Digitising Afrikaans: Establishing a protocol for digitalizing historical sources for Early Afrikaans (1675-1925) as a possible template for indigenous South African languages

Wierenga, Roné
Virtual Institute for Afrikaans, North-West University
Carstens, Wannie
North-West University

Abstract

Afrikaans is one of South Africa's 11 official languages. With the continued rise of English as language of prestige in education, the economy and public sphere a movement towards a monolingual South Africa seems to be underway. This movement results in a loss of interest in the development and use of South Africa's indigenous languages. Another contributing factor is the unavailability or inaccessibility of resources in indigenous South African languages.

This paper reports on the protocol and process of establishing online resources to encourage research in and on Afrikaans.[i] The digitalisation process encompasses establishing a) digital bibliographies for Afrikaans literature and linguistics, where researchers can find links to online resources and pdf documents; b) an online archive where endangered (and often otherwise unobtainable) resources are stored in digital format; c) an online library where digitalized machine readable texts are stored; and d) an online historical corpus for Early Afrikaans (1675-1925). These resources are all angled towards linguistic research, specifically corpus linguistic research. VivA also offers a number of online resources to make research in Afrikaans more accessible – these include a general and school grammar for Afrikaans, a corpus portal for corpus linguists, and a speech atlas to create awareness of linguistic variety in Afrikaans.

The digitisation project is largely based on the Digital Library for Dutch Literature (Dutch: *Digitale Bibliotheek voor de Nederlandse Letteren*) (DBNL)[ii] and is still ongoing with various phases being undertaken simultaneously. *It is our goal to present this project as a template to demonstrate how other indigenous South African languages can go about establishing digital resources and to encourage synergy amongst linguists and linguistic institutions.*

Keywords: digitisation process, digital archive, digital bibliography, historical corpus, Early Afrikaans 1675-1925

1 Background

This paper reports on a project that first began to take shape in 1992. The notion of establishing an online database, similar to the predecessor of the current *Digitale Bibliotheek voor de Nederlandse Letteren* (DBNL), for Afrikaans was first verbalised in 1992. This idea stemmed from an interest amongst Dutch linguists to compare Afrikaans linguistic

phenomena with those identified in Dutch. This form of comparative research was hindered by the lack of Afrikaans resources available to international researchers. No centralised digital platform existed to assist researchers in finding the appropriate resources.

Prof Wannie Carstens of the North-West University was tasked with digitising Afrikaans. In its infancy the process was primarily one of cataloguing all available linguistic publications on Afrikaans linguistics. In 2004 this catalogue was published on the North-West University's website as a digital bibliography for Afrikaans, the *Digitale bibliografie vir Afrikaans* (DBAT).[iii]

Initially DBAT was only a bibliography reference source to inform researchers of the available publications and where they are housed. It has since evolved into a digital repository where researchers can gain access to digital copies of language resources (an overview of DBAT is provided in section 2 of this paper). By 2015 a digital bibliography for Afrikaans literature, DBAL, was also established as a mirror image of its linguistic counterpart.

In 2018 the project began to take on another form. VivA undertook to become the home of and driving force behind the digitisation of Afrikaans, launching two legs of the project, namely a digital archive/library for Afrikaans and a historical corpus for Early Afrikaans (1675-1925). By 2019 a workshop was hosted and key institutions like VivA, SADiLAR, CTexT, the North-West University, University of the Free State and Nelson Mandela University came together to contemplate digitisation and the future of linguistics in South Africa.

Since 2019 immense progress has been made towards establishing a centralized database where researchers can obtain digital resources in and on Afrikaans. Other institutions, like the *Afrikaanse Taalraad*, Afrikaans language council (ATR) and private persons have also shown support for the project and been involved in various aspects of it.

2 The current state of affairs

The project has five branches that each function both as a whole and together as a unit to achieve the same goal, namely establishing a digital library for Afrikaans. The current state of affairs for each of the five parts is outlined in this section.

2.1 DBAT

The digital bibliography for Afrikaans linguistics, DBAT, [iv] currently consists of 17 829 publications about Afrikaans linguistics. These publications are, for the most part, academic in nature and include journal articles, books, and chapters within books, but blog posts, online articles and newspaper or magazine articles are also included when their topics relate to Afrikaans linguistics.

Around 60% of the resources listed on DBAT contain either a pdf copy of the publication or a hyperlink to a digital copy housed by another

institution. The DBAT database is updated weekly, meaning that the newest publications are readily available. The database also has a Google Form that can be filled out by users to let us know of publications that has not been included. We also urge users to check their own publications on the database in order to insure that their entire publication history is available on DBAT. DBAT is currently working with a number of institutions to digitise the publications that do not have digital copies readily available and no longer have copyright restrictions.

2.2 DBAL

The digital bibliography of Afrikaans literature, DBAL,[v] similar to its linguistic counterpart, DBAT, is updated weekly. DBAL currently contains 17 000 publications of which 30% is available in digital format.

Unlike DBAT, DBAL is not yet current and various older publications are still being added to the database. DBAL is therefore in one of its final developmental stages. This is mostly because DBAL was established 11 years after DBAT, and because Afrikaans has vastly more literary resources than linguistic resources. However, once DBAL is current the focus will naturally shift towards the digitisation of the resources.

2.3 Historical Archive

The archive for historical resources is still in its infancy. The development of the archive began in 2018 and is still, partially, in its cataloguing phase. The first round of cataloguing was done alongside project leaders of the envisaged *Database Geschiedenis Nederlandse Taalkunde (*DAGENTA)[vi] Dutch digitalisation project. This is primarily because the project is based on the Dutch model and, due to Afrikaans and Modern Dutch's shared heritage, many Early Afrikaans texts were already known to the DAGENTA team who were interested in digitising older Dutch grammars. The idea was to join forces and share resources in order to digitise both Early Afrikaans and 17th century Dutch simultaneously, but unfortunately this goal has not yet been achieved.

The second round of cataloguing was started in late 2019 in South Africa. This round aimed to catalogue historical texts in and on Afrikaans that are a) no longer in use at South African universities, b) brittle or valuable due to their age, and c) in danger of being lost or damaged.

This process was unfortunately halted by the pandemic and travel restrictions, however a line of communication was established with the relevant universities and in many cases catalogues were provided by the universities which allowed us to pinpoint the text that are of interest to the project. In many cases private persons or institutions contacted us and made historical texts available to us for digitisation. For this reason a small collection of texts where digitised during 2020 and 2021.

The aim of this leg of the project is to preserve historical texts in a digital format and to make them available for research even if they have

not yet been digitalized for corpus linguistic purposes.

2.4 Historical corpus of Early Afrikaans

Early Afrikaans 1675-1925 was a language form that originated from 17th century Dutch but was adapted for use by the Cape Malayan people, various European immigrants and strongly influenced by the Khoi-San community. The result is that this language variety differs from the Dutch used in Europe and shows notable signs of language influence.

Establishing a historical corpus of Early Afrikaans[vii] will allow researchers to explore the developmental phases of Afrikaans, but also to gain a greater understanding of language influence during this phase of South Africa's history. Kirsten (2016) outlines the development of a historical corpus for Afrikaans and list a variety of language changes that would likely not have been observed without a historical corpus. These changes include: a) a decrease in the use of passive constructions; b) the replacement of the preterit forms "had" and "wis" with "het gehad" and "het geweet"; c) the grammaticalisation of "gaan" to a future reference marker; and d) the replacement of the Dutch genitives with the Afrikaans variants "se" and "van".

Many of the texts to be included in the corpus has been catalogued, and the next phase would be to begin the digitisation process. The digitisation process is slower for the development of the corpus than for any other legs of the project. This is primarily because the text that are digitised are valuable, delicate or already partially damaged and therefore require greater care.

The quality of the digital copies also needs to be higher for these text in order for the digitalisation process to be effective and efficient. Low quality digital copies cannot always be digitalized (that is to say made machine readable), and would therefore need to be transcribed manually. Transcriptions can take a long time to complete, they also require a person to devote their time to the transcription, which can be expensive. Both manual transcriptions and automatic digitalisations need to be moderated in order to ensure that no errors occur.

A protocol for the digitisation of historical sources has been established and will be put to the test in 2022. This protocol outlines how to:

- establish networks with libraries and other institutions that house historical resources;
- establish synergy with other digitisation projects (like the Tracing History Trust corpus, [viii] Nuuseum project and DAGENTA project);
- develop an inventory that reflects the quality and scope of the project;
- establish a time line for the project;
- perform the digitisation process in a way that ensure the highest possible quality scans without damaging fragile texts; and
- implement quality control.

2.5 A book on historical sources in Afrikaans

In 1991 Prof Edith Raidt published a book about the origins of Afrikaans, *Afrikaans en sy Europese verlede* ("Afrikaans and its European history"). This publication outlines the landscape in which Afrikaans originated as well as the early development of the language before its standardization in 1925.

In her publication Prof Raidt provides a chapter wherein she names vital sources about the origins of Afrikaans. The idea to write a book that delves deeper into the value of these sources and the insight they give about the origins of Afrikaans. This book will complement the archive for historical Afrikaans resources and the Historical corpus of Early Afrikaans and act as guide for linguists and historians who use these digital resources. The layout and content of the book has already been established and the writing process will commence in 2022.

3 Afrikaans digitisation as a template

Due to commonalities between Afrikaans and Dutch, Dutch language technologies are often easy to adapt for Afrikaans and the needs of the Afrikaans linguistic communities. This does not, however, mean that the adaptation process does not pose any challenges. Often, as is the case in this project, only the basic outline for the DBNL could be used for Afrikaans. The institutional networks, database software, housing and creation, as well as the cataloguing, digitisation and digitalisation processes all needed to be established for Afrikaans. As is often the case, many challenges were only uncovered once a process was already underway and needed to be resolved as they arose.

This is the main benefit of this project for other indigenous languages. The Afrikaans version of the project has already navigated many pitfalls and is now capable of assisting other languages to navigate this terrain.

This project was undertaken with the goal of a centralized platform for digital Afrikaans resources in mind – eventually a comprehensive digital library for Afrikaans – but in a multilingual society and world, a centralized digital library for Afrikaans is of little value if it is not alongside digital libraries for all of South Africa's indigenous languages. Synergy is the key factor needed to undertake the digitisation of an entire language's resources successfully. Synergy on an institutional level, but also on a personal level. In the case of Afrikaans, it is a single person who became the driving force behind the project, motivating and enlisting others, who ultimately become the support structure and current home of the project.

A lot of footwork has been done in order to establish a digital resources for Afrikaans, and this footwork can ultimately benefit a similar project for another South African language.

Acknowledgements

A word of gratitude to everyone who has been involved with, or supported this project in a personal or institutional capacity. Your input,

whether man hours, insight, or willingness to provide access is indispensable.

References

Breed, C.A., Carstens, W.A.M. & Olivier, J. 2016. Die DBAT: 'n Onbekende digitale taalkundemuseum. *Tydskrif vir Geesteswetenskappe*, 56(2-1): 392-409.

Kirsten, J. 2016. Grammstikale verandering in Afrikaans van 1911-2010. Vanderbijlpark: North-West University.

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

Raidt, E. 1991. Afrikaans en sy Europese verlede.

[i] This is an ongoing project.
[ii] This link can be used to access DBNL: https://www.dbnl.org/
[iii] For a detailed overview of the development and functionalities of DBAT refer to Breed *et al.,* (2016).
[iv] This link can be used to access DBAT: https://collections.nwu.ac.za/dbtw-wpd/textbases/bibliografie-afrikaans/dbat.html
[v] This link can be used to access DBAL: https://collections.nwu.ac.za/dbtw-wpd/textbases/bibliografie-afrikaans/dbal.html
[vi] This link can be used to access the DAGENTA repository: https://cls.ru.nl/dagenta/
[vii] Although the historical corpus for Early Afrikaans (1675-1925) has not yet been completed, examples of digital corpora for Afrikaans can be found on VivA's corpus portal: https://viva-afrikaans.org/
[viii] For a detailed overview of the Tracing History Trust corpus and the digitisation projects of the Western Cape Archives refer to Liebenberg (2018). The Tracing History Trust's digitisation products can be accessed at: http://www.tracinghistorytrust.co.za/products.htm