped to allay fears common in humanities researchers when it comes to anything computational. The field of humanities is becoming increasingly digital and there is no going back.

Computational tools open up new doors for humanities researchers and opportunities like the Hundzula Retreat help humanities researchers see that, they do not have to do it all themselves, there are plenty of computer science experts out there to collaborate with. But, for Bukula, it also took away some of the unknown around natural language processing, computer science and machine learning.

'There are courses out there that focus on beginners,' she says. 'Learning these computational skills is just like learning anything else, you start from the beginning, and you work your way up, one step at a time.'

Her advice to humanities researchers and students: 'Do not think this is not for you. Computational skills and programming are becoming essential to every discipline and social sciences and humanities are no different. Universities would be doing their students a great justice if they began teaching these skills from undergraduate level.'

Contact us:

FUNDED BY:

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

COLLABORATIVE DOCUMENTATION OF SIPHUTHI IN LESOTHO: A SUMMARY

- Natalie Simon

Siphuti, a language spoken by ebaPhuti communities in southern Lesotho and in the northern Eastern Cape province of South Africa, is considered an endangered language. The number of siPhuthu speakers in Lesotho is estimated to be around 200 000, but this number is shrinking fast. Intergenerational language transmission – in which children acquire siPhuthi from their parents and grandparents – is confined to two remote river valleys, Daliwe and Sinxondo. Even in these two valleys, when ebaPhuthi marry Basotho or amaXhosa, siPhuthi is often no longer used in the home environment. In these mixed marriages, children grow up speaking Sesotho or isiXhosa as their main languages.

Fortunately, to ensure this language is not completely lost, Matthias Brenzinger of the University of the Free State and Sheena Shah of the University of the Free State and University of Hamburg, Germany, have been collaborating since 2016 to document and revitalise the language.

An important aspect of this work is the close collaboration with the siPhuthi-speaking community, both in Lesotho and South Africa. Brenzinger and Shah believe strongly that the language speakers should have agency to decide how and by whom their languages are being documented.

The project involves the recording of siPhuthi narratives, conversations, interviews, folktales, oral histories and poems for present and future generations. They also support the production of materials such as COVID-19 health awareness posters in siPhuthi as well as the development of a quadrilingual siPhuthi-Sesotho-isiXhosa-English dictionary, which the team hope will serve as a basis for the development of learning and teaching materials for siPhuthi.

Working with the community, building capacity

In November 2021, Brenzinger and Shah hosted one of their regular language documentation workshops in which they trained six young ebaPhuthi from Daliwe and Sinxondo in language documentation methods and techniques. These young fluent siPhuthi speakers might well be considered key guardians of their threatened language and culture.

Contact us:

FUNDED BY:

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

The workshop involved not only recording techniques but also learning about the ethical practices in language documentation, metadata creation, archiving standards and other important skills. The attendees then applied their new skills by recording the most important annual event in the Phuti calendar: the commemoration of the death of their king Murena Moorosi. Among the ebaPhuthi this event is known as Sikhubhuto sa Murena Moorosi.

The training took place at Bethel Business and Community Development Centre (BBCDC), a vocational training centre in one of the most marginalised parts of Lesotho.

Working through COVID-19

A big challenge for the team was to continue their language work through the global COVID-19 pandemic. They were careful to follow the precautions recommended at the local, national and international level. The researchers were PCR-tested before entering Lesotho and all participants in the workshops were tested using rapid lateral flow tests before the day's activities commenced. Daily temperature measurements were taken and documented, and social distancing was practised. Wherever possible, hands-on activities were conducted outside.

Looking forward

After November's successful workshop there are plans to conduct a follow-up workshop in 2022 in which the participants will be taught more advanced techniques to enhance their recording skills and prepare them to carry out their language documentation activities with minimal support and guidance.

What makes siPhuthi special?

This documentation of siPhuti is particularly important as it is a language of historic and linguistic interest. A language that has evolved over 200 years, some scholars consider it to be a hybrid language because of its many shared features with Sesotho. The analysis of siPhuthi – because of its very nature of originating from one language sub-group and changing heavily under the influence of another – will contribute to the ongoing debate about what constitutes a mixed language in comparison to a hybrid language.

For the full version of this summary, please visit this post.

Contact us:

FUNDED BY:

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

DEMOCRACY FROM BELOW PROJECT

- Boitumelo Matlala

Temokerasi ke kgololosego, mme mo aforika borwa ga eyo kgololosego. Ga gona le temokerasi, ke maaka hela (Respondent 60, Democracy from Below Survey 2021)

Go gololosega mo matshelong a rona (Respondent 4, Democracy from Below Survey 2021)

The quotes above are responses to the questions, 'what word or phrase would you use to explain democracy' and 'what does democracy mean to you'. These are two of the foundational questions in the Democracy from Below project. In this project, our objectives are to build a body of knowledge about the conceptualisation and expression of democracy in South Africa's official languages, and to situate the analysis of democratic thought in South Africa within local histories and practices of democracy in everyday life. The project is collaboratively run by the Centre for Social Change, Sociological Research and Practice at the University of Johannesburg and the South African Centre for Digital Language Resources at the North-West University. It is funded by the National Institute for the Humanities and Social Sciences.

In November 2021, we fielded a multilingual survey asking questions about people's understanding of democracy and the language they use to express it. The survey was in 11 of South Africa's official languages and fielded using the Moya App, a zero-rated messenger app used by four million people across the country. For the survey questions, each language had a team of two translators to do back and forward translations in order to carry as much as possible of the crux of each question. Our task in designing this survey was to measure the extent to which tenets of liberal democracy remain important while also constructing survey questions that allow us to measure the strength and extent of alternative expressions of democracy. The survey returned over 2000 responses.

The survey is one of two main research instruments that the Democracy from Below project uses. The second is qualitative research interviews. In December 2021, we started conducting semi-structured interviews. The interviews, still underway, are being done by a team of three researchers. So far, we have had interviews with local activists in Limpopo, North West, Eastern Cape, Kwa-Zulu Natal, Free State and Gauteng. Deepening the survey questions, the qualitative research interviews surface the contexts and histories that shape people's conceptions and expressions of democracy. We conduct them in local languages and then transcribe them in the same languages. The transcriptions are then sent for quality control

Contact us:

FUNDED BY:

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

to ensure that the nuance in the verbal is captured in the written, and then they are translated into English. One intention of the project is to build a collection of linguistic data. Currently, only limited linguistic datasets are available in indigenous languages and none that are specifically about one topic in multiple languages.

Other inputs of the Democracy from Below project include a colloquium and a short film. The colloquium will bring together the different members of the research team, as well as other scholars working in this area, to discuss and debate the initial findings of the research project. The short film will be shot in April 2022 in Duncan Village, in the Eastern Cape. The short film will feature activists we have worked with, and it will express the disjuncture between the meaning and struggle of inkululeko and the limits of liberal democracy. By setting it in Duncan Village, we wish to capture how these ideas are grounded in local histories and struggles. The aim of the film is to help popularise discussions about the meaning, practice and content of democracy and reflects our commitment to decolonial forms of knowledge production that seek to share research in non-conventional forms for the social sciences.

Bodies of knowledge that exist in postcolonial societies about the conceptualisation, practice and struggle for democracy are often excluded from the canon of democratic theory. The Democracy from Below project is designed to overcome this exclusion by advancing scholarship that incorporates the knowledge, history and experiences of ordinary South Africans into the canon of democratic theory to develop democratic thought from below.

We understand democracy and democratic spaces to be a site of contestation that includes key ideas from mainstream democratic theory – such as civil rights and political freedoms – but that such ideas can also be given new meaning when analysed from below.

Contact us:

SADILAR TEAM TEACHES DATA CARPENTRIES AT WITS

- Natalie Simon

The huge advances in digital technologies in the past several years mean researchers, in all disciplines, can now gather data on a scale never previously imagined. However, there is a lack of skills when it comes to the management and analysis of that data, not only in South Africa, but around the world.

FUNDED BY:

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

Data Carpentry is a global initiative, created and sustained by passionate local communities, which offers workshops on the fundamental data skills needed to conduct research. The goal of the Data Carpentries is to 'teach researchers basic concepts, skills and tools for working with data so they can get more done, in less time, and with less pain.'

In February 2022, three members of the South African Centre for Digital Language Resources (SADiLaR), Mmasibidi Setaka, Benito Trollip and Juan Steyn, ran a Data Carpentry workshop held at the University of the Witwatersrand (Wits) attended mostly by postgraduate students in the Master of Arts (MA) National e-Science Postgraduate Teaching and Training Platform (NEPTTP). This programme is part of a national initiative to build data science skills among South Africa's postgraduates. It is a multi-institutional programme championed by a consortium of six South African universities: North-West University, Sol Plaatje University, University of Limpopo, University of Pretoria, University of Venda, and Wits.

While the NEPTTP does offer two streams: a Master of Science and an MA, this workshop was specifically for the MA students. Between 10 and 13 students participated in the three-day workshop that focused on giving an introduction to fundamental data science concepts. The students came from a range of disciplines, including international relations, psychology, media studies and developmental studies.

'The Carpentries is a great initiative for data science training,' says Trollip, one of the SADiLaR instructors at the workshop, 'because the curriculum, the material, all the documents are all openly available through a Creative Commons licence.'

'This means instructors can rely on this well-established curriculum, and if students wish to revise after the workshop, or the group does not get through the full curriculum in the three days, students can access what they missed online.'

In this particular workshop students were introduced to the basics of good data management, including addressing issues like the problem of using proprietary software (like Microsoft Excel) for collaboration or open data sharing, and what alternatives are available.

Participants were also introduced to the programming language R, that is commonly used for both qualitative and quantitative research in the humanities and social sciences.

'Complementary to the Carpentries lessons for R, we also introduced R through a specific R package known as Swirl,' explains Trollip. 'It is a very interactive tool that helps one to learn R in a step by step manner.'

They also introduced the students to OpenRefine, a 'powerful tool for working with messy data' which is an excellent alternative to Excel spreadsheets in that it offers versioning, where even as you change and edit your data, nothing is lost.

Contact us:

FUNDED BY:

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

So it is easy to track the work and to go back a few steps if you make a mistake.

The NEPTTP programme is cross-disciplinary, which means the students are taught data science and then part of the programme is to apply their data science skills and knowledge to their own fields as part of their final research report. For the SADiLaR Data Carpentry instructors there was great satisfaction in helping the students realise the possibilities using R in the students' own research projects.

'If I think of my own postgraduate experience, I know exactly how challenging it is to have a grand idea of the research you want to do but not know how to execute it,' says Trollip.

"The best part of this workshop for me was speaking to the students about their own research projects and together mapping the steps to take in order to get where they want to go.'

For students who have participated in a Data Carpentry workshop and want to expand their knowledge, SADiLaR offers a programme called ESCALATOR, which is designed to help build and support a community of researchers passionate about using digital tools and methodologies in their own research and teaching.

For any feedback on the content of the newsletter, please feel free to contact SADiLaR: info@sadilar.org

For any feedback on the translations, please feel free to address your message to the translator when sending an email to SADiLaR: info@sadilar.org

UPCOMING EVENTS

SADiLaR is planning three major regional events due to take place between September and November 2022 across South Africa. This forms part of the ESCALATOR project and aims is to continue fostering a community of practice in the field of digital humanities and computational social sciences. The format of these events will include formal and informal conversations aimed at sharing ideas, but be sure to keep an eye on our website and social media for invitations to participate.

For more information on ESCALATOR, please visit the website or sign up for the ESCALATOR announcement mailing list.

Click here to subscribe to our newsletter

Contact us: