Proceedings of the Sixth Workshop on Resources for African Indigenous Languages (RAIL)

Co-located with DHASA2025

---

10 November 2025
Pretoria, South Africa

RAIL is proudly supported by

Preface to the Proceedings of RAIL 2025

1 Preface

The sixth workshop on Resources for African Indigenous Languages (RAIL) was held on 10 November 2025 at the CSIR International Convention Centre in Pretoria, South Africa. It was co-located with the Digital Humanities Association of Southern Africa (DHASA) 2025 conference, which took place from 11 to 14 November 2025.

The RAIL workshop series aims to bring together researchers who are interested in showcasing their research and thereby boosting the field of African indigenous languages. This provides an overview of the current state-of-the-art and emphasizes the availability of African indigenous language resources, including both data and tools. Additionally, it allows for information sharing among researchers interested in African indigenous languages and also starts discussions on improving the quality and availability of the resources. Many African indigenous languages currently have no or very limited resources available and they are often structurally different from well-resourced languages, requiring the development and use of specialized techniques. By bringing together researchers from different fields such as computational linguistics, sociolinguistics, and language technology, to discuss the development of language resources for African indigenous languages, we hope to boost research in these fields. The RAIL workshop is an interdisciplinary platform for researchers working on resources such as Natural Language Processing tools, Human Language Technologies, data collections, and annotations, specifically targeted towards African indigenous languages. It aims to create the conditions for the emergence of a scientific community of practice that focuses on data, as well as tools, specifically designed for or applied to indigenous languages found in Africa.

At the sixth edition of the RAIL workshop, we received nineteen high-quality submissions, which were all reviewed by three reviewers. The reviewing process was open, which means that the identity of both authors and reviewers were available to both. Finally, seventeen submissions were accepted (89% acceptance rate). As organizers, we would very much like to thank the programme committee for their in-depth reviewing of the submissions and for providing useful recommendations to the authors.

Note that the aim was to get as many people together to share their ideas and research results, while attaining a high-quality event. This led to a workshop, which consisted of a full day of presentations. This publication adheres to DHET's 60% rule and authors in the proceedings come from a wide range of institutions. In total, the audience listened to sixteen presentations. Each presentation consisted of 25 minutes (including time for discussion). After all the presentations, many interesting discussions took place.

2 Program Committee

Akinola Ayodele James, Michigan Technological University, United States of America

Bekker Martin, University of the Witwatersrand, South Africa

Bombardella Pia, North-West University, South Africa

Calteaux Karen, Council for Scientific and Industrial Research, South Africa

du Plessis Gideon, South African Centre for Digital Language Resources, South Africa

Eiselen Roald, North-West University, South Africa

Gaustad Tanja, North-West University, South Africa

Gouws Rufus, Stellenbosch University, South Africa

Griesel Marissa, South African Centre for Digital Language Resources, South Africa

Groenen Gonneke, North-West University, South Africa

John Henry Chukwudi, Universitat Oberta Catalunya, Spain

Augustine Farinola, University of Alberta, Canada

Josef Langerman, University of Johannesburg, South Africa

Lemeko Papi, Central University of Technology, Free State

Louw Henk, North-West University, South Africa

Louw Johannes Abraham, Council for Scientific and Industrial Research, South Africa

Mabuya Rooweither, South African Centre for Digital Language Resources, South Africa

Madlome Steyn Khesani, Great Zimbabwe University, Zimbabwe

Mahloane Malefu, University of the Free State, South Africa

Marais Laurette, Council for Scientific and Industrial Research, South Africa

Marongedze Reggemore, Nelson Mandela University and University of Zimbabwe

Matfunjwa Muzi, South African Centre for Digital Language Resources, South Africa

Matthew Gordon Derrac, North-West University, South Africa

Naidoo Privolin, Council for Scientific and Industrial Research, South Africa

Ogama Chinedu, MYGEECS, Nigeria

Ogungbemi Olarotimi, University of Texas at San Antonio, USA

Onuoha Chialuka, University of Nigeria, Nigeria

Puttkammer Martin, North-West University, South Africa

Roux Justus C., Stellenbosch University, South Africa

Setaka Mmasibidi, South African Centre for Digital Language Resources, South Africa

Sibeko Johannes, Nelson Mandela University, South Africa

Skosana Nomsa, South African Centre for Digital Language Resources, South Africa

Steyn Juan, South African Centre for Digital Language Resources, South Africa

Trollip Benito, South African Centre for Digital Language Resources, South Africa

Van Zaanen Menno, South African Centre for Digital Language Resources, South Africa

Wilken Ilana, Council for Scientific and Industrial Research, South Africa

Wolff Friedel, South African Centre for Digital Language Resources, South Africa

3 RAIL 2025 organizers and proceedings editors

Rooweither Mabuya

Muzi Matfunjwa

Mmasibidi Setaka-Bapela

Menno van Zaanen

Introducing the Historical African Languages Database: A Translingual Resource of Crosslinked Dictionaries

James Law, Brigham Young University
Daren E. Ray, Brigham Young University
Earl Brown, Brigham Young University

jimlaw@byu.edu

Abstract

The earliest written documentation of most African languages comes in the form of dictionaries and field notes prepared by European missionaries and linguists, with the assistance of African informants, in the nineteenth and early twentieth centuries. These resources have been difficult to access and compare, existing only in either print or unprocessed scans. We present a fully searchable and interconnected online database that makes such resources more easily accessible for study. It currently contains seven bilingual dictionaries, with many more sources to be added as they are processed. We explain the database's design, in which processed entries are separated and their fields tagged according to a consistent structure, maximizing query options and facilitating translingual connections. We describe the functionality of the website through which users can access the data in a variety of ways. We discuss the database's construction process, including particular challenges related to these historical data sources, and outline the development of a scalable procedure for its future expansion. We also present three case studies illustrating potential uses of the database by historians, linguists, and educators. Finally, we identify a roadmap for the resource's continued improvement through additional features.

Keywords: lexicography, historical linguistics, colonialism, database

1 Introduction

This paper presents a new digital resource that makes accessible a significant source of data on the history and diversity of African languages: nineteenth- and early twentieth-century colonial dictionaries and linguistic field notes. Beginning in the nineteenth century, European missionaries, linguists, and lexicographers recruited African informants to prepare descriptions of the grammar and vocabulary of African languages, often in connection with Bible translation efforts (Robinson, 2022). For most African languages, the publications that resulted represent the earliest written documentation of any kind (Nkomo, 2020). Some of these historically important documents have been scanned, with PDFs available on Google Books, HathiTrust Digital Library, Archive.org, or elsewhere on the internet, while others remain only in print form in archives. Our database collects these sources in a searchable, comparable form for the first time. We present this tool in an easy-to-use website for research by historians, linguists, teachers and learners of African languages.

The first dictionaries already added to the database mostly describe Bantu languages in Eastern and Southern Africa. However, the eventual scope of the database will include any language from the continent for which such colonial-era descriptions are available. This could include hundreds of individual sources.

In the following, we describe the historical context in which these dictionaries and notes were produced as well as our process for digitizing and assembling them into a useful database. We also propose three brief case studies illustrating how the database can be used for different research purposes.

2 Historical background

Protestant missionaries began the first coordinated efforts to compile dictionaries of African languages in the mid-nineteenth century. Previous travelers from Europe had recorded samples of African languages as early as the sixteenth century, usually in the form of wordlists and translations of Catholic catechisms (Wonderly and Nida, 1963:123). Protestant missionaries aimed to teach potential African converts in their vernacular languages (Constantine, 2013). They compiled extensive linguistic resources, both to prepare translations of the Bible and train future missionaries.

Missionary lexicographers were also motivated by advances in the study of linguistics and emerging standards for lexicography (Nkomo, 2020). In line with other modern dictionaries that had just begun to appear in Europe, they arranged entries in alphabetical order. Many of them also included parts of speech, definitions, sample sentences, grammatical notes, and etymological information.

Missionary lexicographers in Africa came from nearly every Christian denomination (Wonderly and Nida, 1963; Mkenda, 2018). They relied extensively on African collaborators who provided the raw material of word lists, usage, and grammatical knowledge to make their dictionaries (Robinson, 2022). However, they often worked independently of other missionaries because of their isolation in remote locations. They reported on their work to the Christian mission societies that funded their work, as well as to other missionaries working on similar languages. They often debated how best to represent the African language phonology and elicit vocabulary. Their correspondence included drafts of word lists, dictionaries, and language training manuals. In some cases, they left journals and other records in the archives of missionary societies that give insight into their language work (Paas, 2011; Krapf, 1860).

The Christian mission societies that raised funds for missionary work in Africa also funded editors to prepare language resources for publication. They then formed partnerships with publishing organizations such as the Society for the Promotion of Christian Knowledge. This charity published dictionaries alongside language lessons and pamphlets for distribution in Europe to preachers preparing for mission travel. As such, almost all of the dictionaries of African languages produced in the nineteenth century were bilingual rather than single language dictionaries.

Because of the time in which they were produced, these dictionaries are an invaluable resource for understanding the pre-colonial state of African languages. Colonialism has had a profound effect on the lexicons of African languages, not only through the borrowing of words from European languages but also through the cultural changes imposed or influenced by colonial power structures (Peterson, 1997). Although we are in most cases without written documentation of African languages from a pre-colonial time, the dictionaries included in our database often represent a state of the language at the very earliest stages of colonialism, when European influence was relatively limited.

It is important to acknowledge the significant limitations of these historical dictionaries as a data source. Their compilers had varying degrees of linguistic and anthropological training (Nkomo, 2020). While some dictionaries show remarkable detail and consistency, others are rife with errors or provide scant descriptions. It is especially important to recognize the European and Christian bias displayed in these sources (Peterson, 1997). The African informants who supplied the data for these dictionaries are rarely acknowledged, and in most cases little is known about them (Bank and Bank 2013). This sometimes has the effect of obscuring the particular variety that is being described, since the linguistic background of the informants is unclear and a dictionary may assemble data from multiple sources, some of them second-hand. As an example, Bishop Edward Steere's compilation of the Zanzibar dialect of Kiswahili was collected from students in his school who learned the language as a second or third language (Robinson, 2022). We must exercise caution to evaluate how faithfully the definitions provided represent the usage of African informants and the degree to which European missionaries inserted their own perspective (Peterson, 1997). For some users of the database, this may be an explicit object of inquiry. The content of the dictionaries in our database should not be understood as a fully objective portrait of the languages described, and caution must be exercised in interpretation of the data.

Since the publication of the dictionaries in the time period that is our focus (roughly up to 1930), African lexicography has advanced considerably.

Linguists with a scientific rather than religious objective produced grammars and dictionaries of African languages throughout the twentieth century (Nkomo, 2020). In the twenty-first century, many of these dictionaries have been put online, and new online resources have been developed (de Schryver, 2003). These efforts have naturally been more extensive for those languages that have higher numbers of speakers and politically official status, such as Kiswahili and isiZulu. There have also been projects bringing dictionaries of different languages together for comparison, such as the Comparative Bantu Online Dictionary (CBOLD[1]).

Despite this progress, for some very low-resource languages, dictionaries made in the colonial period may still be among the most detailed documentation yet published. Even for higher-resource languages such as Kiswahili, colonial dictionaries have significant historical and comparative value. While many of these colonial dictionaries have been scanned and put online, none have been digitally transcribed so that their contents can be searched like a modern online dictionary. This makes them largely inaccessible for comparison and research. For these reasons, our database primarily focuses on historical sources from the nineteenth and early twentieth centuries, despite their limitations, setting it apart from comparable databases of more recent sources.

3 Database design

We selected a set of these colonial dictionaries to digitally transcribe, structurally parse, and organize into the initial database. We first prioritized a variety of Bantu languages to ensure that our database structure accounted for representations of Bantu noun classes in dictionary entries. We also prioritized dictionaries that displayed relatively more complexity in the structure of their entries (i.e., including additional fields beyond the standard headword, part of speech, and definition), in order to develop a database framework that could account for this complexity. Additional details regarding the selected dictionaries and parsing process are provided in the next section.

The database is accessible via a website[2] and is regularly updated as additional features and dictionaries are added. The web portal for accessing the database is designed to facilitate comparison across dictionaries and languages, increasing its usefulness for language students, historians, linguists, and other users. The site currently provides three main functions: browsing individual dictionaries, searching entries across multiple dictionaries, and accessing metadata about the dictionaries and their creators.

Individual dictionaries can be browsed by selecting a source and a starting letter. Corresponding entries are then presented in order, with each of their fields (headwords, parts of speech, definitions, example sentences, etc.) presented in a consistent format. Many dictionaries include internal references to related words in the same dictionary; these words are clickable and direct to the corresponding related entry. A button beside each entry displays the image of the PDF page on which it is found, so that users can compare the digitally transcribed text to its original source side-by-side.

The search function permits searching across all dictionaries (by default) or across a subset or a single dictionary. Queries, which can target exact matches or partial matches with the beginning, middle, or end of an expression, can target headwords, particular fields (definitions, example sentences, etc.), or all fields together. Resulting entries matching the query are grouped by dictionary, with the same presentation format as the browse function: consistently structured entries with a button to display the original page side-by-side with the digital text.

Metadata about each dictionary are presented on individual pages of the site. These include the available information about the informants, linguists, lexicographers, and publishers involved in its creation and the relevant historical and ethnographic context. These pages also discuss how we adapted the information and formatting of linguistic data in each dictionary to our database structure. That is, we describe the kinds of information typically found in entries of each dictionary and how they are presented, and we explain how we categorized them to fit into the standard data structure of our database. Any information necessary to interpret a source's entries in the database, such as the meaning of abbreviations used by the lexicographer, is also explained on these pages.

In addition to the entries that comprise the main content of each dictionary, sources generally

[1] http://www.cbold.ddl.cnrs.fr/

[2] http://hald.byu.edu

include prefaces, lists of terms, grammatical descriptions, and other such details in frontmatter or appendices. These contents are provided on these metadata pages in their original PDF forms.

4 Process

The processing of dictionaries for incorporation into the database always begins with an analysis of the document's structure. We examine sample entries closely to identify how they are organized, the kinds of information they contain, and the typographical conventions they follow. These historical documents vary in their formality and level of copyediting, and we must often account for exceptions and mistakes in the layout of dictionary entries, such as inconsistent use of abbreviations or indentation. This analysis informs the rest of the process.

For most of the dictionaries we are interested in, quality scans are already available online; where necessary, we perform our own scans. Scanned PDFs are first preprocessed to maximize their computer readability. Preprocessing steps used vary according to the quality of individual scans, but may include contrast enhancement, binarization, deskewing, or noise removal.

Optical character recognition (OCR) is performed using the Tesseract 5.5 engine (Smith, 2007), which we selected over commercial OCR software because of the ability to fully adjust parameters for each dictionary to produce the best results. In addition to the text itself, our OCR approach outputs data on the position of each word on the page (useful for identifying indentations and other positional characteristics that mark meaningful elements of entry structure) as well as a confidence score, which allows us to automatically delete low-confidence items (typically stray marks) and flag moderate-confidence words for manual verification and correction.

We use Python scripts to convert the raw OCR text output to structured JSON representations of each dictionary. These scripts are individual to each dictionary, as they must account for variations in the fields included in entries and the ways those fields are typographically distinguished. However, because most of the dictionaries follow common formatting patterns (such as indenting the headword of each entry or numbering alternate definitions), we are able to minimize coding time by reusing some core functions and modifying parameters as needed. Unlike some OCR engines, Tesseract 5.5 does not perform font recognition, so data about font style (bold, italics) is unavailable for use in parsing entries; however, positional and character data has been sufficient for parsing entry structure in all dictionaries included so far.

One common element in many of the dictionaries is the inclusion of example sentences in the target language illustrating the use of each word. To identify the example sentences and separate them from their translations and other entry elements such as definitions, we use machine learning models trained with PyTorch (Ansel at al., 2024) for language classification. Of course, low-resource languages can pose a challenge for language classification tools developed using machine learning. This is alleviated by the fact that the included dictionaries pair a (potentially low-resource) African language with a high-resource European language. We can therefore achieve acceptable results simply by distinguishing English/French (for definitions and translations) from not-English/French (for example sentences). Furthermore, for Bantu languages, we have successfully relied on a single Kiswahili language model to distinguish target language text from translation language text, eliminating the need for additional model training resources. For example, in a Mijikenda-English dictionary, strings are classified based only on their similarity to Kiswahili and English. Mijikenda example sentences are classified as Kiswahili by the model due to linguistic similarity between these two Bantu languages, and they can thus be separated from the English definitions and translations.

As we parse each dictionary, its content is converted to a standard JSON structure. Because not all dictionaries contain the same fields, we have opted for a maximalist structure; for example, although most dictionaries do not include etymological information in their entries, this field is available for those that do.

The digital text in the database is primarily a faithful representation of the scanned text. However, there are some exceptions to this. For languages with noun classes (such as Bantu languages), we convert variable information about noun class (for example, some dictionaries list affixes and particles associated with each noun, while others use custom numbering systems) to a maximal Bantu noun class numbering system that includes most scholarly variants (Maho 1999). We

Language Pair	Source
Maa - English	Erhardt & Krapf (1857)
Sotho - English	Kruger (1876)
Mijikenda - English	Krapf & Rebmann (1887)
Yao - English	Maples (1888)
Kiswahili - French	Sacleux (1939 [1888])
Luganda - English	Pilkington (1892)
Kikuyu - English	McGregor (1904)

Table 1: Dictionaries currently in the database.

also expand abbreviated parts of speech (e.g., converting *n.* to *noun*) and standardize them (e.g., converting *s.* for 'substantive', a synonym for noun, to *noun*) so that users can easily target words of a particular part of speech in their search queries. Although such modifications to conform to a consistent structure are minor, they are one reason we prioritized easy access to the original PDF for users of the site. Our standardized representation of dictionary entries can be quickly compared to the original with the click of a button.

Research assistants use a custom software tool to manually verify and correct each dictionary's content before adding it to the database. The software tool displays parsed entries according to the standard JSON structure we have adopted and allows for easy navigation between entries, side-by-side comparison with the PDF, and rapid correction. Errors can arise from mistakes in the OCR or from oversights in the parsing script that neglect to account for entries that have unusual content or formatting. We have found that correction proceeds most efficiently in two stages. During a first pass, assistants correct any errors in parsing (separating erroneously grouped entries, moving incorrectly assigned text to the proper field, etc.). Then during a second pass, they correct OCR errors, focusing on words with low confidence scores that are more likely to contain mistakes (misspellings, cut off words, etc.). They also identify issues that are consistent enough to be resolved through a mass edit, which they pass on to the project directors to implement programmatically. To clarify, the goal at this stage is not to correct any perceived errors on the part of the dictionary's creators, but simply to ensure that the digital representation accurately reflects the content of the document.

As of this writing, seven dictionaries have been processed for inclusion in the database (although final manual corrections are ongoing for some of these). They are listed in Table 1. They represent six Bantu languages and one Nilotic language, with either English or French as a translation language. All derive from data collected in the latter half of the nineteenth century or the early twentieth century.

5 Challenges

We have encountered several challenges specific to our data source. First, because these older dictionaries are printed in a variety of fonts that may not be standard in modern texts, the OCR output contains more errors than would be typical in a modern text. As explained above, manual correction is therefore a crucial step.

Another challenge is the variable entry structure across dictionaries. For example, the Mijikenda dictionary by Krapf and Rebmann (1887) organizes entries hierarchically, with some words subordinate to other words they are derived from, and includes the source language for many words; by contrast, the Yao dictionary by Maples (1888) has a flat entry structure and does not include source language notes. In order to place such varied sources together into a single database, we had to create a standard data structure that would accommodate all of the fields and relationships that the dictionaries include. We did our best at the beginning of the project to anticipate the fields that would be required by dictionaries added later on, designing a data structure that includes a maximal set of fields (many of which can be left blank for simpler dictionaries). However, we have also had to modify the data structure over the course of the project to add fields that we originally missed but that are included in certain dictionaries, such as the etymological notes included in the Kiswahili dictionary by Sacleux (1939 [1888]).

A further set of challenges posed by the data have been the ambiguous intentions of dictionary compilers in their presentation of the dictionaries, which have required interpretation. While most of the dictionaries include at least some frontmatter that explains the abbreviations used and other important context, these introductions are sometimes lacking in detail. Outdated or unusual linguistic terms such as *neuter verb* (generally meaning a kind of stative verb) are often used. In cases such as these, we generally err on the side of a faithful representation of the dictionary's contents, even if those contents may be unclear to modern users of the database; in such cases, we explain our interpretation of the terms used in the

dictionary descriptions included on the website. However, in order to provide a structured and comparable analysis of each dictionary, we have had to commit to certain interpretations of ambiguous notation. For example, in Krapf and Rebmann's (1887) Mijikenda dictionary, similar words are noted in two ways: either the word *See* or an equals sign. Because this notation is not explained in the dictionary's frontmatter, we had to decide, based on consideration of many examples, to interpret these markers in two different ways: *See* indicates related words elsewhere in the dictionary (which are linked in our database for users to easily cross-reference), while an equals sign indicates comparable words that are not necessarily included in the dictionary and may be from other languages. We have no way of knowing if this is exactly the meaning that the compilers intended with this notation, but it is a reasonable interpretation of the data that allows the dictionary to be incorporated smoothly into the rest of the database.

6 Case studies using the database

In this section, we provide three examples of how this database could be used by historians, linguists, and other researchers, or by language learners and teachers. Our intention here is to illustrate the kinds of uses we had in mind as we created the database, although we hope it could be useful in other ways as well.

First, although the database is certainly a valuable resource for information about the languages described, it is even more directly useful for comparison of the dictionaries themselves. The database includes dictionaries that document the same or very similar languages, published at different times and places and compiled by different authors with their own motivations, biases, and sources. As an example of the kind of historical research through dictionary comparison that could be performed with the database, consider how one might compare entries for the same word or cognate words. The word *koma* is defined in the Mijikenda dictionary by Krapf and Rebmann (1887) as follows: "An evil spirit, supposed to be of some dead person. The chief idea of religion among the Wanyika seems to be to appease the koma." The cognate word *k`oma* is defined in Sacleux's (1939 [1888]) Kiswahili dictionary somewhat differently: "Esprit de mort, mânes" ('Spirit of a dead person, ancestral spirit'). The former dictionary thus describes a more pejorative sense, while the latter is more neutral. Supplementary evidence would of course be required to determine whether these differences in definition are due to actual differences in meaning at the time and place when the data was collected or due to bias on the part of the lexicographers (or some combination of these factors). The database facilitates this kind of research by allowing searches to filter results to a specific dictionary or dictionaries and to search for exact or partial matches in various fields. For instance, one could look for related evidence by searching for topical words like *spirit* or *god* or pejorative words like *evil* (and their French equivalents) in the definitions and example sentence translations of these two dictionaries. This kind of evidence provided by the database could be useful for studying the way colonial lexicographers interpreted African linguistic and cultural concepts through a European lens.

A second line of research that could be aided by the use of this database is the investigation of lexical change. The meaning or form of words can change for a variety of reasons, including language-internal factors such as phonological erosion and language-external factors such as contact with other cultures. By comparing the historical dictionaries included in this database to more recent dictionaries of the same languages, it is possible to observe and analyse these kinds of changes. A particularly fruitful line of inquiry might concern the influence of colonialism and related cultural shifts on the lexicon, since many of the dictionaries in this database represent a state of the documented languages prior to large-scale European expansion into Africa. To present just one example of the kind of lexical change we are thinking of, consider the word *chikulundine* from Maples' (1888) Yao dictionary, defined as "The third party who accompanies the bridegroom in asking the bride". Over a century later, this word is defined (written as *cikulundiine*) by Ngunga (2001) as "marriage agreement with the woman's relatives". Of course, there is the possibility that either or both of these definitions misrepresent the full scope of this word's meaning, given the scarcity of other data for confirmation. However, taking these definitions at face value, it appears as though the meaning of this word shifted over time from designating a specific role in a marriage negotiation to the marriage negotiation itself. This

would be a case of broadening or generalization, or simply metonymization in Traugott and Dasher's (2002) categorization of semantic change. The new meaning could be due to natural semantic drift but could also possibly be tied to cultural changes surrounding marriage negotiations. By making the earliest sources for many African languages accessible and searchable, this database allows for diachronic lexical analysis of these languages to be performed at a larger scale than previously possible.

A third potential use of this database is simply as an additional internet source for low-resource languages. Setting aside the specifically historical value of these dictionaries, many of them are one of the only pieces of documentation for certain languages (or at least, once incorporated into our database, one of the only pieces of documentation that is easily accessible on the internet). Despite their flaws, they may also be more thorough in some ways than other resources or include information that is left out elsewhere. Consider Maa, a low-resource language with only one online dictionary of which we are aware (Payne and Ole-Kotikash, 2008). Any dictionary will include gaps that can be filled by others, and two examples will illustrate ways in which the Maa vocabulary in our database (Erhardt and Krapf, 1857) can be a useful supplement to this other online dictionary despite its early publication. First, there are words that appear in the 1857 dictionary and not in the 2008 dictionary, such as *mésera*, a type of tree described in some detail by Erhardt and Krapf. Second, words that appear in both dictionaries may include different information, together providing a fuller picture of the word's meaning. For instance, the verb *nuk* (Erhardt and Krapf's spelling) or *a-nʊ́k* (Payne and Ole-Kotikash's spelling) is defined by both dictionaries as 'bury'. However, the 1857 entry goes on to describe burial customs among the Maasai, while the 2008 entry lists several metaphorical extensions of the word's sense, such as 'to hide or conceal information'. Although separated by 151 years and subject to the same issues of bias and lexical change just discussed, these two dictionaries can still complement one another to provide users with the most complete information possible about this language. Because no dictionary can be entirely exhaustive, the use of multiple sources is often advisable to get the best information about a word; for some low-resource languages, our database of historical dictionaries makes this possible for the first time.

7 Future improvements and discussion

As stated earlier, the database currently contains the seven dictionaries listed in Table 1. An additional 11 sources have already been selected for processing in the database's initial phase of development. Some contain data on multiple languages, and together these will provide full dictionaries of 20 languages and more limited glossaries for more than 60 languages. Once these have all been processed, the database's expansion can continue. There are more than 100 similar sources from the nineteenth and early twentieth centuries that could be added over time. As our processing becomes more streamlined, we anticipate that each new source will be ready to add to the database within two weeks (plus time for manual correction as needed). We envision a database that includes historical dictionaries, glossaries, and field notes for languages from across the continent and representing several different language families.

The current architecture of the database and its website are close to being finalized. However, there are some features that we hope to develop further that would increase the resource's functionality. Currently, entries in search results can be visualized in their original form by displaying an image of the page; we plan to refine this visualization so that just the part of the page containing the entry is displayed. This will increase the ease of use by allowing more rapid comparison of the data and source.

Other planned features relate to the translingual nature of the resource. The dictionaries included in the database (and others that will be added to it in time) use several different translation languages including English, French, and German. We hope to facilitate searching across these different languages by allowing automatic translation of search terms. When this feature is added, a search for the word "book" in the definition field could return not only results such as *chuo* 'a book' from the Mijikenda-English dictionary but also *buku* 'livre' ('book') from the Kiswahili-French dictionary. We likewise intend to provide automated translations of entry contents from any of these European languages to any other to facilitate user access to materials when browsing.

A major goal of the project is to present our data in a way that connects languages together. Our data sources largely present African languages as discrete objects, siloed by a colonial European understanding of language and ethnicity (Makoni, 1998; Chimhundu, 1992; Harries, 1987). Each named language is presented in its own dictionary, a bilingual dictionary that prioritizes its connection to a European translation language over its connection to its sister languages. This presentation inherently creates a linguistic hierarchy, with European languages as a necessary intermediary through which African languages must pass in order to connect with one another. Indeed, the development of these dictionaries was part of a colonial project that imposed a compartmentalizing and hierarchizing conception of language onto African societies (Ndhlovu and Makalela, 2021: 53-54). In contrast, the concept of translanguaging, promoted e.g. by García (2019), reflects the reality that individuals draw on a linguistic repertoire to communicate that may include different languages as traditionally defined but which are not necessarily compartmentalized as such in the minds or in the linguistic outputs of speakers. Translingual competence, fluid movement among multiple languages, has been highlighted in both the African context (Ouane and Glanz, 2010) and elsewhere (Geisler et al., 2007) as an educational priority. If we are to support this new model of language education and use, we need to provide data in formats that allow for such interconnectedness (e.g., Parton et al., 2008).

While our database represents the very dictionaries that historically imposed a monolingualizing view of language, we hope to present the data in a way compatible with translingual philosophy and practice. This conception of language is supported in our database by the ability to simultaneously search across many sources that were previously confined to a single dictionary centering a single named language. This means that words (cognate or non-cognate) that have a formal, semantic or grammatical connection may appear together in search results, regardless of their association to particular named languages. By searching for partial matches, users can identify words with similar forms across different languages, facilitating the study of cognates. However, these connections can go much further. We aim to enrich the database by linking, to the degree possible, words in languages of the Bantu family to their Proto-Bantu roots and borrowed words to the corresponding words in their source languages. This would increase the interconnectedness of the database, allowing users to see cognate words at a glance without the need for more complex searches. It would in effect remove the necessity for a European intermediary language, as users could move from one African language to another directly.

8 Conclusions

We have presented a new online database that assembles fully digitized versions of African language dictionaries, and similar linguistic resources, from the nineteenth and early twentieth centuries. The database is already live and freely accessible in its initial release, with further updates to come.

For many African languages, the earliest linguistic documentation has been difficult to access and compare, locked in print form or in unparsed scans. These historical sources are of great value to historians, linguists, and even language learners and teachers. Although outdated in some ways and influenced by historical biases, they offer a rich data supplement for many low-resource languages and provide a way to study linguistic change in the African colonial context. For some languages, the dictionaries (with example sentences) and linguistic notes made accessible here may be used fruitfully to supplement the training of large-language models, a significant challenge for low-resource languages that currently rely prominently on Bibles and the few available contemporary texts (Alhanai et al., 2025). Our database allows precise search queries and easy visualization and comparison of these lexicographic texts. It is our hope that by making old data available in this new format, this resource can be of use to a wide variety of stakeholders in the research, teaching, and promotion of African languages.

Acknowledgments

The authors would like to express gratitude to the Brigham Young University Office of Digital Humanities for their support of this project, as well as to the Research Development Office for an Interdisciplinary Research Origination Award that funded the development of the database. We would also like to thank our student research assistants

Alek Spackman, Katelyn Shellman, Aaron Shaw, Ben Carson, Emma Law, and Nils Young.

References

Tuka Alhanai, Adam Kasumovic, Mohammad M. Ghassemi, Aven Zitzelberger, Jessica M. Lundin and Guillaume Chabot-Couture. 2025. Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27). 27802–27812.

Jason Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. K. Luk, B. A. Maher, Y. Pan, C. Puhrsch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, and S. Chintala (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '24)*, Vol. 2. 929–947.

Andrew Bank and Leslie Bank. 2013. *Inside African Anthropology: Monica Wilson and Her Interpreters*. New York City, NY: Cambridge University Press.

Herbert Chimhundu. 1992. Early Missionaries and the Ethnolinguistic Factor During the 'Invention of Tribalism' in Zimbabwe. *Journal of African History*, 33(1). 87–109.

Simon Constantine. 2013. Phrasebooks and the Shaping of Conduct in Colonial Africa ca. 1884-1914. *The International Journal of African Historical Studies*, 46(2). 305–28.

James J. Erhardt and Johann Ludwig Krapf. 1857. *Vocabulary of the enguduk iloigob, as spoken by the Masai-tribes in East-Africa*. Ludwigsburg, Germany: F. Riehm.

Ofelia García. 2019. Decolonizing Foreign, Second, Heritage, and First Languages: Implications for Education. In Donaldo Macedo (ed.), *Decolonizing Foreign Language Education: The Misteaching of English and Other Colonial Languages*, 217–235. Oxford: Taylor & Francis.

Michael Geisler, Claire Kramsch, Scott McGinnis, Peter Patrikis, Mary Louise Pratt, Karin Ryding and Haun Saussy. 2007. Foreign Languages and Higher Education: New Structures for a Changed World: MLA Ad Hoc Committee on Foreign Languages. *Profession*. 234–245.

Patrick Harries. 1988. The Roots of Ethnicity: Discourse and the Politics of Language Construction in South-East Africa. *African Affairs*, 87(346). 25–52.

Johann Ludwig Krapf. 1860. *Travels, Researches, and Missionary Labours, during an Eighteen Years' Residence in Eastern Africa*. London: Trubner.

Johann Ludwig Krapf and Johannes Rebmann. 1887. *A Nika-English Dictionary*. London: Society for the Promotion of Christian Knowledge.

Jouni F. Maho. 1999. *A Comparative Study of Bantu Noun Classes* (Orientalia et Africana Gothoburgensia 13). Goteburg, Germany: Acta Universitatis Gothoburgensis.

Sinfree Makoni. 1998. In the Beginning Was the Missionaries' Word: The European Invention of an African Language: The Case of Shona in Zimbabwe. In Kwesi Kwaa Prah (ed.), *Between Distinction and Extinction: The Harmonisation and Standardisation of African Languages*, 157–64. Witwatersrand, South Africa: Witwatersrand University Press.

Chauncy Maples. 1888. *Yao-English vocabulary*. Zanzibar: Universities' Mission Press.

A. W. McGregor. 1904. *English-Kikuyu vocabulary, compiled for the use of the C.M.S. missions in East Africa*. London: Soc. for the prom. of Christ. Knowledge.

Festo Mkenda. 2018. Jesuits, Protestants, and Africa before the Twentieth Century. In Festo Mkenda and Robert Aleksander Maryks (eds.), *Encounters between Jesuits and Protestants in Africa*, 11–30. Leiden, Netherlands: Brill.

Finex Ndhlovu and Leketi Makalela. 2021. *Decolonising Multilingualism in Africa: Recentering Silenced Voices from the Global South*. Bristol: Multilingual Matters.

Armindo Ngunga. 2001. *Yao Dictionary*. Comparative Bantu OnLine Dictionary (CBOLD). http://www.cbold.ddl.cnrs.fr/

Dion Nkomo. 2020. Vernacular Lexicography in African Languages: From Early Days to the Digital Age. *Dictionaries: Journal of the Dictionary Society of North America*, 41(2). 213–43.

Adama Ouane and Christine Glanz. 2010. Why and How Africa Should Invest in African Languages and Multilingual Education. Association for the Development of Education in Africa. Hamburg, Germany: UNESCO Institute for Lifelong Learning.

Steven Paas. 2011. *Johannes Rebmann: A Servant of God in Africa before the Rise of Western*

Colonialism. Nürnberg, Germany: Verlag für Theologie und Religionswissenschaft.

Kristen Parton, Kathleen R. McKeown, James Allan and Enrique Henestroza. 2008. Simultaneous multilingual search for translingual information retrieval. In James G. Shanahan (ed.), *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, 719–728. New York: Association for Computing Machinery.

Doris L. Payne and Leonard Ole-Kotikash. 2008. *Maa Dictionary*. University of Oregon. https://darkwing.uoregon.edu/~maasai/

Derek Peterson. 1997. Colonizing Language? Missionaries and Gikuyu Dictionaries, 1904 and 1914. *History in Africa*, 24. 257–72.

George Lawrence Pilkington. 1892. *Luganda-English and English-Luganda Vocabulary*. London: Society for Promoting Christian Knowledge.

Morgan Robinson. 2022. *A Language for the World* (New African Histories). Athens, Ohio: Ohio University Press.

Charles Sacleux. 1939 [1888]. *Dictionnaire Swahili-Français*. Paris: Institut D'Ethnologie.

Gilles-Maurice de Schryver. 2003. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos 13*.

Ray Smith. 2007. An overview of the Tesseract OCR engine. *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. 629–633.

Elizabeth C. Traugott and Richard B. Dasher. 2002. *Regularity in Semantic Change*. Cambridge, UK: Cambridge University Press.

Compiling Specialised Glossaries: A Case Study of English-isiNdebele Medical Terms

Nomsebenzi Malele
Department of African Languages, University of South Africa
malelnj@unisa.ac.za

Abstract

The inability to access health care because of a language barrier is a matter of concern which this study seeks to address. There is a critical need to access health care communication by amaNdebele, an ethnic group who speak Southern Ndebele known as isiNdebele, one of the official languages of South Africa. IsiNdebele belongs to the Nguni family, alongside isiZulu, isiXhosa and Siswati. The study examines the compilation of a specialised English-isiNdebele glossary of medical terms. It investigates the methodologies and challenges that are involved in the creation of bilingual medical terminologies for English and isiNdebele. A corpus-based approach together with the prescriptive and descriptive lexicographic methods was employed. Medical corpora were compiled, and consultation with the isiNdebele speaking medical community also took place. This was done so that medical terms which are culturally appropriate could be developed. The findings of this study reveal notable gaps in the medical terminology of isiNdebele. The findings further emphasise the significance of term formation in which the medical community (nurses and doctors) and language (isiNdebele) experts are involved. This study contributes to the broader discourse on equity in healthcare. It also contributes to the intellectualisation of African languages in post-apartheid South Africa.

1 Introduction

Owing to the democratic dispensation of South Africa, the constitution mandated that twelve languages be recognised as official. As much as this has created opportunities, it has created challenges also. The challenges are impacting the development of languages particularly in specialised domains like medicine, which is the focus of this research. This study is on the compilation of an English-isiNdebele glossary of medical terms. It also examines the methodological, theoretical, as well as the practical considerations that are involved in bridging the linguistic gaps in healthcare communication, where isiNdebele dominates.

The compilation of specialised glossaries represents a critical convergence between practical language planning and lexicographic theory especially in cases where indigenous languages like isiNdebele require the development of technical vocabulary.

IsiNdebele belongs to the Nguni language group alongside isiZulu, isiXhosa, and Siswati. Although isiNdebele has official status, just like the mentioned Nguni languages of South Africa to which it belongs, this language still remains marginalised, particularly in specialised domains.

Taljard and de Schryver (2002) maintain that during the last decade, South African languages such as isiNdebele have been reduced to writing whereas languages such as isiZulu have a relatively long literary tradition. IsiZulu has more medical terminology than isiXhosa, followed by isiNdebele and Siswati. That is why the researcher compared the isiNdebele medical terminology with that of isiZulu.

The size of medical terminology creates barriers to effective healthcare delivery for amaNdebele. The development of medical terminology for isiNdebele is not just a matter of lexicographic exercise, but a matter of linguistic justice and healthcare equity as well. The research question this study seeks to answer is: *How can specialised medical glossaries be effectively compiled to serve both the linguistic preservation and practical healthcare communication needs, in the isiNdebele-speaking community?*

The significance of this study is not only confined to the compilation of terminology, but it

encompasses broader issues such as the standardisation of healthcare information, the cultural sensitivity in medical discourse, and the intellectualisation of African languages.

2 Literature Review

According to Cabré's (1999) Communicative Theory of Terminology, terms function within specific discourse communities. This theory provides a foundational understanding of how this is realised. According to this theory, the meaning of terminology is determined by the context. Medical terminology is not exempted from this, therefore, cultural appropriateness and precision must cohabit.

According to Temmerman's (2000) socio-cognitive approach, terminological units are shaped by both cognitive and cultural factors. This approach acknowledges that there may be no direct equivalents for certain medical concepts, therefore innovative ways of terminological creation should be sought.

Although the development of specialised terminology in African languages has received increasing scholarly attention, limited research exists on the development of specialised terminology in isiNdebele. Skhosana's (2009) general lexicographic work on isiNdebele served as foundational principles. Mahlangu's (2014) on the other hand focused on isiNdebele language planning.

This research accentuated the need for the development of domain specific vocabulary in the medical field. This is because the isiNdebele medical terminology remains largely unexplored, which is the gap this study is addressing. Madiba's (2001) research on South African languages' scientific terminology served as significant example for the multilingual terminological development.

Angelelli (2008) conducted research on medical terminology and interpreting, emphasising the importance of precise and error free terminology. This kind of terminology ensures desirable treatment outcomes and patients' safety. Karliner et al. (2007) also conducted research in which they demonstrate that language barriers in healthcare can lead to medical errors which may lead to compromised quality of care.

In their research on language barriers in public healthcare, Benjamin et al. (2016) discovered that patients' inability to communicate effectively with healthcare providers in their mother tongue, significantly impacts health outcomes. This they established in the South African context. The present research emphasises the significance of medical terminology in isiNdebele which will be accessible to amaNdebele.

Kilgarriff and Tugwell's (2001) work confirms significant development in corpus linguistics. Their use of sketch engines which showed collocation patterns, helped in identifying terminological patterns and contexts of usage.

Unfortunately, these methodologies introduce distinct challenges that need the adaptation of methodology when it comes to the implementation of under-resourced languages like isiNdebele.

For the identification and validation of terminology, the corpus-driven approach plays a significant role. Modern lexicographic practice heavily relies on methodologies which are corpus-based. Bowker and Pearson (2002) worked with specialised corpora and established methodological standards for the work of terminology which is corpus-based. Their emphasis was on the compilation of a domain-specific corpus and frequency-based term extraction. This provides methodological guidance for this present research.

3 Theoretical Framework

This study operates within a multidisciplinary theoretical framework that integrates insights from terminology science, lexicographic theory, and language planning. The framework is anchored by two primary theoretical constructs:

3.1 Communicative Theory of Terminology

Cabré's (1999) Communicative Theory of Terminology serves as the primary theoretical foundation, emphasising that specialised terms derive their meaning from communicative contexts rather than existing as isolated semantic units. This perspective is particularly relevant to medical terminology compilation, where terms must function effectively within specific healthcare communication scenarios.

3.2 Language Planning Theory

Cooper's (1989) comprehensive treatment of language planning provides the sociolinguistic framework for understanding terminology compilation as a form of corpus planning. This perspective situates the development of isiNdebele medical terminology within broader language planning objectives, including language standardisation, intellectualisation, and status enhancement.

4 Research Methodology

4.1 Research Design

This study employs a mixed-methods approach combining corpus-based quantitative analysis with qualitative consultation processes. The research design integrates descriptive lexicographic methods for identifying existing terminological patterns with prescriptive approaches for developing new terminological solutions where gaps exist.

The methodological approach is inherently participatory, involving isiNdebele-speaking healthcare professionals, isiNdebele interpreters and isiNdebele subject advisors in the terminology development process. This community-driven approach ensures cultural appropriateness and practical applicability of the compiled glossaries.

4.2 Data Collection

4.2.1 Corpus Development

A specialised medical corpus was compiled from multiple sources to provide the empirical foundation for terminology identification. The corpus is made up of the following:

- IsiNdebele medical information for patients.
- Transcribed conversations between isiNdebele-speaking patients and isiNdebele-speaking healthcare providers.
- Training materials for isiNdebele-speaking health workers.
- Government health policy documents translated into isiNdebele.

Regarding copyrights for the medical pamphlets, the process was daunting. It was not clear who holds copyright, this is confirmed by Bowker and Pearson (2002) when they say that the process of getting copyright permission is not a straightforward one. Regarding the transcription of this research spoken corpus, it was done by the researcher herself.

4.2.2 Expert Consultation

Semi-structured interviews were conducted with key stakeholders including medical professionals fluent in both English and isiNdebele (n=12), Parliament of South Africa isiNdebele interpreters (n=2) and isiNdebele-speaking nurses (n=10), and two isiNdebele-speaking medical doctors.

These consultations provided expert validation of terminological choices and cultural insights essential for appropriate term development.

4.3 Data Analysis

4.3.1 Corpus Analysis

For term extraction as well as frequency analysis, computational tools were employed. The AntConc concordance program, 4.3 was used. AntConc is a freeware corpus analysis toolkit which is used for concordancing and text analysis. This program is provided under a licence which is on Windows. It is from translation and language software. It has absolutely no restrictions on usage. As mentioned, this PC software is free, both its download and installation. The WordList, KeyWordList and Key Word In Context of the AntConc tool was used

WordList

This tool counted all the words in the medical corpus and presented them in an ordered list. It enabled the researcher to quickly recognise the most frequent words in a corpus. The WordList shows the terms' rank, frequency (that is, how many times the word appears) and the range of the term (the number of files the particular word appears in).

KeyWord List

Through this tool words which are unusually frequent (or infrequent) were identified in the analysis corpus (which is technical and domain-specific) in comparison with the same words in the reference corpus (which is non-technical and non-domain specific). The running corpus (RC) had 189 517 running words. The size of the RC ensures that the content of the subjects covered is diverse. The analysis corpus had 96 732 running words. This corpus comprises of a variety of medical topics. The size of the corpus provides sufficient data for frequency-based term extraction while remaining manageable for detailed qualitative analysis.

The calculations of Keynes were done in order to recognise the most significant medical terms that needed to be included in the glossary. Those terms that were occurring with frequencies above thresholds that are predetermined, were the ones selected for thorough analysis.

Key Words In Context/ Concordance Tool

This helped with the creation of an in-depth knowledge of the usage of terms in an acceptable way. It enabled the researcher of the current study to find patterns of similarity or contrast in the words surrounding the search term, for example, a term such as *udorhodere* 'doctor' was inserted to be searched.

Through this program, high-frequency medical terms were both identified and analysed in terms of their contextual usage patterns. Collocation analysis helped to reveal terminological patterns and to identify gaps that are in the existing terminology.

4.3.2 Terminological Analysis

Each identified term underwent detailed terminological analysis: In the following paragraphs the isiNdebele equivalents are analysed for instance in terms of what term-formation strategy has been used. See the following:

- **Existing isiNdebele equivalents or related terms**, for example for the English term 'pap-smear', an isiNdebele equivalent existed which is *ukuhlolwa komlomo wesibeletho*. This is translating by paraphrasing, and it literally means 'the examining of the mouth of the cervix'. Paraphrasing is translating by giving a short description or explanation (Taljard, 2008). The English term 'epilepsy' for instance has an existing isiNdebele equivalent which is *isithunthwana* which was created from seemingly nothing.
- **Cultural appropriateness and sensitivity considerations.** According to this sub-topic, culture shapes how terms are formed. It is a fact that there are words that according to culture are deemed sensitive and inappropriate. Instead of such words a more polite one is used. The words that are deemed inappropriate are considered taboo and impolite. Such words are embarrassing to talk about. These are words about sex, private body parts, death, to mention but a few. For example, the isiNdebele equivalent for the English word 'miscarriage' is *ukubuya endleleni* which literally translates as 'to come back from the road'. It is a euphemistic word, used instead of *ukuphunyelwa sidisi* which is deemed insensitive and vulgar according to the culture. Both these isiNdebele examples are paraphrased. The term 'to die' is translated as *ukulala* 'to sleep' or *ukukhamba* 'to go' to avoid *ukufa* which is deemed insensitive.

  The above examples also talk to the register considerations as well as usage contexts. Register refers to how language is used based on social context and audience. Choice of words is evident in the isiNdebele spoken corpus more than its written counterpart, particularly in the case of terms that are regarded as vulgar and inappropriate (as mentioned in the aforegoing paragraph). For instance, sexual intercourse is referred to as *ukuya emsemeni* in the written corpus. In the spoken corpus it is referred to as *ukulalana* 'to sleep with each other'*, ukukha umrorho* and *ikonzo yabantu abadala*; the very act is called *ukuya emsemeni* which loosely translates as 'to go to the reed-mat' or *ukukha umrorho* which loosely translates as 'to pick up the vegetables' or *ikonzo yabantu abadala* which loosely translates as 'old people's service or fellowship'. This is euphemistic. Spoken language is characterised by more euphemism as compared to its written counterpart (Malele, 2021).

- **Morphological and phonological adaptation requirements.** Mahlangu (2007) argues that in the case of isiNdebele, the loan words from English that have consonant clusters such as /sl/, /sq/, /sp/, /st/ and /sch/ insert a vowel. See the following example: The English term 'scanner' is written as *isikena* in isiNdebele (Mahlangu, 2007).

4.3.3 Validation Procedures

Multiple validation procedures were implemented to ensure the quality and appropriateness of compiled terminology. Since human beings will always be final judges in any terminological work, expert reviews by medical and linguistic professionals were implemented. IsiNdebele speaking doctors and isiNdebele Curriculum Implementers (for Further Education and Training) assisted with the verification of terms. For instance, most isiNdebele translations refer to the English 'groin' as *imbilapho*. The experts maintain that *imbilapho* is actually a 'lymph node' in English and not a 'groin'. A 'groin' is not a medical condition, but a body part.

5 Findings and Discussion

5.1 Terminological Gaps and Challenges

The corpus analysis revealed significant gaps in isiNdebele medical terminology, particularly in specialised medical terms. Approximately 70% of identified high-frequency medical terms lacked established isiNdebele equivalents, necessitating new term creation or adaptation. The gaps in the isiNdebele medical terminology were identified by comparing the isiNdebele terms with the isiZulu basic medical terms available at http://hdl.handle.net/10386/3516. The researcher saw it fit to compare isiNdebele

with isiZulu, as it is one of the most developed languages in the Nguni family, to which isiNdebele belongs.

These gaps reflect the historical marginalisation of isiNdebele in formal medical education and practice, creating barriers to effective healthcare communication.

5.2 Terminological Development Strategy

To address the identified terminological gaps, morphological adaptation was used, as a strategy:

The morphology of isiNdebele is agglutinative. Transliteration has proven to be the most preferred method of term formation, amongst amaNdebele. The transliterated term *ihayibhladi* for 'high blood pressure' is used more than its coined counterpart, which is *isigandelelo seengazi*. Following transliteration is paraphrasing. It is the second most preferred method of term formation used by amaNdebele when creating terms. For example, the English term 'dentist' is referred to as *udorhodere wamazinyo* (doctor of teeth) and 'dermatologist' as *udorhodere wesikhumba* (doctor of skin).

The third mostly used term-formation method in isiNdebele is compounding. The challenge with compounding has been the issue of hyphenation. There is a lot of inconsistency as far as compounds are concerned. Mahlangu (2013) maintains that the problem with isiNdebele compound nouns is the issue of hyphenation. Mahlangu (2013) says that some compounds are hyphenated, and some are unhyphenated for no apparent reason. The isiNdebele term *umbulalasihlangu* (AIDS) for instance, is sometimes written as *umbulala-sihlangu* (with a hyphen) and in other instances as *umbulalasihlangu* (without a hyphen) (Mahlangu, 2013).

6 The Study's Contributions

The combination of corpus-based analysis with health community consultation provides a balanced approach that ensures both empirical grounding and cultural appropriateness and acceptance. The validation procedures developed for this study provide templates for quality assurance in specialised terminology compilation for African languages.

The compiled glossaries provide immediate practical resources for healthcare communication in isiNdebele-speaking communities. These materials can support medical interpreter training programs, healthcare provider competency training, and community health worker capacity building.

Ikhwezi National Language Board was established to carry the mandate of isiNdebele terminology development. Although there are terminology lists that have been approved by the board, there is no list on isiNdebele medical terms. This work will contribute and speed-up the development of isiNdebele medical terms and also assist this board in employing computational methods.

The research findings have significant implications for language policy in healthcare settings: They underscore the importance of including indigenous languages in medical education curricula and requirements for healthcare interpreting services in indigenous languages.

7 Future Research Directions

This study opens several avenues for future research: Expansion to other South African indigenous languages, the development of multimedia terminology resources and investigation of terminology adoption and standardisation processes.

8 Conclusion

The compilation of a specialised English-isiNdebele glossary of medical terms promotes equitable healthcare access while it addresses historical marginalisation. The research findings reveal both the possibilities and the challenges that are fundamental in developing specialised terminology for indigenous South African languages particularly isiNdebele, as it is the focus of this paper. While there are significant gaps that exist in isiNdebele medical vocabulary, currently such gaps are gradually bridged by the fact that the isiNdebele speaking medical professionals have the innovative capacity to provide powerful foundations for terminological development.

The theoretical framework developed for this study, combining communicative terminology theory with the socio-cognitive approach, acknowledges that even if there are no direct isiNdebele equivalents for certain medical concepts, innovative ways of terminological creation should be sought and implemented. Language planning approaches offer a vigorous foundation for similar research in other African language contexts.

This research is a testament to the fact that when isiNdebele and other indigenous languages

are given the necessary developmental support, such languages can actually serve a specialised discourse function. This research also contributes to the broader project of African language intellectualisation. The compiled glossary represents concrete resources for improving healthcare communication.

The development of specialised terminology in isiNdebele represents both a constitutional imperative and a practical necessity for achieving a true equitable healthcare delivery.

The success of this terminology compilation project depends ultimately on institutional adoption and community usage. Future efforts must focus on implementation strategies that ensure the developed terminology becomes integrated into actual healthcare practice, serving the communities whose linguistic heritage it seeks to preserve and develop.

Limitations of the Study

Several limitations constrain the generalisability of this research: Geographic concentration in specific isiNdebele-speaking regions, limited representation of specialised medical sub-domains, time constraints preventing longitudinal validation of terminology adoption and resource limitations affecting corpus size and diversity.

Ethical Considerations

This research received ethical approval from the University Ethics Committee and followed established protocols for community-based research. Informed consent was obtained from all participants, and community ownership of developed terminology was established through formal agreements with representative community organisations.

References

Claudia V. Angelelli. 2008. *Medical interpreting and cross-cultural communication*, Cambridge University Press, Cambridge.

Ereshia Benjamin, Leslie Swartz, Linda Hering and Bonginkosi Chiliza. 2016. Language barriers in health: lessons from experiences of trained interpreters working in public sector hospitals in the Western Cape, South African Health Review, pages 73-81. https://hdl.handle.net/10520/EJC189317

Lynne Bowker and Jennifer Pearson. 2002. *Working with specialized language: A practical guide to using corpora*, Routledge. DOI:10.4324/9780203469255

Maria Teresa Cabré. 1999. *Terminology: Theory, methods, and applications*, John Benjamins, Amsterdam/Philadelphia. DOI: https://doi.org/10.7202/004006ar

Robert L. Cooper. 1989. *Language planning and social change*, Cambridge, Cambridge University Press. https://doi.org/10.1017/CBO9780511620812

Leah S Karliner, Elizabeth A Jacobs, Alice Hm Chen and Sunita Mutha. 2007. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature, *Health Services Research*, 42(2):727–754. doi.org/10.1111/j.1475-6773.2006.00629.x

Adam Kilgarriff and David Tugwell. 2001. WORD SKETCH: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*, pages 32-38. Toulouse, Association for Computational Linguistics.

Mbulungeni Madiba. 2001. *Multilingual terminology development in South Africa*. Johannesburg: University of the Witwatersrand Press.

Katjie S Mahlangu. 2007. Adoption of loan words in isiNdebele. Unpublished M.A. Mini Dissertation, Pretoria, University of Pretoria.

Katjie S Mahlangu. 2013. Challenges to the usage of a hyphen in compound nouns in Southern Ndebele, *G.S.T.F. International Journal on Education*, 1(1):107–111.

Katjie S Mahlangu. 2014. Language planning and policy in post-apartheid South Africa: The case of isiNdebele, *Language Matters*, 45(2):223–240.

Nomsebenzi J Malele. 2021. The use of corpora in the compilation of a specialised English-isiNdebele glossary of medical terms. Unpublished DLitt et Phil thesis, Pretoria, University of South Africa. https://hdl.handle.net/10500/28164.

Philemon B. Skhosana. 2009. *The linguistic relationship between Southern and Northern Ndebele*, Pretoria, University of Pretoria Press.

Elsabé Taljard. 2008.Issues in scientific terminology in African Languages. Paper presented at the CEntrePol workshop held at the University of Pretoria on 29 March 2007, Supported by the French Institute of Southern Africa, pp.88-91.https://hal.science/file/index/docid/799553/filename/Cahiers-

Elsabé Taljard and Gilles-Maurice de Schryver. 2002. Semi-automatic term extraction for the African

languages, with special reference to Northern Sotho, *Lexikos*, 12:44-74. https://doi.org/10.5788/12-0-760

Rita Temmerman. 2000. *Towards new ways of terminology description: The sociocognitive approach*, Amsterdam/Philadelphia, John Benjamins. https://doi.org/10.1075/tlrp.3

Mafoko: Structuring and Building Open Multilingual Terminologies for South African NLP

Vukosi Marivate[1,2,3], Isheanesu Dzingirai[1], Fiskani Banda[1], Richard Lastrucci[1], Thapelo Sindane[1], Keabetswe Madumo[1], Kayode Olaleye[1], Abiodun Modupe[1], Unarine Netshifhefhe[1], Herkulaas Combrink[4,5], Mohlatlego Nakeng[1], Matome Ledwaba[1]

[1]Data Science for Social Impact, Dept. of Computer Science, University of Pretoria,
[2]AfriDSAI, University of Pretoria, [3]Lelapa AI,
[4]Economics and Management Sciences, University of the Free State,
[5]Interdisciplinary Centre for Digital Futures, University of the Free State

Correspondence: vukosi.marivate@cs.up.ac.za

Abstract

The critical lack of structured terminological data for South Africa's official languages hampers progress in multilingual NLP, despite the existence of numerous government and academic terminology lists. These valuable assets remain fragmented and locked in non-machine-readable formats, rendering them unusable for computational research and development. *Mafoko* addresses this challenge by systematically aggregating, cleaning, and standardising these scattered resources into open, interoperable datasets. We introduce the foundational *Mafoko* dataset, released under the equitable, Africa-centered NOODL framework. To demonstrate its immediate utility, we integrate the terminology into a Retrieval-Augmented Generation (RAG) pipeline. Experiments show substantial improvements in the accuracy and domain-specific consistency of English-to-Tshivenda machine translation for large language models. *Mafoko* provides a scalable foundation for developing robust and equitable NLP technologies, ensuring South Africa's rich linguistic diversity is represented in the digital age.

1 Introduction

The advancement of Natural Language Processing (NLP) is fundamentally tied to the availability of high-quality language resources. However, the vast majority of the world's languages, including the 12 official languages of South Africa, remain critically under-resourced in this regard (Joshi et al., 2020). This scarcity creates a significant bottleneck for technological development and linguistic preservation. While substantial government and academic initiatives in South Africa have produced multilingual terminology lists over the years (Taljard, 2015), these valuable assets remain largely fragmented, locked in non-machine-readable formats like PDFs (sometimes available as scanned PDFs or flattened text without reusable structure), and lack the standardised structure required for modern computational applications.

To bridge this critical gap, we introduce *Mafoko*[1] the South African curated Terminology, Lexicon, and Glossary Project. The mission of *Mafoko* is not to create new terminology from scratch, but to systematically aggregate, digitise, and standardise these scattered, publicly-funded terminological assets. By transforming them into interoperable, machine-readable formats, we unlock their potential for a new wave of linguistic and computational applications.

This paper presents the foundational work and initial release of the *Mafoko* project. Our primary contributions are threefold: First, we release the first version of the *Mafoko* dataset, a structured, multilingual terminology resource covering key domains for South African languages. Second, we release this dataset under the novel, Africa-centered Nwulite Obodo Open Data License (NOODL) to ensure equitable data governance and local benefit-sharing (Okorie and Omino, 2024). Third, we demonstrate the dataset's immediate practical value by integrating it into a Retrieval-Augmented Generation (RAG) pipeline, which yields substantial improvements in machine translation accuracy and consistency for an English-to-Tshivenda language pair.

Ultimately, *Mafoko* provides both a practical resource and a scalable framework for fostering robust NLP and language technologies that reflect the rich linguistic diversity of South Africa.

2 Motivation

South Africa's official languages, with the exception of English and to a lesser extent Afrikaans, remain critically under-resourced in the digital domain (Joshi et al., 2020). Despite significant investment from state institutions—including the Depart-

[1]Mafoko is a Setswana (TSN) word that means words.

ment of Sports, Arts and Culture (DSAC), the Pan South African Language Board (PanSALB), and Statistics South Africa (StatsSA), in creating terminologies for crucial domains, these valuable assets are largely unusable for modern NLP. The primary barriers are both technical and legal: resources are frequently published as static, non-machine-readable documents (Handbook) and often lack the clear, permissive licensing required for computational reuse and research. This systemic inaccessibility hinders technological development and undermines efforts to achieve linguistic equity in South Africa's digital sphere. We acknowledge the work of the South African Centre for Digital Language Resources (SADILAR) which has worked to enable creation, preserve some these resources or provide ways to keep track of them (for example through the *LwimiLinks*[2] project). *LwimiLinks* still faces the challenge that the data linked or available is mostly in PDF.

South African universities are also key actors in this landscape, developing linguistic resources in response to national policies that mandate the use of indigenous languages in higher education (of Arts and Culture, 2003; of Higher Education and Training, 2020). Language units at institutions like the University of KwaZulu-Natal, the University of Pretoria, and North-West University have produced valuable discipline-specific glossaries and corpora. However, these academic contributions often suffer from the same fate as government resources: they remain siloed within institutional repositories, lacking the standardisation and interoperability required for broad integration into NLP and AI ecosystems. The need for interventions like the Universities South Africa Community of Practice for African Languages (COPAL) highlights this persistent fragmentation, which a systematic project like *Mafoko* is designed to address.

The core motivation for *Mafoko* is to unlock the potential of these dormant linguistic assets. By systematically digitising (Taljard et al., 2022), structuring, and releasing these resources under equitable licenses (Okorie and Omino, 2024; Rajab et al., 2025) that adhere to FAIR principles (Wilkinson et al., 2016), we can directly enhance AI and NLP capabilities for South Africa's indigenous languages. Properly structured terminologies can be ingested to fine-tune large language models (LLMs), improve machine translation, and power a new generation of inclusive technologies like AI-driven spell checkers and voice assistants. This, in turn, empowers linguists, educators, and innovators to build culturally relevant, domain-specific applications, from healthcare diagnostics in isiZulu to financial literacy tools in Setswana. The transition to open, machine-readable resources is therefore a critical step towards ensuring that South Africa's languages not only survive but thrive in the digital era, fulfilling the multilingual promise of both government and higher education policies.

3 Methodology

The methodology for *Mafoko* is centered on the curation, standardisation, and dissemination of existing linguistic resources, rather than the creation of new terminology from scratch. Our approach systematically aggregates terminologies from disparate sources to enhance their accessibility and utility for linguistic research, education, and computational applications.

3.1 Source Identification

The initial phase involved identifying and collating terminological resources created and archived by South African universities, government departments, and research institutions. Universities, often as part of their language policy implementation, develop such resources, though many are in non-machine-readable formats like PDF. We engaged with the DSAC to assess their portfolio of commissioned terminology projects. Furthermore, the extensive terminology repositories maintained by *Statistics South Africa (StatsSA)*[3] and other parastatal bodies were identified as primary data sources.

3.2 Domain Coverage

The scope of *Mafoko* is intentionally domain-agnostic, allowing for the inclusion of terminology lists from a wide array of fields. For instance, the DSAC lists encompass domains such as Information and Communication Technology (ICT), Mathematics, Finance, Health Sciences, and Parliamentary Procedure. The StatsSA collection provides comprehensive multilingual terminology for statistics. Similarly, the Open Educational Resource Term Bank (OERTB) project focused on developing African language terminologies for higher education across multiple disciplines (University of

[2]LwimiLinks is a place for multilingual terminology and other useful language resources. https://lwimilinks.sadilar.org

[3]https://www.statssa.gov.za

Pretoria, 2019; Taljard, 2015). This broad coverage ensures the dataset's utility across diverse research and application contexts. We include the University of Pretoria multilingual glossary, the UNISA Multilingual Robotics Glossary: South African Languages Version[4], Multilingual Linguistic Terminology Project (SAMLT)[5] (Griesel and Mojapelo, 2022) and the AI Terminologies in African Languages (in African Languages Contributors, 2025).

3.3 Challenges in Data Acquisition

A primary challenge was overcoming the fragmented and often inaccessible nature of the source data. This included navigating licensing constraints, which were often unclear or restrictive, and dealing with access limitations, such as portals that only permit single-term queries. The heterogeneity of data formats, ranging from scanned PDFs to structured spreadsheets—required significant and bespoke pre-processing efforts. These hurdles are emblematic of the broader challenges in language resource development for African languages (Taljard et al., 2022). Even within a single source like DSAC, we observed inconsistencies in formatting, such as the representation of part-of-speech, across different terminology lists.

3.4 Data Curation and Structuring

The curation pipeline began with automated data extraction. Since much of the source material was in PDF format, we developed a modular extraction pipeline using Python-based tools. The pipeline required custom adaptations for each document's unique structure, as illustrated by the formatting differences between the DSAC (Figure 1a) and StatsSA (Figure 1b) sources. For some resources, such as the StatsSA list, we were fortunate to be privately provided with a spreadsheet version, which greatly simplified the initial processing.

Automated extraction was followed by extensive manual post-processing to ensure the dataset's quality and utility. A dedicated team member performed detailed cleaning to correct extraction errors, remove artefacts like page headers and garbled characters, and reconstruct table structures to maintain one-to-one alignment between source and target terms. To preserve the authenticity of the original resources, orthographic (use of hyphens, inconsistent capitalisation, or accent marks in lexical entries) and formatting variations from the source documents were retained. Where multiple translations existed for a single term, all variants were included to enable the study of lexical variation, synonymy, and regional differences. The statistics of the datasets are available in Table 1.

Each record was enriched with provenance metadata, including the originating institution, publication date, and contributor information where available. To ensure interoperability, all languages were standardised using ISO 639-3 codes. The data is currently released in CSV and JSON formats, with a TermBase eXchange (TBX) version planned for future releases. This foundational dataset (v0) is designed for iterative improvement, with future work planned to incorporate part-of-speech tags, semantic domain classification, and TEI-compliant lexicographic structuring (Burnard et al., 2014).

3.5 Data Release and Availability

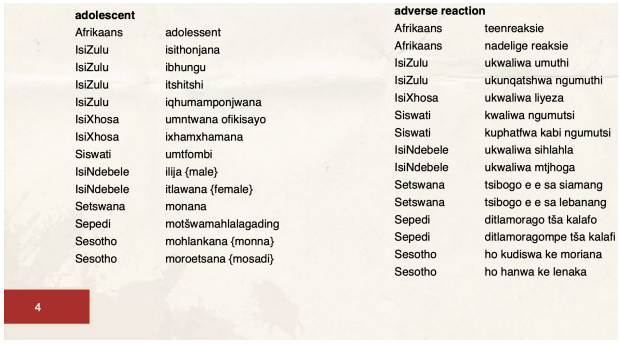
The dataset is openly licensed (see Section 3.6) and accessible on multiple platforms, including GitHub[6], Zenodo[7] (Marivate et al., 2025), and HuggingFace collection[8], to align with FAIR data principles (Wilkinson et al., 2016). We are working with SADILAR to contribute to their *LwimiLinks* initiative (Mafoko is already listed on there), but we also hope to have a mirror of Mafoko available via SADILAR's digital mirrors. We plan to provide both bulk download options and API access. A feedback and validation interface is also under development to enable community-driven refinement by linguists, translators, and other stakeholders. This approach supports a virtuous cycle of continuous improvement, ensuring the resource remains relevant and accurate over time.

3.6 Licensing under NOODL

Standard open licenses, while promoting reuse, often fail to address the power asymmetries and historical contexts inherent in community-generated data. To ensure equitable governance, *Mafoko* adopts the *Nwulite Obodo Open Data License* (NOODL), an African-centered framework designed to protect local agency and mandate fair benefit-sharing (Okorie and Omino, 2024).

In contrast to generic licenses, NOODL differentiates access based on user context, mandates reinvestment from commercial use by entities out-

[4] https://ir.unisa.ac.za/handle/10500/30440
[5] https://hdl.handle.net/20.500.12185/669
[6] https://github.com/dsfsi/za-mafoko
[7] https://doi.org/10.5281/zenodo.17484991
[8] https://huggingface.co/collections/dsfsi/za-mafoko

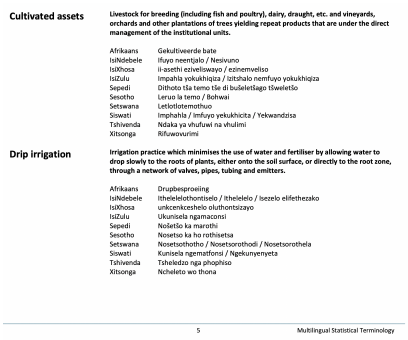


adolescent	
Afrikaans	adolessent
IsiZulu	isithonjana
IsiZulu	ibhungu
IsiZulu	itshitshi
IsiZulu	iqhumamponjwana
IsiXhosa	umntwana ofikisayo
IsiXhosa	ixhamxhamana
Siswati	umtfombi
IsiNdebele	ilija {male}
IsiNdebele	itlawana {female}
Setswana	monana
Sepedi	motšwanahlalagading
Sesotho	mohlankana {monna}
Sesotho	moroetsana {mosadi}

adverse reaction	
Afrikaans	teenreaksie
Afrikaans	nadelige reaksie
IsiZulu	ukwaliwa umuthi
IsiZulu	ukunqatshwa ngumuthi
IsiXhosa	ukwaliwa liyeza
Siswati	kwaliwa ngumutsi
Siswati	kuphatfwa kabi ngumutsi
IsiNdebele	ukwaliwa sihlahla
IsiNdebele	ukwaliwa mtjhoga
Setswana	tsibogo e e sa siamang
Setswana	tsibogo e e sa lebanang
Sepedi	ditlamorago tša kalafo
Sepedi	ditlamoragompe tša kalafi
Sesotho	ho kudiswa ke moriana
Sesotho	ho hanwa ke lenaka

4

(a) DSAC HIV Terminology snippet



Cultivated assets	Livestock for breeding (including fish and poultry), dairy, draught, etc. and vineyards, orchards and other plantations of trees yielding repeat products that are under the direct management of the institutional units.
Afrikaans	Gekultiveerde bate
IsiNdebele	Ifuyo neentjalo / Nesivuno
IsiXhosa	Ii-asethi ezivelisiwayo / ezinemveliso
IsiZulu	Impahla yokukhiqiza / Izitshalo nemfuyo yokukhiqiza
Sepedi	Dithoto tša temo tše di bušeletšago tšweletšo
Sesotho	Lerua la temo / Bohwai
Setswana	Letlotlotemothuo
Siswati	Imphahla / Imfuyo yekukhicita / Yekwandzisa
Tshivenda	Ndaka ya vhufuwi na vhulimi
Xitsonga	Rifuwovurimi
Drip irrigation	**Irrigation practice which minimises the use of water and fertiliser by allowing water to drop slowly to the roots of plants, either onto the soil surface, or directly to the root zone, through a network of valves, pipes, tubing and emitters.**
Afrikaans	Drupbesproeiing
IsiNdebele	Itheleletothontselo / Itheleletlo / Isezelo elifethezako
IsiXhosa	umkcenkcesheloluthontsizayo
IsiZulu	Ukunisela ngamaconsi
Sepedi	Nošetšo ka marothi
Sesotho	Nosetso ka ho rothisetsa
Setswana	Nosetsothotho / Nosetsorothodi / Nosetsorothela
Siswati	Kunisela ngematfonsi / Ngekunyenyeta
Tshivenda	Tsheledzo nga phophiso
Xitsonga	Nchelelo wo thona

5 Multilingual Statistical Terminology

(b) StatsSA Multilingual Terminology snippet

Figure 1: Formatting differences across different terminology lists.

side developing regions, and includes provisions to reinforce community control. For *Mafoko*, this means:

1. South African and other African researchers gain open access with minimal barriers.

2. Community contributors are credited, and downstream use must return value to the originators.

3. Commercial use by external entities requires negotiated terms, correcting historical dataflow asymmetries.

NOODL enables researchers to develop and share with the common agenda, to both promote innovation as well grow the available data for under resourced languages.

4 Terminology Applications

The *Mafoko* datasets are not merely curated artifacts; they are critical assets for both computational evaluation and linguistic inquiry. Their primary applications fall into two key areas: providing a much-needed benchmark for multilingual NLP models and enabling deep analysis of language in a multilingual context.

4.1 A Benchmark for Multilingual NLP Evaluation

A significant bottleneck in developing NLP for African languages is the lack of standardized, domain-specific evaluation benchmarks (Adelani et al., 2023). *Mafoko* directly addresses this gap by providing gold-standard terminologies that can be used to rigorously assess model performance.

One primary use case is in evaluating the cross-lingual consistency of machine translation systems. Given an English term, a model can be prompted to produce translations in various South African languages. These machine-generated outputs can then be quantitatively compared against the ground-truth terms in the dataset, revealing a model's ability to handle domain-specific vocabulary.

Furthermore, the aligned multilingual terms enable the evaluation of multilingual word embeddings (Ruder et al., 2019; Upadhyay et al., 2016). By performing cluster analysis or measuring the cosine similarity of term-pairs (Almeida and Xexéo, 2019), researchers can probe how well semantically equivalent concepts are co-located within a shared vector space (Glavaš et al., 2019). This provides crucial insights into the quality of cross-lingual representations, particularly for low-resource languages. By serving as an evaluation resource, *Mafoko* supports the kind of participatory and community-centered benchmarking necessary for building truly useful technologies (Nekoto et al., 2020).

4.2 A Resource for Linguistic and Sociolinguistic Inquiry

Beyond computational applications, the dataset offers rich opportunities for linguistic and lexicographic research. It serves as a valuable corpus for studying the dynamics of language contact, standardisation, and change in South Africa, mirroring the kind of corpus-driven analysis that has been foundational to modern lexicography for African languages (Prinsloo and De Schryver, 2001).

Since the dataset preserves multiple translations for many terms, it facilitates the study of lexical variation (Freixa, 2022), synonymy, and dialectal preferences. Linguists can use this data to investigate term-formation strategies across languages, examining whether translations are neologisms,

Table 1: Overview of the datasets aggregated in the initial release of *Mafoko* (v0).

Source	Primary Domains/Categories	Languages	Entries
DSAC (Combined)	Multiple (Finance, Health, ICT, Law, Mathematics, Arts, Science, Elections)	11	15,554
AI Terminologies in African Languages Dataset	AI terminologies	4	85
OERTB	Higher Education Terminology	11	5,744
StatsSA	Official Statistics (Demography, Economics, Labour, Health, Geography)	11	1,160
Unisa Robotics*	Robotics Terminology	11	100
Unisa Multilingual	Provide definitions for linguistic terms in nine languages	9	778
UP Glossary	Academic Terminology	3	1,768
Total Entries			**25,189**

* Unisa Robotics glossary is re-released under CC-BY-NC-SA (not compatable with NOODL) but formated like all other Mafoko releases.

calques, semantic extensions, or borrowings. Such analysis can reveal deeper cognitive or conceptual distinctions between languages.

Moreover, the data provides a unique lens for sociolinguistic inquiry into language planning and policy. It captures the outcomes of official terminology development efforts, allowing researchers to analyze the tensions between top-down standardization and organic, community-level usage. This makes the dataset an essential resource for scholars studying the politics of language and curriculum design in multilingual societies, a challenge common across the African continent (Heugh and Stroud, 2019).

5 Improving Translations with Terminology lists and RAG

To demonstrate the practical value of the *Mafoko* terminologies, we conducted experiments to assess their impact on improving machine translation quality for a low-resource language pair. Despite advances in LLMs, their performance often degrades when translating domain-specific or rare terms, especially for languages with limited high-quality parallel corpora like South Africa's (Zhong et al., 2024). This can lead to critical misinterpretations, such as confusing the term *register* in a mathematics context (*ridzhisitara*) versus an electoral one (*redzhistara*) in Tshivenda.

Our experiment investigates whether a Retrieval-Augmented Generation (RAG) pipeline, enriched with our curated terminology, can mitigate these issues. The overall pipeline is visualized in Figure 2.

5.1 Task and Models

We evaluated English-to-Tshivenda translation in two distinct domains: Mathematics and Election, using terminology lists from the DSAC. The datasets that were used for building the RAG terminology were English–Tshivenda Election and Mathematics (Grade R to 6) domain datasets. The election terminology dataset consisted of 576 English–Tshivenda term pairs and their associated context information that defines how a term is used. The mathematics terminology dataset consisted of 966 English–Tshivenda term pairs and their associated context information. The data extraction from PDF files, cleaning and aligning process was auto-

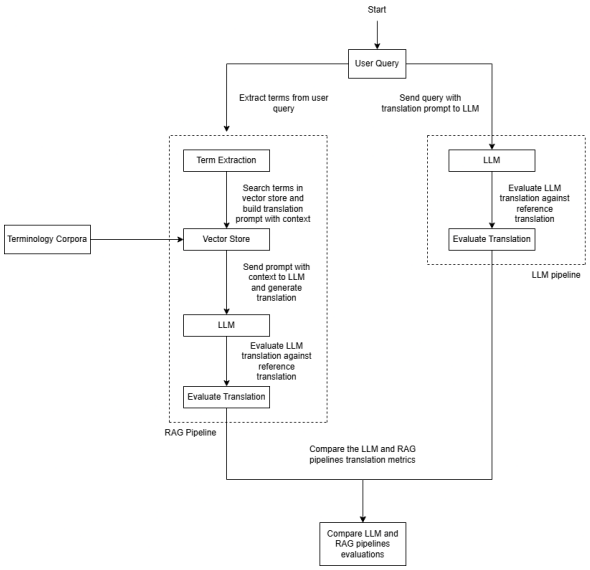


Figure 2: Overview of RAG and LLM Pipelines

mated through a Python script. We used two large language models to assess the impact of our RAG approach: the high-performance GPT-4o-mini and the open-source LLaMA3-8B.

5.2 Experimental Conditions

We tested each model under three conditions to isolate the effect of the RAG pipeline:

1. **No RAG (Baseline):** The LLM was prompted to perform direct translation without any additional context.

2. **RAG with Semantic Terms:** Key terms (nouns, verbs, adverbs) were extracted from the source text using spaCy's `en_core_web_sm` model. These terms were used to retrieve relevant entries from the *Mafoko* vector store to augment the LLM's prompt.

3. **RAG with Rare Terms:** Terms were selected from the source text based on their low frequency in general English corpora (Reuters (Lewis, 1997), Inaugural Speeches(Ahrens, 2021)) using the `wordfreq` library. This strategy focuses the retrieval on the most challenging, domain-specific vocabulary.

In both RAG conditions, the retrieved translations and definitions were appended to the prompt, providing in-context examples to guide the LLM.

5.3 Evaluation Metrics

We evaluated translation quality using 20 elections and 20 mathematics test set pairs using standard automatic metrics: BLEU for n-gram precision, and both chrF and chrF++ for character n-gram recall. Higher scores indicate better translation quality. The test sets were generated using: the 2025/2026 annual teaching plans for grade 4 mathematics for mathematics; the 2016 local government elections post proclamation leaflet and former President Kgalema Motlanthe'S 2009 State of the Nation Address for elections. Results are presented in Table 3 and example sentences used for evaluation are shown in Table 2.

Domain	Example Sentences (English and Tshivenda)
Election	**English:** Local government is in your hands! **Tshivenda:** Muvhuso wapo u zwanani zwavho!
Mathematics	**English:** The difference between numbers. **Tshivenda:** Phambano i re vhukati ha nomboro.

Table 2: Example English and Tshivenda Sentences by Domain

5.4 Results and Analysis

5.4.1 Quantitative Results

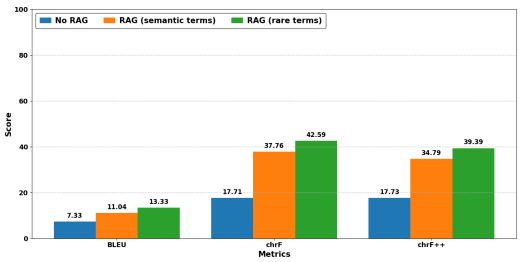
As shown in Table 3, the inclusion of a RAG pipeline with *Mafoko* terminologies leads to substantial improvements in translation quality across all metrics for both models and domains.

For GPT-4o-mini, the gains are significant. In the Mathematics domain, BLEU score improves from 7.33 to **13.33** and chrF++ score from 17.73 to **39.39** using the rare-term RAG. In the Election domain, the semantic-term RAG yields the best results, increasing the BLEU score from 5.87 to **10.41** and chrF++ from 26.99 to **36.85**.

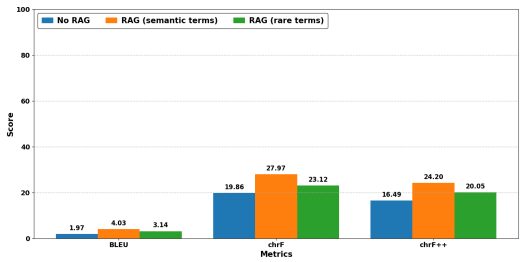
For LLaMA3-8B, while its overall performance is lower than GPT-4o-mini's, it also benefits greatly from RAG. In both domains, the semantic-term RAG provides the best results. For the Election domain, BLEU improves from 1.97 to **4.03** and chrF++ from 16.49 to **24.20**. The performance gains are visualized in Figure 3a and 3b.

Model	Setup	Mathematics			Election		
		BLEU	chrF	chrF++	BLEU	chrF	chrF++
GPT-4o-mini	No RAG	7.33	17.71	17.73	5.87	29.21	26.99
	RAG (semantic terms)	12.44	41.32	38.95	**10.41**	**40.35**	**36.85**
	RAG (rare terms)	**13.33**	**42.59**	**39.39**	9.73	39.88	35.42
LLaMA3-8B	No RAG	2.28	12.43	10.60	1.97	19.86	16.49
	RAG (semantic terms)	**4.54**	**22.66**	**20.29**	**4.03**	**27.97**	**24.20**
	RAG (rare terms)	3.72	20.01	18.56	3.52	26.31	22.89

Table 3: Translation performance of GPT-4o-mini and LLaMA3-8B across different setups with and without RAG. Best scores per model and domain are in bold.



(a) GPT-4o-mini English to Tshivenda Mathematics



(b) LLaMA3-8B English to Tshivenda Election

Figure 3: Translation performance metrics comparison of GPT-4o-mini and LLaMA3-8B models on English to Tshivenda Mathematics and Election datasets.

5.4.2 Analysis

Although the results are low due to the use of early baseline models, the lack of fine-tuning, and a small test set size of 20 samples per domain, the results nonetheless strongly indicate that providing in-context, domain-specific terminology via RAG is a highly effective method for improving LLM translation performance for low-resource languages with the results almost doubling. The fact that this holds true for both a state-of-the-art proprietary model and a smaller open-source model underscores the robustness of this approach.

Interestingly, the optimal RAG strategy differed between the models. For LLaMA3-8B, retrieving based on semantic terms was consistently better. This suggests the model benefits from guidance on a broader range of vocabulary. For the more capable GPT-4o-mini, the rare-term strategy proved superior in the highly specialized Mathematics domain. This may indicate that the model already possesses a strong grasp of common semantic terms, and its performance is most improved by providing context for the most niche and infrequent vocabulary. The overall performance gap between the two models likely reflects differences in their pretraining data and inherent capabilities for handling low-resource languages.

5.5 Discussion

These promising results with Tshivenda open several avenues for future work. First, this evaluation framework should be extended to the other official South African languages to confirm the generalizability of our findings. Second, it would be valuable to investigate why the rare-term RAG strategy was particularly effective for GPT-4o-mini and whether this pattern holds across other domains and models.

6 Future Directions and Call for Open Data

The expansion of *Mafoko* depends on the continued identification and integration of scattered terminological resources. As shown in Table 4, numerous valuable glossaries exist across South African institutions, but their accessibility varies dramatically, from openly licensed datasets like Unisa's robotics glossary to web portals from Stellenbosch University that prohibit bulk download.

Table 4: Examples of Additional Terminological Resources for Integration.

Resource Name	**Institution/Body**	**Accessibility Status**
Termbank[1]	UKZN	Offline (as of 31/07/2025)
Full OERTB[2]	UP	Offline (as of 31/07/2025)
LwimiLinks [3]	SADILAR	Accessible (links to other org resources + PDFs)
Trilingual Wine Industry Dictionary[4]	SA Wine Industry	Accessible (No clear license for reuse)
Multilingual Glossaries[5]	Nelson Mandela Uni.	Accessible (PDFs)
Mechanical Engineering and Statistics and Economics Glossaries [6]	UCT	Accessible (PDF)
Trilingual Terminology Web[7]	Stellenbosch Uni.	Accessible (Web search, no download)
Statistical Terms Glossary[8]	Stellenbosch Uni.	Accessible (PDF)
BAQONDE Resources (Polokelo)[9]	Multiple South African Universities	Multiple formats (PDF, XLS) without clear licensing for reuse
CPUT Multilingual Glossraries [10]	CPUT	Unreliable Access

[1] https://ukzntermbank.ukzn.ac.za may have been replaced by ZuluLex
[2] http://oertb.tlterm.com/ [3] https://lwimilinks.sadilar.org
[4] https://www.sawis.co.za/dictionary/Dictionary_Eng.pdf [5] https://glossaries.mandela.ac.za [6] https://ched.uct.ac.za/multilingualism-education-project/projects/multilingual-glossaries-project
[7] https://www1.sun.ac.za/languagecentre-terminologies/ [8] https://languagecentre.sun.ac.za/wp-content/uploads/2021/01/Stats_Eng_Afr_fin.pdf
[9] https://baqonde.usal.es/polokelo/ [10] http://mlg.cput.ac.za

A significant challenge is the ephemeral nature of digital resources. The impending offline status of both the UKZN Termbank and the Full UP OERTB, for which we fortunately have a partial backup, highlights the critical threat of digital decay and the urgent need for proactive data preservation. Our ability to access The South African Trilingual Wine Industry Dictionary is great, but it does not have a clear license for reuse.

This landscape illustrates the vital need for a structured, centralized effort like *Mafoko*. We acknowledge the institutional constraints that may lead to restrictive access, such as the need to track usage for funding reports. However, we advocate for a collective shift towards open, machine-readable formats under clear, permissive licenses. This not only aligns with Findable, Accessible, Interoperable, and Reusable (FAIR) principles but also empowers researchers, language practitioners, and developers by providing the legal and technical clarity needed to innovate. Ensuring that South Africa's indigenous languages thrive in the digital age requires a concerted effort to make these foundational resources openly and sustainably available.

7 Conclusion

This paper introduced *Mafoko*, a project that directly confronts the critical scarcity of structured, machine-readable terminologies for South Africa's official languages. By systematically aggregating, cleaning, and standardizing fragmented resources from government and academic sources, we have created a foundational, open-access dataset. Our adoption of the Africa-centered NOODL license further ensures that these resources are used in a manner that is equitable and benefits their communities of origin.

We have demonstrated the immediate, practical value of this structured terminology through RAG

experiments, which yielded substantial improvements in English-to-Tshivenda machine translation accuracy and consistency. This result validates our core premise: that well-curated, accessible terminologies are not merely an academic exercise but are essential for enhancing the performance of language technologies for low-resource languages. Ultimately, *Mafoko* serves as both a valuable new resource and a call to action, providing a scalable foundation for developing more inclusive and capable NLP technologies that reflect the rich linguistic diversity of South Africa and the African continent.

8 Limitations

While *Mafoko* successfully structures existing terminological resources into more accessible formats, several limitations frame the scope of this work and offer avenues for future research.

Firstly, the comprehensiveness of our dataset is inherently constrained by the availability and accessibility of source materials. As noted, many valuable terminology and glossary datasets across South Africa's language ecosystem remain difficult to incorporate. This inaccessibility stems not only from resources being unpublished or locked in scanned formats but also from digital decay, where resources like the UP OERTB become permanently offline, or are placed behind restrictive web portals that prevent bulk download. The sustainability of digital language resources in the African context is a significant challenge that affects projects like ours (Taljard et al., 2022).

Secondly, the machine translation experiments, while promising, serve primarily as a proof of concept to demonstrate utility. Our evaluation was limited to English-to-Tshivenda translation in two specific domains. A more exhaustive evaluation is needed to assess the impact of *Mafoko* across all 11 official languages and on a wider array of NLP tasks, such as named entity recognition (NER) or cross-lingual information retrieval. Future work should benchmark performance on diverse tasks and languages to fully understand the resource's capabilities and constraints, following community-driven evaluation standards (Nekoto et al., 2020).

Finally, our adoption of the NOODL license, while principled, may present practical hurdles. As a novel, Africa-centered data governance framework, it may face adoption challenges from institutions or researchers accustomed to more globally recognized licenses like Creative Commons. Educating potential users on its equitable benefit-sharing model is crucial but requires a dedicated effort beyond the scope of this initial project note. The complexities of data governance and licensing for low-resource languages remain a critical area for further exploration (Okorie and Omino, 2024; Rajab et al., 2025).

Acknowledgements

The authors gratefully acknowledge the support of the ABSA Chair of Data Science and the Data Science for Social Impact (DSFSI) Lab at the University of Pretoria. This work was supported by UK International Development and the International Development Research Centre (IDRC), Ottawa, Canada, under the AI4D Africa Program. DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI, and Meta.

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159.

Kathleen Ahrens. 2021. NLTK Inaugural Address Corpus: U.S. Presidential Inaugural Addresses (1789–2021). Originally compiled from C-SPAN sources by Kathleen Ahrens. Available via the NLTK corpora collection.

Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Lou Burnard, Syd Bauman, and 1 others. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Judit Freixa. 2022. Causes of terminological variation. In *Theoretical Perspectives on Terminology*, pages 399–420. John Benjamins Publishing Company.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Marissa Griesel and Mampaka Mojapelo. 2022. Multilingual linguistic terminology. SADiLaR Language Resource Repository, License: Creative Commons Attribution 4.0 license.

Open Data Handbook. Machine readable.

Kathleen Heugh and Christopher Stroud. 2019. *Multilingualism in South African Education: A Southern Perspective*, page 216–238. Studies in English Language. Cambridge University Press.

AI Terminologies in African Languages Contributors. 2025. Ai terminologies in african languages dataset. https://github.com/google-research-datasets/ssa-ai-terminologies/. Version 2025-07-14. Licensed under CC BY-SA 4.0.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

David D. Lewis. 1997. Reuters-21578, distribution 1.0. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html. ApteMod version, widely used benchmark corpus for text categorization.

Vukosi Marivate, Isheanesu Dzingirai, Fiskani Ella Banda, Lastrucci Richard, THAPELO SINDANE, Keabetswe Madumo, Kayode Olaleye, Abiodun Modupe, Unarine Leo Netshifhefhe, Herkulaas MvE Combrink, Mohlatlego Nakeng, and Matome Brilliant Ledwaba. 2025. Mafoko companion dataset for mafoko open multilingual terminologies paper.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Department of Arts and Culture. 2003. National language policy framework. Accessed: 2025-04-19.

Department of Higher Education and Training. 2020. Language policy framework for public higher education institutions. https://www.dhet.gov.za/SiteAssets/Policy%20Frameworks/LanguagePolicyFramework.pdf. Government Gazette No. 43860, 30 October 2020. Effective from 1 January 2022.

C. Okorie and M. Omino. 2024. Licensing African Datasets. [Online]. Available: https://www.licensingafricandatasets.com [Accessed: Jun. 30, 2025].

Daniël Jacobus Prinsloo and Gilles-Maurice De Schryver. 2001. Corpus applications for the african languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies*, 19(1-2):111–131.

Jenalea Rajab, Anuoluwapo Aremu, Everlyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessico Ojo, Atnafu Lambebo Tonja, Maushami Chetty, and 1 others. 2025. The esethu framework: Reimagining sustainable dataset governance and curation for low-resource languages. *arXiv preprint arXiv:2502.15916*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Elsabé Taljard. 2015. Collocations and grammatical patterns in a multilingual online term bank. *Lexikos*, 25:387–402.

Elsabé Taljard, Danie Prinsloo, and Michelle Goosen. 2022. Creating electronic resources for african languages through digitisation: a technical report. *Journal of the Digital Humanities Association of Southern Africa*, 4(01).

University of Pretoria. 2019. Open educational resource term bank (oertb). Accessed: 2025-07-25.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1661–1670.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Tianyang Zhong, Zhenyuan Yang, Zhengliang Liu, Ruidong Zhang, Yiheng Liu, Haiyang Sun, Yi Pan, Yiwei Li, Yifan Zhou, Hanqi Jiang, Junhao Chen, and Tianming Liu. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497. ArXiv preprint.

Bootstrapping Siswati lexical resources from isiZulu

Roné Wierenga
NLP Research Group
CSIR
Pretoria, South Africa
rwierenga@csir.co.za

Laurette Marais
NLP Research Group
CSIR
Pretoria, South Africa
lmarais@csir.co.za

Ilana Wilken
NLP Research Group
CSIR
Pretoria, South Africa
iwilken@csir.co.za

Abstract

IsiZulu and Siswati are closely related languages that share significant morphosyntactic characteristics. Systematic differences between these languages have been identified at the phonological and morphosyntactic levels. Due to the resource-scarce status of these languages, this similarity has led to bootstrapping of computational language resources at the morphological and syntactic levels. In this work, we investigate the feasibility of adapting lexical items in a computational lexicon from isiZulu to Siswati. We use Grammatical Framework resource grammars for both languages to analyse and transform lexical items, which are then evaluated against a parallel term list. An iterative process yields a success rate of 70.5%, indicating that this approach is largely viable as a means of significantly reducing the manual effort needed to develop lexicons for computational resources for Siswati.

1 Introduction

Siswati is a Southern Bantu language of the Nguni language family. Many Nguni languages, including isiZulu and Siswati, exhibit a high degree of mutual intelligibility, which facilitates the adaptation of linguistic resources developed for one language to another with relatively minor modifications (Bosch et al., 2008; Marais et al., 2024).

Both Siswati and isiZulu are recognied as official languages in South Africa; however, census data indicates that isiZulu is the most widely spoken language in the country, with over 15 million speakers, whereas Siswati is a comparatively smaller language group, with approximately 3 million speakers in South Africa and around 1 million speakers in the Kingdom of Eswatini (Moors et al., 2018). Consequently, significantly more linguistic resources have been developed for isiZulu than for Siswati, despite both languages being classified as under-resourced. Marais et al. (2024) demonstrated that syntactic resources can be successfully bootstrapped from isiZulu to Siswati due to their shared linguistic features, such as conjunctive orthography and morphophonological affixing (refer to Taljaard and Bosch (1988) for an outline of isiZulu grammar and Taljaard et al. (1991) for Siswati). Building on this research, the present study employed semi-automatic methods to derive Siswati computational lexicons from isiZulu ones. We show how parallel Grammatical Framework Resource Grammars for isiZulu and Siswati enable the bootstrapping process by enabling morphosyntactic analysis of isiZulu terms, as well as generation of parallel Siswati terms. An evaluation is performed by comparing the automatically generated terms with those found in parallel dictionaries for isiZulu and Siswati.

2 Morphophonological and Lexical Differences Between isiZulu and Siswati

As previously mentioned, Siswati and isiZulu are closely related Bantu languages. The grammars of both languages are governed by two fundamental principles: (i) nominal classification, which involves a system of noun classes, and (ii) the conjugation of various word categories to ensure concordial agreement. Typically, nouns consist of a noun class prefix and a root or stem. Noun prefixes determine the noun class, which functions as a grammatical gender system, grouping nouns into specific classes. Each noun class is associated with systematic parameters that regulate grammatical agreement. Additionally, noun classes are numbered, and the noun class systems of isiZulu and Siswati are largely similar.

Both languages also employ a structural element often referred to as an agreement morpheme, which formally marks the relationship between the noun and all other words that share a semantic-syntactic relationship with it, ensuring grammatical (or con-

cordial) agreement. Consider the following isiZulu example:

(1) *Ukudla kumnandi.*
'The food is delicious.'

In example (1), the noun class prefix *uku-* establishes agreement with the verbal prefix *ku-*, which expresses third-person singular in the present tense. This type of morpheme is referred to as subject concord. These concord morphemes can mark subject-verb agreement and express semantic functions such as possession, vocation, negation, and plurality. Marais et al. (2024) summarise key systematic morphophonological differences between isiZulu and Siswati as outlined in existing literature. For clarity, a brief overview is provided below:

Although the lexicons of isiZulu and Siswati are very similar, certain lexical items, including loanwords, exist in one language without existing in the other. These differences in the two languages are not necessarily systematic and it is not necessarily possible to identify which loan words or lexical items are likely to occur in both languages or where the lexicons of these languages diverge.

3 Methodology

Figure 1 shows an overview of our methodology, which consists of three distinct phases. These phases are indicated by arrows of different colours: light green for Phase 1 (Term list compilation), light blue for Phase 2 (Lexicon adaptation) and dark blue for Phase 3 (Evaluation).

3.1 Phase 1: Term list compilation

The term list was compiled by manually capturing lexical items from *The Official Foundation Phase CAPS English-isiZulu Dictionary* and *The Official Foundation Phase CAPS English-Siswati Dictionary*, which are picture dictionaries developed by the National Lexicography Units of South Africa for Grades R to 3. A parallel list of terms represented by single words in both isiZulu and Siswati was extracted. An overview of the work done in the phase is given in Section 4.

3.2 Phase 2: Lexicon boostrapping

Phase 2 involved the the utilisation of the isiZulu Resource Grammar (RG) to facilitate the identification of roots and stems within the isiZulu term list by parsing the isiZulu words. Having identified systematic phonological and orthographic differences between isiZulu and Siswati from the literature, a Python script was developed that adapts isiZulu roots and stems according to these differences. Finally, the Siswati RG was used to generate (via linearisation) Siswati terms using the adapted roots and stems. This process is discussed in Section 5.

3.3 Phase 3: Evaluation

The final phase was to evaluate the lexical items generated via the Siswati resource grammar by comparing them to the manually captured term list developed during Phase 1. The main aim of this phase was to quantify the success of the bootstrapping approach via adaptation, as well as to categorise and quantify the errors produced. In this way, insights could be gained about the feasibility of the bootstrapping approach for development of Siswati lexica from isiZulu sources in future, with reference to the amount and the nature of post-editing required to obtain a high quality resource. The analysis and evaluation of the process is discussed in Section 6.

4 Phase 1: Term list compilation

One of the parallel resources used for this project is the Official Foundation Phase Curriculum and Assessment Policy Statement's (CAPS) bilingual picture dictionary series published by the South African National Lexicography Units (NLUs) for each of South Africa's official languages. For this project, the English-isiZulu (Mbatha et al., 2018) and English-Siswati (Lubisi et al., 2018) dictionaries were used. The picture dictionaries use English as the pivot language to facilitate lexical acquisition in Grades R, 1, 2, and 3, collectively known as the foundation phase in the South African education system. Children in this phase are typically between the ages of six and nine. The picture dictionaries constitute a parallel term list in 11 of South Africa's official languages (excluding South African Sign Language). With English as the pivot language, each term corresponds to a single English word, but is often expressed in the other languages using multi-word terms.

Since these dictionaries were developed by the NLUs under the Pan South African Language Board (PanSALB), we can assume that the lexicons were compiled by (i) native speakers of each language and (ii) teams of linguistic and language-practice experts specialising in various sub-fields. Given that many South African Bantu languages,

Morphophonological difference	**isiZulu**	**Siswati**
Alphabet and click omission	/c/, /q/ and /x/	/c/
Consonant substitution or addition	/z/ /th/ and /t/ /th/ and /t/ /d/ /d/ /mp/ /nk/	/t/ /tf/ before /o/, /u/ and /w/ /ts/ before /a/, /e/ and /i/ /df/ before /o/, /u/ and /w/ /dz/ before /a/, /e/ and /i/ /mph/ /nkh/
Pre-prefix vowel deletion, addition and substitution	Noun class prefix consists of a consonant-vowel sequence (also referred to as the basic prefix), preceded by a so-called augment (also referred to as class pre-prefix), a preceding copy vowel that fulfils different grammatical functions, e.g., definiteness and specificity, and is subject to morphophonological processes such as vowel deletion and coalescence.	This augment is only present in classes 1, 3, 4, and 6 and in class 9 where it precedes a nasal consonant.
Relative construction and concords	a-	la-
Orthography	Demonstratives are written disjunctively from the noun that follows.	First position demonstrative ('this/these') is written conjunctively with the following noun.
Imperative	Monosyllabic verb stems, yi- or i- are prefixed or -na suffixed to the stem for the imperative directed at one person.	-ni is suffixed to monosyllabic verb stems for imperatives directed at one person.

Table 1: Systematic morphophonological differences between isiZulu and Siswati.

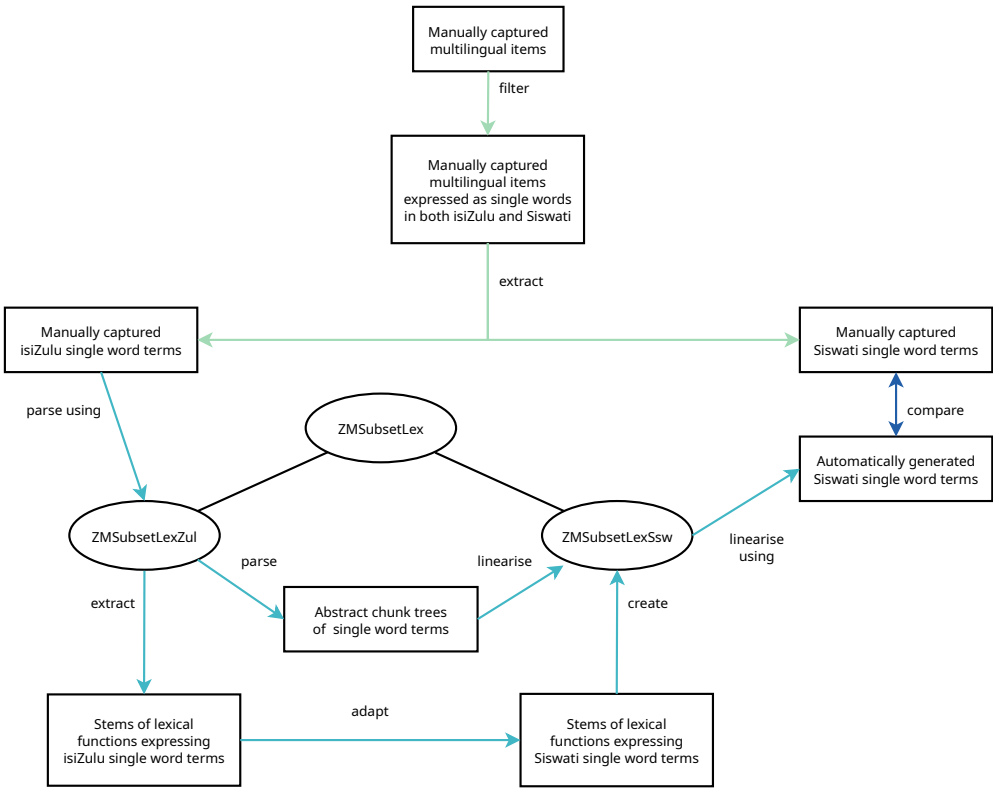


Figure 1: Methodology overview

including Siswati, remain under-resourced, certain terminologies do not yet exist (refer to Khumalo and Nkomo (2022) for a discussion on the challenges of developing terminology in isiZulu). Thus, generally speaking, there are often uncertainties regarding the validity of specific translations due to dialectal variation or the absence of standardised forms for specific terms (Khumalo and Nkomo, 2022). However, these particular picture dictionaries can be considered verified and relatively reliable, since any standardisation issues or dialectal disparities would have been resolved by the respective NLUs. The process to capture terms from the dictionaries for each language involved intensive manual work and was done on a Microsoft Excel spreadsheet. First, all English terms were captured by hand and categorised according to the preset themes and categories as set out in the dictionaries. For each English term, a tag was added to indicate if the term is a noun, verb, adjective, adverb or a sentence, as indicated in the dictionaries. The isiZulu and Siswati translations of each term were added next to the English. For several English terms, more than one translation was provided in either isiZulu or Siswati. In those instances, cells were added to the spreadsheet to ensure the translations still aligned across the three languages. For the English-isiZulu list, a total of 1990 terms were recorded, and 1917 terms for the English-Siswati list. These terms were mostly nouns, followed by verbs, adjectives, sentences, and adverbs.

From this term list, all terms that comprised single words in both isiZulu and Siswati was extracted. The reasons for this is that in order to develop parallel Grammatical Framework (GF) lexicons, a reliable list of meaning-equivalent isiZulu and Siswati words is required for the bootstrapping process.

5 Phase 2: Lexicon bootstrapping

The isiZulu resource grammar has a large lexicon with several thousands of lexical items, and the work described in this paper is aimed at the eventual development of a similarly sized Siswati lexicon.

The artifact in view, therefore, is a GF lexicon module that can be used in conjunction with the Siswati resource grammar. Such a lexicon is based on specifying the noun stems and verb roots, with the grammar itself handling all the applicable morphology. Therefore, a list of Siswati stems and roots must be developed to serve as the basis for a GF lexicon.

In this section, we describe the process of leveraging the parsing and linearisation capabilities of

Lexical category	Occurrences	%
Nouns	756	73%
Verbs	239	23%
Adjectives	30	3%
Adverbs	13	1%
Total	1038	

Table 2: Lexical categories of words in final term list

the GF runtime (Angelov and Ljunglöf, 2014), in conjunction with the isiZulu and Siswati resource grammars, to extract stems and roots from the isiZulu single word terms, to adapt these to Siswati, and then to automatically generate Siswati single word terms using a script based on the known systematic differences set out in Section 2.

5.1 Parsing isiZulu words

A list of about 1500 manually captured isiZulu-Siswati single word term pairs served as the starting point, including nouns, verbs and so-called adjectives. In reality, true adjectives in isiZulu and Siswati are a small, closed class of which the roots can easily be enumerated. English adjectives are almost always translated to isiZulu (and Siswati) using predicate-based qualificatives. In fact, an analysis of the isiZulu Wordnet (Griesel et al., 2019) revealed that 98% of terms captured by translating English adjectives to isiZulu were morphologically complex, predicate-based constructions involving a verb or noun (Marais and Pretorius, 2023).

The isiZulu resource grammar (ZRG) was used to perform a morphosyntactic analysis of each isiZulu word. This was done by employing the ZRG as a morphosyntactic parser. Around 400 items could not be parsed, mostly due to being proper nouns or loan words that are not included in the large isiZulu lexicon. However, a total of 1138 lexical items were successfully parsed, representing a success rate of about 75% and providing a suitably representative basis for the adaptation experiment. A breakdown of the composition of the final term list according to lexical category is given in Table 2.

As an example of the output of the parsing process, and to illustrate how English adjectives that were translated as qualificatives were handled, consider the syntax tree obtained by parsing the isiZulu token *okuthambile*, as shown in Figure 2. Apart from the leaf nodes, each node label consists of a function name and a syntactic category, separated by a colon.

The tree represents a relative sentence chunk involving a so-called stative verb, which may be translated to English in a somewhat syntactically equivalent way as '(that) which has become soft', but is more naturally translated simply as '(that) which is soft'. The presence of the pronoun (`Pron`) in the tree is due to the fact that the initial morpheme *oku-* is the relative agreement morpheme of class 17, which is often used to indicate indefinite agreement in the absence of a specific noun. The English adjective was therefore captured by the dictionary developer using a relative construction with a stative verb and a particular choice of agreement information (it should be noted that in other cases, different choices for agreement were made). It is this stative verb that our process is aimed at adapting from isiZulu to Siswati.

The leaf nodes of the tree represent the morpheme-based substrings that make up the word. In fact, one of them is the root of the verb that we are seeking to extract in order to adapt to Siswati, namely *thamb*, which means 'to become soft'. This information is captured in the ZRG in its lexicon module in the following way:

```
become_soft_V : V ;           -- abstract
become_soft_V = mkV "thamb" ; -- concrete
```

In this code, `become_soft_V` is the name of an abstract lexical function[1] that produces an expression of type `V`, representing a verb. This is defined in the abstract module. The concrete module contains the logic for how this verb should be represented in isiZulu. Hence, `mkV` is an operation for generating a record that represents an isiZulu verb, and it uses the string "`thamb`", the root, to do this.

The Siswati resource grammar (SRG) also has an operation `mkV`, which similarly accepts a root string as argument and produces the correct record representing a Siswati verb. We know that a systematic change that replaces /th/ with /ts/ in this context would yield the correct root *tsamb*. Hence, our goal for this example would be to produce the following line of code to be included in a Siswati lexicon module:

```
become_soft_V = mkV "tsamb" ; -- concrete
```

This process is discussed in the next section.

[1]While the function name has been chosen to simplify discussion of the GF code for English readers, it has no computational significance.

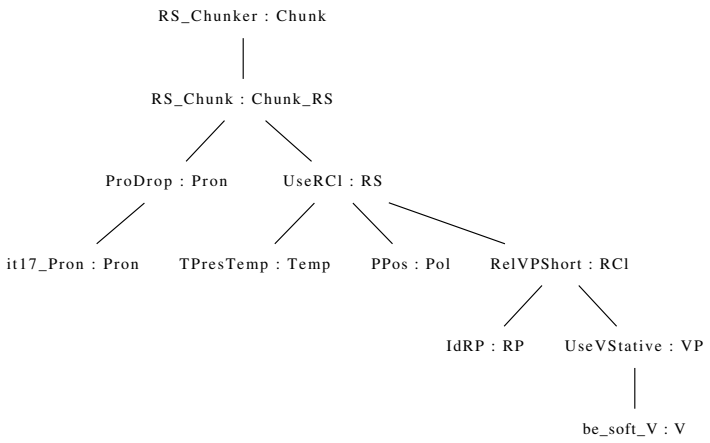


Figure 2: Abstract tree obtained by parsing the isiZulu token *okuthambile*

5.2 Automatic adaptation of stems and roots

Once the morphosyntactic trees were obtained and the roots and stems within them identified, a systematic adaptation could be implemented to facilitate the compilation of a Siswati lexicon module. Since the adaptation approach was applied directly to roots and stems, there was no need to account for additional contextual factors. Consequently, the adaptation solution functioned effectively as a find-and-replace mechanism, as root- or stem-adjacent morphemes had already been stripped from the input. In fact, our process captures roots and stems in their lemmatised forms. The replacement rules shown in Table 3 were applied in the specified order.

This solution has certain limitations. In its current form, it does not distinguish between loanwords and other lexical items. Since loanwords do not always conform to the same phonological rules as native lexical items, it is possible that the substitution rules may not apply consistently to them. However, no existing module or dictionary systematically differentiates loanwords from other lexical items in either isiZulu or Siswati. As a result, this distinction would have to be made manually.

Additionally, the current approach does not account for exceptions to the defined substitution rules. While efforts were made to identify systematic exceptions or deviations, incorporating such exceptions led to unintended errors in the generation of other lexical items. Consequently, no systematic exceptions could be handled without compromising the overall accuracy of the transformation process.

Both of these limitations present opportunities for further research. Future work could explore

isiZulu form	**Siswati form**
q	c
x	c
z	t
tho	tfo
thu	tfu
thw	tfw
to	tfo
tu	tfu
tw	tfw
tha	tsa
the	tse
(thi)[a-zA-Z]	tsi
do	dvo
du	dvu
dw	dvw
da	dza
de	dze
di	dzi
(d)	dz
(th)[ou]	tf

Table 3: Replacement rules for converting isiZulu to Siswati

methods for distinguishing loanwords, potentially through corpus-based analyses or machine learning approaches that focus on identifying phonological patterns. Similarly, refining the substitution rules to accommodate exceptions without introducing errors could enhance the accuracy and applicability of this approach.

5.3 Linearisation of Siswati surface forms

Once the stems were adapted automatically, a Siswati lexicon module was generated as described in Section 5.1. This allowed us to use the SRG, in conjunction with the new lexicon module, to re-linearise the tree shown in Figure 2. The result for our example is the string "lokutsambile", adapted, via the parallel grammars, from "okuthambile".

At this point, it is possible to perform a direct comparison between the generated string and the string found in the original dictionary. If it is identical, we know that the stem was adapted successfully. However, in the case of our example, the Siswati dictionary contains "tsambile", instead of "lokutsambile". The difference is that the generated string matches the morphosyntactic form of the isiZulu dictionary term, while the same term was included in the Siswati dictionary in a slightly different form. The isiZulu term represents a fully realised relative construction using the unspecified agreement value of class 17, while the Siswati term represents a somewhat lemmatised form that lacks any verb prefixes. However, it is clear that the adaptation of the stem succeeded, and that the correct entry in the GF Siswati lexicon was generated, and hence the evaluation process should reflect this.

To account for these differences in morphosyntatic form between the two dictionaries, in all cases where generated Siswati strings failed a direct comparison with the isiZulu term, the Siswati dictionary term was itself parsed using the newly created lexicon. In our example, this means parsing "tsambile" using the SRG along with the newly created Siswati lexicon. The result is shown in Figure 3, which shows the correct lexical function `be_soft_V` used to parse the string as a chunk. This confirms that the adaptation process was successful for the relevant verb root.

6 Phase 3: Evaluation/analysis

As mentioned, the third phase of the experiment focused on evaluating the accuracy of the adaptation solution by comparing the automatically generated

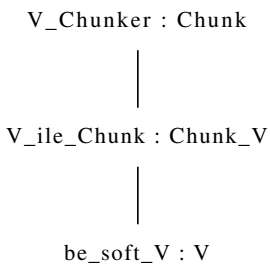


Figure 3: Abstract tree obtained by parsing the Siswati token *tsambile*

Measurement	Number of items
Total isiZulu items	1 038
Successes	488
Reparses	161
Failures	389

Table 4: Overview of bootstrapping results

Siswati terms with those manually captured from the Siswati dictionary.

The success of the boot strapping solution was evaluated by taking several factors into account. These are outlined in this section. The overall quantitative results are provided in Table 4.

The total number of isiZulu items refers to the number of lexical items consisting of a single word extracted from the isiZulu picture dictionary. As mentioned, only single word lexical items were included in the experiment. "Successes" refer to the number of isiZulu lexical items that were successfully converted to Siswati lexical items using the bootstrapping solution. A lexical item was marked as a "reparse" when the isiZulu and Siswati do not constitute the same construction. As set out in Section 5, Siswati terms failing an initial comparison were reparsed to determine whether the reason for the failure is the choice of different forms of the same root or stem by the different developers of the isiZulu and Siswati dictionaries. The reparses are therefore also considered as successfully bootstrapped, because the Siswati root or stem was accurately adapted.

Lastly, Table 4 indicates the number of failures. These were instances where the bootstrapping solution could not automatically generate the correct Siswati lexical item. In order to identify why the Siswati lexical item could not be predicted successfully, the failures were manually categorised into the subgroups shown in Table 5.

Failure classification	Number of items
Total Failures	389
Capitalisation errors	3
Class errors	21
Unique lexical item	229
Phonological differences	39
Exception errors	81
Dictionary errors	16

Table 5: Overview of bootstrapping results

6.1 Analysis of failures

Capitalisation errors: Three Siswati verbs were wrongly categorised as failures because the capitalisation of the words are misaligned. For example, the word Chinese (Eng), *iShayina* (zul) was capitalised as *liShayina* when the term in the dictionary is given as *Lishayina* (ssw). We therefore do not consider this kind of inconsistency as a true error.

Class errors: In these instances the lexical items were generated correctly but the wrong class morpheme was added. For instance *iNgisi* (zul) was used to generate *liNgisi* (ssw); however the correct form included in the Siswati picture dictionary was *Umngisi* (ssw). Only 21 class errors were made.

Unique lexical items and words sourced from other languages: Cases where the isiZulu and Siswati lexical items are not derived from the same root and the Siswati counterpart can in no way be derived from the isiZulu were classified as unique lexical items or words sourced from other languages. The isiZulu and Siswati translations for cellphone are *iselula* and *makhalekhikhini*, respectively. The isiZulu form is derived from the English whilst the Siswati form cannot be derived from either the English or isiZulu, rendering these lexical items impossible to predict using the bootstrapping solution.

Phonological differences: Failures where the Siswati deviates from the isiZulu due to a pronunciation difference that is orthographically respresented, for example language is *ulimi* in isiZulu and *lulwimi* in Siswati, were categorised as phonological differences. Many of the words with phonological differences are loan words from isiZulu in Siswati.

Exception errors: Instances where the bootstrapping solution was applied to lexical items that can be regarded as exceptions to the rules and where the changes made by the bootstrapping solution resulted in the incorrect Siswati form being generated were categorised as exception errors. In these instances the Siswati form is derived from the same root as the isiZulu form but deviates from the isiZulu in an unpredictable or unsystematic way, rendering it difficult to account for these derivations in the Siswati bootstrapping solution without causing a higher failure rate. Some of these deviations are small, for instance an <h> is added to the isiZulu word for ox *inkabi* to form *inkhabi* in Siswati, while others show a higher level of deviation, like the words for mouse *igundane* (zul) and *ligundvwane* (ssw) where the prefix could be correctly generated by the RG but the <vw> in <dvwane> could not be anticipated by or accounted for in the bootstrapping solution. In instances like *ligundvwane* (ssw) the simplicity of the rules stipulated in the bootstrapping solution prevent the correct form from being generated, however, adding layers of nuance to these rules could have rendered the correct form. The refinement of these rules to accommodate more exceptions is reserved for future research.

Dictionary errors: These instances were suspected to result from incorrect dictionary entries, dialectal variants, or human inconsistencies in the Siswati picture dictionary.

Taken together, these results indicate an overall success rate of 62.5% when including capitalisation errors and failures due to the uniqueness of a lexical item - especially those sourced from other languages. If those, along with suspected dictionary errors, are excluded, the bootstrapping solution achieved a refined accuracy of 70.5%. This aligns with other findings in computational Bantu linguistics that show success rates between 65–75% for rule-based transfers between closely related languages (Marais et al., 2024), reinforcing the potential of cross-linguistic bootstrapping in under-resourced contexts.

7 Conclusion

This paper demonstrated the potential of bootstrapping a computational lexicon for Siswati using existing resources in isiZulu, leveraging their close linguistic affinity within the Nguni language cluster. Through the use of the Grammatical Framework resource grammars, we were able to evaluate the application of systematic phonological and morphological transformations to generate parallel isiZulu and Siswati terms.

Findings indicate a success rate of 70.5%, ren-

dering this approach largely viable as a means of significantly reducing the manual effort needed to develop lexicons for computational resources for Siswati. This result is comparable to, and in some cases exceeds, the success rates reported in related works that applied bootstrapping and transfer learning across closely related Bantu languages (Marais et al., 2024), and specifically isiZulu and Siswati. However, the remaining error margin underscores persistent challenges, including phonological exceptions, irregular morphological mappings, and incomplete alignment in lexical categories.

A persistent theme has been the challenge presented by words incorporated into isiZulu and Siswati from other languages, often English. Phonologically focused methods may provide a way of addressing this shortcoming of the current bootstrapping methodology.

These findings affirm the broader linguistic claim that resource sharing within language families can be productive, yet they also support earlier observations in the literature that standardisation discrepancies and dialectal variation must be addressed to optimise outcomes.

Acknowledgments

References

Krasimir Angelov and Peter Ljunglöf. 2014. Fast statistical parsing with parallel multiple context-free grammars. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 368–376.

Sonja Bosch, Laurette Pretorius, and Axel Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2):23.

Marissa Griesel, Sonja Bosch, and Mampaka Lydia Mojapelo. 2019. Thinking globally, acting locally–progress in the African wordnet project. In *Proceedings of the 10th global wordnet conference*, pages 191–196.

Langa Khumalo and Dion Nkomo. 2022. The intellectualization of african languages through terminology and lexicography: Methodological reflections with special reference to lexicographic products of the university of kwazulu-natal. *Lexikos*, 32:1817–1853.

M. Lubisi, South African National Lexicography Units, and Siswati National Lexicography Unit. 2018. *Official foundation phase CAPS English-Siswati picture dictionary*. South African National Lexicography Units.

Laurette Marais and Laurette Pretorius. 2023. Extending the usage of adjectives in the Zulu AfWN. In *Proceedings of the 12th Global Wordnet Conference*, pages 303–314.

Laurette Marais, Laurette Pretorius, and Lionel Clive Posthumus. 2024. Bootstrapping syntactic resources from isiZulu to Siswati. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 77–85.

M.O. Mbatha, South African National Lexicography Units, and IsiZulu National Lexicography Unit. 2018. *Official foundation phase CAPS English-isiZulu picture dictionary*. South African National Lexicography Units.

Carmen Moors, Ilana Wilken, Karen Calteaux, and Tebogo Gumede. 2018. Human language technology audit 2018: analysing the development trends in resource availability in all South African languages. In *In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists, SAICSIT '18*, page 296–304.

P.C. Taljaard and Sonja Bosch. 1988. *Handbook of IsiZulu*. J.L. Van Schaik, Johannesburg.

P.C. Taljaard, J.N. Khumalo, and Sonja Bosch. 1991. *Handbook of Siswati*. J.L. Van Schaik, Johannesburg.

An exploration of the computational identification of English loan words in Sesotho

Mmasibidi Setaka
South African Centre
for Digital Language Resources
North-West University
Potchefstroom
South Africa
Mmasibidi.Setaka@nwu.ac.za

Menno van Zaanen
South African Centre
for Digital Language Resources
North-West University
Potchefstroom
South Africa
Menno.vanZaanen@nwu.ac.za

Abstract

South Africa, with its twelve official languages, is an inherently multilingual country. As such, speakers of many of the languages have been in direct contact. This has led to a cross-over of words and phrases between languages. In this article, we provide a methodology to identify words that are (potentially) borrowed from another language. We test our approach by trying to identify words that moved from English into Sesotho (or potentially the other way around). To do this, we start with a bilingual Sesotho-English dictionary (Bukantswe). We then develop a lexicographic comparison method that takes a pair of lexical items (English and Sesotho) and computes a range of distance metrics. These distance metrics are applied to the raw words (i.e., comparing orthography), but using the Soundex algorithm, an approximate phonological comparison can be made as well. Unfortunately, Bukantswe does not contain complete annotation of loan words, so a quantitative evaluation is not currently possible. We provide a qualitative analysis of the results, which shows that many loan words can be found, but in some cases lexical items that have a high similarity are not loan words. We discuss different situations related to the influence of orthography, phonology, syllable structure, and morphology. The approach itself is language independent, so it can also be applied to other language pairs, e.g., Afrikaans and Sesotho, or more related languages, such as isiXhosa and isiZulu.

1 Introduction

In multilingual environments, speakers from different languages are regularly in direct contact with each other, which often influences their languages. In a multilingual country such as South Africa, where there are twelve official languages (Afrikaans, English, Sepedi, Sesotho, Setswana, Siswati, Tshivenḓa, Xitsonga, isiNdebele, isiXhosa, isiZulu, and, recently added, South African Sign language) and several other non-official languages, language change through a process of borrowing is inevitable (Campbell, 2013).

Language contact can have a wide range of influences on the languages. For example, borrowing of various aspects such as lexical items, morphological, or phonological aspects (Hock, 1991), but also code switching or code mixing are possibilities. Of these aspects, borrowing lexical items typically happens most easily (Bergerson, 2011). The degree to which lexical items of one language can be adapted is influenced by various factors, such as social status of the language and usability in daily life (among others).

Given that borrowing of lexical items occurs most frequently, it will be useful to have an automatic, objective way of identifying items that may have been borrowed. Although it may be impossible to decide in an automated way whether lexical items are borrowed or not, or from which language the items are borrowed, having an automated way to identify potentially borrowed lexical items will greatly help with the identification of such items.

As borrowed lexical items show similarities on the orthographic or phonemic level, it may be possible to identify them using a computational similarity measure. We take a similar approach as that of computing text similarity. Whereas text similarity measures compute similarity between texts (see, for instance, Majumdar (2025)) by considering lexical overlap between two texts, we are interested in the overlap between lexical items. In other words, we will use similarity measures to compute the overlap of letters between two lexical items.

If we want to identify the extend of lexical borrowing between two languages, we need to compare the lexicons of the languages. Ideally, we would like to know which lexical items in the two languages are related to each other (most likely semantically as loan words typically have a very similar meaning). For this, a bilingual dictionary

is useful, because the lexical items in the two languages are linked based on their meaning. In this research, we use the Sesotho-English Bukantswe online dictionary (Olivier, 2016), which contains over 10,000 entries. Lexical similarity between pairs of Sesotho-English lexical items can then be computed using a range of similarities measures.

The aim of this article is to investigate the feasibility of identifying loan words from English into Sesotho using computational means. Even though we do not expect the computational approach to provide final decisions on whether lexical items are loan words (as that most likely requires further linguistic investigation), the approach should help identify potential loan words, which will allow for a more efficient approach to a more complete investigation of loan words from English into Sesotho.

This research takes place in the context of the South African languages, which, given the intense interaction between speakers of the different languages, form a perfect environment for the investigation into language contact and language change. Having automated ways to help identify language change, such as lexical borrowing or loan words, will allow for a more focused and time-efficient research on these matters. For instance, Zulu (2013) mentions that limited work has been done on distance measures and this is even more pressing for the South African languages. Even though the focus here is on Sesotho-English interaction, this work can also be applied to other (South African) language pairs.

The article is structured as follows. We start with a short background on language contact and lexical similarity. Next, we provide the methodology where we discuss the data collection used as well as the proposed metrics to compute lexical similarity. We then apply these metrics to the data collection and provide an overview of the results. This is followed by a discussion of the results and finally we provide a conclusion.

2 Background

In this article, we will look at the identification of loan words in Sesotho from English. Sesotho is one of the official languages of South Africa, Lesotho, Zimbabwe, Zambia, and Botswana and is closely related to Sepedi (Northern Sotho) and Setswana. In contrast, English, a part of the Germanic (Gmc) language group, is an official language in many countries, including South Africa, Lesotho, and many others. English is considered a global language (Eventhough, 2012; El Garras, 2025; Mnguni, 2021; Kamwangamalu and Moyo, 2003; Kula and Marten, 2008).

Sesotho and English have been in close contact in South Africa for a while. In South Africa, English was introduced during the 19th century, and by the end of the century, it was used in many black communities as well (Silva, 1997). Against this background, language contact between English and Sesotho was unavoidable. Due to the power dynamics between the two languages, Basotho (Sesotho-speaking communities) had to learn English, which led to Sesotho borrowing words from English. Additionally, modifications (phonological, morphological, grammatical) of English loan words took place to properly fit the English loan words in Sesotho (Kunene, 1963). Due to the nature of the relationship between English and Sesotho, it was mostly Sesotho borrowing from English (and not vice versa). Note, that similar processes took place for Afrikaans (Majola and Lemeko, 2024).

In *Southern Sotho Words of English and Afrikaans Origin*, Kunene (1963) investigates, compares, and discusses methods of adapting words of English and Afrikaans origin to the grammatical and phonemic structures of Southern Sotho (Sesotho). A number of examples are provided that illustrate the influence of English (and Afrikaans) on the Sesotho lexicon. This shows the extensive lexical borrowing between the languages even though the languages are very different.

The examples in Kunene (1963) are hand picked to illustrate the different types of adaptations of loan words in Sesotho. Such an approach is suitable for describing the variety of loan words and their modifications, but it does not allow for a more quantitative analysis of loan words. For this, a most automated way to identify possible loan words is required.

2.1 Language Contact

Language contact is a phenomenon that occurs when speakers of different languages interact and their languages influence one another (Matras, 2020). It may lead to language change, for instance, through the borrowing of words, but it may even lead to new languages. In general, the interactions have the ability to influence languages as we known them, resulting in observable additions,

subtractions, changes, and similarities in their lexicons. As such, information on how the interaction between the languages took place may also teach us about the history of these languages.

Historically, in Africa, Southeast Asia, and Oceania, Germanic languages came into contact with other languages, particularly as a result of colonization and trading activities. Riehl (2025) has seen many languages adapting and adopting certain sections of one language and then assimilating it into their own. In the context of the influence of the Germanic languages on South African languages, it is useful to have tools that can help identify these influences (semi-)automatically.

2.2 Language similarity

There are several ways to measure the impact of language contact on the different languages. With language contact come different types of similarities and measurements, which can be used on languages, such as lexical similarities, dialectometry, lexicostatistics, and many others. For the sake of this article, we will briefly share a few of them.

Lexical similarity is a natural language processing concept that focuses on similarities that exist between words and texts of different languages. According to Ethnologue[1], it is the percentage of lexical similarity between two linguistic varieties, which is determined by comparing a set of standardized word lists and counting those forms that show similarity in both form and meaning. This is particularly interesting for many low-resource languages (such as Sesotho) that share certain lexical similarities with some high-resource languages (such as English) (Maurya et al., 2023).

Many systems have been proposed to build resources automatically based on lexical similarity (Jadi et al., 2016). To provide a few examples of previous research in the area of lexical similarities, we mention work comparing German and Afrikaans (Bergerson, 2011), Khelobedu, Tshivenḓa, and Sepedi languages (Tebogo and Mandende, 2023), English, Sesotho, and Afrikaans (Kunene, 1963), English and Portuguese (Günther et al., 2019), and English and German (García and de Souza, 2014).

Related work can also be found in the field of dialectometry (which aims to measure the distance between dialects of a language). Dialects of a language are often mutually intelligible, meaning that speakers of one dialect can understand speakers of the other dialect (Gooskens and Van Heuven, 2022). This means that the research in this field can focus on the linguistic variations between the dialects. The lexical variation is often clear between dialects and can be measured. As such, there has been a wide range of work in this field, see for example, Heeringa and Nerbonne (2001); Nerbonne and Kretzschmar (2003); Nerbonne and Kretzschmar Jr (2013); Wieling and Nerbonne (2015) for an overview of the field. If we look at such research on the official South African languages, we find that some work has been done comparing isiZulu and isiXhosa (Zulu, 2013).

Phrasal similarities check the extent to which phrases in different languages are similar. Various data models have been developed such as the Document Index Graph, which indexes web documents based on phrases, rather than single terms only (Hammouda and Kamel, 2002). As such, this measure calculates the similarities based on the ratio of how much they overlap and rewards phrases with high frequency, significance level and length (Momin et al., 2006).

Semantic similarity is a relation between concepts on the level of meaning (Kolb, 2009). In semantic similarity, the likelihood of occurrence of a concept is determined by its frequency in a corpus (Li et al., 2003). Though it has been a part of natural language processing and information retrieval discussions for many years, it is has been a problem for many applications of computational linguistics and artificial intelligence (Li et al., 2003).

On a more general level, the field of lexicostatistics measures the similarity or difference between languages which shows a degree of relatedness and this provides a value that describes the distance between the languages. Based on these distances, family tries of languages can be constructed, which indicate relatedness between languages and may even be used to create proto-languages, which indicate historical relationships (Dyen, 1964).

Throughout the previous research described in this section, several computational techniques are used regularly. In particular, measures that describe the similarity or distance between lexical items, such as the edit-distance or Levenshtein distance (Wagner and Fischer, 1974) or Damerau distance (Damerau, 1964), are used. Here we also use the Levenshtein distance, but also consider a number of additional metrics coming from the field of in-

[1] See https://www.ethnologue.com/methodology/.

formation retrieval, where such measures are used to evaluate the set of documents found by the information retrieval system.

3 Methodology

To identify which English words may have influenced or are incorporated in Sesotho, we need access to lexical items in both languages. Ideally, the lexical items should be translations of each other. Bilingual dictionaries offer exactly this: the lexical items in the dictionary are paired based on their meaning. Given pairs of lexical items, we can then investigate whether the lexical items in both languages are similar enough.

Starting with a pair of lexical items we need to decide if the words are similar enough that they could be related. To decide whether a word may be a loan word, we assume that, in such a case, the words should be orthographically or phonologically similar (and at the same time that words that are not loan words are orthographically and phonologically dissimilar). We implement a range of distance and similarity metrics to help identify the loan word pairs.

Even though the methodology proposed is language independent, in this article, we will focus on Sesotho-English word pairs. We already know that Sesotho has borrowed from English as several lexical items are marked as such in the data collection that we are using (see below). Note that in this article, we assume that there are words that were incorporated into Sesotho from English, but it may also be the case that some Sesotho words have been taken over in (South African) English. This method does not identify the direction.

3.1 Data collection

To investigate which words are (potentially) loan words, we require lexical items in the two languages. These lexical items could be extracted from (large) corpora in the languages, which can provide lists of lexical items. However, in such a case it is unclear which words in the two languages have similar meaning. (We assume that loan words have similar meaning.) Aligning words with similar meaning in two languages is challenging. However, bilingual dictionaries provide lexical items in two languages where the lexical items are directly linked to each other based on their meaning.

Here, we make use of the Bukantswe Sesotho-English Bilingual Dictionary, which is developed by Olivier, Masilo, and Thejane and is publicly available[2]. In this data collection, each line contains a pair of lexical items (separated by a tab) with the Sesotho lexical item in the first column and corresponding English in the second. The English column sometimes also contains some additional information in brackets, such as part of speech (e.g., adj. v., n.) and whether the Sesotho lexical item is borrowed from English (<Eng) or Afrikaans (<Afr). In total, the dictionary contains 10,084 entries, but some of the entries contain multiple (comma separated) lexical items. After splitting these items and removing duplicate entries, the data collection contains 12,958 entries.

3.2 Metrics

To compare how similar two lexical items are, we compute their distance. There are many metrics that can be used for this purpose. Table 1 provides an overview of the metrics used in this article. Note that the metrics used here focus on the orthography (and through Soundex on the pronunciation). For example, metrics that take into account contexts of the use of the lexical item, such as word vector models (Turney and Pantel, 2010), are not used here. Several of the used metrics come from the field of information retrieval (Manning et al., 2008) to compare search results of different search systems. In Table 1 we provide the name of the metric, the implementation, whether the metric is a distance or similarity metric, whether the metric takes the ordering of letters in a word into account, and the range of the values of the metric.

To compute word similarity or distance[3], we compare words by looking at their individual letters. A relatively simple metric to compute letter overlap is the Jaccard metric. This metric does not take the order of letters into account. The formula is applied after the lexical items are converted into a bag (which allows for letters to occur multiple times), which is slightly different from the regular definition of the Jaccard metric, which works on sets. Similarly, the Euclidean and Cosine metrics are computed based on vectors that describe the bag of letters of the lexical items. This essentially compares the number of occurrences of the letters

[2]This resource can be downloaded from SADiLaR's repository: https://hdl.handle.net/20.500.12185/419.

[3]Similarity metrics are have high values when words are similar and low when words are dissimilar, whereas distance metrics have low values when words are similar and vice versa.

in the lexical items one by one. The Hamming distance, Levenshtein distance, Levenshtein ratio, and the Jaro metric do take the order of the letters of the words into account. The Hamming distance requires lexical items of the same length, which is attained by padding the shorter lexical item. For the Levenshtein distance and ratio, we use the standard weights for the edit operations (one for insertion, deletion, and substitution).

Several other metrics could be implemented and experimented with as well, such as the Damerau–Levenshtein distance (Damerau, 1964) or the Jaro-Winkler metric (Winkler, 1990), just like a number of variations of the current metrics (e.g., the standard set-based Jaccard metric, different weights for the Levenshtein distance). However, due to space constraints we will focus on the metrics in Table 1 only here.

The metrics described so far are applied to the orthographic representation of the lexical items. However, some words may have a similar pronunciation even if their orthographic representation is different. The Soundex algorithm (Knuth, 1998) maps phonetically similar orthographic letters onto common symbols. (Note, however, that Soundex is developed for English.) The metrics described in Table 1 can be applied to the Soundex converted sequences as well.

The different metrics and Soundex conversion are implemented in Python. We used the Levenshtein[4] package for most of the different metrics, and the Soundex[5] package to map orthography to Soundex sequences.

4 Results

For all the lexical items that we find in the Bukantswe bilingual dictionary, we compute the distance metrics as mentioned in Table 1. We first look at the correlations between the different metrics to understand the general behavior of the different metrics with respect to each other.

Following that, we look at a sample of the lexical items in the dictionary. We can try to identify loan words by sorting the items based on the computed distance values. The assumption is that lexical items that have a low distance are very similar (in orthography or in phonetic representation in the Soundex case). These words are possibly lexical items that have gone from one language to the other. We provide examples where indeed loan words are identified, but also cases where the metrics do not provide the right information. Unfortunately, the information in Bukantswe on whether lexical items come from English is not complete, so a fully automated evaluation is currently not possible.

4.1 Correlations

To investigate how the different metrics behave, we first investigate the correlations between the calculated values. In total we implemented seven metrics (Jaccard, Euclidean, Cosine, Levenshtein, Levenshtein ratio, Hamming, and Jaro), but these are also applied to the Soundex representation, so in total we have fourteen variants to compare.

Creating a correlation matrix shows that all correlations are significant ($\alpha = .05$), except for the Levenshtein and Cosine metrics ($p = .058$) and Hamming and Jaro ($p = .072$). The correlation matrix (shown in Table 2) illustrates that there is quite some variation between the results of the different metrics. Depending on the metrics, the correlation can range from very weak (e.g., Levenshtein versus Cosine with $r = -.02$) to very strong (e.g., Levenshtein versus Hamming with $r = .97$ and $r = .94$ with Soundex, Jaccard versus Cosine with $r = .92$ and $r = .93$ with Soundex, or Jaccard versus Levenshtein ratio with $r = .92$ and $r = .95$ with Soundex). This means that we should consider the effectiveness of the different metrics separately.

4.2 Analysis of lexical items

As mentioned before, the information on English loan words in Sesotho as marked in Bukantswe is not complete. As such, we cannot perform a computational evaluation of the entire data collection. Here we will look at a sample of the lexical items that show low distance (based on a few metrics).

Words that are very different, e.g., “bonolo” versus “ease” get the lowest value for the similarity metrics (0), and high values (e.g., 4.24 for Euclidean, 6 for Hamming and Levenshtein) for the distance metrics. Similar values are found for the Soundex versions of the metrics (2.24 for Euclidean, 3 for Hamming and Levenshtein Soundex). In contrast, words that are exactly the same, e.g., “radar” have the highest value for similarity metrics (1) and the lowest value for distance metrics (0). The Soundex versions are exactly the same for these.

[4] See https://pypi.org/project/python-Levenshtein/.

[5] See https://pypi.org/project/soundex/.

Table 1: Overview of metrics to compute lexical distance. Where A and B are bags of letters of the lexical items, whereas $\vec{A}$ and $\vec{B}$ describe sequences of letters. We provide the implementation, type (distance or similarity), whether the order of letters is taken into account and the range of the values.

Metric	Implementation	Type	Order of letters	Range
Jaccard	$\frac{A \cap B}{A \cup B}$	Similarity	Unordered	$[0, 1]$
Euclidean	$\sqrt{\Sigma_{i=A \cup B}(A_i - B_i)^2}$	Distance	Unordered	$[0, \infty)$
Cosine	$\frac{A \cdot B}{\|\|A\|\| \, \|\|B\|\|}$	Similarity	Unordered	$[0, 1]$
Hamming	Hamming (1950)	Distance	Ordered	$[0, \infty)$
Levenshtein	Wagner and Fischer (1974)	Distance	Ordered	$[0, \infty)$
Levenshtein ratio	$1 - \frac{Levenshtein(\vec{A}, \vec{B})}{\|\vec{A}\| + \|\vec{B}\|}$	Similarity	Ordered	$[0, 1]$
Jaro	Jaro (1989)	Similarity	Ordered	$[0, 1]$

For words like "diesel" (English) and "disele" (Sesotho), which have the same letters, but in a different order, we see that for the unordered metrics, the values are optimal, but, as expected, the ordered metrics show slight variation (4 for Hamming, 2 for Levenshtein, 0.83 for Levenshtein ratio, and 0.89 for Jaro).

The Soundex algorithm is effective in words such as "afrika" (Sesotho) versus "africa" (English). The orthography is very similar (e.g., Jaccard 0.71, Euclidean 1.41, Cosine 0.88, Levenshtein 1, Levenshtein ratio 0.83, Hamming 1), we have optimal scores for the Soundex versions of the metrics as the "k" and "c" are mapped to the same symbol. In situations like these, the Soundex algorithm has a major impact on the identification of similar words. Similar situations happen for words like "leaforikanere" versus "afrikaner", "disetatemente" versus "statement", "faele" versus "file", and "yunifomo" versus "uniform" (all Sesotho versus English words).

Given these results, it seems that the metrics work well, with the Soundex having advantages where the orthography is slightly different. However, there are also situations where the approach may not work. For example, according to the Bukantswe dictionary, "sekete" or "dikete" (Sesotho) comes from "circuit" (English). If we look at the metrics, we find low scores for the similarity (0.08 Jaccard, 0.09 Cosine, 0.15 Levenshtein ratio, and 0.44 Jaro) and high values for distance (4.58 Euclidean, 7 Hamming and Levenshtein) with somewhat better scores for the Soundex variants (0.4 Jaccard, 0.58 Cosine, 0.57 Levenshtein ratio, and 0.72 Jaro, 1.73 Euclidean, 4 Hamming and 2 Levenshtein) with slighly higher values for "dikete".

Looking at the words that have a high similarity (or low distance), especially using the Soundex metrics, we can identify a number of reasons for differences in the words. First, there are relatively simple orthographic differences. For example, "histori" (Sesotho) versus "history" (English), or "mengo" (Sesotho) versus "mango" (English). Second, there are words with orthographic changes based on the phonology or the structure of Sesotho syllables in words such as "boroto" (Sesotho) versus "board" (English), "keresemese" (Sesotho) versus "christmas" (English), or "potefoliyo" (Sesotho) versus "portfolio" (English). Third, there are morphological influences in cases like "dibanana" (Sesotho) versus "banana" (English), or "ditonki" (Sesotho) versus "donkey" (English). (Note that some words may show a combination of these, such as "okesejene" (Sesotho) from "oxygene" (English) as Sesotho does not use "x" in its orthography and the "kese" is added due to syllable structure.)

Whereas many "modern" words (like "visa", "radio", "bikini", "sms") are borrowed from English as well as names of countries (e.g., "Zambia", "Zimbabwe"), we also see names that may have been borrowed by English (e.g., "Kgalahadi" (Sesotho) and "Kalahari" (English).

Unfortunately, there are also a number of words that have a high similarity (or low distance) even though the Sesotho words do not come from the English words. For example, "ba batle" (Sesotho) versus "beautiful" with (all Soundex metrics) a Jac-

Table 2: Correlation matrix (r) of all metrics.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Jaccard (1)	1.00	-.42	.92	-.18	.92	-.05	.75	.71	-.46	.66	-.30	.68	-.18	.56
Euclidean (2)	-.42	1.00	-.29	.90	-.44	.85	-.24	-.23	.71	-.06	.74	-.22	.73	-.05
Cosine (3)	.92	-.29	1.00	-.02	.85	.12	.77	.62	-.30	.63	-.13	.61	-.02	.53
Levenshtein (4)	-.18	.90	-.02	1.00	-.28	.97	-.10	-.14	.72	.06	.83	-.14	.83	.03
Levenshtein ratio (5)	.92	-.44	.85	-.28	1.00	-.13	.79	.71	-.50	.64	-.38	.71	-.26	.56
Hamming (6)	-.05	.85	.12	.97	-.13	1.00	.02	-.05	.66	.15	.78	-.04	.82	.11
Jaro (7)	.75	-.24	.77	-.10	.79	.02	1.00	.58	-.30	.56	-.19	.59	-.09	.52
SOUNDEX														
Jaccard (8)	.71	-.23	.62	-.14	.71	-.05	.58	1.00	-.64	.93	-.41	.95	-.28	.79
Euclidean (9)	-.46	.71	-.30	.72	-.50	.66	-.30	-.64	1.00	-.47	.90	-.60	.83	-.36
Cosine (10)	.66	-.06	.63	.06	.64	.15	.56	.93	-.47	1.00	-.22	.93	-.06	.84
Levenshtein (11)	-.30	.74	-.13	.83	-.38	.78	-.19	-.41	.90	-.22	1.00	-.43	.94	-.22
Levenshtein ratio (12)	.68	-.22	.61	-.14	.71	-.04	.59	.95	-.60	.93	-.43	1.00	-.27	.84
Hamming (13)	-.18	.73	-.02	.83	-.26	.82	-.09	-.28	.83	-.06	.94	-.27	1.00	-.10
Jaro (14)	.56	-.05	.53	.03	.56	.11	.52	.79	-.36	.84	-.22	.84	-.10	1.00

card and Cosine values of 1, and Levenshtein value of 2, Levenshtein ratio of 0.75 and Jaro of 0.92. Similarly, it is unlikely that "nko" (Sesotho) comes from the English word "nose" (English), similarly for "amohetse" (Sesotho) versus "accommodated" (English).

We also see a number of words that have a high similarity and low distance that most likely come from Afrikaans, although the English word is similar as well. Examples of these are "tanki" (Sesotho) versus "thanks" (English) and "dankie" (Afrikaans), or "sekere" (Sesotho), "shears" (English) and "skêr" (Afrikaans).

5 Discussions

This article provides an automatic way to identify similar words in bilingual word lists. The underlying assumption is that such words are likely to represent words that are borrowed from one language to the other. Different metrics have been proposed that rely on orthographic or (through Soundex) phonologic similarities. These metrics have been applied and investigated on Bukantswe, a Sesotho-English dictionary.

This approach illustrated that many loan words can automatically be identified. The data collection contains a number of word pairs where the words are exactly the same, but also many words that are similar in the sense that orthographic, phonologic or morphologic differences can be filtered out.

This research extends the work by Kunene (1963) in the sense that the proposed approach can (semi-)automatically identify words that are borrowed from another language. Whereas Kunene (1963) focused on Sesotho words from either English or Afrikaans, we have focused on English only, although (as illustrated in Section 4.2) some loan words of Afrikaans origin were also identified. Since some Afrikaans and English words are quite similar (due to Afrikaans and English being members of the Germanic language group), it is not always clear what the original language is.

The work by Kunene (1963) also shows that a human, manual analysis of the results is still needed. Even though the proposed techniques identify similar words, it does not explicitly indicate borrowings. Furthermore, even in the case of borrowings, it does not indicate the direction of borrowing. While for the Sesotho-English case (in particular for "modern" words) this is often clear, when applying these techniques to other language pairs (e.g., isiZulu-

isiXhosa), the picture may be much more difficult.

The results also indicated that there are some structural differences as Sesotho has more complex morphological structures when compared to the English words. For example, Sesotho's noun class information is encoded as prefixes on nouns (Moloi and Thetso, 2014), which has no clear equivalent in English. These structural differences can potentially be handled in a consistent way. The current metrics do not take this explicitly into account. Such an approach would make the method language dependent, however.

6 Conclusion

In this article, we have described an approach to identify a computational approach to identify lexical items that are similar in two languages. We have applied this to lexical items from a Sesotho-English bilingual dictionary with this assumption that such similar words are most likely words that have been incorporated in Sesotho from English.

We have implemented seven distance metrics with different properties (such as taking the order of letters in the word into account or not) and we have applied the same metrics to a Soundex representation of the lexical items. The Soundex representation is an approximate phonological representation, contrasting from the orthographic dictionary entries. The metrics showed varied correlations, which indicates that they do describe alternative distances.

Unfortunately, we could not find a data collection that provides information on which words are actually incorporated from English into Sesotho, so a structural, computational analysis of the different metrics was not possible. However, the metrics did help to identify words that are quite similar, which may potentially save much time when trying to find borrowed words. Further (linguistic, historic) research needs to take place to properly identify those words that are incorporated into Sesotho from English.

The computational approach described in this article shows promising results, but a more controlled evaluation is needed. This, however, requires annotated data (e.g., containing word pairs in addition to a label that indicates whether that the lexical items are related).

The currently investigated metrics do not take any morphological information into account. The Sesotho lexical items show a different morphological structure, for instance, including morphemes that describe the noun class. English and Sesotho morphological structure is different, leading to (almost) regular differences which could be implemented as part of the distance metrics.

Limitations

The research in this article has a number of limitations. First, the data collection used (Bukantswe's bilingual Sesotho-English dictionary) contains not only words, but also some multi-word expressions. Even though these are interesting to investigate, comparing these as lexical items is challenging as some words may be loan words, but the entire expression does not necessarily have to.

Second, Bukantswe does not have complete tagging of English loan words. Even though some lexical items have a tag indicating this, there are many lexical items that are loan words from English without the tag. Complete information allows for a more robust evaluation of the methodology used.

Third, we assume that if the lexical items are similar enough (according to a particular metric), there is a relation between the words. As the lexical items are orthographically similar (or potentially similar on the phonologic level when using Soundex), this does not necessarily mean that these words are loan words. Even if they are, the direction of the relationship is not identified. Even though in this case it is more likely that words are taken from English and incorporated in Sesotho, there may be words taken from Sesotho and incorporated in South African English.

References

Jeremy Bergerson. 2011. *Apperception and linguistic contact between German and Afrikaans*. Ph.D. thesis, University of California, Berkeley.

Lyle Campbell. 2013. *Historical linguistics*, 3rd edition. Edinburgh University Press.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Isidore Dyen. 1964. Lexicostatistics in comparative linguistics. *Lingua*, 13:230–239.

Hassan El Garras. 2025. English, the global language: Its strength, status, and future. *International Journal of Linguistics, Literature and Translation*, 8:122–130.

Ndlovu Eventhough. 2012. Mother tongue education in the official minority languages in zimbabwe. *South African Journal of African Languages*, 31:229–242.

Maria Isabel Maldonado García and Ana Maria Borges de Souza. 2014. Lexical similarity level between english and portuguese. *Elia*, 0(14):145.

Charlotte Gooskens and Vincent Van Heuven. 2022. *Mutual intelligibility*, pages 51–95. Cambridge University Press.

Fritz Günther, Eva Smolka, and Marco Marelli. 2019. 'understanding' differs between english and german: Capturing systematic language differences of complex words. *Cortex*, 116:168–175.

Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160.

Khaled M Hammouda and Mohamed S Kamel. 2002. Phrase-based document similarity based on an index graph model. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings*, pages 203–210. IEEE.

Wilbert Heeringa and John Nerbonne. 2001. Dialect areas and dialect continua. *Language variation and change*, 13(3):375–400.

Hans Henrich Hock. 1991. *Principles of historical linguistics*, 2nd edition. Mouton de Gruyter.

Grégoire Jadi, Vincent Claveau, Béatrice Daille, and Laura Monceaux-Cachard. 2016. Evaluating lexical similarity to build sentiment similarity. In *Language and Resource Conference, LREC*.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical association*, 84(406):414–420.

Nkonko Kamwangamalu and Themba Moyo. 2003. Some characteristic features of englishes in lesotho, malawi and swaziland. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 19(1_2):39–54.

Donald E. Knuth. 1998. *The Art of Computer Programming—Sorting and Searching*, 2nd edition, volume 3. Addison-Wesley Publishing Company, Reading, MA, USA.

Peter Kolb. 2009. Experiments on the difference between semantic similarity and relatedness. In *Proceedings of the 17th Nordic conference of computational linguistics (NODALIDA 2009)*, pages 81–88.

Nancy Kula and Lutz Marten. 2008. *Central, East and Southern African Languages*, chapter 4. University of California Press.

DP Kunene. 1963. Southern Sotho words of English and Afrikaans origin. *Word*, 19(3):347–375.

Yuhua Li, Zuhair A Bandar, and David McLean. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882.

Yanga L. P. Majola and Papi Lemeko. 2024. The influence of afrikaans on naming among the basotho of south africa. *Southern African Linguistics and Applied Language Studies*, 42(3):357–365.

Dattatreya Majumdar. 2025. Text analysis and distant reading using r. Https://ladal.edu.au/tutorials/lexsim/lexsim.html.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Infomation Retrieval*. Cambridge University Press, New York.

Yaron Matras. 2020. *Language contact*. Cambridge University Press.

Kaushal Kumar Maurya, Rahul Kejriwal, Maunendra Sankar Desarkar, and Anoop Kunchukuttan. 2023. Charspan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. *arXiv preprint arXiv:2305.05214*.

Aaron Mnguni. 2021. Dreams and realities for south africa: Use of official languages act, 2012. *Studies in Media and Communication*, 9:1.

Francina L Moloi and Madira L Thetso. 2014. The morphology of the sesotho form/bo-: An exploratory study. *Journal of Linguistics and Language in Education*, 8(1):65–79.

BF Momin, PJ Kulkarni, and Amol Chaudhari. 2006. Web document clustering using document index graph. In *2006 International Conference on Advanced Computing and Communications*, pages 32–37. IEEE.

John Nerbonne and William Kretzschmar. 2003. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37(3):245–255.

John Nerbonne and William A Kretzschmar Jr. 2013. Dialectometry++. *Literary and Linguistic Computing*, 28(1):2–12.

J. A. K. Olivier. 2016. Bukantswe Sesotho-English bilingual dictionary. SADiLaR Language Resource Repository, License: Creative Commons Attribution 3.0 South Africa (CC BY 3.0 ZA): https://creativecommons.org/licenses/by/3.0/za/.

Claudia Maria Riehl. 2025. *Germanic Languages in Contact in Africa, Asia, and Oceania*. Oxford University Press.

Penny Silva. 1997. South African English: oppressor or liberator. *The major varieties of English*, 97:1–8.

Rakgogo J Tebogo and Itani P Mandende. 2023. Lexical similarities between Khelobedu dialect and Tshivenda and sepedi languages. *Literator-Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 44(1):1910.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the Association for Computing Machinery*, 21(1):168–173.

Martijn Wieling and John Nerbonne. 2015. Advances in dialectometry. *Annu. Rev. Linguist.*, 1(1):243–264.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods. American Statistical Association*, pages 354—359.

Peleira Nicholas Zulu. 2013. Classification of south african languages using text and acoustic based methods: A case of six selected languages. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 280–287.

Stop words in Khoekhoe

Menno van Zaanen
South African Centre for Digital Language Resources
North-West University
Potchefstroom, South Africa
menno.vanzaanen@nwu.ac.za

Alena Witzlack-Makarevich
University of Bremen
Bremen, Germany
witzlack@uni-bremen.de

Abstract

Stop word lists are useful resources that allow for the filtering of words in texts that typically do not carry (much) content. Filtering stop words can improve the efficiency and accuracy of data processing. Stop words are typically short and occur very frequently in texts. Stop word lists are language dependent and many low-resource languages currently do not have (accurate) stop word lists. In this article, we look at how we can create, based on word frequency, a stop word list for Khoekhoe, which is a low-resource language spoken in Southern Africa. Given that stop words do not carry much content, they can be expected to occur consistently across different texts. We compare lists of most frequent words between texts in different genres and which words feature in these lists consistently. We look at the overlap of frequent words in English texts and compare these to a known English stop word list as well, and compare the results with the overlap of frequent words in Khoekhoe texts. The results show that there is a high overlap between genres for English, but the overlap between the Khoekhoe genres is lower. This may be due to a different typological profile of Khoekhoe. This means that creating a stop word list for Khoekhoe is more complicated and most likely requires other techniques to produce a useful stop word list.

1 Introduction

The term *low-resource language* refers to languages that have limited or no digital resources. Low-resource languages present distinct challenges in the area of natural language processing due to both quality and quantity of both annotated and unannotated datasets and other linguistic resources needed for the development of NLP applications (see Pakray et al. 2025; Pava et al. 2025 for recent overviews). Even though some of these languages are often of limited or no commercial interest, this does not mean that they are less important or even that have fewer speakers speak them than high-resource languages (e.g., Persian with over 90 million speakers according to Eberhard et al. 2025 is among the low-resource languages). On the other hand, developing practical NLP applications for low-resource languages is central for preserving linguistic diversity, fostering inclusion within the digital world, and providing a robust foundation for expanding our understanding human linguistic capacity and linguistic diversity (Pakray et al., 2025).

Many of the languages spoken in Africa, in particular Southern Africa, are considered low-resource. One of these languages is Khoekhoe, which also lacks many resources. Khoekhoe does not have a strong presence online (e.g., on websites or social media), and (partially due to the lack of available computational linguistic training material) there is no support in many tools, such as in word processors, although an online crowd-sourced dictionary is currently being compiled[1] and a Universal Dependencies Treebank of 27,000 tokens has been recently released (Tulchynska et al., 2025b,a).

One of the basic NLP resources for a language is a list of stop words. Such a list provides the words that in general do not carry (much) meaning (see e.g., Luhn 1957; Lo et al. 2005; Manning et al. 2008). We discuss specifics of stop words in detail in Section 2.1. To tackle the scarcity of digital resources for Khoekhoe, in this article, we investigate the compilation of a Khoekhoe stop word list. To identify Khoekhoe stop words, we rely on the idea that stop words occur frequently in texts (Fox, 1989). If, however, we use texts on a particular topic, those texts will have a different distribution of word frequencies compared to texts on another topic (see e.g., Kilgarriff 1997). To identify stop words, we can compare the frequency lists of texts

[1] See https://lexikhoe.com/.

in different genres. Words that occur frequently even if the topics of the texts are different are likely to be stop words.

To test the idea of consistent high word frequencies of a set of words between texts of different genres, we investigate the consistency of the high frequency words in English texts across different genres first. We can compare these words against a known stop word list for English. Next, we apply the same approach to Khoekhoe texts in different genres.

This article is structured as follows. First we provide some background on stop words as well as the Khoekhoe language in Section 2. Next (Section 3) we present the methodology used, which includes a brief description of the corpora used and the practical approach taken. After applying the approach to the texts in the corpora, we present the results in Section 4 and discuss them in Section 5. Finally, in Section 6 we conclude.

2 Background

2.1 Stop words

Stop word lists are lists of words that do not carry (much) meaning and thus have low discrimination power, for instance, when indexing documents or processing queries (see e.g., Luhn 1957; Lo et al. 2005; Manning et al. 2008) and consequently may be of little help with information retrieval. These words typically occur very frequently (Fox, 1989) in all kinds of texts and are also typically very short, such as *the*, *of*, *and*, and *to*.

Stop word lists have traditionally been developed in the context of the field of information retrieval (van Rijsbergen, 1979; Baeza-Yates and Ribeiro-Neto, 1999). To search for documents in a large document collection based on a search query, one relies on indices that efficiently identify documents containing query terms. Including words in the index that do not describe content do not help with identifying relevant documents simply takes up memory and slows down the search. In particular, this holds for frequently occurring words that do not contribute to the identification of relevant documents. As such, removing words that do not typically describe content was expected to benefit the entire process (Huston and Croft, 2010).

Inspired by Luhn's (1957) work, several precompiled stop word lists have been generated. The earliest is the Van Rijsbergen stop word list (or "Van list", (van Rijsbergen, 1979)), which consists of 250 stop words based on word frequency. The Brown stop word list (Fox, 1989), which we use in this study, consists of 421 stop words and is also partially based on word frequency. The most frequent words include *the*, *and*, *a*, *that*, and *was*. A few of the frequent words were removed from the list because they were considered too important as potential index terms (hence, too useful for searching). They include *business*, *family*, and *house*. Furthermore, 26 words were manually added to the list as they were expected to be frequent in some types of genres. They include *sure*, *behind*, and *whether*. These lists are known as the *classic* or the *standard* stop word lists. These classical stop word lists have been criticized for being too generic and outdated, and various alternatives to the compilation and curation of stop word lists have been proposed including e.g., the ones based on TF-IDF (Term Frequency-Inverse Document Frequency) technique, see e.g., (Saif et al., 2014) for an overview and Chavan et al. (2024) for an example of its application in the low-resource language Marathi.

Several stop word lists exist for English. These may be relatively generic lists that focus on words that are not helpful in general for identifying relevant documents, but, additionally, depending on the task, some lists may also contain content carrying words are not useful in the identification of relevant documents in a particular subject area (Hvitfeldt and Silge, 2021).

A stop word list is an important digital resource which has an enabling function and as such is essential for many other digital resources. For example, stop words are useful in practical situations, such as information retrieval (van Rijsbergen, 1979; Baeza-Yates and Ribeiro-Neto, 1999), text classification (Uysal and Gunal, 2014), sentiment analysis (Saif et al., 2014), and even authorship attribution (Zhao and Zobel, 2005). Though in modern information retrieval systems, the use of stop lists is less common (Jurafsky and Martin, 2025), stop word removal remains useful in various NLP tasks.

2.2 The Khoekhoe language

Khoekhoe, also known as Nama-Damara (`ISO 639-3: naq`; `Glottocode: nama1264`), is a Khoe language of the Khoe-Kwadi family (Güldemann, 2014). Khoekhoe represents a dialect continuum (Haacke, 2018). It includes two northern varieties (Hai‖om and ǂĀkhoe) that differ considerably from the central and southern varieties, which are closer

to what is considered as standardized Khoekhoe. This article is based on a corpus which contains texts written in the standardize Khoekhoe and includes spoken data only from the latter varieties.

Khoekhoe is spoken mainly in central and southern Namibia. With approximately 245,000 speakers (11.8% of the total population of Namibia), Khoekhoe is the second most spoken language in the country after Oshiwambo (Namibia Statistics Agency, 2011). Although Khoekhoe is well-documented and receives official recognition as a language of instruction in Namibia's educational system (Brenzinger, 2013), and despite being the largest non-Bantu click language in Africa, Haacke and Eiseb (2002) consider it to be an "endangered language". Khoekhoe has a limited literary tradition, historically confined to school readers and religious texts compiled by missionaries.

Other Khoisan languages, including other thirteen Khoe-Kwadi languages have even few digital resources than Khoekhoe. To our knowledge no accessible digital corpus of substantial size exists for these languages, not even a Bible translation, one of the most widely translated works for low-resource languages (Pava et al., 2025).

The phonemic inventory of Khoekhoe includes 20 click phonemes, featuring four primary click articulations (dental, alveolar, palatal, and lateral) and five secondary articulations (Haacke, 2013, pp. 54–55). As in other Khoisan languages (cf. Nakagawa et al. (2023)), Khoekhoe roots underlie the following phonotactic restrictions: The canonical root shape is $C_1V_1C_2V_2$ (e.g., *ǂkhoro* 'bottle'), where only four consonants can occur as $C2$: *w*, *r*, *m* and *n*. The other two possible root shapes are $C_1V_1V_2$ and C_1V_1N (e.g., *xai* 'kudu' and *xam* 'lion' respectively). These structures resulted historically from the elision of the intervocalic consonant C_2. N in the latter structure represents the nasal consonants *m* or *n*. When $V1$ and $V2$ are identical, they result in a long oral or nasal vowel. Long oral vowels are represented by the vowel character with a macron symbol, e.g., <*ā*>. Long nasal vowel are represented by the vowel character with a circumflex symbol, e.g., <*â*>. The corpus used in this study follows the standard orthography of Khoekhoe as outlined in Curriculum Committee for Khoekhoegowab (2003). Though Khoekhoe is a tone language, exhibiting four distinctive register tones (Haacke, 1999), tones are not marked in the standard orthography and are also absent in the Khoekhoe corpus presented here.

3 Methodology

In general, as mentioned in Section 2, stop words are highly frequent and are not content words. This gives us a means to identify them automatically. First, just like Fox (1989), we can take the most frequent words of the language (based on a corpus) as our stop word list. Second, based on the notion of context mentioned by Hvitfeldt and Silge (2021), we may improve on the frequency-only approach by taking word use in different text genres into account.

To identify potential stop words, we investigate which words are both highly frequent and, at the same time, genre-independent, by taking a data-driven approach. Starting from a corpus with texts in different genres, we can easily rank words based on their frequency per genre. Stop words are then those highly ranked that overlap between the different genres.

We apply this approach to a Khoekhoe corpus consisting of texts with corresponding genre information. We then compare the Khoekhoe results against those of an English corpus (with genre information). For English we can also compare the most frequently occurring genre-independent words against a known stop word list.

We first describe the corpora and the stop word list that we use (Section 3.1), followed by a description of the approach taken (Section 3.2).

3.1 Data collections

In the current study, we used two corpora. The first is a corpus of Khoekhoe texts, which is the main focus of the research. The second is the Brown corpus, which we use as an English reference corpus, which allows us to investigate whether the approach applied to Khoekhoe behaves in a similar way to that applied to English. Furthermore, we can compare the behavior of most frequent English words against a known stop word list. These resources are described here in more detail.

The Khoekhoe corpus used in this study was compiled from a range of sources. It includes texts of six genres as described in Table 3. The spoken genre (Conversation) includes primarily dyadic conversations between friends, relatives, and colleagues and was collected in Windhoek in 2023. The conversations were transcribed into Khoekhoe by students of the University of Namibia. The written genres include several books of fiction (Fiction), school books (Learned), newspaper articles of the

Khoekhoe edition of Namibian newspaper New Era (Press), a translation of the Bible into Khoekhoe (Religion), as well a few other text types collapsed to Misc (e.g., several procedural texts and translations of film subtitles). The corpus is processed in R and tokenized using `unnest_tokens()` from the `tidytext` package (Silge and Robinson, 2016).

The Brown corpus (Kučera and Francis, 1967) is a collection of text samples (500 samples of American English with each having just over 2000 tokens).[2] The text samples are divided into different genres: Belles-Lettres, Fiction (with subdivisions: general, mystery, science, adventure, romance), Humor, Learned, Miscellaneous: government & house organs (Misc), Popular lore, Press (with subdivisions: reportage, editorial, and reviews), Religion, and Skill and hobbies. An overview of the genres of the Brown corpus can be found in Table 4.

In the appendix (in Table 3 and 4) we provide an overview of the distributions of words for both corpora where we combine the subdivisions. For the analysis, however, in this article, we only focus on the five genres that can be found in both corpora. The distribution of words in these genres of both corpora can be found in Table 1. Note that both corpora are approximately of the same size and the sample used in the experiments is also of similar size, although the distribution of tokens is a bit different.

Table 1: Distribution of words per genre used in the approach from the Khoekhoe and Brown corpora.

Genre	Khoekhoe # tokens	Brown # tokens
Fiction	37,513	235,489
Learned	18,083	159,936
Misc	75,153	61,143
Press	76,328	177,177
Religion	656,808	34,308
Total	863,885	668,053

To be able to compare the most frequent words in the different English genres, we take the stop word list for English from the NLTK Stopwords Corpus (Bird et al., 2009).[3] This stop word list consists of 198 English words.

3.2 Method

To investigate whether we can use the top N words in Khoekhoe to identify stop words, we take the corpus as described in Section 3.1. We subdivide the corpus in the different genres and use the genres as described in Table 1. For each genre, we then rank the words based on frequency and take the top N words. Then we measure the overlap of the top N words pair-wise for all genres in the corpus. We also apply the same approach to the Brown corpus. This allows us to compare the behavior of the most frequent words across genres between Khoekhoe and English.

We expect stop words to occur in the top N words, but the ranking of these words (within the top N) is irrelevant. We are simply interested whether the same words occur at the top of the frequency lists. A metric that measures the overlap of elements between two sets is the Jaccard coefficient. Applying this metric to our frequency lists, computes the proportion of the number of words of the intersection of the two sets over the number of words in the union of the two sets. Jaccard coefficients range between 1 (complete overlap) and 0 (no overlap at all). When A and B are the sets of the top N most frequent words in two genres, the Jaccard coefficient can be calculated as follows.

$$Jaccard(A, B) = \frac{A \cap B}{A \cup B}$$

Although we do not know how many words should be contained in a stop word list, the English stop word list from the NLTK gives us a rough estimate. Although different stop word lists may have different lengths, given that the NLTK English stop word lists contains approximately 200 words, we investigate the Jaccard coefficients by varying N from 1 to 200.

Note that for us to be able to compare the English stop word list against the English genres, we need to make sure that we have a frequency ranking for the words in the NLTK stop word list as well. Unfortunately, the NLTK stop word list is alphabetically ordered. We compute the ranking of these stop words based on the frequency of occurrence in the complete Brown corpus.

Summarizing, for each of the genres, we first rank all words based on their frequency. Next, we select the top N words (where N ranges from 1 to 200) and compute the Jaccard coefficients by

[2] The Brown corpus can be downloaded from https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/brown.zip.

[3] See https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/packages/corpora/stopwords.zip.

comparing genres pairwise. To get an overview of these coefficients, they can be plotted. Note that in addition to the pair-wise comparisons between the genres, for English we also compute the Jaccard coefficients between the genres and the NLTK stop word list (Stop words) and the entire corpus (Total).

4 Results

For each pair of genres, we compare the overlap of the top N words using the Jaccard coefficients. We can now investigate the statistical significance of the effect of the genre pair under consideration and the source (either Khoekhoe or Brown) on the Jaccard coefficients. We see that both variables have statistically significant impact on the Jaccard coefficients ($p < .0001$). The independent variable genre pair accounted for 15% of the variance in the overlap of the top N words ($\eta_p^2 = .15$), whereas the corpora variable accounts for 26% ($\eta_p^2 = .26$) of the variation. Performing a TukeyHSD analysis on the statistical model indicates that there are statistically significant differences between the two corpora ($p < .0001$) and that there are significant differences ($p < .05$) between most genre pairs. We provide an overview of the comparisons between genre pairs that are not statistically significant (with $\alpha = .05$ in Table 2). Where the comparisons between the pairs are insignificant, this corresponds to those plots in Figure 1 whose the shapes are the same.

A more in-depth exploration can be done by visualizing the Jaccard coefficients for both the corpora and genre pairs, as in Figure 1. Each individual plot represents the Jaccard coefficients (represented on the y-axis) for the top N words (represented on the x-axis). All plots contain information on the Khoekhoe (dashed line) and Brown (solid line) corpora. High Jaccard coefficients indicate a large overlap and low coefficients a small overlap. Additionally, the bottom row represents the Jaccard coefficients for the different genres of the Brown corpus and the ranked English stop word list from the NLTK. A similar comparison is not possible for Khoekhoe as we do not have a Khoekhoe stop word list available.

Figure 1 suggests a number of observations. First, we see that with increasing N (i.e., the top most frequent words used for the comparison), the Jaccard coefficients go down, but stabilize at around 100 words for English and at less than 50 words for Khoekhoe. This indicates that when N is very small, many of the most frequent words between the genres are the same. Increasing the number of words under consideration leads to larger differences between the genres.

Second, overall, the Khoekhoe Jaccard coefficients are typically lower (except for the comparison between the Fiction and Learned genres and with low N for Fiction and Religion). This indicates that most of the compared genres do not share many words in the top N most frequent words.

Third, the Khoekhoe Jaccard coefficients typically do not show a consistent peak with low N. This is not what we expected. It means that even if we consider only the really most frequent words (taking N very small), the overlap between the words in the different genres is very small. In other words, Khoekhoe does not seem to have consistent frequent words or stop words like English does.

Fourth, looking at the Jaccard coefficients of the English stop word list and the different genres, we see that the coefficients (for most genres) go down after approximately 75 words, but overall remain relatively high (between 0.5 and 0.25). This indicates that there are relatively many stop words in the most frequent English words in the different genres. The lower Jaccard coefficients with higher N indicate that at that point frequent words are found that are not contained in the stop word list.

Fifth, we see that when we compare the stop word list against the entire Brown corpus, up to about $N = 75$ the Jaccard coefficients are extremely high. After that, the Jaccard coefficients go down, which indicates that non-stop words are found in the most frequent words of the corpus.

5 Discussion

The idea behind the study described here is that stop words are frequently occurring words that occur (frequently) in all types of text. As such, a comparison between the most frequently occurring words in different genres should allow one to identify these stop words. To investigate this, we compared the most frequent words between pairs of genres using the Jaccard coefficient and we did this for both Khoekhoe and English corpora.

The results show that the approach yields different results for the Khoekhoe and English corpora. Not only are the Jaccard coefficients for Khoekhoe mostly lower than those for English, the (small) peaks that can be seen in the English results are not present in the Khoekhoe results.

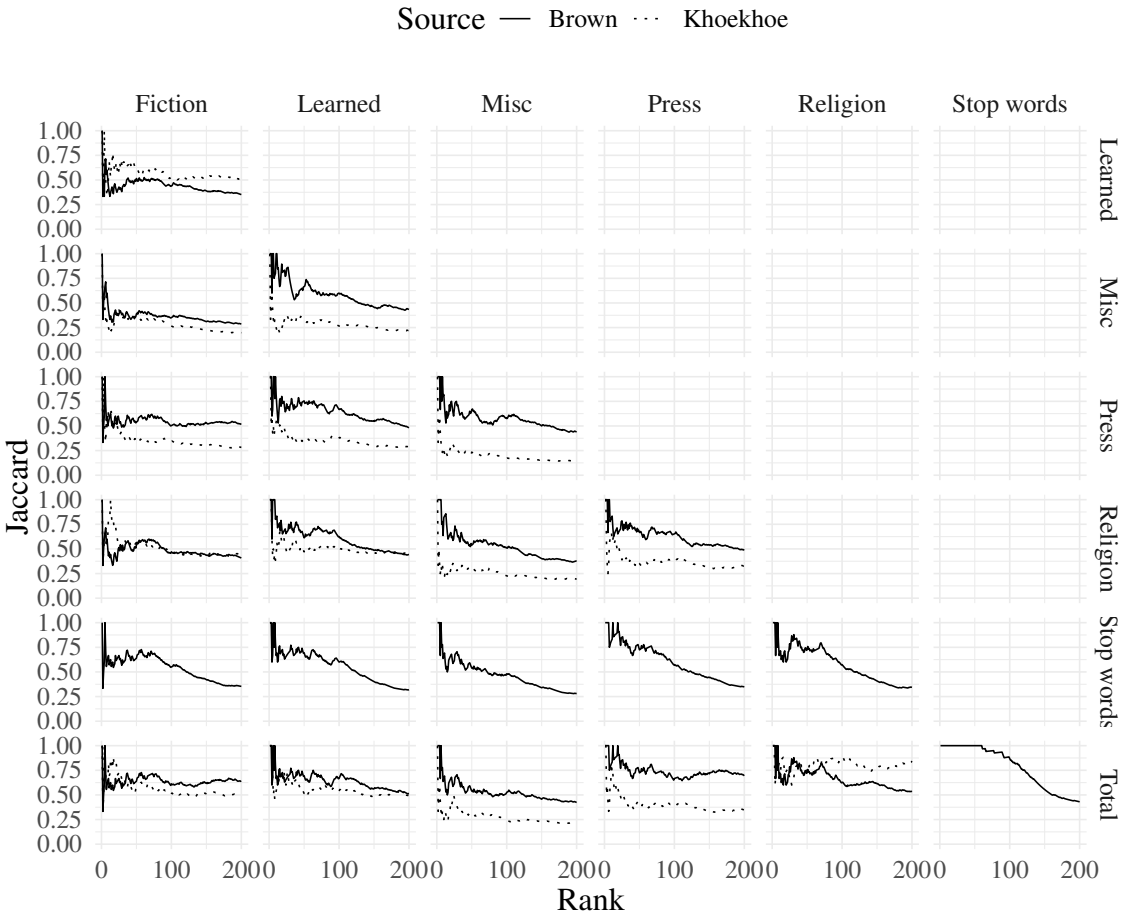


Figure 1: Overview of pairwise comparisons between genres in the Brown (solid line) and Khoekhoe (dotted line) corpora. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes). Note that the stop word data is only available for the Brown data.

Table 2: Overview of pair-wise comparisons that are not significantly different from other pair-wise comparisons.

Genre pair		Genre pair	p
Fiction and Religion	vs.	Fiction and Learned	1.000
Fiction and Stop words	vs.	Fiction and Learned	.145
Fiction and Stop words	vs.	Fiction and Religion	.054
Learned and Misc	vs.	Fiction and Press	.921
Learned and Press	vs.	Fiction and Learned	1.000
Learned and Press	vs.	Fiction and Religion	1.000
Learned and Press	vs.	Fiction and Stop words	.124
Learned and Religion	vs.	Fiction and Stop words	.998
Learned and Stop words	vs.	Fiction and Stop words	.999
Learned and Stop words	vs.	Learned and Religion	1.000
Learned and Total	vs.	Fiction and Total	1.000
Misc and Religion	vs.	Misc and Press	1.000
Misc and Stop words	vs.	Fiction and Learned	.166
Misc and Stop words	vs.	Fiction and Press	1.000
Misc and Stop words	vs.	Fiction and Religion	.355
Misc and Stop words	vs.	Learned and Misc	.315
Misc and Stop words	vs.	Learned and Press	.192
Misc and Total	vs.	Learned and Misc	.101
Misc and Total	vs.	Misc and Press	.979
Misc and Total	vs.	Misc and Religion	.986
Press and Religion	vs.	Fiction and Learned	1.000
Press and Religion	vs.	Fiction and Religion	1.000
Press and Religion	vs.	Fiction and Stop words	.114
Press and Religion	vs.	Learned and Press	1.000
Press and Religion	vs.	Misc and Stop words	.207
Press and Stop words	vs.	Fiction and Total	1.000
Press and Stop words	vs.	Learned and Stop words	.087
Press and Stop words	vs.	Learned and Total	1.000
Press and Total	vs.	Fiction and Stop words	.111
Press and Total	vs.	Fiction and Total	.271
Press and Total	vs.	Learned and Religion	.742
Press and Total	vs.	Learned and Stop words	.982
Press and Total	vs.	Learned and Total	.375
Press and Total	vs.	Press and Stop words	.819
Religion and Stop words	vs.	Fiction and Stop words	.305
Religion and Stop words	vs.	Fiction and Total	.707
Religion and Stop words	vs.	Learned and Religion	.936
Religion and Stop words	vs.	Learned and Stop words	.996
Religion and Stop words	vs.	Learned and Total	.795
Religion and Stop words	vs.	Press and Stop words	.956
Religion and Stop words	vs.	Press and Total	1.000

These results are somewhat unexpected; we expected the Khoekhoe stop words to behave like the English stop words. The linguistic properties of Khoekhoe, however, may provide an answer to this. Though Khoekhoe is comparable to English in terms of morphological complexity, its frequent grammatical words (e.g., particles and auxiliaries) only occasionally have more than one word form, whereas in English many frequent non-content words have frequent word forms, which

jointly make up a substantial proportion of the top-200 list. For instance, there are only two tokens of the negation marking in Khoekhoe (the highly frequent *tama* in the present and past tenses and the less frequent *tide* in the future tense). In the NLTK stop word list for English, on the other hand, there are about 30 items with the negative marker including such orthographic forms as *aren*, *aren't*, *didn*, *didn't*, etc. (which are essentially related to variation in tokenization). Similar abundance of word forms is found with pronouns, e.g., in addition to *you*, the list includes *you'd*, *you'll*, *you're*, and *you've*, as well as *your* and *yours*. No comparable variation in frequent pronominal forms exists in Khoekhoe, though there are many more personal pronoun types including very infrequent ones, such as *sāse* for the first person inclusive plural feminine (i.e., "we" in the sense of the female speaker and female listeners including).

Essentially, whereas the range of high-frequency grammatical words is covered with some 50 words in Khoekhoe, after this various content words (nouns, verbs, adverb) start to show up in the top-200 list. They in turn vary more substantially between genres. On the other hand in English, there are many more high-frequency word forms of some of the most ubiquitous grammatical words, which are more uniformly distributed across genres.

If we look at the Jaccard coefficients between the English stop words from the stop word list and the different genres, we also see that the coefficients relatively quickly stabilize. Depending (a bit) on the genre under consideration, the coefficients are relatively high up to approximately $N = 75$. After that the Jaccard coefficients decrease. This indicates that around that point, content words are becoming more frequent. This corresponds to the fact that stop words are not always in the top most frequent words. From this we can learn that the development of a stop word list cannot only rely on the word frequency. An in-depth analysis of stop words is required.

6 Conclusion

Stop word lists are useful resources for a wide range of natural language processing tasks. For many low-resource languages, however, these lists are not available. In this article, we look at whether we can develop stop word lists based on word frequency information from corpora in different genres.

Previous work in identifying lists of stop words relied on the assumption that stop words are highly frequent and consistently highly frequent across documents. This means that by looking at the overlap of highly frequent words between documents (or documents from different genres), we should be able to identify stop words. Here, we apply this approach to an English (Brown) and Khoekhoe corpus that contains documents classified in various genres. We use the Jaccard coefficient to measure the overlap between the words in the different genres.

The results show that the English corpus has relatively high Jaccard coefficients for the comparison of different genres, which indicates that the stop words (taken from a known stop word list) are indeed often present in the top of frequency-ranked word lists. However, for Khoekhoe, the Jaccard coefficients are much lower, which means that there is more variation in the words that occur frequently in different genres.

Limitations

The research presented in this article has a number of limitations. First, the choice of genres and (amount of) texts may have an influence on the results. However, given that both corpora contain approximately 1 million tokens, the influence of this may be limited. On the other hand, it should be noted that overall fewer texts were included in the Khoekhoe corpus (e.g., only a couple of long texts in Fiction) and some genres (Religion) were massively overrepresented, this might have skewed both similarities and differences between some of the genres and between specific genres and the total frequencies.

Second, depending on the task, a slightly different stop word list may be required. Whereas for some tasks, for example, modal verbs or frequent adverbs are not relevant, for others they may be. This means that a generic stop word list is not always useful.

Finally, for practical reasons, we chose to use the existing stop word list for English as a baseline, however, as we discuss above, the English stop word list contains a large number of word forms and spelling variants (e.g., related to tokenization decisions) for many non-content items. A comparison of Khoekhoe high-frequency items with other languages might result in a more similar picture.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. Addison-Wesley Publishing Company, Reading, MA, USA.

Steven Bird, Ewan Klein, and Edward Loper. 2009. NLTK: The natural language toolkit. In *Proceedings of the ACL 2009 Interactive Demonstrations*, pages 66–71. Association for Computational Linguistics.

Matthias Brenzinger. 2013. The twelve modern Khoisan languages. In Alena Witzlack-Makarevich and Martina Ernszt, editors, *Khoisan languages and linguistics: Proceedings of the 3rd International Symposium, July 6-10, 2008, Riezlern/Kleinwalsertal*, Research in Khoisan Studies. Cologne: Rüdiger Köppe.

Rohan Chavan, Vishal Madle, Gaurav Patil, and Raviraj Joshi. 2024. Curating stopwords in Marathi - A TF-IDF approach. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, page 1–6. IEEE.

Curriculum Committee for Khoekhoegowab. 2003. *Khoekhoegowab: 3||î Xoaigaub*. Gamsberg Macmillan, Windhoek.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas, TX, USA.

Christopher Fox. 1989. A stop list for general text. *ACM SIGIR Forum*, 24(1-2):19–35.

Tom Güldemann. 2014. 'Khoisan' linguistic classification today. In Tom Güldemann and Anne-Maria Fehn, editors, *Beyond 'Khoisan': Historical relations in the Kalahari Basin*, pages 1–41. John Benjamins, Amsterdam.

Wilfrid H. G. Haacke. 1999. *The tonology of Khoekhoe (Nama/Damara)*. Rüdiger Köppe, Cologne.

Wilfrid H. G. Haacke. 2018. Khoekhoegowab (Nama/Damara). In *The social and political history of Southern Africa's languages*, pages 133–158. Palgrave Macmillan, London.

Wilfrid H. G. Haacke and Eliphas Eiseb. 2002. *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Gamsberg Macmillan, Windhoek.

Wilfrid H.G. Haacke. 2013. Phonetics and phonology: Namibian Khoekhoe (Nama/Damara). In Rainer Vossen, editor, *The Khoesan languages*, pages 51–56. Routledge, London.

Samuel Huston and W. Bruce Croft. 2010. Evaluating verbose query processing techniques. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–298, New York, NY, USA. Association for Computing Machinery.

Emil Hvitfeldt and Julia Silge. 2021. *Supervised Machine Learning for Text Analysis in R*. Chapman and Hall/CRC.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Rachel T.-W. Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. *Journal of Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR '05)*, 3(1):3–8.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press, Cambridge.

Hirosi Nakagawa, Alena Witzlack-Makarevich, Daniel Auer, Anne-Maria Fehn, Linda Ammann Gerlach, Tom Güldemann, Sylvanus Job, Florian Lionnet, Christfried Naumann, Hitomi Ono, and Lee J. Pratchett. 2023. Towards a phonological typology of the Kalahari Basin Area languages. *Linguistic Typology*, 27(2):509–535.

Namibia Statistics Agency. 2011. Namibia household income & expenditure survey 2009/2010. Technical report, Namibia Statistics Agency, Windhoek.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. White paper, Stanford Institute for Human-Centered AI (HAI) / The Asia Foundation / University of Pretoria. Available online.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 810–817, Reykjavik, Iceland. European Language Resources Association (ELRA).

Julia Silge and David Robinson. 2016. Tidytext: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, 1(3):37.

Kira Tulchynska, Sylvanus Job, and Alena Witzlack-Makarevich. 2025a. Universal Dependencies treebank for Khoekhoe (KDT). In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 119–128, Ljubljana, Slovenia. Association for Computational Linguistics.

Kira Tulchynska, Alena Witzlack-Makarevich, Sylvanus Job, and Michael Hahn. 2025b. Universal dependencies treebank for khoekhoe (KDT). In Daniel Zeman and et al., editors, *Universal Dependencies 2.16*. Universal Dependencies Consortium.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112.

C. J. van Rijsbergen. 1979. *Information Retrieval*, 2nd edition. University of Glasgow, Glasgow, UK. Printout.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*, pages 174–189. Springer.

A Visualization of full data collections

In the article, we have selected only those genres from the Khoekhoe and Brown corpus that occur in both corpora. Here we provide similar information to Table 1, but now for the full Khoekhoe corpus in Table 3 and the full Brown corpus (after removing non-word tokens) in Table 4. Plots similar to that of Figure 1 can be found in Figure 2 for all Khoekhoe genres and in Figure 3 for all Brown genres.

Table 3: Distribution of words per genre in the complete Khoekhoe corpus.

Genre	# tokens
Conversion	155,292
Fiction	37,513
Learned	18,083
Misc	75,153
Press	76,328
Religion	656,808
Total	1,019,177

Table 4: Distribution of words per genre in the complete Brown corpus. The Press and Fiction genres are combined at that level. All non-word tokens are removed.

Genre	# tokens
Belles-Lettres	151,548
Fiction	235,489
Humor	18,265
Learned	159,936
Misc	61,143
Popular lore	96,691
Press	177,177
Religion	34,308
Skill and hobbies	71,531
Total	1,006,088

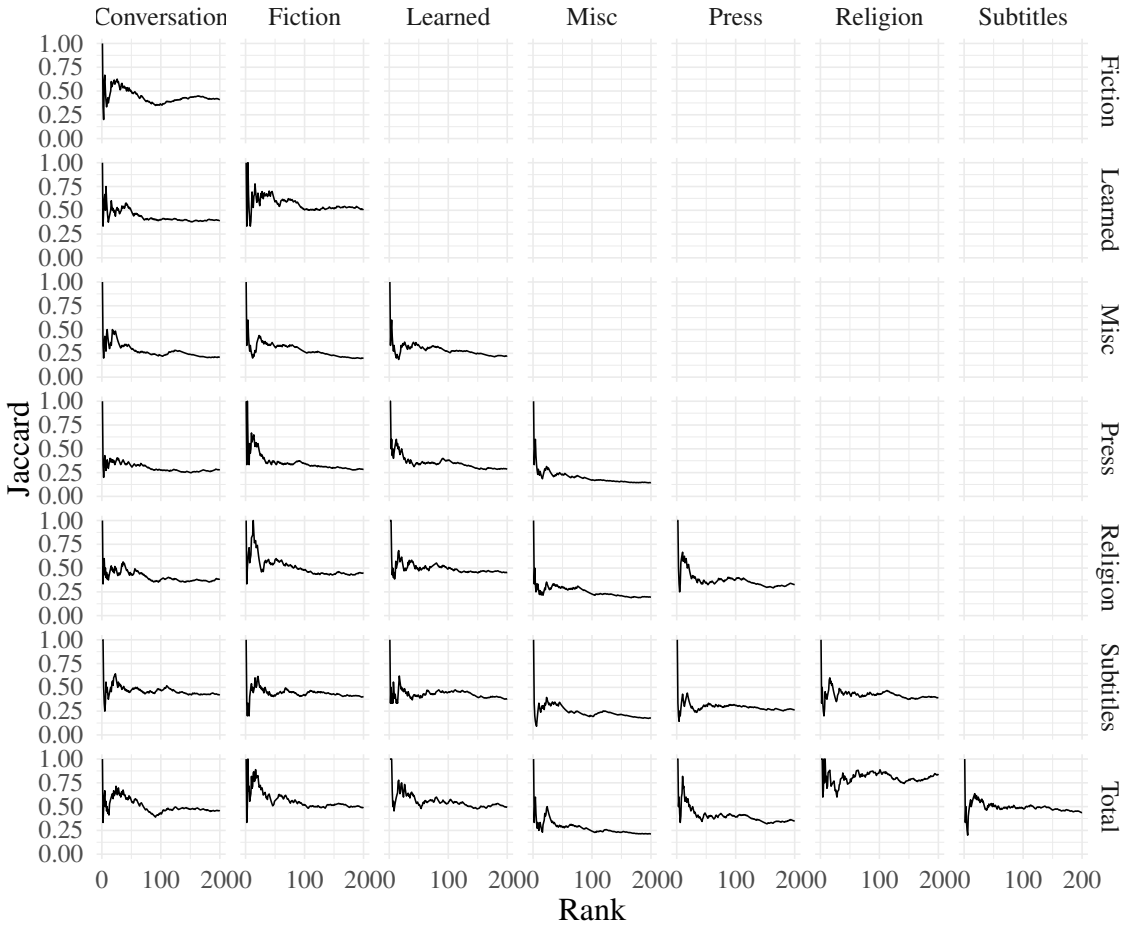


Figure 2: Overview of pairwise comparisons between genres in the complete Khoekhoe corpus. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes).

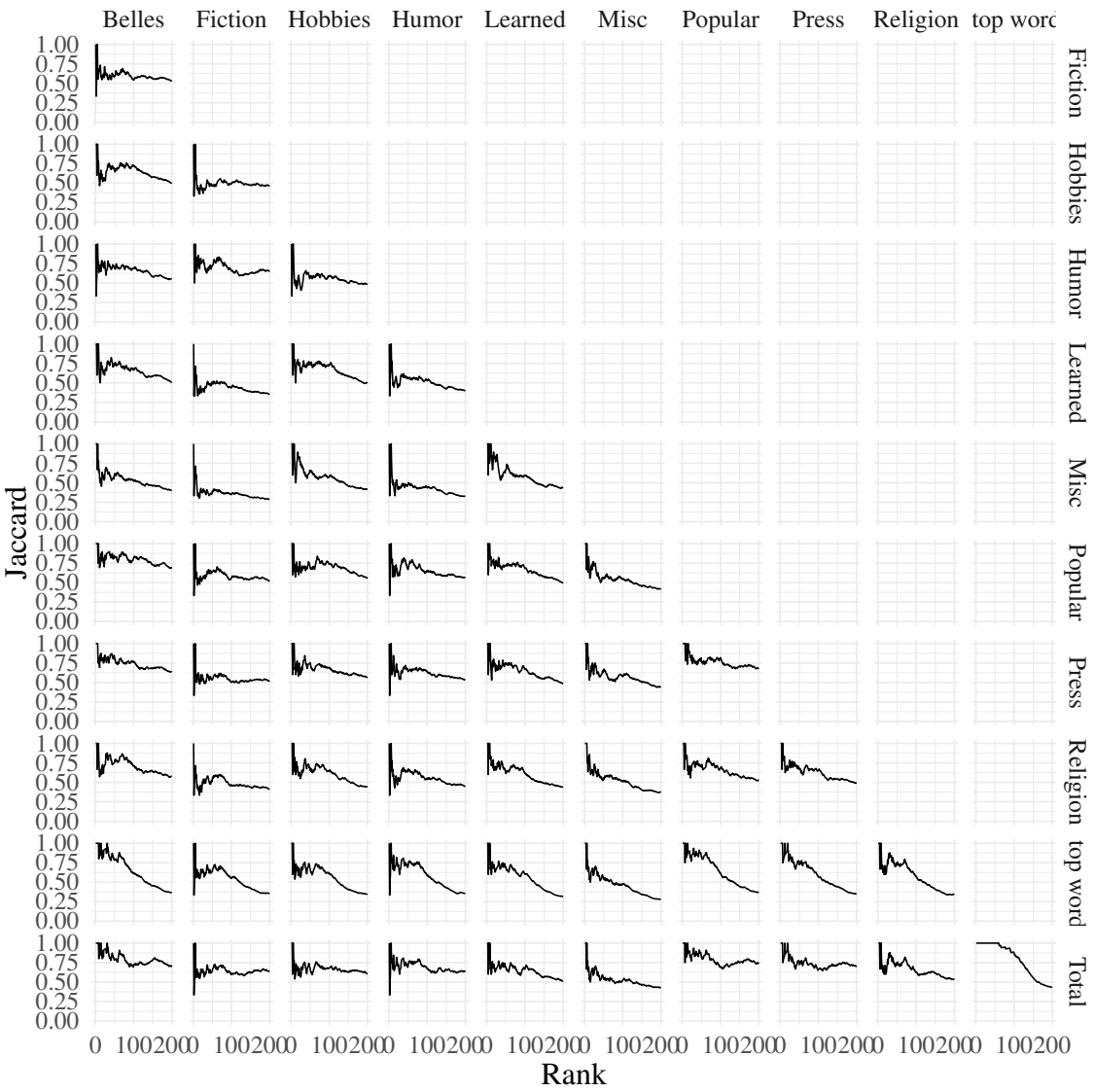


Figure 3: Overview of pairwise comparisons between genres in the complete Brown corpus. The Jaccard coefficients (represented on the y-axes) are computed for top N words where N ranges from 1 to 200 (represented on the x-axes). Additionally the stop word data is provided.

Using the isiZulu GF Resource Grammar for morphological annotation

Laurette Marais
Natural Language Processing
CSIR
Pretoria, South Africa
lmatais@csir.co.za

Laurette Pretorius
Dept of Mathematical Sciences
Stellenbosch University
Stellenbosch, South Africa
lpretorius@sun.ac.za

Abstract

The isiZulu GF Resource Grammar (ZRG) enables syntactic parsing using the GF runtime system. In order to perform this task, the ZRG implicitly encodes rich morphosyntactic information about isiZulu. In this paper we show how such information can be made explicit by adapting the way the grammar linearises GF abstract syntax trees. The result is annotated text, which can be utilised in various ways for supporting natural language processing of an under-resourced, morphologically complex language like isiZulu.

1 Introduction

Large quantities of high quality data is essential for building language technologies (LTs) that are useful. This lack of data is often one of the most limiting factors in developing such technologies for languages such as isiZulu, one of the official Bantu languages of South Africa.

The most basic form of data may be thought of as language texts in digital, machine processable form, drawn from any source or genre and often collected by digitising existing written texts. In an age where LLMs are often perceived as the end goal of LT, the development of LLM models for the Bantu languages remains an open challenge due to the lack of sufficient data. Moreover, the situation is exacerbated by the morphosyntactic complexity of the languages, which increases the size requirements for datasets (Hussen et al., 2025). One way of extracting the most out of available datasets is via annotation.

Broadly speaking, linguistic annotation involves the association of descriptive or analytic notations with language data (Ide and Pustejovsky, 2017, p.2). Aside from providing richer data on which to train LTs, linguistically annotated data also provide a resource for corpus linguistics, a growing field of study within the Bantu languages (Prinsloo and de Schryver, 2001; Taljard and de Schryver, 2016).

There are many types of linguistic annotation. Well-known examples are morphosyntactic tagging (e.g. lemmatisation, part of speech tagging and morphological tagging), syntactic analyses, a range of semantic analyses (e.g. semantic roles, named entities, sentiment and opinion), time and event and spatial analyses, and discourse level analyses including discourse structure, co-reference, etc.

The two essential components of a linguistic annotation project are firstly, the *annotation scheme* that defines the labels or tags, and secondly, an *annotation tool* that supports accurate and fast annotation (Ide and Pustejovsky, 2017, p.3).

In this paper we show how morphological annotation of isiZulu, based on the *ZulMorph tagset*[1], is accurately and efficiently performed by adapting the existing *GF resource grammar for isiZulu* (ZRG) (Marais and Pretorius, 2023) as a tool.

For morphological annotation, two approaches could be considered, namely surface form (morph) annotation and canonical (morpheme) annotation[2]. This distinction has been noted especially with regards to morphological segmentation for isiZulu and the other languages in the Nguni language group (von der Wense et al., 2016; Moeng et al., 2021).

We start with a discussion of isiZulu in Section 2, followed by a brief overview of Grammatical Framework (GF) as an annotation tool in Section 3. In Section 4, we highlight the most relevant aspects of the isiZulu GF resource grammar (ZRG), before discussing our adaptation of it for morphological annotation in Section 5. Finally, in Section 6, we contrast the adapted ZRG with two state-of-the-art tools for isiZulu morphological annotation.

[1] https://portal.sadilar.org/FiniteState/demo/zulmorph

[2] A morpheme is the smallest meaningful unit in the grammar of a language and morph is the phonetic realization of a morpheme.

2 IsiZulu morphosyntactic features

IsiZulu morphosyntax is essentially based on two principles, viz. nominal classification (the system of noun classes) and concordial agreement (the system of concords). IsiZulu has a complex agglutinative morphology (Poulos and Msimang, 1998, p.6). Moreover, isiZulu exhibits a high degree of morphophonological alternation (Poulos and Msimang, 1998, pp.515-534). The language has a conjunctive orthography.

System of noun classes. The noun in isiZulu consists of two main parts, viz. a noun prefix (preprefix and basic prefix) and a noun stem[3]. Furthermore, every noun belongs to a so-called noun class by virtue of the form of its prefix, also referred to as its class gender. This notion of class gender is significant since it generates grammatical agreement by means of these class prefixes. The noun classes are numbered. IsiZulu has 18 noun classes. Generally speaking the nouns occur in singular/plural pairs[4]. Commonly found pairs are 1/2; 1a/2a; 3/4; 5/6; 7/8; 9/10; 11/10; 14/6. Nouns in classes 15, 16, 17 and 18 do not usually have a plural form. The noun classes 16, 17 and 18 are so-called locative classes (Poulos and Msimang, 1998, Chapter 1).

System of concords. A concord is a structural element (agreement marker/morpheme) which formally marks the relationship between a noun and other words in a sentence. This class gender agreement (see above) must be observed in all parts of the utterance which are linked to the noun. Therefore, we say that word categories such as verbs, pronouns, adjectives, relatives, possessives etc. are brought into concordial (i.e. grammatical) agreement by means of these concords (Kosch, 2006, p.90).

The verb. It is the morphologically most complex word category, consisting of multiple morphemes, viz. a root, the morpheme that carries the basic meaning of the verb, and affixes (prefixes and suffixes), morphemes added to the root to give the verb its required functional value, expressing a variety of moods, forms, tenses, aspects and polarity. The prefixes that may occur are, in order, a negative morpheme, subject concord, negative morpheme[5], temporal morpheme, aspectual morpheme, object concord and reflexive morpheme. Possible suffixes include verbal extensions, a verb terminative, a relative suffix and an imperative suffix. Filling these various slots for the affixes depends of the required functional value of the verb, with the root as only obligatory morpheme.

The need for computational morphological analysis as a first step in LT for isiZulu is illustrated by means of the sentence *izalukazi azizukuziphekela imifino* (the women will not cook vegetables for themselves) in Figure 1, analysed by means of ZulMorph (Pretorius and Bosch), a state-of-the art *finite-state* morphological analyser for isiZulu (Pretorius and Bosch, 2010).

The verb *azizukuziphekela* consists of seven morphemes that have to be identified and annotated in order to fully understand its meaning. It is also linked to the subject of the sentence (the class 8 noun) by means of the (class 8) subject concord, *zi*. From our example it is clear that once the sentence has been morphologically annotated as in (1), all its essential morphosyntactic information is known. In particular, it encompasses other kinds of annotation such as part-of-speech tags (e.g. `NOUN` and `VERB` resp.) and lemmatisation (as stems) (`alukazi`, `pheka` and `fino`).

3 Using GF for annotation

A *resource grammar* is a computational grammar that models the morphology and syntax of a language via a set of rules. It is specifically aimed at being precise and comprehensive in its coverage of the linguistic structure of a language, and is used for both parsing and generation.

Grammatical Framework (GF) (Ranta, 2011) is a grammar formalism for multilingual grammars. It provides a functional programming language for defining reversible mappings from interlinguas to concrete languages. GF has been used for building comprehensive resource grammars for over 40 languages. It is considered the state of the art in multilingual grammar engineering.

GF draws a distinction between abstract (language independent) syntax and concrete (language specific) syntax. An abstract syntax is defined by a set of categories and a set of functions by means of which *abstract syntax trees* are constructed. A concrete syntax, on the other hand, linearises these concepts and relations in a particular language in accordance with the linguistic requirements of the specific language and gives rise to *parse trees*. For

[3] Optional suffixes include the diminutive, augmentative, deverbative, feminative etc. They are not discussed further.

[4] Meinhof's numbering system

[5] At most one of the negative morphemes may occur. They originate from different verbal constructions.

(1) *izalukazi*
i[NPrePre][8]zi[BPre][8]alukazi[NStem][7-8]
azizukuziphekela
a[NegPre]zi[SC][8]zuku[FutNeg]zi[ReflPre]phek[VRoot]el[ApplExt]a[VT]
imifino
i[NPrePre][4]mi[BPre][4]fino[NStem][3-4]
'The old women will not cook vegetables for themselves.'

Figure 1: Illustrating the morphological complexity of an isiZulu sentence

each abstract category (`cat`) and function (`fun`) there is a corresponding linearisation type definition (`lincat`) and linearisation rule (`lin`), respectively, in the concrete syntax. A multilingual GF grammar consists of a single abstract syntax and a concrete syntax for every supported language.

As programming language, GF has, apart from its compiler, also a run-time system for performing parsing and linearisation. Translation can be achieved by parsing a sentence using the concrete syntax of one language and then linearising the resulting abstract syntax tree using the concrete syntax of a different language. *Annotation* using the GF runtime system can be achieved by parsing sentences using an RG and linearising the resulting abstract syntax trees using an adapted version of the RG that inserts annotation. The ZRG has been used in this way for morphological surface segmentation (Mkhwanazi and Marais, 2024).

Since an RG is designed to model deep linguistic structure, it encodes complete morphosyntactic information of a sentence. The deep structure is encoded explicitly in the abstract syntax tree, while surface structure is implicitly encoded in the concrete syntax and used to realise the correct surface form of the tree. Adapting an RG for morphological annotation involves adapting the strings of the grammar to make the linguistic structure explicit in the linearisation.

One advantage that this approach has over classical finite-state morphological analysis is the ability to perform sentence-internal morphological disambiguation. In classical finite-state morphological analysis, disambiguation between multiple possible analyses is seen as a next step in the NLP pipeline. The reason for this is that (finite-state) morphological analysis is performed on linguistic words and may therefore produce multiple plausible analyses that can only be disambiguated in the context in which the word occurs. When using the RG to annotate a sentence, the abstract syntax tree provides a syntactic context. Of course, when parsing a sentence, syntactic disambiguation is still required, since a sentence may have multiple possible parses. The GF parser can be utilised in a probabilistic way to assist with disambiguation at this level.

Once the morphosyntactic annotation has been done, shallow forms of annotation such as part-of-speech tagging and lemmatisation can be derived easily and accurately from it by discarding some aspects of the detailed annotation.

4 The isiZulu Resource Grammar

The ZRG is an implementation of isiZulu morphosyntax using GF (Marais and Pretorius, 2023). Apart from `cats`, `funs`, `lincats` and `lins`, two core constructs in GF that we briefly discuss by means of the example in Figure 4, are parameter (`param`) and table (`table`) types.

In linguistics, parameters are multi-valued features that represent the core grammatical properties of a language. In GF such parameters form part of the concrete syntax since they are specific to a language. Indeed "designing the parameter system ... is one of the main tasks in GF grammar writing" (Ranta et al., 2020, p.8). The parameter system of the ZRG consists of 25 parameter type definitions.

Tables operate on parameter types and are used, amongst others, to house morphological variation and inflection. For example, in the ZRG the parameter `RInit`, which differentiates between different possible initial letters (a vowel or any consonant) of a root, is defined by listing its possible values separated by 'l':

```
param RInit = RA | RE | RI | RO | RU | RC ;
```

Figure 4 shows a `table` operating on `RInit` by assigning to each value of `RInit`, the required variant (of type `Str`) of the instrumental prefix *nga-*, thereby ensuring the correct morphophonological alternation between the prefix and the root to which it is prefixed.

When developing a GF RG, it is possible either to include all the full forms of words as strings

in the grammar, and to combine these words into phrases and sentences at run time, or to include meaningful, useful subwords, i.e. morphs, as strings that are bound together at run time into words, which in turn are combined into phrases and sentences. Due to the significant reduction in compiled grammar size, the latter approach has become standard practise in GF RG development for morphologically complex languages.

5 Adapting the ZRG

In this section we discuss the chosen annotation approach for the work described in this paper, then we consider from a practical perspective how the implementation of a morphological adaptation of the ZRG (MZRG) would be done, and finally we discuss the adaptation itself in terms of the principles and design decisions involved.

5.1 Morphological annotation approach

In surface form annotation, the task is to identify where morpheme boundaries occur and to insert the correct morphological tag at these boundaries. For languages with a high degree of morphophonological alternation, this is not a trivial task. In fact, in the case of morpheme fusion in isiZulu, the question of where to mark the boundary often has no clear correct answer from a linguistic or computational point of view. However, a consistent, systematic strategy is essential. In canonical annotation, the task is to identify the morphemes that have given rise to a specific surface form and to reproduce the morphemes in their canonical form, along with the appropriate tags.

As has been noted for the case of morphological segmentation of isiZulu (Mkhwanazi and Marais, 2024), surface form morphological annotation and canonical morphological annotation typically serve different use cases. In particular, part-of-speech tags can be derived trivially from surface form annotation, while lemmatisation can be derived from canonical annotation.

In this paper, we focus on surface form annotation. Consider the syntax tree in Figure 2 that linearises to the isiZulu sentence *umfazi uyapheka* (‘the woman cooks’) using the standard ZRG. Note that the tree shows how the grammar utilises subword strings. These strings, representing morphs, bind together in specified ways at runtime to produce the correct surface forms.

Our goal in adapting the ZRG for surface form morphological annotation is to generate a syntax tree like that shown in Figure 3.

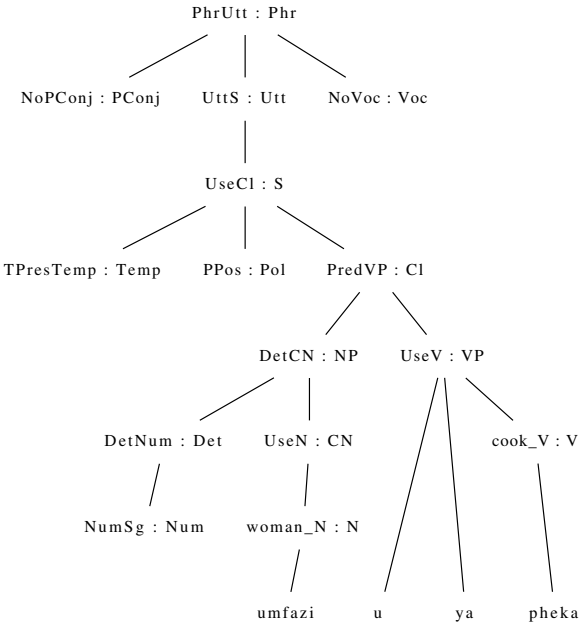


Figure 2: Syntax tree of an isiZulu sentence.

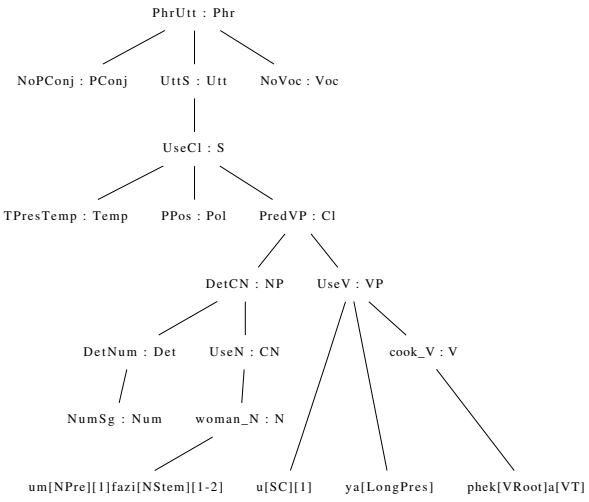


Figure 3: Syntax tree of an annotated isiZulu sentence.

5.2 Implementation of the MZRG

Since all strings contained in the ZRG are defined in the `ResZul` resource module, from a practical perspective, adapting the grammar involves making changes to a single module. With the *ZulMorph* tagset as our guide, this module was systematically changed by inserting the appropriate tags into the string elements. Figure 4 shows such a change, which involved inserting the morpheme tag at the end of each string element. In effect, the morphs

that constitute the strings of the grammar are simply decorated with the appropriate tags. Figure 5 shows what the adaptation of the same operation would have looked like, if canonical annotation were in view. Here, the dimensionality of the table is reduced to remove the mechanism for dealing with morphophonological alternation. Note that although this represents a more substantial change, it is straight forward to implement, since it consists of simplifying the grammar in systematic and predictable ways.

Other changes required for surface annotation were more complex, such as the one shown by the comparison in Figure 6. In the original ZRG, the operation `prefix_nasal` performs pattern matching on the root string in a case statement and generates the required substring, namely a root prefixed with the appropriate form of the nasal sound associated with morphemes for classes 9 and 10. For example, the first branch of the case statement models the case of roots that start with "ph", to which the nasal sound is prefixed as "m", while the "h" is dropped from the root. In the MZRG, the `[NPre]` tag is inserted, along with a tag that indicates the class associated with the prefix. This is generated by a helper operation called `noun_prefix_tag`, to which the class gender value and number are supplied as arguments. These in turn must be supplied to `prefix_nasal`. Given that this information always forms part of the syntactic context within which `prefix_nasal` is called, this change is easily integrated into the grammar.

In a small number of cases, finer distinctions had to be implemented in order to differentiate cases where different morphemes result in the same surface string. This involved adding conditional branches to existing operations.

5.3 Designing a tagset and annotation strategy

In order to retain consistency with previous work, the ZulMorph tagset and analyser served as our starting point, taking into account that it produces canonical morphological annotations, while our adaptation is aimed at surface form morphological annotation.

The main difference in surface form morphological annotation is the identification of morpheme boundaries. Due to the high degree of morphophonological alternation, and particularly morpheme fusion, that isiZulu exhibits, a strategy for systematically identifying such boundaries was required.

Since they provide the main semantic content, roots and stems are given priority. If the root stays in tact during morpheme fusion, the adjacent morpheme is considered to have absorbed the change. For example, when the class 9 relative concord (with canonical form *e-* is prefixed to the verb root *eq*, while the ZulMorph annotation is `e[RC][9]eq[VRoot]a[VT]`, the ZRG annotation is `[RC][9]eq[VRoot]a[VT]`. As in this case, this sometimes leads to "empty" morphemes. The tags are nevertheless inserted, since the annotation must still indicate which morphemes contribute to the final surface form. Apart from cases where there are differences in the tagsets of ZulMorph and the MZRG (discussed later in this section), this ensures that morpheme tag sequences are identical between the two systems for any given word. Table 1 lists the examples discussed in this section, with the ZulMorph annotation alongside the MZRG annotation in blue.

Sometimes, both the root/stem and an adjacent morpheme undergo changes, as in the case of locative palatalisation that occurs in nouns when adding the locative suffix. For example, for the noun stem *-chopho* ('pole (of the earth/magnetic)'), its singular locative form is *echosheni*. We annotate this as shown in example 2 of Table 1, where *-eni* is regarded as the locative suffix, despite its canonical form being *-ini*. The chain of phonetic processes that cause this kind of change has been documented by Poulos and Msimang (1998), showing that the initial *i* of *-ini* suffix undergoes vowel lowering to become an *e*, which then triggers further processes within the noun stem. In cases of stems ending in *e*, the suffix is considered to be *-ni*. Therefore, the annotation for *ezweni* ('in/to the country') would be `e[LocPre]zwe[NStem][5-6]ni[LocSuf]` (see example 3 in Table 1).

For other morphemes, any parts of a morpheme that are retained after a sound change are considered part of the morpheme. For example, the locative for the class 1a noun *kudadewenu* ('to your sister') is annotated as shown in example 4 of Table 1.

Generally, if no other principle can be applied, syllable boundaries guide morpheme boundary identification in cases of morpheme fusion.

One major difference between ZulMorph and the MZRG is the way noun prefixes are handled. ZulMorph consistently distinguishes the pre-prefix

```
-- with
instrPref : RInit => Str = table {
  RU => "ngo" ;
  RI => "nge" ;
  RO => "ngo" ;
  _  => "nga"
} ;
```

```
-- with
instrPref : RInit => Str = table {
  RU => "ngo[AdvPre]" ;
  RI => "nge[AdvPre]" ;
  RO => "ngo[AdvPre]" ;
  _  => "nga[AdvPre]"
} ;
```

Figure 4: Original and adapted code snippets for generating variants of the instrumental prefix.

```
-- with
instrPref : Str = "nga[AdvPre]" ;
```

Figure 5: Adapted code for canonical annotation.

(with tag `[NPrePre]`) and the base prefix (with tag `[BPre]`) which together constitute the noun prefix. Due to the frequency with which at least of these morphemes becomes "empty" due to morphophonological alternation, the MZRG tags the noun prefix as a single morpheme with tag `[NPre]`.

Another significant difference is the lack of annotation for verb root extensions. This was a design decision of the ZRG, in which such extensions were assumed to form part of the extended verb roots in a lexicon, instead of being handled productively within the morphosyntactic implementation of the ZRG. Consequently, *ushisa* is annotated by ZulMorph as `u[SC][3]sh[VRoot]is[CausExt]a[VT]`, while the MZRG annotation is `u[SC][3]shis[VRoot]a[VT]` (see example 5 of Table 1).

6 Applying the MZRG

The goal of this section is to show the MZRG in action as a morphological analyser, and to contextualise it as a tool among other state-of-the-art morphological analysers for isiZulu. We therefore contrast the MZRG to both ZulMorph (as another rule-based tool) and the CTexT Core Technologies (a data-driven tool).

The corpus used as our basis is a treebank of 100 sentences taken from an isiZulu textbook (Taljaard and Bosch, 1988). The sentences were chosen to represent a linguistically diverse set of sentences, originally used to demonstrate and teach a variety of linguistic constructions in isiZulu. It therefore represents a suitable set of sentences for comparing the ZulMorph analysis with that of the MZRG. The 100 sentences comprise 243 tokens, of which 184 are unique, although they appear in different syntactic contexts (and may therefore have different analyses).

A GF treebank can be obtained via direct engineering or by parsing a corpus of sentences using the probabilistic parsing functionality of the GF runtime. This corpus of 100 sentences was obtained using a combination of both (Marais and Pretorius, 2023). The sentences were therefore known to fall within the definition of the original ZRG.

6.1 MZRG compared to a rule-based tool

When comparing two rule-based tools, similar assumptions about the performance of the tools exist, namely that rule-based tools are engineered for correctness and may fail entirely on ungrammatical input or input that falls outside of its definition. In comparing the MZRG to ZulMorph, the aim is not to evaluate accuracy, but rather to discuss the effects of the different designs of the two systems on a set of well-formed sentences.

While our corpus was chosen to be linguistically diverse, it cannot serve as a suitable basis for drawing statistical conclusions about the differences that ZulMorph and the MZRG would produce on a typical isiZulu text. Nevertheless, some insight may still be gained by quantifying the differences within this set.

For our comparison, we use both the original ZRG and the MZRG to linearise all 100 sentences. We then use ZulMorph to obtain analyses for each token in the ZRG linearisations, which are then manually disambiguated. We also perform a simplification of the ZulMorph analyses in order to discount the major known differences between the two sets, namely the consolidation of the noun preprefix and base prefix, as well as the extensions into the verb roots.

We therefore have a parallel list of 243 analysed tokens from a corpus of 100 sentences, resulting in an average sentence length of 2.4 tokens. We

```
prefix_nasal : Str -> Str = \root -> case root of {
  "ph"+x => "mp" + x ;
  "Ph"+x => "mP" + x ;
  "bh"+x => "mb" + x ;
  "Bh"+x => "mB" + x ;
  -- ...

prefix_nasal : Str -> ClassGender -> Number -> Str = \root,classgender,number -> let
  class_tag = noun_prefix_tag classgender number ;
in
  case root of {
  "ph"+x => "m[NPre]"+class_tag+"p" + x ;
  "Ph"+x => "m[NPre]"+class_tag+"P" + x ;
  "bh"+x => "m[NPre]"+class_tag+"b" + x ;
  "Bh"+x => "m[NPre]"+class_tag+"B" + x ;
  -- ...
```

Figure 6: Original and adapted code snippets for roots prefixed with the nasal sound of with classes 9 and 10.

	isiZulu word	Analyses
1	eqa	e[RC][9]eq[VRoot]a[VT]
		[RC][9]eq[VRoot]a[VT]
2	echosheni	e[LocPre]i[NPrePre][5]li[BPre][5]chopho[NStem][5-6]ini[LocSuf]
		e[LocPre]chosh[NStem][5-6]eni[LocSuf]
3	ezweni	e[LocPre]u[NPrePre][14]bu[BPre][14]zwe[NStem][14]ini[LocSuf]
		e[LocPre]zwe[NStem][5-6]ni[LocSuf]
4	kudadewenu	ku[LocPre]u[NPrePre][1a]dadewenu[NStem][1a-2a]
		k[LocPre]u[NPre][1a]dadewenu[NStem][1a-2a]
5	ushisa	u[SC][3]sh[VRoot]is[CausExt]a[VT]
		u[SC][3]shis[VRoot]a[VT]

Table 1: A comparison of annotations by ZulMorph and MZRG (blue)

first determine the unique analyses in each list: the ZulMorph list contains 187 unique analyses, while the MZRG list contains 188. The discrepancy is due to the word *ukudla* being used both as a noun meaning 'food' and as an infinitive meaning 'to eat' in different trees. In simplified form, ZulMorph always provides the deverbative analysis, namely `uku[NPre][15]dl[VRoot]a[VT]`, while in different syntactic contexts, depending on the composition of the abstract syntax tree, the MZRG provides `uku[NPre][15]dl[VRoot]a[VT]` (in the sentence meaning 'To eat is good') and `uku[NPre][15]dla[NStem][15]` (in the sentence meaning 'The woman cooks the food.').

Among the 243 tokens in each list, 127 analyses (52.2%) are identical, of which 101 are unique. These words represent those that have undergone no morphophonological alternation, and hence not only the tag sequence, but also the surface forms are the same. On the other hand, 208 analyses (85.6%) share identical tag sequences, of which 159 are unique. In such cases, ZulMorph and the MZRG produces the same tag sequence (apart from the simplification mentioned above), although the surface form of the token exhibits morphophonological alternation.

The differences in tag sequences for the remaining tokens are the result of different design decisions taken in the development of the two systems, such as differences in how the subject concord is tagged in different contexts, cases where nouns belong only to one class and not to a pair of classes (this is not encoded in the ZRG), a distinction in the ZRG between locative nouns and other adverbs, and slight differences in how class information is associated with certain tags, such as the copulative prefix. However, both sets of analyses are completely accurate with respect to these design decisions.

6.2 MZRG compared to a data-driven tool

Having shown how the MZRG compares to ZulMorph, we now contrast it to a data-driven tool. Arguably the state-of-the-art data-driven morphological analyser that is freely available for isiZulu is the CTexT Core Technologies (CTT) tool bundled with the *ctextcore* Python package (Centre for Text Technology). It produces an analysis for any input, in contrast to the ZRG and ZulMorph. Gener-

Measure	Count	Percentage
Correct analysis	103	42.4%
Correct stem/root	130	53.5%
Plausible sequence	166	68.3%

Table 2: Preliminary comparison of accuracy

ally speaking, rule-based tools prioritise accuracy over coverage, while data-driven tools prioritise coverage, often resulting in lower accuracy. This is especially the case in resource-scarce environments.

Our aim in this comparison (Table 2) is to give some indication of the difference in accuracy of the CTT analyser compared to the MZRG. We therefore run the CTT analyser on the same set of linguistically diverse sentences. The results are manually evaluated and successes and errors categorised: for each analysis of a token, we indicate if it is correct, whether the stem or root of the token is correctly identified, and whether the morpheme sequence is at all plausible. Note that the final category does not consider whether the analysis of any given token is plausible for that word, but only whether the tag sequence is one that *could* exist in isiZulu. Typical errors are cases where two roots or stems are present in the analysis, where verb prefixes and a noun stem or (incorrect) noun prefixes and a verb root are present, or where prefixes appear at the end of a sequence.

One concern with the CTT analyser is its inconsistency with regards to surface form or canonical annotation. The data on which it was trained comprises canonical annotations (Gaustad and McKellar, 2024), but in many cases, the CTT failed to provide canonical forms of the morphemes. Our analysis in Table 2 did not penalise the CTT for this, but it may prove problematic for using the CTT within a larger NLP pipeline.

The three measures revealed subsets within the analyses: the set of tokens correctly analysed is a subset of those for which the stem or root is correctly identified, which in turn is a subset of those analyses that represent a plausible morpheme sequence. Viewed from a user perspective, on a well-behaved set of sentences with average length of 2.4 tokens per sentence, a user can expect around a third of analyses to be implausible, with the root or stem incorrectly identified. Moreover, a user can expect just less than half of analyses to incorrectly identify the root or stem and just below 60% of analyses to be erroneous in some way.

A full comparison of the accuracy and coverage, with reference to precision and recall, of the MZRG and the CTT on a larger, less curated and more natural corpus is beyond the scope of this paper and forms part of future work. The comparison given here should be regarded as a preliminary result indicating that a more in-depth comparative evaluation may be worthwhile.

6.3 Obtaining accurate annotations

We end this section by briefly sketching the workflows required for using the MZRG, ZulMorph and the CTT analyser to obtain accurate morphological annotations for an isiZulu corpus.

ZulMorph is highly accurate (Bosch, 2020) and with a lexicon of several thousand items, provides excellent coverage of the isiZulu language. However, additional manual effort is required to disambiguate all the possible analyses for each token. This kind of approach was followed in developing the annotated corpora on which the CTT analyser was trained (Gaustad and McKellar, 2024).

CTT analyser will produce output on any input, but requires additional manual effort in correcting erroneous analyses. Our analysis shows that around 60% of tokens may need to be corrected in this way, or around 50% if only correct identification of stems or roots is required.

MZRG depends on the availability of a GF treebank. The GF runtime in conjunction with the original ZRG provides a probabilistic parser, which mitigates the effort involved, namely that of manually disambiguating between possible trees. Once the treebank is obtained, highly reliable, syntactically contextualised annotations may be generated via the MZRG.

7 Conclusion

We presented an adapted version of the ZRG for morphological annotation (MZRG). The string elements of the ZRG were decorated with morphological tags, resulting in contextually appropriate morphological analyses. The adaptation makes explicit the morphosyntactic model of the ZRG during the process of linearisation. We contrasted the MZRG with two other state-of-the-art tools to show how it could be used on a linguistically diverse set of isiZulu sentences.

References

Sonja Bosch. 2020. Computational morphology systems for Zulu–a comparison. *Nordic Journal of African Studies*, 29(3):28–28.

Centre for Text Technology. CTexT Core Technologies for South African languages.

Tanja Gaustad and Cindy A McKellar. 2024. Updated morphologically annotated corpora for 9 south african languages. *Journal of Open Humanities Data*, 10(1).

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The state of Large Language Models for African languages: Progress and challenges.

Nancy Ide and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*, 1st edition. Springer Publishing Company, Incorporated.

Inge M. Kosch. 2006. *Topics in morphology in the African language context*. Unisa Press.

Laurette Marais and Laurette Pretorius. 2023. Parsing IsiZulu Text Using Grammatical Framework. In *Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference*, pages 167–177, Cham. Springer Nature Switzerland.

Sthembiso Mkhwanazi and Laurette Marais. 2024. Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and surface morphological segmentation for Nguni languages. In *Southern African Conference for Artificial Intelligence Research*, pages 125–139. Springer.

George Poulos and Christian T. Msimang. 1998. *A linguistic analysis of Zulu*. Via Afrika Ltd.

Laurette Pretorius and Sonja Bosch. ZulMorph: Finite state morphological analyser for Zulu (Version 20190103).

Laurette Pretorius and Sonja Bosch. 2010. Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues. In *Finite-State Methods and Natural Language Processing*, pages 123–130, Berlin, Heidelberg. Springer Berlin Heidelberg.

DJ Prinsloo and Gilles-Maurice de Schryver. 2001. Corpus applications for the African languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies*, 19(1-2):111–131.

Aarne Ranta. 2011. *Grammatical Framework: Programming with multilingual grammars*. CSLI Publications, Stanford, California.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract syntax as interlingua: Scaling up the Grammatical Framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2):425–486.

P.C. Taljaard and S.E. Bosch. 1988. *Handbook of IsiZulu*. J.L. Van Schaik.

Elsabé Taljard and Gilles-Maurice de Schryver. 2016. A corpus-driven account of the noun classes and genders in Northern Sotho. *Southern African Linguistics and Applied Language Studies*, 34(2):169–185.

Katharina von der Wense, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 961–967.

Siswati Part of Speech Tagger: A Quantitative Evaluation

Muzi Matfunjwa

South African Centre for Digital Language Resources, North-West University

Muzi.Matfunjwa@nwu.ac.za

Abstract

This article evaluates the performance of the Siswati Text Annotation Tool part of speech (STAT POS) tagger using Recall, Precision and F1 score metrics. A quantitative research design was adopted for analysis, and data was collected through purposive sampling. Python 3 was utilised to calculate the Recall and Precision of the STAT POS tagger outputs. The results show that the Recall for nouns was 0.761, Precision 0.417, with an F1 score of 0.54; for verbs, the Recall was 0.756, Precision 0.798 and F1 score 0.54; for adverbs, the Recall was 0.571, Precision 0.8, and F1 score 0.67; for possessives, the Recall was 0.963, Precision 0.813 and F1 score 0.88. For relatives (REL), the Recall was 0.706, Precision 0.523, and the F1 score 0.60; for class-indicating demonstratives, the Recall was 0.333, Precision 0.25, and the F1 score 0.29; and for copulatives (COP), the Recall was 0.75, Precision 0.75, and the F1 score 0.75. For conjunctions, the Recall was 0.85, the Precision was 0.68, and the F1 score was 0.76; for pronouns, the Recall was 0.563, the Precision was 1.0, and the F1 score was 0.72; for adjectives, the Recall was 0.75, the Precision was 0.75, and the F1 score was 0.75. However, question words, interjections and ideophones received 0.0, highlighting the need for refinement of the STAT POS tagger.

Keywords: Siswati, Part of speech tagger, Recall, Precision, F1 score

1 Introduction

Siswati is one of the under-resourced languages in South Africa, especially in Human Language Technology (HLT) tools (Mlambo and Matfunjwa 2025). Although HLT has existed in South Africa for the past two decades, Siswati still does not receive sufficient attention in this field compared to other Nguni languages such as isiZulu and isiXhosa. In recent years, there has been notable progress in the development of HLT for Siswati, including machine translation, morphological analysers and decomposers, lemmatisers, and part of speech taggers created by various entities such as the Centre for Text Technology (CTexT) at North-West University. Typically, these tools' accuracy and efficacy are lower than those of well-resourced languages because they are developed using inadequate data, mainly from government documents (Bosch, Pretorius & Fleisch 2008; Grover et al. 2011; Heeringa, De Wet and Van Huyssteen 2015; Koehn & Knowles 2017; Mlambo and Matfunjwa 2024). This makes it imperative to evaluate the HLT tools to determine the quality of their output using a different corpus. Therefore, this article evaluates the performance of the Siswati Text Annotation Part of Speech (POS) tagger using Recall, Precision and F1 score.

2 Related Works

Scholars have developed and assessed a variety of POS taggers for the official languages of South Africa using various metrics. Eiselen and Puttkammer (2014) developed and evaluated POS taggers for ten South African languages, excluding English, as part of the National Centre for Human Language Technology project. The taggers were evaluated, revealing superior performance for Afrikaans, Xitsonga, Tshivenda, Sesotho sa Leboa, Setswana, and Sesotho, which use disjunctive writing systems, in contrast to the Nguni languages that employ an agglutinative writing system. This was attributed to the morphological complexity of the Nguni languages and the need for more data to mitigate this. In contrast, using annotated corpora, Du Toit and Puttkammer (2021) developed four text annotation language tools for the Nguni

languages: isiNdebele, Siswati, isiXhosa, and isiZulu. These tools included POS taggers, lemmatisers, morphological analysers, as well as morphological decomposers. Previously created tools as part of the NCHLT project were improved upon by this.

Heid, Prinsloo, Faaß, and Taljard (2009) created a noun POS tagger for Northern Sotho, also referred to as Sesotho sa Leboa. The tagger underwent testing, yielding results that indicated a 90% accuracy rate in recognising nouns along with their corresponding noun-class numbers. Mathe and Eiselen (2021) then assessed the performance of two Sesotho sa Leboa POS taggers created by NCHLT and CTexT. The authors determined that the NCHLT tagger exhibited difficulties in accurately tagging noun classes, specifically failing to differentiate between nouns in Class 9 and those in Class 10, resulting in an accuracy of 88.40%. The CTexT tagger achieved an accuracy of 94.18%, indicating minimal errors in the tagging of noun classes. The errors identified in the CTexT tagger were linked to foreign words and proper nouns, which were categorised as either Class 9 nouns or foreign words.

Malema, Okgetheng and Motlhanka (2017) created and assessed a Setswana POS tagger for the identification of relatives. A set of ten pages of Setswana text, comprising 77 relatives, was analysed. The tagger successfully identified 65 of these relatives, resulting in an overall performance rate of 84%. It was determined that the language's disjunctive orthography made it difficult to carry out precise automatic POS tagging and tokenisation of the relatives. Dibitso, Owolawi, and Ojo (2019) employed annotated corpora comprising 65,784 tokens derived from governmental documents to create and train a Setswana POS tagger. The corpus was annotated with tagsets developed following EAGLES guidelines and the tagsets proposed by Taljard, Faaß, Heid, and Prinsloo (2008). The utilised tagsets comprised 128 tags formulated to satisfy the morphosyntactic needs of Setswana text. The developed POS tagger underwent an accuracy assessment, revealing that its accuracy improved with an increase in the number of training words.

Matfunjwa (2025) evaluated and compared the CTexT NCHLT Web Service POS tagger with the Text Annotation Tools POS tagger for Siswati based on accuracy using data sourced from a novel titled *Tinyembeti*. The author found that the overall accuracy of the Text Annotation Tools' POS tagger was higher than that of the CTexT NCHLT Web Service POS tagger in all word categories tagged. It was also determined that these taggers struggled to identify interjections, ideophones, and Class 1a nouns.

From the consulted work, it is evident that no study has evaluated the performance of the Siswati Text Annotation Tools POS tagger based on the metrics: Recall, Precision and F1 score.

3 Methodology

This study is quantitative and uses data obtained from the Siswati drama book titled *Hawu Babe* written by Malindzisa (2005). A sample of 389 tokens, including words, punctuation, and numbers, was purposively collected in a PDF format from the book and converted into a text document. These words were selected because they represent the POS found in Siswati. This data was then manually cleaned to remove numerals, English words, and punctuation, resulting in a clean corpus of 345 Siswati words. This was then uploaded to the Siswati Text Annotation Tool (henceforth, STAT POS) tagger for automatic tagging, as presented in Figure 1. This POS tagger in Figure 1 was obtained from the South African Centre for Digital Resources website at https://repo.sadilar.org/handle/20.500.12185/548. The results received from the STAT POS tagger were then exported to an Excel spreadsheet and stored as a CSV file. A gold standard was created, assigning accurate tags to each word against the predicted POS tags by the tagger. The tagsets used in the POS tagger are presented in Table 1. This resulted in three columns: pos, which contained the words, prediction, which consisted of the tagging by the tagger, and the truth, which was the gold standard being created. Following that, the CSV file was loaded into Jupyter Notebook and then imported into Python 3 using the Pandas package. The Recall and Precision were calculated using Python 3. The formulas for $Recall = \frac{TP}{TP+FN}$ and for $Precision = \frac{TP}{TP+FP}$ were utilised. Thereafter, F1 scores for the POS

were calculated using the formula $\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

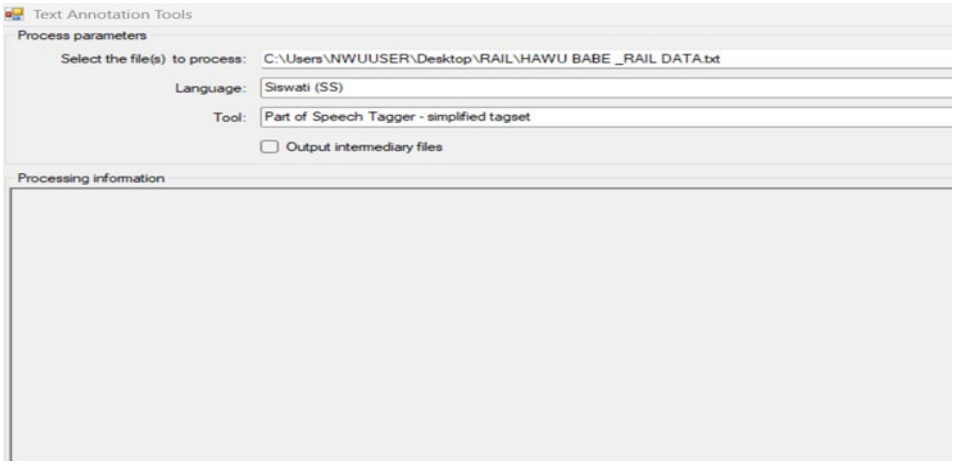


Figure 1: Siswati Text Annotation POS tagger

Part of Speech	Tagsets
Noun	N
Verb	V
Adjective	ADJ
Adverb	ADV
Class-indicating demonstrative	CDEM
Conjunction	CONJ
Copulative	COP
Interjection	INT
Question word	INTER
Place and brand name	NPP
Possessive	POSS
Pronoun	PRO
Quantitative pronoun	PROQUANT
Relative	REL
Abbreviation (incl. acronyms)	ABBR

Table 1: Tagsets used in the STAT POS tagger

4 Results

4.1 Recall

Table 2 shows the results for the calculated Recall for each POS using Python 3.

POS	Recall
N	0.761
V	0.756
ADV	0.571
POSS	0.963
REL	0.706
CDEM	0.333
COP	0.75
CONJ	0.85
PRO	0.563
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 2: Recall for all POS in the data

In Table 2**,** the Recall for N (nouns) is 0.761, V (verbs) 0.571, ADV (adverbs) 0.571, POSS (possessives) 0.963, REL (relatives) 0.706, CDEM (Class-indicating demonstrative) 0.5, COP (copulatives) 0.75, CONJ (conjunctions) 0.85, PRO (pronouns) 0.563, ADJ (adjectives) 0.75, INTER (question words) 0.0, INT (interjections) 0.0 and IDEO (ideophone) 0.0.

4.2 Precision

Table 3 shows the results for the calculated Precision for each POS using Python 3.

POS	Precision
N	0.417
V	0.798
ADV	0.8
POSS	0.813
REL	0.523
CDEM	0.25
COP	0.75
CONJ	0.68
PRO	1.0
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 3: Precision for all POS in the data

In Table 3, the Precision for N is 0.417, V 0.798, ADV 0.80, POSS 0.813, REL 0.523, CDEM 0.25, COP 0.75, NPP 0.0, CONJ 0.68, PRO 1.0, ADJ 0.75, INTER 0.0, INT 0.0 and IDEO 0.0.

4.2 F1 Score

Table 4 shows the results for the calculated F1 scores for each POS.

Part of Speech	F1 score
N	0.54
V	0.78
ADV	0.67
POSS	0.88
REL	0.60
CDEM	0.29
COP	0.75
CONJ	0.76
PRO	0.72
ADJ	0.75
INTER	0.0
INT	0.0
IDEO	0.0

Table 4: F1 score for POS

In Table 4, the calculated F1score for N is 0.54, V 0.78, ADV 0.67, POSS 0.88, REL 0.60, CDEM 0.29, COP 0.75, CONJ 0.76, PRO 0.72, ADJ 0.75, INTER 0.0, INT 0.0 and IDEO 0.0.

5 Discussion

5.1 Nouns

The Recall of 0.761 for nouns means that out of all the true nouns in the data, the STAT POS tagger accurately tagged 76.1% of them, showing that it identified most nouns in the data.

The Precision of 0.417 reveals that out of all the words the tagger labelled as nouns, only 41.7% were true nouns, demonstrating that the tagger is also tagging a lot of words that are not nouns as nouns. The F1 score of 0.54 shows that the tagger achieved a moderate overall performance, even though there is a disparity between its Recall and Precision.

5.2 Verbs

The Recall of 0.756 for verbs reveals that out of all the actual verbs in the data, the tagger correctly identified 75.6% of them, with the remaining 38.5% being false negatives, which are the real verbs missed or misclassified by the tagger. The Precision of 0.798 means that out of all the words the tagger labelled as verbs, 79.8% were indeed verbs, with the remaining 20.2% being false positives, that is, words the tagger incorrectly tagged as verbs when they were other POS. The F1 score of 0.78 indicates that the tagger attained a strong and balanced performance, assigning verb tags with high accuracy.

5.3 Adverbs

The Recall of 0.571 for adverbs means that out of all the adverbs that appear in the dataset, the POS tagger correctly identified 57.1% of them as adverbs, and the remaining 42.9% are false negatives, which are the real adverbs that the tagger missed or classified as other POS. The Precision of 0.80 shows that of all the words the tagger labelled as adverbs, 80% were indeed adverbs, with the remaining 20% being false positives, that is, words the tagger incorrectly tagged as adverbs when they were something else. The F1 score of 0.67 shows a moderate overall performance of the tagger, with a fairly accurate prediction of adverbs, but missing many of them. This F1 score reveals an imbalance between the low recall and high precision.

5.4 Possessives

The Recall of 0.963 means that out of all the actual possessives in the dataset, 96.3% were correctly identified by the POS tagger, with the remaining 3.7% being false negatives or real possessives missed by the tagger. The Precision of 0.813 reveals that out of all the words the tagger

labelled as possessive, 81.3% were real possessives, and the remaining 18.7% were words it labelled as possessive but were not. The F1 score of 0.88 indicates a strong and very good overall performance of the tagger, with a high, well-balanced accuracy.

5.5 Relatives

The Recall of 0.706 for relatives means that out of all the relatives in the data, the POS tagger correctly tagged 70.6% of them. The Precision of 0.523 means that out of all the words the tagger labelled as relatives, only 52.3% were correct, demonstrating that the tagger is producing a lot of false positives, that is, many words it marks as relatives are not relatives. The F1 score of 0.60 indicates that the POS tagger has moderate performance, successfully identifying many correct relatives with frequent tagging errors. This reveals an imbalance as the tagger has a strong recall at the expense of precision.

5.6 Class indicating demonstratives

The Recall of 0.333 for demonstratives means that out of all the actual demonstratives in the data, the POS tagger correctly identified 33.3% of them and missed 66.7%, meaning that many of the demonstratives were tagged as something else. The Precision of 0.25 shows that out of all the words the tagger labelled as demonstratives, only 25% were real demonstratives. The 75% were false positives; words incorrectly tagged as demonstratives. The F1 score of 0.29 indicates a poor performance of the tagger, showing failure in consistently and accurately tagging demonstratives.

5.7 Copulatives

The Recall of 0.75 for copulatives means that out of all the actual copulatives in the dataset, the POS tagger correctly identified 75% of them. This implies that it missed about 25% of the copulatives, and these were false negatives. The Precision of 0.75 shows that out of all the words the tagger predicted as copulatives, 75% were demonstratives, implying that 25% of its copulative predictions are false positives (words incorrectly tagged as copulatives). The F1 score of 0.75 indicates a balanced and strong overall performance of the tagger.

5.8 Conjunctions

The Recall of 0.85 for conjunctions means that out of all the actual conjunctions in the dataset, the tagger found 85% of them, missing about 15% of conjunctions (false negatives). The Precision of 0.68 means that out of all the words the tagger labelled as conjunctions, only 68% were real conjunctions, suggesting that 32% of the tagger's 'conjunctions' predictions are wrong (false positives). The F1 score of 0.76 indicates that the tagger has good performance overall, accurately identifying most conjunctions, despite a considerable number of tagging errors.

5.9 Pronouns

The Recall of 0.563 on pronouns means that out of all actual pronouns in the data, the tagger correctly identified only 56.3% of them. The Precision of 1.0 means that out of all the words the tagger predicted as pronouns, every single one was correct. This shows that there were no false positives, and it never tagged a non-pronoun as a pronoun. The F1 score of 0.72 indicates that the tagger has good performance but lacks comprehensiveness, resulting in perfect precision but average recall.

5.9.1 Adjectives

The Recall of 0.75 for adjectives means that out of all actual adjectives in the dataset, the POS tagger correctly identified 75% of them. The Precision of 0.75 shows that out of all the words the tagger predicted as adjectives, 75% were true adjectives, implying that 25% of its adjective predictions were false positives. The F1 score of 0.75 indicates a balanced and strong overall performance of the tagger.

5.9.2 Question words, interjections, and ideophones

The Recall of 0.0 for question words (INTER), interjections (INT) and ideophones (IDEO) means that the POS tagger failed to identify any of these POS in the dataset correctly. The Precision of 0.0 also means that none of the words the tagger labelled as INTER, INT and IDEO were in these word categories. The F1 score of 0.0 therefore, shows that the STAT POS tagger underperformed in the tagging INTER, INT and IDEO.

6 Conclusion

The results show that the Recall for nouns was 0.761, Precision 0.417, with an F1 score of 0.54; for verbs, the Recall was 0.756, Precision 0.798 and F1 score 0.54; for adverbs, the Recall was 0.571, Precision 0.8, and F1 score 0.67; for possessives, the Recall was 0.963, Precision 0.813 and F1 score 0.88. For relatives, the Recall was 0.706, Precision 0.523, and the F1 score 0.60; for class-indicating demonstratives, the Recall was 0.333, Precision 0.25, and the F1 score 0.29; and for copulatives, the Recall was 0.75, Precision 0.75, and the F1 score 0.75. For conjunctions, the Recall was 0.85, the Precision was 0.68, and the F1 score was 0.76; for pronouns, the Recall was 0.563, the Precision was 1.0, and the F1 score was 0.72; for adjectives, the Recall was 0.75, the Precision was 0.75, and the F1 score was 0.75. Meanwhile, the Recall, Precision, and F1 score for question words, interjections and ideophones were 0.0. These results show that the STAT POS performed better in tagging possessives, verbs, conjunctions, copulatives, adjectives, relatives, and nouns. However, it performed poorly on demonstratives, question words, interjections and ideophones, highlighting the need for refinement of the STAT POS tagger to enhance its tagging, especially when data from a non-governmental domain is used.

In future, the researcher will use data from Wikipedia and other sources that demonstrate multifaceted use of Siswati to evaluate the performance of the STAT POS tagger using the metrics accuracy, Recall, Precision and F1 score. For improving the performance of the STAT POS tagger, this study recommends that collaboration between Siswati POS tagger developers, language speakers, and linguists is essential for enhancing the efficiency of the tagger. Speakers of the language are required to provide extensive corpora that can be derived from Siswati literature and Wikipedia. The corpora would include all POS, including ideophones and interjections, utilised in various contexts within this language for training, refining and testing of the POS tagger.

7 Limitations of the study

This study utilised a limited corpus from a single genre to evaluate the performance of the STAT POS tagger.

References

Sonja Bosch, Laurette Pretorius and Axel Fleisch. 2008. Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, 17(2), 66–88.

Jakobus S. Du Toit and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce Nguni languages. *Information*, 12(12), 520. https://doi.org/10.3390/info12120520.

Roald Eiselen and Martin J. Puttkammer. 2014. *Developing text resources for ten South African languages*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3698-3703, Reykjavik, Iceland. European Language Resources Association.

Aditi Sharma Grover, Gerhard B. Van Huyssteen, and Marthinus W. Pretorius. 2011. The South African human language technology audit. *Language Resources and Evaluation* 45(3): 271–288. https://doi.org/10.1007/s10579-011-9151-2.

Wilbert Heeringa, Febe De Wet, and Gerhard B. Van Huyssteen.2015. Afrikaans and Dutch as closely related languages: A comparison to West Germanic languages and Dutch dialects. *Stellenbosch Papers in Linguistics Plus* 47: 1–18. https://doi.org/10.5842/47-0-649.

Ulrich Heid, Danie J. Prinsloo, Gertrud Faaß, and Elsabé Taljard. 2009. Designing a noun guesser for part of speech tagging in Northern Sotho. *South African Journal of African Languages*, 29(1): 1–19.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872.

Gabofetswe Malema, Boago Okgetheng, and Moffat Motlhanka. 2017. Setswana part of speech tagging. *International Journal of Natural Language Computing*, 6(6): 15–20.

Gubudla Aaron Malindzisa. 2005. *Hawu babe!* Shuter & Shooter, Pietermaritzburg.

Dimakatso S. Mathe and Roald Eiselen. 2021. Quantitative analysis of Sesotho sa Leboa part of speech tagger. South African. *Journal of African Languages, 41*(3): 259-269 https://doi.org/10.1080/02572117.2021.2010921.

Muzi Matfunjwa (2025). The efficacy of a Siswati part-of-speech tagger. In H. Ekkehard Wolff and Justus C. Roux, editors, Contextualising African Language Dynamics of Change, pages 251-268. African Sun Media, Stellenbosch, https://doi:10.52779/9781991260581.

Respect Mlambo and Muzi Matfunjwa. 2025. Human language technology tools for indigenous South

African languages and their potential use. *Literator* 46(1), a2049. https://doi.org/10.4102/lit.v46i1.2049.

Respect Mlambo and Muzi Matfunjwa. 2024. The use of technology to preserve indigenous languages of South Africa. *Literator 45*(1), a2007. https://doi.org/10.4102/lit. v45i1.2007

Siswati Text Annotation Tools Part of Speech Tagger. Available: https://repo.sadilar.org/handle/20.500.12185/548. Accessed on 5 August 2025.

Multilingual Data from the Agricultural Domain: Presenting the NWU-Pula/Imvula Corpora

Tanja Gaustad and Cindy McKellar and Martin J. Puttkammer
Centre for Text Technology (CTexT)
North-West University
Potchefstroom, South Africa
{tanja.gaustad|cindy.mckellar|martin.puttkammer}@nwu.ac.za

Abstract

This paper presents new multilingual corpora from the agricultural domain for seven South African Languages, namely Afrikaans, English, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, and Setswana, based on the Pula/Imvula magazine. After pre-processing, the data has been automatically sentencized, tokenized, lemmatized and annotated with part-of-speech information using the services available at https://v-ctx-lnx7.nwu.ac.za/. The final resources comprising between 774k and 1,38M tokens per language are included on the Corpus Cooperative at North-West University (COCO@NWU) corpus platform at https://coco.nwu.ac.za/ as searchable corpora. In addition, the data can be made available as text files for research purposes upon request. To highlight the value of this agricultural domain-specific data collection in relation to more general data, we also include some corpus-based statistics and comparisons with previous research.

1 Introduction

As has already been pointed out by Hanks (2000), language usage is dependent on domain. For instance, lexicography (Drouin, 2004), language learning (Barrière, 2009) and machine translation (van Noord et al., 2022) benefit greatly from domain-specific corpora. Domain-specific datasets are collections of data customized to a particular field or application. These datasets focus on distinct types of content, information, formats, or use cases relevant to a domain, such as healthcare, finance, or—in our case—agriculture. For example, a financial dataset could contain transaction records flagged for fraud, or a question-answer dataset could contain data tailored to medical records and linked diagnoses. Unlike general-purpose datasets, domain-specific datasets address niche problems: Their value lies in capturing real-world patterns which are pertinent to a field and which will not be found in generic data, enabling models to perform tasks particular to that field with higher accuracy.

Currently, in the age of Large Language Models (LLMs), there is a reliance on large amounts of general heterogeneous data implicitly assuming that different domains will be covered due to sheer volume. For languages with less data, however, coverage of different domains will generally be (very) low. As most of the South African languages are considered under-resourced, domain-specific corpora can hopefully add value to develop machine learning and AI solutions, e.g. for fine-tuning LLMs, in areas of importance for everyday life.

One such area is agriculture: The livelihood of billions of people worldwide depends on agriculture and AI-enabled solutions show promise to contribute to solutions geared at reaching sustainable development goals (see https://sdgs.un.org/goals). However, often the necessary training or information is not available in the native language of the farmers and therefore does not reach the work force. With the multilingual corpora presented in this paper, we hope to contribute a new valuable resource for the development of tools aimed at agriculture.

The rest of the paper is structured as follows: After a description of some background on domain-specific research for low-resource settings in section 2, we discuss the source data used (section 3) as well as the pre-processing and automatic annotations applied (section 4). We then describe where the corpora can be accessed (section 5), followed by some corpus-based statistics and lexicon comparisons to give the reader a better idea of the value of this agricultural domain-specific data in relation to more general data (section 6). We end with conclusions (section 7) and limitations.

2 Background

One of the aims of UNESCO's initiative "Language Technologies for All" (LT4All) is to advance language technologies in order to (help) preserve linguistic heritage and promote multilingualism. An indispensable building block to achieving this aim are language corpora as they are essential for language research, language learning, and the development of language technologies.

Echoing some of the aims of LT4All, the recent white paper by Pava et al. (2025) focuses on the challenges and strategies for developing LLMs for low-resource languages. These languages face limitations due to data scarcity as well as data that is not representative of the socio-cultural context. One of the recommendations of the paper is to invest in research that increases the availability and quality of low-resource language data. We believe this includes domain-specific data.

Lately, investigations into using domain-specific data have become popular. For instance, Edwards et al. (2020) show that using unlabeled domain-specific data in supervised text classification delivers more robust results than models trained on more but only general data, such as BERT (Devlin et al., 2019), even when BERT is pre-trained on domain-relevant data.

Experiments conducted by Singh et al. (2022) also indicate that models trained with domain-specific implicit reasonings significantly outperform domain-general models in both automatic and human evaluations.

Tang and Yang (2024) investigate if the development of domain-specific embedding models is necessary and even useful, specifically for the financial domain. They come to the conclusion that general-purpose models struggle to capture domain-specific linguistic and semantic patterns whereas using domain-specific data delivers better performance.

The main advantages of domain-specific AI based on domain-specific data is that it is more relevant to the task, more reliable, and more easily scalable. The major downside is the limited adaptability of a domain-specific system.

3 Source Data and Languages Included

The data included in the corpora presented here issue from Pula/Imvula, a South African magazine focusing on the developing farmer and published by Grain SA.[1] The main aim of the magazine is to support developing farmers in becoming sustainable commercial farmers. The magazine is distributed on a monthly basis and currently only available in English. However, until September 2024 it was published in five languages (English, isiXhosa, isiZulu, Sesotho, and Setswana), and previously also included Afrikaans and Sesotho sa Leboa (discontinued due to lack of funding). The Centre for Text Technology (CTexT) has a long-standing working relationship with the editors of the magazine and has been receiving and archiving the original MS Word documents since 2007.

For the seven corpora, we have combined files from editions acquired between 2007 and 2024 for English, isiXhosa, isiZulu, Sesotho, and Setswana, and between 2007 and 2019 for Afrikaans and Sesotho sa Leboa (see Table 1).

4 Data Pre-Processing and Automatic Annotations

To ready the data orginally received from the editors for deployment as corpora, pre-processing steps as well as automatic sentencization, tokenization, lemmatization and annotation with part-of-speech (POS) tags were required. These steps will now be described in more detail.

4.1 Pre-Processing

In a first step, all data was extracted from the original MS Word documents to UTF-8 encoded text files. To ensure the accuracy and quality of our corpora, all text was then run through a language identifier (Hocking, 2014; Puttkammer et al., 2018) to ensure the correct language is contained in the respective corpora since South African documents sometimes contain mixed languages. Data was also checked for encoding problems caused by the incorrect use of diacritics as this is often a problem found in South African texts. Lastly, some English headers with instructions to the editors and translators of the magazine were stripped out.

Each file in the corpora contains metadata detailing the source of the data, the language it is in as well as the publication date. See Table 1 for the total amount of tokens contained in the final data for each language as well as Section 6 for more detailed information on the data.

[1] https://www.grainsa.co.za/farmer-development

Language	Years	# Tokens	# Types	TTR
Afrikaans	2007–2019	798,067	33,047	0.041
English	2007–2024	1,148,871	31,266	0.027
isiXhosa	2007–2024	825,364	113,613	0.138
isiZulu	2007–2024	774,124	120,335	0.155
Sesotho	2007–2024	1,376,801	30,672	0.022
Sesotho sa Leboa	2007–2019	1,014,905	28,357	0.028
Setswana	2007–2024	1,335,512	38,746	0.029

Table 1: Overview of final data included in the NWU-Pula/Imvula Corpora, including number of tokens, number of types and type-token ratio (TTR).

4.2 Automatic Annotations

All the non-English data has been automatically sentencized, tokenized, lemmatized and annotated with part-of-speech (POS) information using the CTexT NCHLT Web Services available at https://v-ctx-lnx7.nwu.ac.za/.[2] An in-depth explanation of the principal technologies used as well as evaluation results for isiXhosa and isiZulu can be found in (du Toit and Puttkammer, 2021).

The English data has been processed using the Stanza pipeline (Qi et al., 2020), also resulting in sentencized, tokenized, lemmatized and POS-tagged text.

5 Availability of the Corpora

The Corpus Cooperative at North-West University (COCO@NWU) is an initiative of the North-West University's (NWU) Faculty of Humanities. The broad aim of the initiative is to advance corpus-based research, in particular in the digital humanities. More specifically, it also aims to make corpora developed at the NWU available to all researchers and students at the NWU, their collaborators, or researchers outside the NWU.

The corpora available on the COCO@NWU platform have mostly been developed in collaboration with various corpus suppliers, such as publishing houses, news websites, (literary) blog sites, libraries, etc. Therefore, these corpora may be used for academic research purposes only.

In addition to the Afrikaans corpora already available on the platform, the multilingual Pula/Imvula data has been added as a searchable corpus for each language at https://coco.nwu.ac.za/ where researchers can carry out (corpus) linguistic and statistical analyses of the data. The platform includes the possibility for simple word-based searches as well as more advanced queries using information on words, POS, lemmas, etc. and allows for exports of the query results. Figure 1 contains a screenshot as an example of how the platform can be used for queries. The corpora can also be made available as UTF-8 encoded text files for research purposes upon request.

[2]The underlying python packages are available at https://pypi.org/project/ctextcore/.

6 Corpus-based and Lexicon Statistics

Table 1 contains an overview of relevant statistics for the presented corpora detailing the years of data included, the number of tokens, number of types and the associated type-token ratio for each language. As expected, the type-token ration for isiXhosa and isiZulu, two Nguni languages, is higher on account of their agglutinative nature and conjunctive writing style. The three Sotho languages, Sesotho, Sesotho sa Leboa and Setswana, all have lower type-token ratios as a result of being written disjunctively. The higher value for Afrikaans compared to English is due to its abundant use of compounding.

In 2002, Prinsloo and de Schryver (2002) presented an instrument to measure the degree of conjunctivism/disjunctivism of the South African languages. Building on this work, they later proposed multidimensional lexicographic rulers for all South African languages, which were "prediction instruments aimed at assisting the South African lexicographers with the compilation of their national dictionaries" (Prinsloo and de Schryver, 2005). These rulers drew on electronic corpora available at the time as well as on dictionary data, and their goal is to help predict what the distribution of dictionary entries per letter should be for a given language. Looking at dictionary entries for English, for example, it is immediately obvious that there are far less entries starting with X, Y an Z than there are entries starting with S. The distribution is, of course,

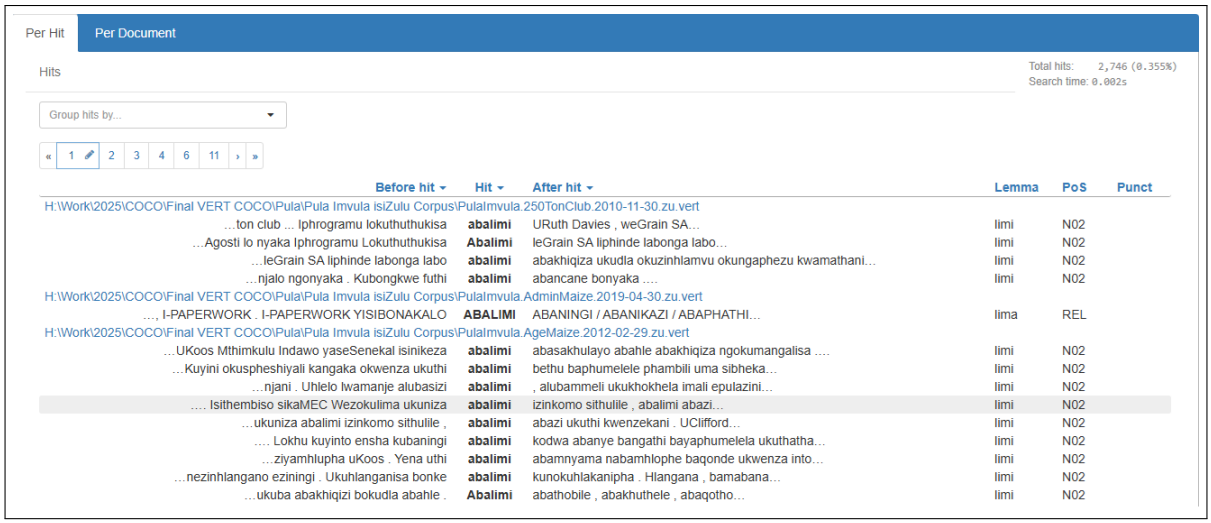


Figure 1: Screenshot of the COCO@NWU platform with a search result in the isiZulu Pula/Imvula corpus.

different for different languages. The authors were able to establish that so-called 'stretches', i.e. sections containing all lemmas starting with the letter A, then B, etc. could be modelled from corpora available for the language of interest.

To give the reader a better of idea of the valuable content of this agricultural domain-specific data in relation to more general data (albeit from 2005 and before), we compare the Pula/Imvula corpora to these rulers. This will help to identify any deviation from the expected distribution, possibly indicating new terms, loan words or spelling variations not present in the more general corpora used by (Prinsloo and de Schryver, 2005).

Figures 2 and 3 show the distribution of Afrikaans and English stretches compared to (Prinsloo and de Schryver, 2005). For Afrikaans, we see an increase of 7.55% in 'D' as well as an increase of 3.43% in 'I', both explained by the presence of agriculture-related terms such as *dier* ('animal'), *droog* ('dry'), *droogte* ('drought'), *insek* ('insect'), *inset* ('contribution') in the Pula/Imvula data, while e.g. 'S' is 5.31% lower compared to the data used by (Prinsloo and de Schryver, 2005). For English, only 'T' shows a large difference with an increase of 10.5%. Looking at tokens starting with 'T', we find several agricultural terms like *ton*, *tractor* or *tillage* that explain the increase.

For Figure 4, we calculated and plotted the difference between the ruler values in (Prinsloo and de Schryver, 2005) and in the lemmatised Pula/Imvula corpora for Sesotho (ST), Sesotho sa Leboa (NSO) and Setswana (TN). It is interesting to note that the variations are very similar for all three languages, with the exception of 'G' where the difference is only present for Sesotho sa Leboa and Setswana (8.88% and 10.29% respectively) and on 'H' where the difference is only present for Sesotho (10.44%).[3] The higher occurrence of 'G' in the Sesotho sa Leboa and Setswana data can be attributed to the higher than expected use of particles, for example *go* ('to/it') that is also the most frequent word in the Setswana corpus with 95,725 occurrences, *ga* ('I/it/he/she') and *gore* ('so that'). The Sesotho data shows similar observations with *ho* ('to/it') as the most frequent token occurring 71,868 times, followed by *ha* ('I/it/he/she') and *hore* ('so that'). Echoing the findings for 'G' and 'H', the increase for 'T' is mainly due to the high frequency of *tse/tše* ('this') and *tla* ('shall/will'). The higher use of particles can possibly be attributed to the instructional nature of the corpus.

The marked differences at 'K', 'L' and 'Y' can be explained by the use of specialised vocabulary such as *khemikale* ('chemical'), *kgwedi* ('month'), *kgwebo* ('business'), *laola* ('manage'), *laodi* (lemmatised form of *molaodi/balaodi* 'manager(s)') or agriculture-specific loanwords such as *yunite* ('unit') or *yield* ('yield'). The noticeable drop in lemmas starting with 'M' is most likely due to the deletion of class prefixes *mo-*, *ma-*, and *me-* (used in 5 different noun classes) during lemmatisation.

When examining the graphs for the isiXhosa (XH) and isiZulu (ZU) Pula/Imvula data in Figure 5, the differences compared to the rulers are quite similar for the two languages. The biggest

[3]This is due to a sound shift between 'g' and 'h' for these languages.

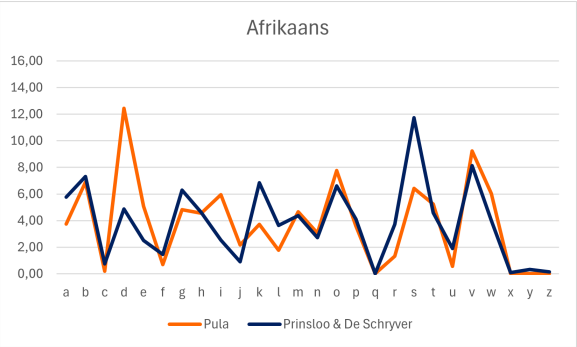


Figure 2: Graph showing the distribution of stretches for Afrikaans in Pula/Imvula compared to rulers in (Prinsloo and de Schryver, 2005).

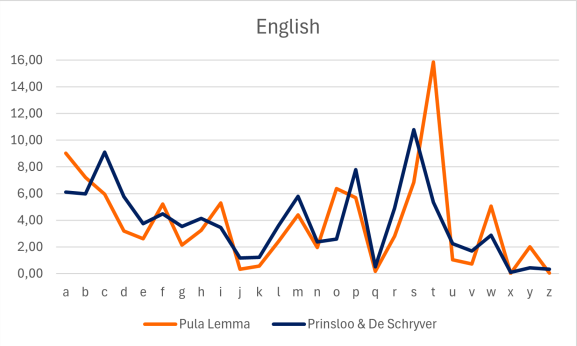


Figure 3: Graph showing the distribution of stretches for English in Pula/Imvula compared to rulers in (Prinsloo and de Schryver, 2005).

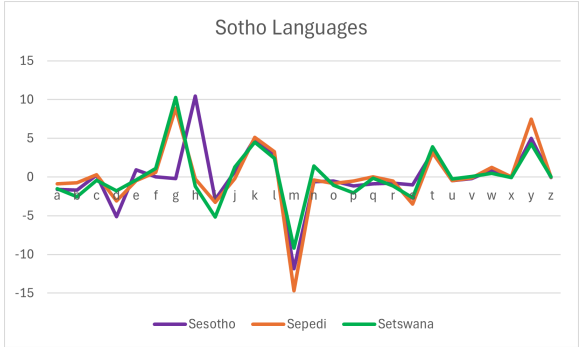


Figure 4: Graph showing the (calculated) differences in distribution for Sesotho, Sesotho sa Leboa and Setswana in the Pula/Imvula corpora.

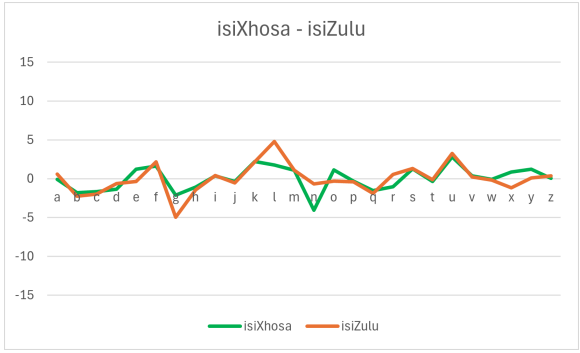


Figure 5: Graph showing the (calculated) differences in distribution for isiXhosa and isiZulu in the Pula/Imvula corpora.

dissimilarities are on 'F', 'K', 'L' and 'U'. Looking at examples from these four letters, we find that the large deviance is due to two reasons. On the one hand, words related to farming, learning or trade partners are prevalent, e.g. *fama* ('farm'), *fundo/a* ('study/learn'), *khemikali* ('chemical'), *khiqiza/o* ('produce/product'), *kg* ('kilogram'), *USA*, or *Ukraine*. On the other hand, like for the Sotho languages discussed above, the data contains a lot of particles used in instructions: *lokho/u* ('these'), *le/o* ('this'), *ukuthi* ('that'), or *ukuze* ('so that'). These particles together with lemmatised forms of 'farmer' and 'crop(s)' (*limi* and *limo*) also explain the bigger difference for isiZulu for 'L'.

In addition to the lexicographic rulers, we use AntConc (Anthony, 2025) to identify possible domain-specific terms from the Pula/Imvula corpora based on keyness calculated using log likelihood (Pojanapunya and Todd, 2018). The top 10 terms for each language are presented in Table 2.

As would be expected, the terms (based on full word forms, not on lemmas) are mostly of agricultural origin as can be seen from their translations. Namely the words *farmer* and *farmers* appear in the top ten selection for every language (Afrikaans: *boer/boere*; isiXhosa: *umlimi/abalimi*; isiZulu: *abalimi*; Sesotho/Sesotho sa Leboa: *molemi/balemi*; Setswana: *balemirui*). There are also many occurrences of other domain-specific terms across languages in just these top ten examples, such as *soil* (Afrikaans: *grond*; Sesotho: *mobu*; Sesotho sa Leboa: *mmu*) or *maize* (Afrikaans: *mielies*; isiZulu: *ummbila*; Sesotho: *poone*; Sesotho sa Leboa: *lehea*).

This further shows that the corpora are indeed a source that can be used to identify new terms specific to the agricultural domain and could be a valuable source for further linguistic research. It may also be possible, with a little help from a native speaker, to identify the terms along with their translations in other languages thereby creating translation lists of these domain-specific terms.

AF	Keyness	EN	Keyness	NSO	Keyness	ST	Keyness
graan (grain)	4046.24	crop	15112.39	balemi (farmers)	5416.89	poone (maize)	3850.75
boer (farmer)	3737.71	maize	14137.04	dibjalo (crops)	5410.29	mobu (soil)	3409.38
boere (farmer)	3566.30	grain	13348.70	ye (this one)	4434.32	dijothollo (cereal)	3258.05
grond (soil)	3238.18	soil	13049.94	lehea (maize)	4261.97	tlhahiso (production)	2585.10
mielies (maize)	3052.16	your	13047.17	tše (these)	3018.34	temo (farming)	2537.92
plant (plant)	2874.41	you	12350.11	puno (crop)	2891.40	balemi (farmers)	2514.88
sa (South Africa)	2361.17	farmers	11995.19	molemi (farmer)	2852.89	grain (grain)	2479.66
ha (hectare)	2290.34	farmer	9740.30	mme (madam)	2774.93	jala (sow)	2415.53
baie (much/very)	2077.75	production	8232.20	grain (grain)	2633.38	peo (seed)	2322.45
oes (harvest)	1970.48	can	8072.30	mmu (soil)	2283.31	molemi (farmer)	2217.17

TN	Keyness	XH	Keyness	ZU	Keyness
bokana (amount)	4539.75	abalimi (farmers)	3584.61	abalimi (farmers)	3825.05
go (to)	3949.11	sa (South Africa)	2298.85	sa (South Africa)	2718.37
balemirui (farmers)	3941.43	lokulima (of farming)	2150.72	isilimo (crop)	2333.58
jwala (sow)	3605.10	izityalo (plants)	1772.43	izilimo (crops)	2224.48
bolemirui (farming)	3445.27	ngehektare (hectare)	1750.17	ummbila (maize)	1997.58
tlhotlhwa (cost)	3216.86	zesoya (soya)	1667.59	isivuno (harvest)	1996.37
kumo (produce)	3190.06	isityalo (plant)	1547.00	ha (hectare)	1864.51
dijwalwa (seeds)	3129.57	ukuze (so that)	1542.46	ukhula (weed)	1725.21
kgono (skill)	3011.36	ngokunjalo (accordingly)	1532.64	isoya (soya)	1656.42
tlaa (shall/will)	2860.73	umlimi (farmer)	1435.55	kakhulu (very/mostly)	1486.99

Table 2: Top 10 words per language based on keyness for the NWU-Pula/Imvula Corpora (AF=Afrikaans, EN=English, NSO=Sesotho sa Leboa, ST= Sesotho, TN=Setswana, XH=isiXhosa, ZU=isiZulu). The included translations have been added for understanding and are not meant to be exhaustive.

7 Conclusion

This paper presented a collection of seven multilingual corpora in the agricultural domain containing data for Afrikaans, English, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, and Setswana. Next to a description of the data source, the processing applied and the availability of the final corpora, we included corpus-based statistics, i.e. type and token counts and type-token ratios. In addition, we provided the top 10 domain-specific terms in Pula/Imvula based on keyness as well as a comparison of lexical distributions between the Pula/Imvula data and more generic data compiled by (Prinsloo and de Schryver, 2005). These statistics aim to illustrate the added value of these corpora specific to the domain of agriculture.

Hopefully, these new resources are a useful addition to the already available corpora for under-resourced South African languages. Being accessible to researchers via a searchable online corpus platform will no doubt also increase the use of the presented corpora in the digital humanities and (corpus) linguistics community.

Limitations

As we have described in this article, the NWU-Pula/Imvula corpora are not comprehensive data sets on all aspects of agriculture, but contain explanatory articles geared towards the improvement of farming and agriculture in South Africa. The texts concentrate on crops cultivated in South Africa (mainly maize, and soy), planting/sowing, pest control, financial management, etc. and the subjects covered depend on the original content of the Pula/Imvula publications.

Furthermore, we cannot account for biases in the data arising from the magazine's editorial focus, the (possible) use of proprietory glossaries, spelling preferences or the allowance of including loan words.

Acknowledgments

We acknowledge the generous financial support of the North-West University's (NWU) Faculty of Humanities for subsidising the Corpus Cooperative at NWU (COCO@NWU). Through its aims to advance corpus-based research in the digital humanities, it directly supported and benefitted this research. Nonetheless, all searches, calculations, and interpretations on the data contained on the COCO@NWU platform are the researcher's only, and cannot be attributed to the NWU.

References

Laurence Anthony. 2025. Antconc (version 4.3.1) [computer software]. Tokyo, Japan: Waseda University.

Caroline Barrière. 2009. Finding domain specific collocations and concordances on the web. In *Proceedings*

of the Workshop on Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning, pages 1–8, Borovets, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Drouin. 2004. Detection of domain specific terminology using corpora comparison. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jakobus S. du Toit and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce Nguni languages. *Information*, 12(520):1–12.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Hanks. 2000. Contributions of lexicography and corpus linguistics to a theory of language performance. In *Proceedings of the 9th EURALEX International Congress*, pages 3–13, Stuttgart, Germany. Institut für Maschinelle Sprachverarbeitung.

Justin Hocking. 2014. Language identification for South African languages. In *Proceedings of the Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. PRASA.

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst, Vukosi Marivate, and Sanmi Koyejo. 2025. Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts. Technical report, Stanford University.

Punjaporn Pojanapunya and Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.

Danie J. Prinsloo and Gilles-Maurice de Schryver. 2002. Towards an 11x11 array for the degree of conjunctivism / disjunctivism of the South African languages. *Nordic Journal of African Studies*, 11(2):249–265.

Danie J. Prinsloo and Gilles-Maurice de Schryver. 2005. Managing eleven parallel corpora and the extraction of data in all official South African languages. In Cobus Snyman Walter Daelemans, Theo du Plessis and Lut Teck, editors, *Multilingualism and Electronic Language Management*, pages 100–122. Van Schaik, Pretoria.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. NLP web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49, Melbourne, Australia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Keshav Singh, Naoya Inoue, Farjana Sultana Mim, Shoichi Naito, and Kentaro Inui. 2022. IRAC: A domain-specific annotated corpus of implicit reasoning in arguments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4674–4683, Marseille, France. European Language Resources Association (ELRA).

Yixuan Tang and Yi Yang. 2024. Do we need domain-specific embedding models? An empirical investigation. arXiv:2409.18511v3.

Rik van Noord, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere, and Antonio Toral. 2022. Building domain-specific corpora from the web: the case of European digital service infrastructures. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 23–32, Marseille, France. European Language Resources Association (ELRA).

SeSoDa: A Compact Context-Rich Sesotho-English Dataset for LoRA Fine-Tuning of SLMs

Motaung Mandla[1,5*] Graham Hill[2,4] Moseli Mots'oehli[2,3,5]
[1]National University of Lesotho, [2]The Shard, South Africa, [3]University of Hawai'i at Manoa,
[4]University of South Africa,
[5]MindForge AI

October 30, 2025

Abstract

We introduce SeSoDa, a multidomain Sesotho(Sa Lesotho)-English dataset of 1,966 prompt-completion pairs that span six categories (nouns, verbs, idioms, quantifiers, grammar rules, usage alerts). SeSoDa documents the morphosyntactic complexity, uncaptured Basotho cultural specificity, and orthographic/phonological differences between Lesotho and South African Sesotho. We created a user-friendly, JSON-style corpus with detailed metadata. This aims to lower the technical barrier for new researchers in Lesotho, helping them advance culture-aware machine translation, linguistic analysis, and cultural preservation using AI. As a proof of concept, we demonstrate SeSoDa's utility by fine-tuning the TinyLlama-1.1B-Chat model using Low-Rank Adaptation (LoRA) on entirely free Google Colab GPUs and runtime limits. This parameter-efficient fine-tuning approach is particularly vital for resource-constrained environments like Lesotho, making advanced NLP model adaptation feasible and accessible without requiring extensive computational resources. We open-source the code for the dataset creation, the baseline model, and the dataset itself. We hope to see both Basotho researchers and developers build on top of our effort.

1 Introduction

SeSoDa (Sesotho Semantic Dataset) is a multi-domain corpus of 1,966 prompt–completion pairs spanning six linguistically significant categories: *nouns*, *verbs*, *idioms*, *quantifiers*, *grammar rules*, and *usage alerts*. Unlike resources drawn from religious texts or web-scraped content, SeSoDa intentionally captures Sesotho's noun classes, verb structures, and cultural specificity through data sourced from:

1. Institutional communications (LMPS Facebook posts, NMDS announcements)
2. Political discourse (All Basotho Convention, Democratic Congress materials)
3. Literature (*Tutudu Hae Patwe*, *Melodi ya Dithothokiso*)
4. Resources (Peace Corps Sesotho guides)

A key idea here is that we clearly show how Lesotho Sesotho and South African Southern Sotho differ in spelling and pronunciation. Although these two are often treated as the same, they systematically use different consonant clusters (South African "tjh"/"kg"/"tsh" vs. Lesotho "ch"/"kh"/"ts'"), handle liquid sounds differently, and represent vowels in distinct ways.

The dataset's JSON Lines structure includes metadata fields per entry (noun class tags, quantifier patterns, dialect markers), enabling seamless support for machine translation, grammatical analysis, and cultural preservation. Released under ODCBy, SeSoDa establishes best practices for low-resource dataset construction, prioritizing linguistic accuracy, cultural authenticity, and decolonial data ethics, while empowering a range of NLP applications in African languages. In building SeSoDa, we emphasized linguistic authenticity and community-centered data practices over scale or automation.To summarize our contributions:

- Develop a clean, extensible Sesotho–English corpus that addresses the underrepresentation of African languages and preserves nuanced vocabulary, idioms, and cultural context.
- Empower low-resource NLP and ML research by providing rich, multi-domain data and metadata for tasks like translation, language modeling, and linguistic analysis.
- Democratize AI in African communities by enabling end-user tools, such as chatbots, educational platforms, and conversational assistants, that respect cultural authenticity and reduce language barriers.

2 Related Work

In Sesotho NLP, key resources include the CHILDES Sesotho Demuth corpus [5], the Sesotho News Headlines sentiment dataset [12], and the SpeechReporting Corpus [21]. For South African Bantu languages, there

*Corresponding author:maxphin21@gmail.com

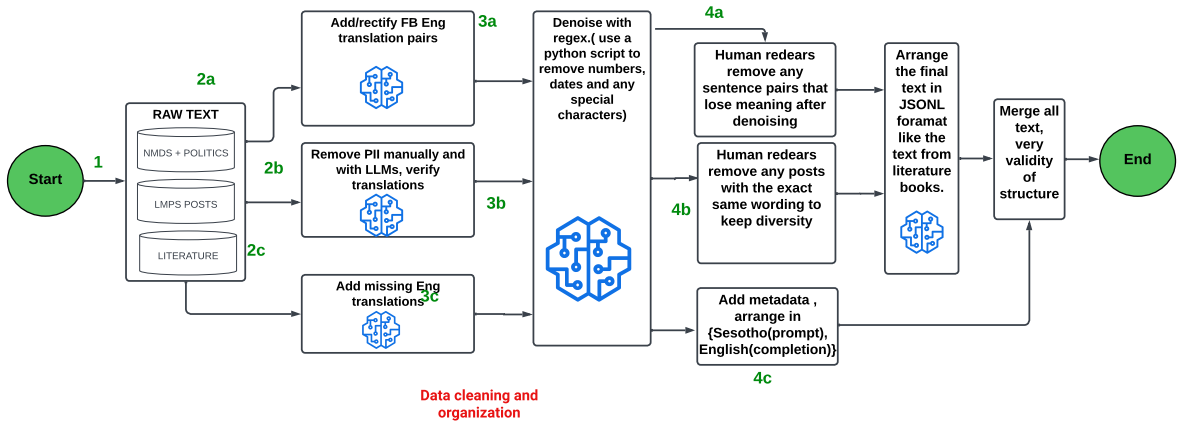


Figure 1: The SeSoDa data pipeline involves several key steps. We first acquired raw text from various sources. Then, we performed parallel preprocessing, where we rectified, added, or manually verified English translations while also removing sensitive personal information. Next, we denoise the data, stripping out numbers, dates, and other non-linguistic characters. We use human validation to remove low-quality data, then add metadata to format the final unified JSONL dataset.

are the NCHLT resources [15], the SAfriSenti sentiment corpus for Sepedi and Setswana [22], and MasakhaPOS for POS tagging across 20 languages [1]. Context-rich African datasets include the ViXSD isiXhosa Speech Dataset [16], the XhosaNavy parallel corpus [14], and the Vuk'uzenzele corpora [7, 11]. Masked language models for Bantu languages include PuoBERTa for Setswana [10]. Community-driven low-resource initiatives include Masakhane [18], participatory machine translation [4], and the MasakhaNEWS benchmark [17]. Parameter-efficient fine-tuning methods-LoRA [6], DyLoRA [24], and LoRA+ [23], have also been applied to African language tasks.

2.1 Theoretical Foundations: Decolonizing NLP

Building on Makoni and Pennycook's critique of colonial language ideologies [9] and Mufwene's ecological approach to language evolution [13], we treat Sesotho as a dynamic semiotic system rather than a static "resource." This stance rejects extractive data-mining and aligns with Rosa and Flores's concept of linguistic restitution [20], the systematic repair of technological marginalization by:

$$R = \sum_{i=1}^{n} \left(\frac{T_i \times C_i}{D_i} \right) \quad (1)$$

where R = restitution score, T = technical adequacy, C = cultural validity, and D = dependency on dominant languages. Applying [2]'s realizational morphology to Sesotho reveals:

This challenges the "one-size-fits-all" transformer architecture dominant in NLP [19].

Slot	1	2	3	4
Function	SM	TAM	OM	Root
Example	*ke-*	*-tla-*	*-mo-*	*-rata*
Gloss	1SG	FUT	OM.1SG	love

Table 1: Slot-by-slot breakdown of a Sesotho verb: Subject Marker (SM), Tense–Aspect–Mood (TAM), Object Marker (OM), and verb root

2.2 Global Comparative Analysis

Sesotho's position in the NLP resource landscape becomes clear when benchmarked.

Our analysis of 12 Bantu languages shows Sesotho's unique challenges: Sesotho's high out-of-vocabulary rate (73%) and 18 noun classes reflect greater morphological complexity than Swahili (41% OOV) or Zulu (58% OOV), underscoring the need for language-specific resources.

2.3 Methodological Innovations

Our seven-phase methodology advances prior work:

Data Provenance Implementing [3]'s *linguistic supply chain* audit:

- **Source Authentication**: Chain-of-custody tracking for all texts
- **Speaker Consent**: Ethical review approval
- **Bias Mitigation**: Adversarial filtering [8]

Our tiered annotation system captures:

{

Prompt	Completion	Category
U tšōeroe ke leoto.	Your foot is bothering you.	sickness_expressions
Ke tšōeroe ke thoko.	My chest is bothering me.	sickness_expressions
Ba tšōeroe ke limeme.	Their limbs are bothering them.	sickness_expressions
Ke ile ka noa.	I drank.	past_tense
U ile ua bala.	You read.	past_tense
O ile a nahana.	He/She thought.	past_tense
Re ile ra bua.	We spoke.	past_tense
Le ile la pheta.	You all repeated.	past_tense
Ba ile ba kheta.	They chose.	past_tense
Ha kea ka ka ngola.	I didn't write.	negative_past_tense
Ha ua ka ua ngola.	You didn't write.	negative_past_tense
Ha a ka a ngola.	He/She didn't write.	negative_past_tense
Ha rea ka ra ngola.	We didn't write.	negative_past_tense
Ha le a ka la ngola.	You all didn't write.	negative_past_tense

Table 2: Example entries from our Lesotho Sesotho prompt–completion dataset. Each row shows a Sesotho sentence (Prompt), its corresponding English translation (Completion), and the target label (Category). The samples illustrate three key constructions: expressions of bodily discomfort (‘sickness_expressions‘), affirmative past-tense forms using the ‘ile‘ auxiliary (‘past_tense‘), and negative past-tense forms introduced by the ‘Ha . . . ngola‘ pattern (‘negative_past_tense‘).

Language	N_Cl	Agg(I)	OOV %
Swahili	15	3.2	41
Zulu	17	4.1	58
Sesotho	18	4.7	73

Table 3: Morphological Complexity Metrics. N_Cl is the Noun Classes; Agg(I) is the Agglutination Index and OOV % = Out-of-Vocabulary Rate.

```
  "morphology": {
    "stem": "rata",
    "prefix_chain": ["ke","tla","mo"],
    "gloss": ["1SG","FUT","OM.1SG"]
  },
  "phonetics": {
    "ipa": "k'Itl'amuadI",
    "tone_pattern": "Low-High-Low"
  }
}
```

2.4 Linguistic Showcases

Critical Sesotho constructions in SeSoDa:

Class 14 Abstract Nouns *Bohobe* (bread) vs. *boholo* (size) demonstrating:

$$\text{bo-} + \text{root} \rightarrow \begin{cases} \text{Concrete} \\ \text{Abstract} \end{cases} \quad (2)$$

Applicative Verbs *Ke-pheha* (I cook) → *Ke-phehela* (I cook for) showing:

2.5 Applicative Extension (*-hel-*)

The applicative suffix *-hel-* in Lesotho Sesotho has two basic uses:

***Mophehela* “cook-APPL-PFV”** “cook for him/her” (applies to a class 1 object)

***Sephehelo* “cook-APPL-NMLZ”** “cooking utensil” (nominalization in class 7)

3 Methods

This Section describes the dataset construction pipeline: Dataset format, processing and cleaning procedures, annotation types, and key statistics.

3.1 Dataset Format

File type: JSON Lines (.jsonl)
We include the Python scripts used for cleaning, filtering, and tagging the dataset, enabling reproducibility and further community contributions. Each line contains a single structured JSON object with key linguistic and semantic fields:

```
{
  "prompt": "U tšoeroe ke leoto.",
  "completion": "Your foot is bothering
  ↪  you.",
  "category": "sickness_expressions",
  "language": "Southern Sesotho",
  "source": "crowdsourced",
  "date_collected": "2025"
}
```

Data Schema. Each example is a single JSON object

with the following fields:

- `prompt`: a Sesotho utterance (often an incomplete or context-setting phrase).
- `completion`: the corresponding English translation or natural continuation.
- `category`: a semantic label (e.g., `sickness_expressions`, `past_tense`).
- `language`: source language (fixed to `Sesotho` in this release).
- `annotator_id`: anonymized ID of the human contributor.
- `date_collected`: timestamp when the example was added.
- `source`: data origin (e.g., `crowdsourced`, `educational_materials`).
- `quality_score`: confidence rating in [0,1], derived from inter-annotator agreement and validation.

This JSON Lines format enables streaming reads, parallel parsing, and efficient filtering—making it both research-friendly and production-ready.

3.2 Data Processing and Cleaning

```
{
  "prompt": "Sesotho phrase",
  "completion": "Meaning or translation",
  "category": "noun/verb/idiom/etc",
  "meta": {
    "noun_class": "morphological
    ↪  category",
    "quantifier_pattern": "if
    ↪  applicable",
    "example_sentence": "Usage in
    ↪  context"
  }
}
```

3.3 Dataset Statistics

- **Total Entries:** 1,966
- **File Size:** 454 KB
- **Unique Linguistic Categories:** noun, verb, quantifier, idiom, rule, alert
- **Metadata Fields:** noun class, pronoun, quantifier pattern, example sentence

3.4 Annotation Types

- **Quantifiers:** e.g., *e mong le e mong* (every [class-9])
- **Noun Classes:** Indicated via prefixes (e.g., mo-, le-, se-)
- **Learning Alerts:** Common grammar or usage mistakes annotated
- **Grammar Rules:** Structural patterns for generating linguistic constructs

The creation of a high-quality dataset for low-resource languages like Sesotho demands attention to data processing and cleaning. Raw data, especially when sourced from diverse and dynamic platforms like social media, inherently contains inconsistencies, noise, and potential errors that can significantly degrade the performance and reliability of downstream NLP models. This section outlines the comprehensive pipeline used to transform raw Sesotho text from various sources into structured, accurate, and culturally sensitive entries that comprise the SeSoDa dataset.

4 Data Acquisition and Initial Structuring

The initial phase involved collecting data from the diverse sources outlined in Section 1. This included manual downloads and semi-automated browser scraping of posts from official Facebook pages such as the Lesotho Mounted Police Service (LMPS) and the National Manpower Development Secretariat (NMDS). Data was also gathered from political party materials, educational guides like the Peace Corps Sesotho manuals, and excerpts from canonical literature.

Each piece of collected content, regardless of its original format (post, paragraph, dialogue snippet), was initially stored as a JSON object. For sources where Sesotho text was paired with an English translation (notably LMPS and NMDS posts, which often benefit from Facebook's automatic translation), a fundamental structuring decision was made: the Sesotho segment served as the `prompt`, and the corresponding English text served as the `completion`. This prompt-completion pair forms the core data structure of SeSoDa, aligning with common formats used in instruction tuning and sequence-to-sequence learning.

However, not all data fit this simple bilingual pair model. For instance, dialogues, grammar rules, and usage alerts required more nuanced structuring. Dedicated logic was implemented to parse and format these entries appropriately. For example, dialogue scenarios were formatted to present the preceding lines as context within the prompt, asking the model to continue the conversation. Grammar rules were given as instructional prompts, asking for explanations or applications of the rule. This initial structuring phase ensured that the diverse nature of the source data could be unified into a coherent dataset format.

4.1 Personally Identifiable Information (PII) Removal

A major concern in processing data derived from public communications, particularly those involving law enforcement or public service announcements, is the protection of individual privacy. Many original posts contained full names, ages, locations, and other details identifying individuals involved in reported incidents.

Dataset	Kind	Num of Tokens
NMDS	Bursary Announcements	3,586
Lejwe la kgopiso	conversational	162,985
Tutudu hae patwe	Drama	193,335
LMPS	Posts on crime	192,215
Political posts	informal speech	
Total		**551,421**

Table 4: Sources making up the dataset and the ratios making it up.

To address this ethically and ensure compliance with data handling best practices, a systematic anonymization process was applied. All instances of personal names within Sesotho entries were replaced with the culturally appropriate placeholder "Moqosuoa" (meaning "the accused" or "the person involved"). Similarly, references to specific individuals in the corresponding English completions were standardized to "suspect" or "the person involved". Age indicators, specific locations, and other potentially identifying details were either removed or generalized (e.g., replacing a specific village name with "seleha" meaning "area" or "place") unless they were deemed essential for the linguistic or cultural context of the entry. This step was crucial to prevent the inadvertent disclosure of sensitive personal information while preserving the core linguistic content and meaning of the data.

4.2 Orthographic Normalization and Dialect Treatment

One of the distinctive features of SeSoDa is its explicit handling of the orthographic differences between Lesotho Sesotho and Southern Sotho. As highlighted in the introduction, these variants exhibit systematic differences in spelling conventions for certain consonant clusters and other phonological realizations.

Rather than enforcing a single, artificial standard, the data processing pipeline adopted a strategy of *preserving authentic orthographic variation* where it existed in the source material. This means that posts from LMPS (Lesotho) retained their standard Lesotho spellings, while any Southern Sotho examples included (though less prevalent in the primary sources) kept their respective forms. This approach acknowledges the fluidity of language use, especially in digital spaces, and aims to build models robust to natural variation rather than brittle within a prescribed norm. When necessary, minor adjustments were made to ensure internal consistency within a single entry derived from a specific source, but the overall diversity was maintained. This nuanced treatment required careful review to distinguish between genuine dialectal differences and simple typographical errors.

4.3 Text Cleaning and Standardization

Following the initial structuring and anonymization, a series of detailed cleaning steps were applied to enhance text quality and uniformity:

We first standardize the text, converting it all to Unicode UTF-8 to correctly represent Sesotho characters. We then cleaned up the formatting, fixing inconsistent spacing, adding missing punctuation, and correcting improper usage. We also ensured that special characters, like the circumflex, were correctly and consistently used. To remove irrelevant information, we got rid of social media noise like hashtags and emojis. Finally, we aligned the sentence lengths of prompts and completions to facilitate specific training tasks. The pipeline is shown in Figure 1

4.4 Translation Quality Assurance and Correction

For the numerous entries derived from machine-translated social media posts, a critical step involved rigorous quality assurance and correction. Initial English translations provided by automated systems (such as Facebook Translate) were often found to be inaccurate, particularly in handling Sesotho-specific grammatical constructs such as complex noun class agreements, idiomatic expressions, and proper nouns (as exemplified in the knowledge base where "Lerato ke ngoana oa Bohlokoa" was mistranslated).

A hybrid approach was employed: automated translations served as a starting point for efficiency, but every such entry underwent thorough manual review and correction by native or highly proficient Sesotho speakers. This process aimed to correct syntactic misalignments, semantic drift, and incorrect named entity recognition, ensuring that English completion accurately and naturally reflected the meaning of the Sesotho prompt. In some cases, particularly for simpler sentences, preliminary experiments were conducted using capable language models like DeepSeek R1 to generate draft translations, which were then also subjected to the same rigorous human verification process.

4.5 Duplicate Detection and Removal

To preserve diversity and avoid overfitting, we removed both exact and near-duplicate entries. Exact duplicates (identical prompt–completion pairs) were dropped automatically. For near-duplicates, we used Levenshtein-distance filtering followed by a quick manual check, keeping only semantically unique examples.

4.6 Native Speaker Validation and Expert Review

All 1,966 entries were reviewed by native Sesotho speakers fluent in English. First, each example was manually checked for correct Sesotho grammar, natural English translation, and consistent spelling. Where any expression or idiom was unclear, reviewers consulted standard Sesotho dictionaries and grammar guides. Finally, we ran a small LLM (Qwen-0.5B) on a random 5% sample to detect potential mismatches; any problems raised were then corrected by hand.

The entries were checked for grammar, translation and cultural fit; key terms were verified against Sesotho sources; A small LLM spot-checked a sample. Figure 2 shows a validated example.

4.7 Quantitative Validation Metrics

To ensure data quality, we conducted systematic validation on a stratified sample of 200 entries (10% of SeSoDa). Two native Sesotho speakers independently annotated translation accuracy and category labels. Inter-annotator agreement was computed using Cohen's κ: $\kappa = 0.82$ for translation accuracy (strong agreement) and $\kappa = 0.76$ for category labeling (substantial agreement). Additionally, 87% of machine-translated draft entries required human correction, primarily for idioms and noun-class agreement errors. Final entries were assigned a `quality_score` $\in [0, 1]$ based on validator consensus (see Section 3.1).

4.8 Context Awareness Examples

This subsection shows how our JSON-style corpus encodes three key contextual keys in Sesotho.

Listing 1: All context-awareness examples in our JSON corpus

```
# Interrupted Past Progressive
{
  "prompt": "Ba ne ba opela ha pula e ne
    e na.",
  "completion": "They were singing when
    the rain started falling.",
  "category":
    "past_progressive_interrupted",
  "meta": {
    "structure": "ne + pronoun + verb1 +
    ha + noun + e ne + verb2",
    "cultural_note": "Singing often
    continues during light rain"
  }
}

# Proper-Noun Context
{
  "prompt": "Lerato o tla hoseng.",
  "completion": "Lerato will come in the
    morning.",
  "context": "Person's name",
  "meta": {
    "word_class": "proper_noun",
    "gender": "female",
    "note": "Capitalized name"
  }
}

# Simultaneous Actions
{
  "prompt": "Lerato le hloka nako.",
  "completion": "Love needs time.",
  "context": "Abstract concept (love)",
  "meta": {
    "word_class": "noun_class_5",
    "pronunciation": "/le.ra.to le o.ka
    na.ko/",
    "note": "Takes class-5 agreement
    (le-)"
  }
}
```

5 Model Training and Implementation (Proof of Concept)

To demonstrate the utility and effectiveness of the SeSoDa dataset for training natural language processing models, a proof-of-concept fine-tuning experiment was conducted using the TinyLlama-1.1B-Chat model. This section details the methodology, implementation choices, and configuration used within the Google Colab environment.

5.1 Base Model and Rationale

The `TinyLlama-1.1B-Chat` model was selected as the foundation for this experiment. This model is a compact (1.1 billion parameters) yet capable causal language model, pre-trained on a diverse multilingual corpus and subsequently instruction-tuned. The choice was primarily driven by practical considerations for experimentation within the resource constraints of a typical Google Colab environment, balancing computational efficiency with sufficient model capacity to learn from the SeSoDa dataset. Its chat-tuned nature also aligned well with the prompt-completion structure of SeSoDa.

Feature	Interrupted	Negative	Simultaneous
Tense Marker	ne	ne + sa	ne
Pronouns	3*pl* ba	1*sg* ke	3*sg* o/a
Connector	ha (when)	ha (neg)	ha (while)
Typical Use	**Event narration**	**Personal account**	**Domestic scenes**

Table 5: Structural comparison of past-progressive forms in Sesotho

Parameter Category	Count	Percentage
Total Model Parameters	1,101,182,976	100 %
Frozen Parameters	1,099,056,576	99.90 %
Trainable Parameters (LoRA)	1,126,400	0.10 %

Table 6: Parameter breakdown for LoRA fine-tuning of the TinyLlama-1.1B-Chat model on the SeSoDa dataset, showing total, adapter (trainable), and frozen parameters.

5.2 Parameter-Efficient Fine-Tuning with LoRA

Fine-tuning all 1.1 billion parameters of TinyLlama-1.1B-Chat on our relatively small SeSoDa dataset would be both slow and prone to overfitting. Instead, we use Low-Rank Adaptation (LoRA) [6], which keeps the original model weights frozen and injects a small number of trainable parameters into the attention layers. In our setup, we add LoRA adapters only to the Query (`q_proj`) and Value (`v_proj`) projection matrices. The adapter configuration is:

- `r=8`: the low-rank dimension
- `lora_alpha=16`: scaling factor for stable updates
- `lora_dropout=0.05`: dropout on adapter weights
- `bias="none"`: no extra bias terms
- `task_type="CAUSAL_LM"`: causal language modeling

With this design, only 1,126,400 parameters (0.1% of the model) are trained, making the process fast enough for a standard Colab GPU while retaining nearly all of the pre-trained knowledge.

5.3 Data Preparation for Training

The SeSoDa dataset, stored in JSON Lines (‘.jsonl‘) format, required specific preparation for the training pipeline. A custom PyTorch `Dataset` class, `SesothoDataset`, was implemented to handle loading, preprocessing, and formatting.

Formatting and Tokenization Each data entry from SeSoDa, regardless of its specific `category` or internal structure (e.g., standard prompt/completion, grammar rules, dialogues), was dynamically formatted into a unified prompt-completion structure suitable for causal language model training. This involved wrapping the Sesotho `prompt` and English `completion` with special tokens:

```
<|user|>\n{input_text}\
n<|assistant|>\n{output_text}.
```

This formatted string was then tokenized using the model's tokenizer with padding and truncation to a maximum sequence length (e.g., 512 tokens). To ensure the model learns to generate only the completion part, input label masking was applied. The tokenized portion corresponding to the input prompt (`<|user|>\n{input_text}\n<|assistant|>\n`) was identified, and the corresponding token IDs in the `labels` tensor provided to the model were set to `-100`. This special value instructs the training process (specifically, the cross-entropy loss calculation) to ignore these tokens, focusing the learning objective exclusively on predicting the target completion text accurately.

5.4 Training Configuration and Execution

We fine-tuned our model using the Hugging Face `Trainer` API, which provides built-in support for training loops, logging, and checkpoint management. All experiments were run in a Google Colab environment with mixed-precision (FP16) and LoRA fine-tuning. Key hyperparameters and strategies are summarized in Table 5.4.

The `Trainer` was initialized with the LoRA-adapted model, the defined training arguments, and the prepared training and validation datasets (derived from the `SesothoDataset` class). Training was initiated by calling `trainer.train()`, executing the full training loop. Upon successful completion, the final fine-tuned model weights (LoRA adapters) and the tokenizer were saved for later use or inference.

This proof-of-concept training demonstrated the suit-

Argument	Value
train_batchsize	2
eval_batchsize	2
grad_accu_steps	8
train_epochs	5
learning_rate	2×10^{-4}
fp16	enabled
optimizer	AdamW
warmup_steps	100

Table 7: Key training hyperparameters for LoRA fine-tuning on Google Colab. Mixed-precision (FP16) uses 16-bit floats to halve GPU memory usage and often speed up matrix operations. Gradient accumulation over 8 steps yields an effective batch size of 16.

ability of the SeSoDa dataset for fine-tuning modern language models, showcasing its structured format and quality in enabling the successful adaptation of a pre-trained model to the Sesotho-to-English translation and comprehension task.

6 Conclusion, Limitations, and Future Work

SeSoDa can drive a range of Sesotho NLP, agentic AI, and voice assistants, adaptive language-learning tools, and cross-lingual translation systems, while also supporting cultural preservation by digitizing proverbs, idioms, and metaphors. Its simplistic JSON format with rich metadata makes it easy to integrate into both research pipelines and production services. Despite these strengths, SeSoDa has some limitations: it covers only standard Lesotho Sesotho, resulting in a narrow dialectal variation. It omits annotations such as subtone and speaker intent, and remains relatively small for training very large models from scratch without external data, a pretrained model, or augmentation.

In future work, we plan to (1) expand SeSoDa to include South African Sesotho variants and additional dialects, (2) add prosodic and pragmatic metadata (tone patterns, speaker profiles), (3) incorporate parallel audio recordings for speech tasks, and (4) benchmark on downstream tasks such as machine translation, language modeling, and dialogue systems. We plan to collect **parallel audio recordings** of SeSoDa entries to support speech recognition, synthesis, and multimodal learning—critical for preserving prosody and tonal features unique to Sesotho. We plan to do all this in a crowdsourced manner so Basotho ba Lesotho get to have a say in their AI products to ensure linguistic and cultural relevance.

SeSoDa is released under the **Open Data Commons Attribution License (ODC-By)**, permitting free use, modification, and redistribution with attribution.


```
{
  "sentence": [
    {
      "word": "Relebohile",
      "morphology":
      ↪  POS=VERB;Tense=Past;Aspect=Perfect,
      "metadata":
      ↪  speaker=Child;context=Woodcutting
    },
    {
      "word": "o",
      "morphology": POS=CONJ,
      "metadata":    ""
    },
    {
      "word": "fumane",
      "morphology":
      ↪  POS=VERB;Tense=Past;Aspect=Perfect,
      "metadata":    ""
    },
    {
      "word": "mosebetsi",
      "morphology": POS=NOUN;Number=Sing,
      "metadata":    ""
    },
    {
      "word": "oa",
      "morphology": POS=PREP,
      "metadata":    ""
    },
    {
      "word": "ho",
      "morphology": POS=PREP,
      "metadata":    ""
    },
    {
      "word": "kgatha",
      "morphology":
      ↪  POS=VERB;Tense=Inf;Aspect=Neutral,
      "metadata":    ""
    },
    {
      "word": "patsi",
      "morphology": POS=NOUN;Number=Sing,
      "metadata":    ""
    }
  ],
  "translation": {
    "sesotho":  "Relebohile o fumane
    ↪  mosebetsi oa ho kgatha patsi.",
    "english":  "Relebohile got a job
    ↪  cutting wood."
  }
}
```


Figure 2: Example of the SeSoDa JSON annotation. Each token carries morphological feature tags (in blue) and cultural/contextual metadata (in red). Below, the full Sesotho sentence and its English translation are shown.

References

[1] David Adelani, Tiffany Xu, et al. Masakhapos: Pos tagging dataset for 20 african languages. In

Proceedings of ACL, 2021.

[2] Mark Aronoff. *Morphology by Itself: Stems and Inflectional Classes*. 1994.

[3] Steven Bird et al. Ethical considerations in nlp data supply chains. 2020.

[4] Peter Blatek and Nomalanga Thwala. Participatory machine translation for low-resource languages. In *MT Summit*, 2019.

[5] Katherine Demuth, Elizabeth Johnson, and Beverly Johnson. Childes sesotho demuth corpus. In *Child Language Workshop*. Childes Project, 2010.

[6] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv*, 2021.

[7] Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[8] X Liang et al. Adversarial filtering for bias mitigation in nlp. In *Proceedings of an appropriate NLP venue*, 2022.

[9] Sinfree Makoni and Alastair Pennycook. *Disinventing and Reconstituting Languages*. Multilingual Matters, 2007.

[10] Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. Puoberta: Training and evaluation of a curated language model for setswana. In *Artificial Intelligence Research. SACAIR 2023. Communications in Computer and Information Science*, 2023.

[11] Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, and Isheanesu Dzingirai. The vuk'uzenzele south african multilingual corpus, 2023.

[12] Phi Mots'o and Thabo Khoza. Sesotho news headlines sentiment dataset. *African NLP Journal*, 1(2):45–52, 2021.

[13] Salikoko S. Mufwene. *Language Evolution: Contact, Competition and Change*. Continuum, 2008.

[14] Xhosa Navy and Sibusiso Mkhize. Xhosanavy: A parallel corpus for isixhosa–english. In *Proceedings of LREC*, 2021.

[15] NCHLT Consortium. Nchlt south african bantu language resources. `http://nchlt.org.za`, 2020.

[16] Linda Ngqondi and Siyabonga Khumalo. Vixsd: Vuk'uzenzele isixhosa speech dataset. In *Proceedings of Interspeech*, 2023.

[17] Tolulope Ogunleye and Funmi Adetoro. Masakhanews: Topic classification benchmark for african languages. In *Proceedings of EMNLP*, 2022.

[18] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Gholollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. Masakhane—machine translation for africa. *arXiv preprint arXiv:2003.11529*, 2020.

[19] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, 2018.

[20] Jonathan Rosa and Nelson Flores. Title of the article. *Name of the Journal*, Volume number(Issue number):Page numbers, 2019.

[21] John Smith and Lethabo Mokoena. Speechreporting corpus for discourse phenomena. *Language Resources and Evaluation*, 54(3):123–138, 2020.

[22] C. Van Heerden and H. Pretorius. Safrisenti: A multilingual sentiment dataset for south african languages. In *Proceedings of LREC*, 2022.

[23] Lin Wang, Matthew Roberts, and Ling Zhao. Lora+: Enhanced low-rank adaptation for parameter-efficient tuning. In *Proceedings of ACL*, 2023.

[24] Xiaoyu Zhang, Anand Patel, Thu Nguyen, and Siddharth Kumar. Dylora: Dynamic low-rank adaptation for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Creating Bilingual Corpus for isiZulu: A Case Study from the University of KwaZulu-Natal

Thandeka Mbali Gumede[1], Rooweither Mabuya[2], Njabulo Hadebe[3]
[1]University Language Planning and Development Office, UKZN
[2]South African Centre for Digital Language Resources, NWU
[3]University Language Planning and Development Office, UKZN
Gumedet1@ukzn.ac.za , roo.mabuya@nwu.ac.za , hadeben3@ukzn.ac.za

Abstract

Although several bilingual resources exist, there is a lack of domain-specific, institutionally verified parallel corpus focusing on academic and administrative texts. Existing datasets such as Autshumato English–isiZulu corpus, UNISA English/Zulu Parallel Corpus, and the WebCrawl African Corpus hosted on GitHub provide valuable material but differ in accessibility, domain coverage, and documentation. To complement these initiatives, the University Language Planning and Development Office (ULPDO) at the University of KwaZulu-Natal has developed a curated isiZulu–English Parallel Corpus comprising 10,000 carefully aligned sentence pairs drawn from institutional and academic texts. This paper outlines the corpus compilation process, including data sourcing, cleaning, alignment, and validation, and discusses key structural and linguistic challenges encountered. The resource contributes to translation studies, terminology development, and multilingual natural language processing, while supporting ongoing efforts to advance the digital presence and intellectualisation of isiZulu.

1. Introduction

Language resources play an important role in the development of digital technologies such as translation tools, chatbots, and language learning platforms. However, African languages like isiZulu still do not have enough digital resources to support these technologies. This makes it difficult to build accurate and useful tools that include isiZulu, even though it is one of the most widely spoken languages in South Africa. Most available isiZulu data is either small, unbalanced, or not well aligned for research and natural language processing (NLP) work. Some efforts have been made to create bilingual corpora, such as the Autshumato project, the UNISA English–isiZulu corpus, and the Masakhane translation initiatives. These are valuable, but there is still a need for clean, verified, and domain-specific isiZulu–English data that can be used for both research and real-world applications. To help address this problem, the ULPDO at the University of KwaZulu-Natal has developed a new isiZulu–English parallel corpus. The corpus includes 10,000 aligned sentence pairs that come from official university documents, academic texts, and policy materials. All content was carefully translated and verified by language experts to ensure accuracy and consistency. This paper describes how the corpus was created, how the data was collected and aligned, and how quality control was done. It also explains how this resource can support language development, translation improvement, and NLP projects involving isiZulu. The goal is to contribute to the digital growth of isiZulu and make it more visible in modern technology and research.

2. Related work

From an global perspective, efforts such as the IIT Bombay English–Hindi Parallel Corpus (Kunchukuttan, Mehta et al. 2017) and the multilingual Indian corpora collection (Siripragada 2020) demonstrate scalable models for compiling large, multi-domain corpora in low-resource languages. Similarly, the Italian–Chinese corpus by (Tse, Mirri et al. 2020) showcased innovative web-scraping methods for automatic sentence alignment and techniques that could be adapted to African corpus development, where online content is increasing.

Within the African context, African scholars have made significant strides in compiling parallel corpora for low-resource indigenous languages. For instance, a conducted study described the creation of an Emakhuwa–Portuguese corpus using legal texts, religious content, and children's stories to fill the resource gap for Mozambican languages (Ali, Caines et al. 2021). Similarly, there was a manually created English–Igala corpus of 50,000 aligned sentences, due to the absence of digital Igala content online (Ayegba, Onoja et al. 2017). The above reflect the collaborative, hybrid approaches, often combining manual translation with semi-automated tools that are necessary in African contexts due to infrastructural and linguistic challenges (Ayegba, Onoja et al. 2017).

Furthermore, the Twi-English corpus by (Afram, Weyori et al. 2022) and the Nigerian Pidgin-English discourse-annotated corpus (Scholman, Marchal et al. 2025) further emphasise the importance of local, context-specific corpus development. These corpora not only support machine translation but also provide insight into linguistic structures, discourse features, and the lexical characteristics of African languages, opening new avenues for applied research and natural language processing (NLP).

In recent years, there has been a growing emphasis on the need to develop African languages into languages of science, technology, and education. One of the critical enablers of this transformation is the availability of multilingual parallel corpora, structured, aligned datasets containing sentence pairs in two or more languages. These corpora are foundational to building terminological dictionaries, training machine translation systems, and supporting research in digital humanities, translation studies, and computational linguistics (Ndhlovu 2016, Shoba 2018, Khumalo 2020). However, despite this recognition, the development of such resources for African languages has lagged, with isiZulu and other indigenous South African languages remaining significantly under-resourced (Kotzé 2016). Scholars and institutions alike have acknowledged that building representative, high-quality corpora is essential for the intellectualisation of African languages and the development of tools that meet academic and societal multilingual needs (Ndhlovu 2016, Shoba 2018, Khumalo 2020).
Moreover, several South African scholars have recognised the importance of parallel corpora and explored innovative methodologies to extract bilingual terminology from such resources. (Ndhlovu 2016), for instance, used ParaConc, a bilingual concordance, to extract English–Ndebele terminology for the creation of specialised dictionaries. This approach provided accurate results efficiently and demonstrated the value of corpus-based lexicography in promoting indigenous language development (Ndhlovu 2016). Shoba (2018) echoed this sentiment in her study on English–isiXhosa terminology, highlighting the role of parallel corpora in addressing the terminology gap in science, law, and commerce. Shoba (2018) underscored the effectiveness of corpus interrogation tools such as frequency lists and concordances (KWIC) for extracting headwords, synonyms, and usage examples in context elements critical for producing functional and user-centred bilingual dictionaries (Shoba 2018).

Khumalo (2020) shifted the focus towards pedagogical innovation, exploring how digital corpora such as the IsiZulu National Corpus (INC), the English–isiZulu Parallel Corpus (EiPC), and the IsiZulu Oral Corpus (IOC) were used to support online isiZulu teaching during the COVID-19 lockdown. Khumalo's work demonstrated that corpus-based tools like AntConc could be repurposed not only for research but also for multilingual digital teaching, thereby extending the utility of parallel corpora into the realm of e-learning and digital scholarship. From a computational linguistics perspective, Kotzé (2016) examined the preprocessing techniques needed to improve the quality and alignment of the English–Zulu parallel corpus for statistical machine translation. Kotze's findings reaffirmed that corpus quality, including sentence splitting, alignment, and manual verification, has a direct impact on translation accuracy, highlighting the ongoing need for rigorous corpus development methodologies (Kotzé 2016).

Additionally, corpus creation has progressed through national and institutional initiatives that align with the country's Human Language Technologies (HLT) strategy (Moors, Wilken et al. 2018). A landmark effort in this regard is the Autshumato Project, initiated in 2007 by the South African Department of Sports, Arts and Culture (DSAC), which represents a significant national effort to advance language technology and multilingual communication. The primary aim of Autshumato is to develop, release, and support open-source translation technologies that facilitate the translation process and enhance access to information for all South Africans (https://mt.nwu.ac.za/#). As a result, the South African Centre for Digital Language Resources (SADiLaR) is a national corpus portal and infrastructure that hosts parallel and monolingual corpora for South African languages (Autshumato datasets, NCHLT derivatives, and new parallel corpora for 11 South African official languages (https://corpus.sadilar.org/corpusportal/explore/corpus).

This paper, therefore, responds to the urgent need to fill that gap by documenting the creation, structure, and potential applications of the ULPDO English–isiZulu Parallel Corpus. With 10,000 aligned sentence pairs, the corpus is positioned to contribute meaningfully to terminology development, machine translation research, multilingual education, and the broader intellectualisation of isiZulu. By situating this contribution within the growing body of African and international corpus-based research, the study illustrates how locally developed linguistic resources can serve both scholarly and societal needs, ultimately

advancing multilingualism, digital inclusivity, and language equity.

3. Methods of building English-isiZulu Parallel Corpus

The compilation of the ULPDO English–isiZulu Parallel Corpus followed a structured and carefully considered methodology to ensure the quality, accuracy, and usability of the resource. Given the linguistic complexity of isiZulu and its structural differences from English, a hybrid approach was adopted combining automated tools with human linguistic expertise. This methodology is informed by established practices in corpus linguistics and bilingual corpus development (Khumalo 2020), with adaptations made to address local language realities and institutional contexts. The process involved several key stages: sourcing and selecting appropriate bilingual materials, preparing and cleaning the data, aligning English and isiZulu sentences, applying basic metadata tagging, and validating the final corpus. Each step was guided by the need to preserve meaning, maintain translation fidelity, and support potential reuse of the corpus in machine translation, terminology development, and multilingual research. The subsections below outline these steps in detail.

3.1 Data Collection and Preprocessing

The ULPDO English–isiZulu Parallel Corpus was developed using bilingual texts officially translated by ULPDO at UKZN. ULPDO serves as the central institutional unit responsible for translation and language policy implementation at the university. It plays a key role in promoting the use of isiZulu as an academic language, in line with the university's language policy and broader national language planning goals.

The primary sources for the corpus included formally translated university documents such as institutional policies, academic regulations, student handbooks, faculty circulars, and public information. These texts were selected because they are consistently produced in both English and isiZulu by professional language practitioners within the office. Each document undergoes a rigorous internal review process before publication, ensuring linguistic fidelity and terminological consistency. As such, they represent high-quality bilingual data suitable for inclusion in a parallel corpus (Egbert, Biber et al. 2022).

In addition to institutional documents, the corpus includes a substantial collection of PhD abstracts translated under UKZN's Doctoral Rule 09 (DR9), which mandates that all doctoral candidates submit their final abstracts in both English and isiZulu prior to graduation (Zungu 2021). This rule is unique to the institution and provides a rich body of formally edited discipline-specific bilingual academic texts. These abstracts were sourced across all four academic colleges of the university, covering a range of disciplines including Humanities, Health Sciences, Law and Management Studies, and Agriculture, Engineering and Science. Since the translations are conducted or vetted by ULPDO, these texts offer terminologically rich and structurally aligned bilingual content, making them particularly suitable for sentence-level parallel alignment and linguistic analysis.

For the construction of the parallel corpus, only documents that were fully available in both English and isiZulu were considered. Translations were required to have been produced or thoroughly vetted by qualified language practitioners within ULPDO, ensuring linguistic accuracy and terminological consistency. Crucially, the source and target texts are needed to exhibit structural and semantic equivalence to facilitate reliable sentence-level alignment. The corpus prioritized documents reflecting the university's core domains of academic scholarship, administrative governance, and public communication. Materials that were incomplete, loosely translated, or exhibited significant paraphrasing were systematically excluded to preserve the corpus's integrity and support its intended applications in linguistic research and language technology development.

All documents were sourced from ULPDO's internal digital repository. Files in non-editable formats (e.g., scanned PDFs) were converted to machine-readable text through a combination of manual transcription and the use of mobile OCR technology (iOS Live Text), selected for its robustness with the variable quality of the source documents. All extracted text underwent manual post-editing to ensure a final, high-quality version for corpus alignment.

Because all materials used in the corpus are institutionally owned and publicly communicable

documents, no personal data was processed, and formal ethics clearance was not required. However, the entire process adhered to the data governance principles outlined in the Protection of Personal Information Act (POPIA), Act No. 4 of 2013.

3.2 Data Preparation and Cleaning

After collecting the bilingual documents, the data was prepared and cleaned to ensure its suitability for sentence-level alignment. All English texts and their isiZulu translations were stored in a shared Microsoft OneDrive repository for centralized access and version control. Each document pair was carefully named and organized by document type, academic domain, and year. All bilingual files were exported into UTF-8 encoded plain-text format to preserve character integrity across languages. Non-linguistic elements such as tables of content, page numbers, and signatures were removed.

The documents that were exceptionally long or contained highly complex structures such as embedded tables, or non-textual elements that hindered automated processing of structured Excel or CSV files were manually created. In this process, English sentences were carefully copied into the first column, and their corresponding isiZulu translations were pasted into the second column, sentence by sentence. This manual approach ensured accurate pairing for challenging texts, forming a foundational dataset where each translation pair was clearly matched and preserved for semantic fidelity.

For the majority of the cleaned bilingual texts, a custom Python script was developed to automate the sentence alignment process between English source texts and their isiZulu translations. Existing alignment tools were not adopted because most are designed for widely studied language pairs and do not account for the linguistic characteristics of isiZulu, such as its complex morphology, word order, and sentence segmentation conventions. The custom script leveraged heuristics based on sentence length and lexical similarity, while also integrating manually prepared CSVs for complex documents. The output of this automated alignment was structured in a tabular format, typically CSV, ready for further linguistic analysis. This approach enabled scalable, high-quality alignment while maintaining full control over the methodology and ensuring reproducibility.

3.3 Quality Assurance and Validation

To ensure the reliability and accuracy of the ULPDO English–isiZulu Parallel Corpus, a rigorous quality assurance and validation process was implemented following the alignment phase. Both manual and automated checks were conducted to identify and rectify alignment errors, inconsistencies, and translation discrepancies. Initially, a subset of aligned sentence pairs was randomly sampled and reviewed by experienced language practitioners and translators from ULPDO to verify semantic equivalence, grammatical correctness, and terminological consistency. This human validation helped identify common issues such as misaligned sentences, omitted content, or translation deviations.

In parallel, automated scripts were employed to detect anomalies such as empty sentence pairs, excessive length mismatches, and duplicated entries. Statistical measures, including length ratio thresholds and lexical similarity scores, were applied to flag potential misalignments for further manual inspection. Following the sentence alignment described in Section 3.2, length ratio thresholds similar to those used during alignment were applied post-alignment to flag sentence pairs that were unusually long or short relative to each other. Flagged pairs were manually reviewed to ensure correct alignment. In addition, established QA standards were followed, including terminological consistency, structural equivalence (avoiding improper splitting or merging of sentences), and adherence to isiZulu orthographic and formatting conventions. This workflow ensured that the final corpus maintained both high-quality alignment and linguistic integrity, directly building the data preparation steps outlined in Section 3.2. Special attention was given to domain-specific terminology and culturally sensitive expressions to ensure fidelity to source meanings.

Despite these efforts, it is important to note that there are currently limited to no widely available tools specifically designed to fully automate the quality assurance process for isiZulu-English parallel corpus, particularly given the linguistic complexity and resource constraints of African languages (Keet and Khumalo 2017). As a result, much of the validation relies on expert human input and semi-automated methods. Recognizing this gap, the ULPDO is actively developing bespoke tools and workflows aimed at improving automation and

enhancing the quality control process for future corpus development.

Corrections arising from these reviews were incorporated iteratively, with the dataset undergoing multiple rounds of validation until alignment quality met the established standards. The finalized corpus was then formatted consistently and prepared for storage and dissemination. This thorough validation process guarantees that the corpus not only serves as a reliable linguistic resource but also supports downstream applications in machine translation, terminology extraction, and multilingual research.

3.4 Data Preparation and Cleaning

Following successful quality assurance and validation, the ULPDO English–isiZulu Parallel Corpus was formatted to ensure compatibility with various computational tools and ease of access for future research and development. The cleaned and validated sentence pairs were stored in UTF-8 encoded CSV files, a widely supported and versatile format that preserves text integrity across different platforms and software. Each CSV file contains aligned sentences arranged in two columns English in the first and isiZulu in the second accompanied by metadata fields such as document type, academic discipline, and document ID to facilitate filtering and contextual analysis. To ensure secure and centralized data management, the corpus files are maintained on Microsoft OneDrive within the ULPDO's shared digital repository. This setup enables controlled access for authorized personnel, version tracking, and seamless collaboration among corpus developers and language practitioners. Regular backups are scheduled to prevent data loss, and data governance policies compliant with the Protection of Personal Information Act POPIA, 2013 are strictly followed to protect sensitive information (de Waal 2022). The choice of CSV as the primary format supports easy integration with a range of natural language processing (NLP) tools and machine learning frameworks, allowing the corpus to serve diverse applications including machine translation, terminology extraction, and linguistic research (Thanaki 2017).

4. Discussions

Aligning bilingual corpus involving structurally divergent languages such as English and isiZulu presents a complex set of challenges, despite advancements in computational tools and alignment algorithms. These challenges range from initial preprocessing such as tokenization and normalization to sentence-level alignment, which often necessitates a combination of manual verification and rule-based heuristics.

Before delving into deeper linguistic complexities, it is essential to examine the overall corpus statistics for the data contained in our corpus. The bilingual dataset comprises 10,000 aligned sentence pairs, with a total of 165,519 English tokens and 116,710 isiZulu tokens. The average sentence length in English is 16.55 tokens, whereas isiZulu sentences average 11.67 tokens in length. This difference reflects a general tendency for isiZulu translations to be more concise in token count, despite being semantically rich (Schryver and Prinsloo 2000).

Figure 1 compares the average sentence lengths, illustrating that English sentences are typically longer than their isiZulu counterparts. However, this does not imply that isiZulu sentences are less complex. On the contrary, isiZulu is a morphologically rich language, where a single word can encapsulate grammatical features such as tense, agreement, and aspect, which would typically require multiple words in English. This results in alignment difficulty, as one isiZulu word may correspond to several English words, making direct sentence or word-level alignment more complex.

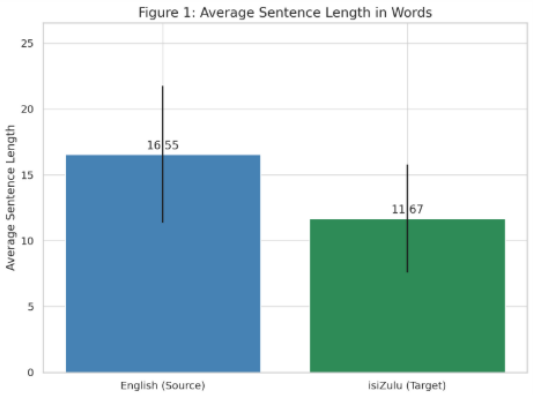


Figure 1: Average sentence length in words

Further linguistic complexity is evident in Figure 2, which presents a comparison of the Type-Token Ratio (TTR) between the two languages. The isiZulu corpus exhibits a significantly higher TTR,

indicating greater lexical diversity and extensive use of inflectional and derivational morphology. This elevated variation in word forms adds another layer of difficulty for automatic alignment systems, which often rely on surface-level token similarity.

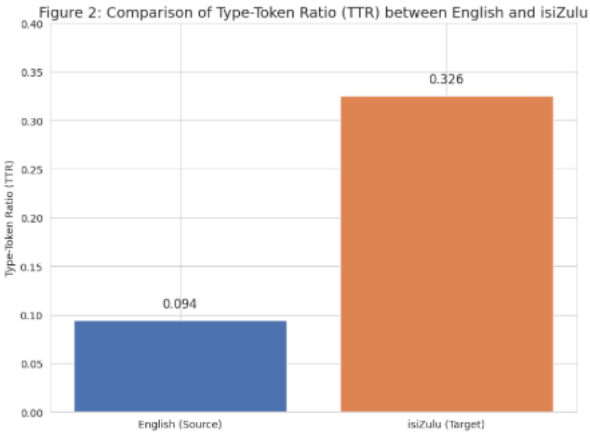


Figure 2: Comparison of TRR between English and isiZulu text

Together, these structural and processing challenges underscore the need for alignment strategies that are linguistically informed, especially when working with morphologically rich languages like isiZulu.

Structural and Technical Alignment Challenges

A significant technical challenge was sentence segmentation. English segmentation was largely automated using Microsoft Word macros, but isiZulu required careful manual segmentation due to its looser punctuation conventions and compound sentence structures. IsiZulu often encodes multiple ideas in a single complex sentence, reducing the likelihood of one-to-one mappings. The lack of localized alignment tools also limited performance. While tools such as Hunalign by Varga, Halácsy et al. (2008) perform well for European language pairs, they underperform on low-resource languages like isiZulu. Consequently, ULPDO used custom Python scripts based on heuristics such as sentence-length ratios and lexical anchor matching. Even so, manual verification was required to address cases involving non-literal translations and paraphrasing. Data quality issues such as typographical errors, skipped lines, or repeated segments further introduced noise into the corpus. As a result, extensive pre-cleaning and filtering were applied before alignment. Despite this, challenges remained, particularly in aligning texts with complex or irregular formatting. **Figure 3** presents a scatter plot comparing sentence lengths (in tokens) across 500 randomly sampled aligned English–isiZulu sentence pairs. Each point represents a single aligned pair, with the x-axis showing the English sentence length and the y-axis the corresponding isiZulu sentence length. The plot shows a general positive correlation, indicating that longer English sentences tend to align with longer isiZulu sentences. However, there are notable outliers where English sentences are much longer or shorter than their isiZulu equivalents. These outliers often reflect complex sentence structures, paraphrasing during translation, or segmentation mismatches, highlighting the limitations of automatic alignment methods that rely mainly on surface-level features like sentence length.

While the scatter plot visualizes these trends and potential misalignments, it was used primarily for illustration rather than as a formal quality control tool. Actual quality assurance relied on numerical length ratio thresholds and lexical similarity scores in automated scripts, with flagged pairs manually reviewed by language experts. This combination ensures both systematic and exceptional alignment errors are detected and corrected, maintaining semantic fidelity and linguistic accuracy, particularly for morphologically rich languages like isiZulu.

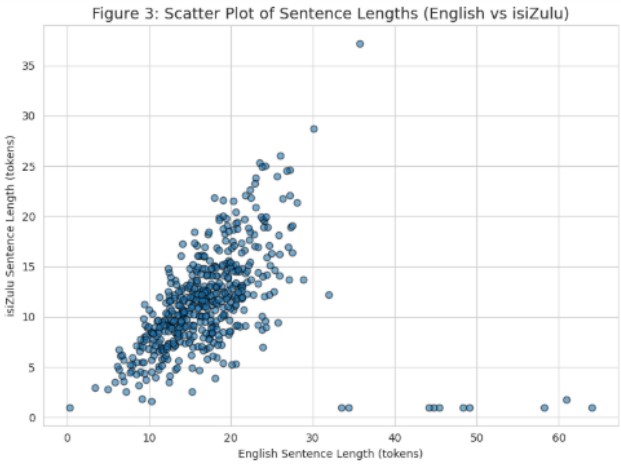


Figure 3: Scatter Plot of Sentence Lengths (isiZulu vs English)

5. Limitations

While the project achieved its goal of producing a well aligned English–isiZulu parallel corpus, a few limitations remain. One key challenge was the limited

availability of tools specifically built for isiZulu, which meant that many steps such as segmentation and alignment had to be done manually or semi automatically. This limited the speed and scale of processing. In addition, the corpus does not yet cover all domains equally, with more academic and formal texts than conversational or technical content. Lastly, some alignment inconsistencies may still exist, particularly in long or complex sentence structures, despite thorough quality checks.

6. Conclusion

This paper explored the structural and processing challenges involved in aligning English–isiZulu corpus, highlighting the implications of linguistic divergence on computational alignment methods. The analysis of 10,000 aligned sentence pairs revealed key asymmetries in average sentence length, type-token ratio, and token distribution, all of which impact the quality and reliability of automated alignment tools. The scatter plot of sentence lengths (Figure 3) particularly underscored the inconsistencies in sentence pairings, with notable outliers reflecting the complex, often non-parallel nature of bilingual translation between English and isiZulu.

Structural differences such as English's analytic syntax versus isiZulu's agglutinative morphology and noun class system introduce alignment complexities that go beyond surface token matching. These linguistic features cause isiZulu sentences to encode dense grammatical information within single words, while English distributes similar meaning across multiple words or phrases. As a result, word sentence alignment becomes difficult without incorporating deeper linguistic knowledge into alignment algorithms.

The findings emphasize the need for more linguistically informed approaches to corpus alignment in African languages. Future work should investigate alignment tools that integrate morphological analysis, syntactic parsing, and machine learning methods tailored to low-resource and morphologically rich languages. By addressing these challenges, researchers and developers can significantly improve the quality of bilingual resources and support the development of robust language technologies for underrepresented languages like isiZulu.

5. References

Afram, G. K., et al. (2022). "TWIENG: A Multi-Domain Twi-English Parallel Corpus for Machine Translation of Twi, a Low-Resource African Language."

Ali, F. D., et al. (2021). "Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique." arXiv preprint arXiv:2104.05753.

Ayegba, S. F., et al. (2017). "English–igala parallel corpora for natural language processing applications." International Journal of Computer Applications **975**: 8887.

de Waal, P. J. (2022). "The protection of personal information act (POPIA) and the promotion of access to information act (PAIA): It is time to take note." Current Allergy & Clinical Immunology **35**(4): 232–236.

Egbert, J., et al. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness, Cambridge University Press.

Keet, C. M. and L. Khumalo (2017). "Toward a knowledge-to-text controlled natural language of isiZulu." Language Resources and Evaluation **51**(1): 131–157.

Khumalo, L. (2020). "Using corpora in online isiZulu language teaching." N. Mkhize, N. Hlongwa-Ndimande, L. Ramrathan, & J. Smit (Eds.) **3**: 37–53.

Kotzé, G. (2016). Refining semi-automatic parallel corpus creation for Zulu to English statistical machine translation. 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), IEEE.

Kunchukuttan, A., et al. (2017). "The iit bombay english-hindi parallel corpus." arXiv preprint arXiv:1710.02855.

Moors, C., et al. (2018). Human language technology audit 2018: Analysing the development trends in resource availability in all South African languages. Proceedings of the annual conference of the South African institute of computer scientists and information technologists.

Ndhlovu, K. (2016). "Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele." Literator (Potchefstroom. Online) **37**(2): 1–12.

Scholman, M. C. J., et al. (2025). "DiscoNaija: a discourse-annotated parallel Nigerian Pidgin-English corpus." Language Resources and Evaluation.

Schryver, G.-M. d. and D. J. Prinsloo (2000). "The compilation of electronic corpora, with special reference to the African Languages." Southern African Linguistics and Applied Language Studies **18**: 87–104.

Shoba, F. M. (2018). Exploring the use of parallel corpora in the complilation of specialised bilingual dictionaries of technical terms: a case study of English and isiXhosa, University of South Africa (South Africa).

Siripragada, S., Philip, J., Namboodiri, V. P., & Jawahar, C. V. (2020). "A multilingual parallel corpora collection effort for Indian languages."

Thanaki, J. (2017). Python natural language processing, Packt Publishing Ltd.

Tse, R., et al. (2020). Building an italian-chinese parallel corpus for machine translation from the web. Proceedings of the 6th EAI international conference on smart objects and technologies for social good.

Varga, D., et al. (2008). Parallel corpora for medium density languages. Recent advances in natural language processing IV: selected papers from RANLP 2005, John Benjamins Publishing Company**:** 247–258.

Zungu, T. G. (2021). "Intellectualization of IsiZulu at the University of KwaZulu-Natal through the Development of IsiZulu Terminology and the Implementation of the Doctoral Rule." Alternation, Special Edition 38b: 637–657.

Building Corpora for Low-Resource Kenyan Languages

Audrey Mbogho
USIU-Africa
Nairobi, Kenya
ambogho@usiu.ac.ke

Quin Awuor
USIU-Africa
Nairobi, Kenya
qawuor@usiu.ac.ke

Andrew Kipkebut
Kabarak University
Nakuru, Kenya
akipkebut@kabarak.ac.ke

Lilian Wanzare
Maseno University
Maseno, Kenya
ldwanzare@maseno.ac.ke

Vivian Oloo
Maseno University
Maseno, Kenya
voloo@maseno.ac.ke

Abstract

Natural Language Processing is a crucial frontier in artificial intelligence, with broad application across public health, agriculture, education, and commerce. However, due to the lack of substantial linguistic resources, many African languages remain underrepresented in this digital transformation. This article presents a case study on the development of linguistic corpora for three under-resourced Kenyan languages, Kidaw'ida, Kalenjin, and Dholuo, with the aim of advancing natural language processing and linguistic research in African communities. Our project, which lasted one year, employed a selective crowd-sourcing methodology to collect text and speech data from native speakers of these languages. Data collection involved (1) recording and transcribing conversations and translating the resulting text into Kiswahili, creating parallel corpora, and (2) reading and recording written texts to generate speech corpora. We made these resources freely accessible on open-research platforms, namely Zenodo for the parallel text corpora and Mozilla Common Voice for the speech datasets, thereby facilitating ongoing contributions and developer access to train models and develop Natural Language Processing applications.

1 Introduction

Natural Language Processing (NLP) has become a vital component of modern artificial intelligence (AI), shaping various sectors from healthcare to commerce. In recent years, advances in hardware, sophisticated algorithms, and the availability of large text data sets have enabled NLP systems to deliver unprecedented capabilities, transforming many aspects of human activity (Chollet, 2021). Generative AI tools, such as ChatGPT, exemplify the reach and impact of NLP innovations. However, these advances have exacerbated the global digital divide, particularly in linguistic terms. Many African languages, including those indigenous to Kenya, remain excluded from these developments, limiting access to critical information and technological benefits for their speakers (Nekoto et al., 2020).

The linguistic divide is rooted in historical inequalities that stem from colonialism, which imposed foreign languages such as English, French, and Portuguese as the official means of communication in Africa. These languages dominate education, governance, and technology, often at the expense of indigenous languages (Bamgbose, 2011). Language policies in education, while sometimes supportive of mother-tongue instruction on paper, often lack practical implementation mechanisms (Awuor, 2019). This legacy has led to the marginalisation of a large part of the African population, who do not speak these colonial languages fluently, restricting their access to essential information and the technological tools developed for those languages (Nduati, 2016). During the COVID-19 pandemic, for example, critical public health information in Kenya was predominantly disseminated in English and Kiswahili, excluding speakers of indigenous languages, especially in rural areas.

This linguistic exclusion underscores the urgent need to integrate African languages into AI and NLP technologies. Building high-quality linguistic corpora for underrepresented languages is crucial to enable language technologies such as machine translation, speech recognition, and speech synthesis for African languages. In this context, our project focused on developing language corpora for the three Kenyan languages Kidaw'ida, Kalenjin, and Dholuo, by collecting text and speech data and translating the text data into Kiswahili to generate parallel corpora. These resources are intended to facilitate the development of NLP applications that meet the needs of African communities, thereby promoting linguistic diversity in AI development.

In the last five years, there has been a surge in NLP activity for African languages. There is a

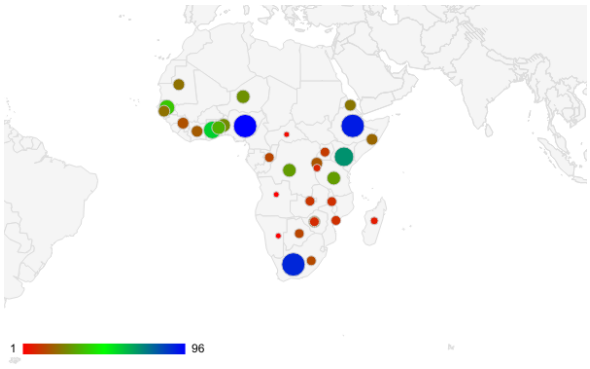


Figure 1: Distribution of NLP activity in Africa based on the number of NLP terms appearing in relation to each country in public sources

need to get an overall picture of what is being done and where. A search for "African language NLP" reveals a concentration of activity in only a handful of countries, among them South Africa, Nigeria, Ethiopia, Kenya, and Ghana. This state of affairs is depicted in Figure 1 based on the number of mentions of specific African languages in publicly available Internet sources. Although our project focused on only three Kenyan languages, one objective of this article is to inspire others to initiate similar projects across the African continent and to ensure that no language is left behind.

This research aimed to answer the following questions:

1. What methods are effective in acquiring data for building language corpora for low-resource languages?

2. What can motivate members of target language communities to contribute language data?

3. What methods can be applied to ensure the quality of the language data collected?

2 Background on Kidaw'ida, Kalenjin and Dholuo

In this section, we provide background information on each language and describe any language data available for it. We also describe the status of the language with respect to NLP resources.

2.1 Kidaw'ida (ISO 639-3 Code: dav)

Kidaw'ida is a Bantu language spoken Kenya by approximately half a million people in Taita Taveta County in Southeastern. Apart from the Bible, and the Anglican Hymn Book, few publications in Kidaw'ida can be found. Frank Mcharo and Peter Bostock published, in Kidaw'ida, a small 71-page book on the customs and traditions of the W'adaw'ida (Mcharo, 1995). This is a copyrighted resource and can only be used with permission. Furthermore, it is only available in print. Although online articles, social media posts and self-published story books written in Kidaw'ida are known to exist, they are hard to discover and consolidate. Indeed, social media platforms like Facebook, X and WhatsApp could be a rich source of Kidaw'ida text, but to extract it from the mixture of languages used on such platforms presents a significant challenge.

Kidaw'ida faces the threat of glottophagy from the more dominant Kiswahili due to their close geographical proximity and the intermingling of speakers of different Kenyan languages. Glottophagy is a term coined by the French linguist Louis-Jean Calvet (Calvet, 1974) to describe the "eating" of a language by a more dominant one. It perfectly captures what is arguably the greates current threat to Kidaw'ida. This threat is witnessed in real time as Kiswahili words are rapidly replacing Kidaw'ida words in day to day conversations. Under normal circumstances, borrowing of words is a boon to a language as it gives the language words that it lacks, thus strengthening the expressive power of the language. But the replacement of words when not necessary is very concerning as it could mean the erosion of the culture and the loss of crucial indigenous knowledge. This is not to deny the inevitable evolution of all languages, but rather to raise an alarm about the rapidity of the change affecting languages like Kidaw'ida that are spoken by a small population and that are threatened by a dominant language.

2.2 Kalenjin (ISO 639-3 Code: kln)

Kalenjin is a Nilotic language, belonging to the Eastern Nilotic branch of the Nilo-Saharan language family. It is primarily spoken by the Kalenjin people, who predominantly reside in the Rift Valley region of Kenya, particularly in counties such as Uasin Gishu, Elgeyo-Marakwet, Bomet, Baringo, Nandi, and Kericho. The language is also spoken by smaller populations in parts of Uganda and Tanzania (Naibei and Lwangale, 2018). The Kalenjin people are part of the larger Nilotic group that migrated from the Nile Valley to the East African Great Lakes region thousands of years ago (Chelimo and Chelelgo, 2016). The migration process,

which began around the fifteenth century, saw the Kalenjin spread across the Rift Valley and into the surrounding areas. As a member of the Eastern Nilotic group, Kalenjin shares linguistic features with other languages such as Turkana, Pokot, and Maasai, all of which are spoken by ethnic groups in the same region (Naibei and Lwangale, 2018). Kalenjin holds significant importance in the cultural and social life of its speakers. It is used for daily communication, in traditional ceremonies, and in storytelling, with oral literature being a vital part of Kalenjin cultural identity. However, similar to many African languages, Kalenjin use in formal domains like education, governance, and media is limited. In Kenya, English and Kiswahili dominate in these sectors, relegating Kalenjin primarily to informal communication (Ogot, 2002). Kalenjin, like other indigenous languages in Kenya, faces the challenge of language shift, where younger generations tend to prefer English and Kiswahili for social mobility and economic opportunities. Despite this, the language remains a central part of Kalenjin identity and continues to be passed down through generations, particularly in rural areas.

2.3 Dholuo (ISO 639-3 Code: luo)

Dholuo is a Nilotic language belonging to the Western Nilotic branch of the Nilo-Saharan language family. It is primarily spoken by the Luo ethnic group, which resides predominantly in the western parts of Kenya, along the shores of Lake Victoria, and parts of Tanzania and Uganda (Omulo and Williams, 2018). The language traces its origins to the migration of Nilotic peoples from southern Sudan, who gradually settled in the Great Lakes region of East Africa around the 15th century (Heine and Nurse, 2000). As a Western Nilotic language, Dholuo shares linguistic similarities with other languages in that family, such as Acholi and Lango spoken in Uganda and South Sudan.

Dholuo plays an important role in the cultural and social life of the Luo people, serving as a medium of communication in daily life, traditional practices, and artistic expressions such as music and oral literature. However, its usage is restricted to informal settings, as English and Kiswahili dominate formal domains such as education, governance, and commerce in Kenya (Nduati, 2016).

2.4 Challenges Facing Low-Resource African Languages

The challenges faced by the three languages in this study exemplify common barriers that prevent low-resource African languages from becoming digitally viable for NLP applications. Drawing from the experiences with Kidaw'ida, Kalenjin, and Dholuo, we identify the following critical gaps (Mbogho et al., 2025):

1. Insufficient Digital Resources: Comprehensive text corpora, speech recordings, and annotated datasets are scarce or non-existent. Without substantial linguistic data, it is challenging to develop effective NLP applications such as speech recognition, machine translation, and language generation.

2. Limited Research Capacity: Low-resource language research typically requires speakers of the language with relevant technical training. Smaller language communities face challenges in finding researchers with both linguistic knowledge and NLP expertise to spearhead such initiatives.

3. Lack of Institutional Support: Formal domains such as education, governance, and media predominantly use dominant languages (English and Kiswahili in Kenya), relegating indigenous languages to informal communication. Language policies, while sometimes supportive on paper, often lack practical implementation mechanisms (Awuor, 2019).

4. Community Engagement Challenges: While there is initial excitement, sustaining long-term community participation in language data collection projects requires consistent engagement and appropriate remuneration. Volunteers need motivation beyond linguistic preservation alone.

5. Digital Marginalisation Risk: Without substantial efforts to build linguistic resources and NLP tools, low-resource languages face digital extinction, where speakers are excluded from AI-powered communication technologies and digital services.

6. Resource Scalability: Current African language projects generate modest datasets (tens of thousands of tokens or hundreds of hours of

speech), while state-of-the-art language models require massive datasets in the order of trillions of tokens (Liu et al., 2024). Methods must be scalable and sustainable.

3 Related Work

Many African languages are primarily oral, with limited written materials and even less available electronically. This means that standard approaches to the harvesting of online data, such as web crawling, are excluded in many cases. Therefore, for many African languages, researchers must rely on other methods to collect data, such as fieldwork, community involvement, and collaborations with native speakers. By actively seeking and collecting data through these means, researchers can ensure that their work is inclusive and representative of the diverse linguistic landscape of Africa. This section reviews a few examples of other researchers' approaches to building African-language corpora from which we drew inspiration for our own project.

Kencorpus (Wanjawa et al., 2023) is a corpus of Kiswahili, Dholuo, and Luhya, three of the most widely spoken languages in Kenya. The corpus includes text documents, voice files, and a question-answering dataset. Data collection was conducted by researchers who were deployed to communities, schools and media houses. Kencorpus collected 4,442 text sentences and 177 hours of recorded speech. From the raw data, the project annotated Dholuo and Luhya texts with part-of-speech tags and created question-answer pairs for the Kiswahili corpus. In addition, a parallel corpus was created by translating the Dholuo and Luhya datasets into Kiswahili.

Adelani et al. (2021) used online news sites to develop named entity recognition (NER) datasets for ten African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian-Pidgin, Swahili, Wolof, and Yorùbá. The data consisted of about 30,000 sentences that were annotated for the NER task. The population sizes of the speakers of the ten languages range from 4 million for Dholuo to 98 million for Kiswahili. Volunteers annotated the raw data with entity labels for person, location, organisation and date. The annotators were part of the Masakhane community and were not paid but were included as co-authors on the paper. Multiple annotators of the same entities provided reliability and quality assurance. This approach offers a viable alternative to the more common paid or unpaid crowd-sourcing approach. Volunteers who are part of a community such as Masakhane that has been organised for language work are already motivated and knowledgeable and may be better prepared to participate.

Nakatumba-Nabende et al. (2024) developed text and speech resources for four Ugandan languages and Kiswahili. The four Ugandan languages are Luganda, Runyankore-Rukinga, Lumasaba, and Acholi. Some of the text data were created by translating English text into the five African languages. The English text was obtained from various published sources, including English Wikipedia, social media, online local newspapers, story books, novels and human rights charters. In addition to generating data for the African languages, this process also generated parallel corpora for English and the African languages. Additional text was obtained by recruiting native speakers through crowd-sourcing to write sentences and review them for quality. Due to a lack of access to computers, some contributors wrote their sentences by hand, which had to be typed later. Text sentences were later read by native speakers, again through crowd-sourcing, and recorded to generate speech corpora. Five parallel text corpora were created consisting of 40,000 sentence pairs each. The project also recorded 582 hours of Luganda speech and 1100 hours of Swahili speech.

Ogayo et al. (2022) built a speech synthesis dataset for 11 African languages: Dhouo, Lingala, Kikuyu, Yoruba, Hausa, Luganda, Ibibio, Kiswahili, Wolof, Fongbe and Suba. The data consisted of a total of 65,537 utterances distributed unevenly across languages, ranging from 125 to 11,971 utterances. These utterances corresponded to a total of 113.84 hours of recordings, ranging from 0.33 to 24.82 hours. Religious texts and other online and print sources were used as data sources. The data were supplemented with contributions from recruited participants. To help encourage community participation, comprehensive guidance on data collection methods and data licensing was provided. Due to the quality of speech required for speech synthesis, voice talents were carefully vetted and remunerated in cash and kind.

The examples discussed in this section give an idea of the kinds of activities that are going on in corpus building for African-language NLP, and suggest approaches that have been shown to work well and that others can emulate. It is essential to note that these examples focus on specific re-

gions and languages and only partially represent Africa's linguistic diversity. They also offer only a glimpse into the significant amount of ongoing NLP work (including corpus building) in Africa. As momentum in African language technologies builds, researchers need to be intentional in ensuring that no language is left behind and that global inequalities are not replicated locally.

As these examples show, current African language projects, regardless of the methods used, only generate modest amounts of data, typically tens of thousands of tokens or hundreds of hours of recorded speech. On the other hand, the latest advances in NLP are in large language models (LLMs), which require massive datasets. Llama 2 from Meta was trained on 2 trillion tokens (Touvron et al., 2023); OpenAI's latest LLM was trained on even more data than this, although the specifics have not been officially disclosed. Therefore, it is important that the methods employed for data collection are scalable and sustainable to support more expansive research efforts across the continent. This is a significant challenge that requires innovative solutions and we hope that our work makes a positive contribution to the available approaches.

4 Our Methodology

Our main objective was to collect text and speech data for three indigenous Kenyan languages, namely Kidaw'ida, Kalenjin, and Dholuo. These languages were chosen because they are the home languages of the research team. The project proposed to collect data from members of the language communities through crowd-sourcing. The data collection was realized through a grant from the Lacuna Fund, which made it possible to pay small stipends to compensate language data contributors for their time, effort and the knowledge shared.

4.1 Text Data Collection

Two approaches were used to generate text data:

1. The contributors wrote down sentences in the three languages covered by this project. These could be made-up sentences or sentences from public-domain materials. Some contributors wanted to use the bible, but most local-language bibles in Kenya are copyrighted by the Bible Society of Kenya and the British and Foreign Bible Society. However, we advised that they could use such copyrighted materials for inspiration. For example, a Bible story could serve as a prompt and remind them of something else they could write about. The issue of cultural appropriateness of data is often cited as militating against using texts of foreign origin, but we propose that such texts can catalyse ideas.

2. Contributors recorded conversations in the three languages, and the conversations were later transcribed. Participants in conversations were informed of the recording in advance and asked to sign a consent form. During transcription, words that were in a language other than the target language were replaced, as code-switched speech was outside the scope of the project.

The sentences collected were translated into Kiswahili to generate three parallel corpora. Microsoft Excel and Google Sheets were used to compose and translate sentences. The resulting spreadsheets were stored on Google Drive and GitHub throughout the duration of the project.

Quality assurance was achieved by identifying contributors with high levels of language proficiency who were designated as Data Collection Leads (DCLs) and paid a monthly salary. The DCLs identified language contributors they knew were qualified and recommended them for recruitment. DCLs also contributed data and checked the contributor data for correct spelling, grammar, fluency, and proper translation into Kiswahili. The DCLs corrected any errors they found in the data.

To increase inclusion in NLP work, 7 of the DCLs selected were female and 5 were male, and 5 of the researchers were female and 1 was male. Women were also well represented among the contributors, as can be seen in Table 1. It is important to have gender balance in language projects to ensure that applications work for everyone and also to avoid bias, for example, in generative AI.

4.2 Voice Data Collection

To enable our geographically distributed contributors to easily make voice data contributions easily, we determined that an open-access online platform would be ideal. Quite surprisingly, we were only able to find two options: Living Dictionaries (`https://livingdictionaries.app`) and Mozilla Common Voice (`https://commonvoice.mozilla.org`).

Initially, it seemed that Living Dictionaries would be easier to use. Mozilla Common Voice involves a rather steep on-ramp, which we explain below. Living Dictionaries, although easy to get started with, turned out not to be fit for purpose as the main aim of that platform is language documentation and it is tailored for linguistic work rather than NLP projects. For example, when recording voice for NLP, the same phrase is required to be read by multiple people. This is not well supported in Living Dictionaries as the interest there is in capturing how a phrase is spoken and having one person speak it suffices. On the other hand, in NLP and, more specifically, in automatic speech recognition (ASR), the different characteristics of people's voices while speaking the same phrase must be captured, so that applications built on those data can work for everyone.

Thus, we ultimately settled on Mozilla Common Voice. The first step was to make a language request. Once this request was approved for each of our three languages, we had to localize the pages associated with our languages using a platform provided by Mozilla known as Pontoon (https://pontoon.mozilla.org). This involved translating more than one thousand technical strings into each language, a task requiring both technical knowledge and language skills, including the ability to create suitable terms for new technical concepts such as "database", "download" or "speech recognition". This process was quite time-consuming and was a necessary step before launching the language on Mozilla Common Voice. After launch, we uploaded the text sentences collected earlier and once enough text was available on the platform, we could start reading and recording. Our DCLs recruited more volunteers to read and record, ensuring the diversity of voices needed for speech data.

The DCLs provided continuous quality assurance by reviewing and correcting, as necessary, both the text and the audio data simultaneously with the data contribution.

5 Results

5.1 Text Data

We collected 30,000 text sentences for each of the three languages and had contributors proficient in both their mother tongue and Kiswahili provide translations. This resulted in the creation of three parallel corpora: Kidaw'ida-Kiswahili, Kalenjin-Kiswahili and Dholuo-Kiswahili. These parallel corpora are freely available for download from https://zenodo.org/records/13355021 (Mbogho et al., 2025). The same repository remains on Github, where it can be updated and reuploaded to Zenodo. This is important not only for the continual expansion of the datasets, but also for making any needed corrections and quality improvements.

To analyze the translation consistency of the corpora, we used two measures. The first is average Sentence Length Ratio that computes the ratio of average sentence length of the target language to the average sentence length of the source language. This measure shows whether translations expand or contract compared to the originals. A number between 0.8 and 1.3 is generally considered indicative of good-quality translations. For the three corpora in this study, the average sentence length ratio falls in the range 1.0-1.2, indicating that target sentences are generally longer by up to 14% for Dholuo-Kiswahili and 13% for Kidawida-Kiswahili. For Kalenjin-Kiswahili, the expansion is minimal (the source and target sentences are generally of the same length) with a percentage less than 1% (see Table 2).

Table 2 also shows the sentence length correlation computed using Pearson's r measure. This shows how strongly sentence lengths correspond across the two languages in a parallel dataset. It measures translation consistency where values above 0.8 show a likelihood of good alignment and consistent translation length patterns while values lower than 0.5 show possible alignment problems or major structural differences between languages. The three corpora show consistent alignment patterns between Kiswahili and the corresponding local languages with values over 0.8, except Kalenjin, which has about 0.75.

The above insights will be useful when machine translation models are trained and the corresponding applications are developed.

5.2 Voice Data

The collection of voice data is ongoing, but currently the speech data on Mozilla Common Voice is as shown in Table 2.

Like gender, age is an important consideration for voice data collection as a person's voice normally changes with age. For this reason, age categories should span a wide range and should match the distribution in the general population as much

Measure	Dholuo-Kiswahili	Kalenjin-Kiswahili	Kidaw'ida-KIswahili
Average sentence length ratio	1.148	1.002	1.136
Sentence length correlation	0.893	0.747	0.842

Table 1: Translation consistency analysis

Language	Hours	Speakers	Female	Male
Kidaw'ida	56	24	60%	40%
Kalenjin	92	41	70%	30%
Dholuo	120	44	58%	42%

Table 2: Speech Data on Mozilla Common Voice

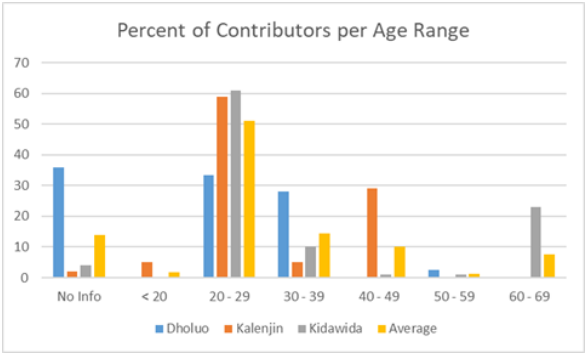


Figure 2: Age range of voice contributors

as possible. Figure 1 shows the age distribution of the voice contributors in our project. Most of the contributors were between 20 and 29 years of age. Thus, there is room for improvement in this regard as we plan for expansion of the work.

6 Discussion

This work was guided by three research questions. We aimed to find out what methods are effective in acquiring data for building language corpora for low-resource languages; how to motivate communities to contribute to language data; and how to ensure quality in the collected data.

We have referred to the type of crowd-sourcing we used as "selective" because the contributors were known to members of the project team. The researchers recruited DCLs they knew and, in turn, DCLs recruited contributors from among their close contacts. This was important because language proficiency was a key consideration, and using a "random" crowd could have compromised data quality. We found selective crowd-sourcing to be a useful approach for developing African language corpora, balancing rapid data generation with quality assurance. However, this approach may not scale well to larger projects. In that case, a project may have to accept data of varied quality and then apply additional quality control measures more aggressively in a subsequent phase.

With respect to motivating communities to participate in language projects like ours, the lesson from this project is that remuneration is critical. Some of our contributors were going through periods of unemployment, and this is what allowed them the freedom to take part. It is important to provide reasonable compensation for contributors' time, effort and linguistic expertise. We also found that many people with whom we spoke did not have a sense of the importance of their own language. Through our project, we have anecdotal evidence that there is increased awareness of the value of African languages and of the ways communities can participate in corpus building efforts.

Quality in the collected data remains a challenge. Although most of the data we have produced in this project is of good quality, we do sometimes find low quality contributions in the datasets. The work of reviewing and correcting is thus ongoing. We attribute this challenge to a lack of seriousness towards the task among a minority of the contributors. We speculate that the necessary compensation we have emphasised above can be a double-edged sword; a project may attract participants because of the pay, rather than any commitment to providing quality language contributions. However, we are encouraged by the fact that the majority of participants were serious about the task and felt they had a personal stake in promoting their own language. Educating African communities about the value of their languages and their potential to make important contributions should be a priority in language projects.

7 Conclusions and Future Work

The project ran for a year and generated approximately 90,000 sentence pairs for the parallel text corpora Kidaw'ida-Kiswahili, Kalenjin-Kiswahilli and Dholuo-Kiswahili. We also collected 268 hours of recorded speech for Kidaw'ida, Kalenjin and Dholuo by the time of the project's conclu-

sion. The researchers are actively seeking more funding to continue the work and extend it to additional African languages. We recommend that the datasets be used in their current state to train models to establish a baseline that can serve as an indication of the level of accuracy currently achievable. It is also important to document what is possible with small datasets as not all applications require massive amounts of data.

The datasets are available on open-access platforms, namely, Zenodo and Mozilla Common Voice. Anyone who wishes to download the data from either platform can do so at no cost and with minimal barriers, as the licences on both platforms are highly permissive. Developers are encouraged to take advantage of this unrestricted access to the data to train models and develop applications in these three languages. Language communities are encouraged to continue adding to the repositories to achieve greater accuracy in future models.

The project demonstrates how grassroots efforts in corpus building can support the inclusion of African languages in artificial intelligence innovations. In addition to filling resource gaps, these corpora are vital in promoting linguistic diversity and empowering local communities by enabling Natural Language Processing applications tailored to their needs. As African countries like Kenya increasingly embrace digital transformation, developing indigenous language resources becomes essential for inclusive growth. We encourage continued collaboration from native speakers and developers to expand and utilise these corpora.

Limitations

One limitation of this project is that we did not make sufficient effort to recruit participants in all age groups. Most of them fell in the 20-29 age range. This would limit the performance of automatic speech recognition models trained on the data. However, we expect that continued contribution from the wider community will give rise to greater representation with respect to age. Still, it is important to pay attention to the age aspect in future work.

Licensing issues are a major concern in corpus building for low-resource languages. A greater involvement of intellectual property specialists is essential to ensure that language communities are not disadvantaged but rather stand to benefit from the resulting data and associated applications. This consideration was precluded by the available funds.

As mentioned earlier, even though we made sure that each contributor was known to someone in the project team, we still had participants who did not take the work seriously and contributed low-quality data. In future projects, we will prioritise stronger vetting of contributors and recruitment of a larger quality assurance team.

Acknowledgments

This work was carried out with support from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, Canada's International Development Research Centre, and GIZ on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ). The authors also wish to thank the Kidaw'ida, Kalenjin and Dholuo language communities for sharing their languages with the world.

Declarations

The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Quin Elizabeth Awuor. 2019. Language policy in education: The practicality of its implementation and way forward. *Journal of Language, Technology & Entrepreneurship in Africa*, 10(1):94–109.

Ayo Bamgbose. 2011. African languages today: The challenge of and prospects for empowerment under globalization. In *Selected proceedings of the 40th annual conference on African linguistics*, pages 1–14. Cascadilla Proceedings Project Somerville.

Louis-Jean Calvet. 1974. Linguistique et colonialisme. *Petit traité de glottophagie, Paris, Payot*.

Florence J. Chelimo and Kiplagat Chelelgo. 2016. Precolonial political organization of the kalenjin of kenya: An overview. *International journal of innovative research and development*, 5.

François Chollet. 2021. *Deep learning with Python*. Manning.

Bernd Heine and Derek Nurse. 2000. *African languages: An introduction*. Cambridge University Press, United Kingdom.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Audrey Mbogho, Quin Awuor, Andrew Kipkebut, Lilian Wanzare, and Vivian Oloo. 2025. Building low-resource african language corpora: A case study of kidawida, kalenjin and dholuo. *arXiv preprint arXiv:2501.11003*.

A.F. Mcharo. 1995. *Mizango na Maza ra Kidawida ra Kufuma Kokala*. Peter Bostock, Kenya.

Faith K Naibei and David Lwangale. 2018. A comparative study of the kalenjin dialects. *International Journal of Academic Research in Business and Social Sciences*, 8(8):476–503.

Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukiibi, Eric Peter Wairagala, Chodrine Mutebi, Tobius Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, and 1 others. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: Experiences and lessons. *Applied AI Letters*, 5(2):e92.

Rosemary N Nduati. 2016. *The post-colonial language and identity experiences of transnational Kenyan teachers in US Universities*. Ph.D. thesis, Syracuse University.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Perez Ogayo, Graham Neubig, and Alan W Black. 2022. Building tts systems for low resource languages under resource constraints. In *Proc. S4SG 2022*.

Bethwell A Ogot. 2002. Historical portrait of western kenya up to 1895 bethwell a. ogot. *Historical Studies and Social Change in Western Kenya: Essays in Memory of Professor Gideon S. Were*, 13(4):99–120.

Albert Gordon Omulo and John James Williams. 2018. A survey of the influence of 'ethnicity', in african governance, with special reference to its impact in kenya vis-à-vis its luo community. *African Identities*, 16(1):87–102.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2023. Kencorpus: A kenyan language corpus of Swahili, dholuo and luhya for natural language processing tasks. In *Journal for Language Technology and Computational Linguistics, Vol. 36 No. 2*, pages 1–27, unknown. German Society for Computational Linguistics and Language Technology.

Exploring Zoonotic Disease Knowledge Through AI for Enhanced Risk, Prevention, and Response Awareness in Low-Resource Languages

Armand De Wet[1], Reolin Govender[1], Tedson Nkoana[2], Wanda Markotter[2], Abiodun Modupe[1,3], Vukosi Marivate[1,3,4,5]

[1]Department of Computer Science, University of Pretoria,
[2]Future Africa Institute, University of Pretoria,
[3]Data Science for Social Impact,
[4]AfriDSAI, University of Pretoria, [5]Lelapa AI,

Correspondence: reolin.govender@tuks.co.za

Abstract

Limited linguistic inclusivity in public health communication leaves many South African communities underserved, particularly regarding critical information on zoonotic diseases such as rabies. This pilot study addresses this gap by developing and evaluating AI-driven methods for delivering reliable rabies information to Sepedi speakers, a low-resource language group. The study presents a novel, curated Sepedi dataset of 60 question–answer pairs, created through a systematic pipeline: thematic analysis of authoritative English sources guided the synthetic generation of QA pairs, which were then translated and manually verified by a native-speaking expert. This dataset was used to compare two large language models, GPT-4o and Gemini-1.5 Flash, under both base and fine-tuned conditions. Evaluation used a human-centred rubric assessing fluency, accuracy, and cultural appropriateness. The findings reveal a key nuance in applying LLMs to low-resource domains. The base GPT-4o model, with strong foundational multilingual capabilities, outperformed all other configurations, including its own fine-tuned variant. In contrast, fine-tuning provided a marked improvement for the less capable base Gemini model. This result indicates that fine-tuning can enhance weaker models; its benefits are not universal and may be outweighed by the strong zero-shot performance of state-of-the-art architectures when training data is scarce. The curated Sepedi rabies QA dataset will be released under an open licence to support future work in low-resource public health communication. The data can be found at this GitHub link.

1 Introduction

Linguistic barriers persistently hinder effective public health communication in South Africa. While critical health information is widely available in English, it often fails to reach communities where indigenous languages are predominantly spoken. This gap is perilous in the context of high-stakes diseases like rabies, a preventable yet fatal zoonotic illness that continues to pose a threat in rural provinces (World Health Organization, 2024; National Institute for Communicable Diseases, 2025). The scarcity of accessible, culturally relevant health resources in languages such as Sepedi, spoken by over 4.6 million people, has been linked to reduced disease awareness and delayed treatment, with severe consequences (Makungo et al., 2025; Cotoc and Notaro, 2022).

The emergence of large language models (LLMs) presents a novel opportunity to bridge this linguistic divide. This study investigates that potential by developing and evaluating AI-driven methods for delivering reliable rabies information in Sepedi. At the core of the study is the creation of a novel, thematically structured question–answer (QA) dataset in Sepedi. To build this resource, authoritative English-language public health sources were first analysed to identify key thematic areas. These themes were then used to guide the generation of 60 QA pairs, which were translated into Sepedi and rigorously validated by a native-speaking expert to ensure linguistic and cultural accuracy. This dataset served as the basis for fine-tuning and evaluating two leading LLMs: OpenAI's GPT-4o and Google's Gemini-1.5 Flash.

This paper makes two primary contributions. First, it introduces what, to the authors' knowledge, is the first curated Sepedi QA dataset for a specific public health domain. The dataset is intended for public release under an open licence to enable external inspection and reuse, and can be found at this GitHub link. Second, it presents a comparative evaluation of base and fine-tuned LLMs for low-resource health communication. The experiments yielded a significant, counterintuitive result: the base GPT-4o model, without any fine-tuning, outperformed all other configurations, including

its own fine-tuned variant. This finding challenges the common assumption that fine-tuning is always optimal and highlights the impact of strong foundational multilingual capabilities when training data is scarce. By contrast, fine-tuning provided a marked improvement for the Gemini model, indicating that targeted adaptation remains valuable for models with weaker baseline performance in this setting.

The remainder of this paper is organised as follows. Section 2 surveys related work on low-resource NLP and health communication. Section 3 details our data curation process and the analysis that guided it. Section 4 describes our modelling approach and results. Section 6 discusses the study's limitations and outlines future research directions. Finally, Section 5 summarises our key findings and broader implications.

This paper addressed that gap in two ways. First, it constructed a curated Sepedi question–answer dataset on rabies, derived from authoritative South African public health material and verified by a native Sepedi speaker. Second, it evaluated base and fine-tuned variants of two large language models, GPT-4o and Gemini-1.5 Flash, to test whether they can generate reliable, culturally appropriate health guidance in Sepedi. This is not a general study of "AI in health," but a targeted examination of Sepedi rabies communication using large language models.

2 Literature Review

Developing effective AI tools for public health communication in under-resourced languages, such as Sepedi, requires navigating the intersection of natural language processing (NLP), multilingual AI, and health informatics. This review synthesises existing research in these areas to contextualise the study's challenges and contributions. The discussion focuses on three aspects: the scarcity of linguistic resources, the use of large language models in specialised and low-resource settings, and the methods guiding data curation and evaluation.

This scarcity is even sharper in specialised domains like public health. Institutions such as the National Institute for Communicable Diseases (NICD) produce many English-language resources on topics like rabies. However, similar materials in Sepedi are almost nonexistent. This linguistic gap hinders effective health communication and increases the need for domain-specific datasets that are both medically relevant and culturally sound. This study directly addresses this by creating a curated Sepedi QA dataset for rabies. It provides a foundational resource for developing and evaluating specialised conversational AI.

2.1 The Data Scarcity Challenge in Sepedi

A lack of digital resources has long hampered the development of NLP for African languages. Sepedi, in particular, is consistently classified as a low-resource language (Marivate et al., 2020). Foundational work by the National Centre for Human Language Technology (NCHLT) and projects like Atshumato have highlighted the scarcity of structured corpora and annotated data. While researchers have demonstrated the feasibility of training models for Sepedi, they consistently note severe constraints related to corpus size and quality (Makgatho et al., 2021). The Pedi corpus remains one of the few publicly available benchmarks for the language, underscoring the resource gap (Marivate et al., 2020).

2.2 Large Language Models for Low-Resource Health Communication

Recent advancements in multilingual LLMs offer a promising pathway for bridging information gaps in healthcare. Research has shown that fine-tuning models on specialised, domain-specific data can significantly improve performance. For instance, studies using LLM derivatives in Vietnamese health communication demonstrated that targeted fine-tuning enhanced factual accuracy and user trust (Bui et al., 2025). Similarly, work adapting models for biomedical and ophthalmology question-answering has confirmed that domain-specific fine-tuning improves both fluency and factual correctness (Tan et al., 2024; Christophe et al., 2024).

These findings directly motivated our decision to experiment with fine-tuning for Sepedi, rather than relying solely on zero-shot prompting. Furthermore, this body of work highlights the inadequacy of standard automated metrics (e.g., BLEU, ROUGE) for evaluating performance in specialised, low-resource settings, thus justifying the human-centred evaluation strategy employed in our study.

2.3 Methodological Considerations for Analysis and Evaluation

The dataset was constructed using established exploratory NLP techniques. Topic modelling, specifically Latent Dirichlet Allocation (LDA) and Non-

negative Matrix Factorisation (NMF), is effective for discovering latent themes in public health corpora (Murel and Kavlakoglu, 2024; Murakami et al., 2017). This approach was used to identify the core informational categories in the source documents. Similarly, Named Entity Recognition (NER) is a standard method for extracting structured information, such as organisational and geographic references, from text (Li et al., 2022). These methods provided an empirical foundation for the thematic structure of the generated QA pairs.

For evaluation, our human-centric approach aligns with a growing consensus that qualitative, expert-led assessment is essential in high-stakes domains. Unlike automated metrics, a human-led rubric can assess dimensions such as cultural appropriateness, contextual relevance, and the subtle nuances of factual accuracy, all of which are paramount for building trustworthy health information systems. This qualitative approach ensures that our assessment of model performance is grounded in real-world communicative effectiveness rather than abstract statistical similarity.

3 Data Curation and Exploratory Analysis

A significant challenge in developing NLP tools for Sepedi is the scarcity of structured, domain-specific digital text. To address this in the context of rabies, the study adopted a two-stage approach. First, an exploratory data analysis (EDA) was conducted on a curated English-language corpus to identify key thematic and terminological patterns. Second, these insights were used to guide the creation of a synthetic, high-quality Sepedi question–answer (QA) dataset for fine-tuning.

3.1 Source Corpus and Analysis

Our initial corpus was compiled from authoritative English-language sources, including public awareness materials from the National Institute for Communicable Diseases (NICD), regional surveillance reports, and relevant peer-reviewed studies. This corpus was subjected to a suite of NLP techniques, including word frequency analysis, n-gram extraction, topic modelling, and Named Entity Recognition (NER), to systematically map the rabies information landscape in the South African context. The primary goal was to establish an evidence-based blueprint for our Sepedi dataset.

3.2 Key Findings from the English Corpus

The analysis revealed a strong focus on public health and clinical response. Word frequency analysis (Figure 2) and n-gram extraction (Figure 3) highlighted the centrality of terms such as *rabies*, *animal*, *vaccination*, and phrases like "post exposure prophylaxis" and "rabies south africa" (Table 3).

Table 1: Topic descriptions and key themes using NMF

Topic	Keywords
Dog Health	dog, africa, south, province, species, wildlife, domestic, disease, table, areas
Specimen Collection	case, specimen, biopsy, mortem, csf, nicd, saliva, skin, post
Post-Exposure Prophylaxis	rig, exposure, animal, vaccine, dose, pep, vaccination, day, wound, doses
Transmission Risk	animals, animal, gov, infected, contact, state, humans, prevent, symptoms
Disease Reporting	conditions, notifiable, fever, medical, laboratories, list, public, private

Table 2: Topic descriptions and key themes using LDA

Topic	Keywords
Notifiable Diseases	medical, disease, conditions, notifiable, fever, category, list, laboratories, public, private
NICD & Animal Exposure	nicd, exposure, case, post, animals, csf, saliva, prophylaxis, africa, south
Reporting Animal Contact	animal, contact, gov, state, wildlife, com, must, veterinarian, suspect
Animal Vaccination	animal, exposure, vaccine, vaccination, may, south, africa, bat, dogs
Dog Surveillance	africa, south, dog, species, disease, province, domestic, wildlife, surveillance

Topic modelling, using both Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorisation (NMF), further distilled these patterns into distinct thematic clusters (Tables 2 and 1). These clusters consistently aligned with six core public health domains:

1. Epidemiology and surveillance
2. Prevention and control
3. Virology and diagnosis
4. Clinical presentation

5. Geographical risk areas

6. Public awareness and education

These six domains were not chosen ad hoc. They were derived directly from authoritative South African public health sources, including National Institute for Communicable Diseases (NICD) rabies guidance, provincial surveillance reports, and clinician-facing bite management and post-exposure protocols. Topic modelling, frequency analysis, and named entity recognition over these materials confirmed that these domains captured core informational needs around exposure, vaccination, reporting, symptom escalation, and local geographic risk.

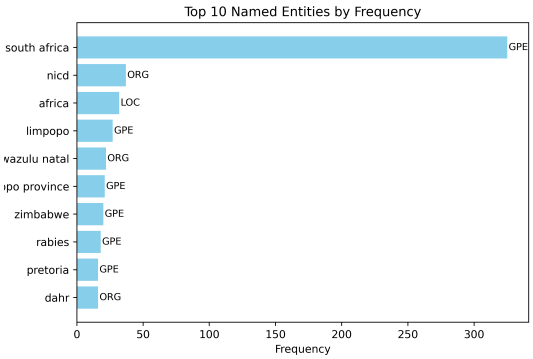


Figure 1: Top named entities identified using NER. Note: Some geographic entities were misclassified due to model limitations (e.g., "KwaZulu-Natal" was classified as ORG rather than GPE).

Additionally, NER confirmed the corpus's regional relevance, identifying key entities such as *South Africa*, *Limpopo*, *NICD*, and specific animal vectors (Figure 1). While the NER tool occasionally misclassified some entities (e.g., treating provinces as organisations), it provided a useful high-level confirmation of the corpus's focus.

Table 3: Top 10 Trigrams with Frequency Counts from the source corpus.

Freq.	Word 1	Word 2	Word 3
51	post	exposure	prophylaxis
42	rabies	south	africa
42	significant	disease	clusters
34	human	rabies	cases
34	trop	med	infect
24	pre	exposure	vaccination
23	world	health	organization
22	bat	eared	foxes
21	black	backed	jackals
21	positive	rabies	cases

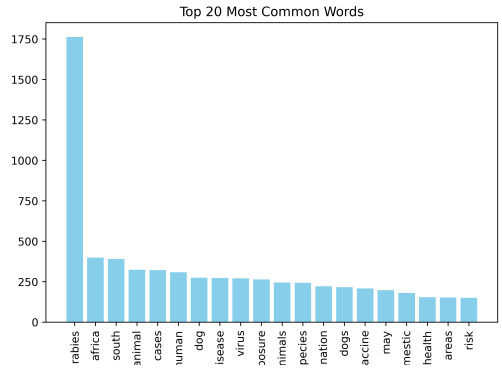


Figure 2: Top 20 most frequently occurring words in the English source corpus.

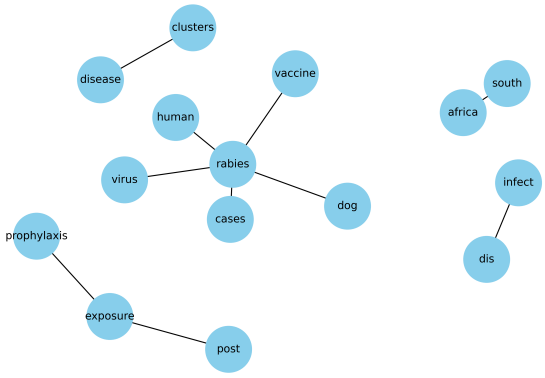


Figure 3: Bigram network visualising core word pairings in the source corpus.

3.3 Guided Creation of the Sepedi Dataset

The six themes identified in the EDA served as a structured framework for generating our fine-tuning dataset. This process involved two key steps:

1. **Synthetic QA Generation:** The initial English question–answer pairs were produced using OpenAI's GPT-4o in an interactive browser session. The model was instructed to act as a public health explainer and to generate clear, non-technical answers about rabies for a general South African audience. The instructions emphasised three constraints: (i) provide medically responsible guidance consistent with recognised South African practice, (ii) describe immediate actions after potential exposure (for example wound cleaning and urgent reporting) without giving alarmist or shaming language, and (iii) avoid inventing unavailable services or phone numbers. The prompting was iterative rather than fully scripted: the model was asked to generate

question–answer pairs for each high-priority theme, the output was inspected, and prompts were refined until all required themes were covered. No custom decoding parameters (for example, temperature or maximum tokens) were manually set in that interface, and exact provider defaults from that session are not recoverable. "Comprehensive coverage", therefore, meant enforced inclusion of each of the six domains identified in Section 3.2, not blind acceptance of the model's first attempt.

The 60 English question–answer pairs were generated so that all six public health domains identified in Section 3.2 (epidemiology and surveillance, prevention and control, virology and diagnosis, clinical presentation, geographical risk areas, and public awareness and education) were represented. At least one QA pair was created for each subtopic within these domains (e.g., vaccination, bite response, and reporting requirements), ensuring no domain was left empty. Every core domain defined from authoritative sources was present in the draft set before translation, and no clinically meaningful theme was omitted.

2. **Translation and Expert Verification:** The English pairs were then translated into Sepedi using GPT-4o and subjected to a manual review by a native Sepedi speaker with practical familiarity in health communication. The reviewer examined each question–answer pair for (i) grammatical correctness and natural phrasing in Sepedi, (ii) preservation of factual meaning from the English source, (iii) cultural and situational relevance (for example, how care-seeking or animal exposure would realistically be described), and (iv) medically safe advice. Corrections were applied at the sentence level, and terminology for key rabies concepts (for example, vaccination, post-exposure treatment, symptom onset, and reporting urgency) was standardised across all pairs for internal consistency. This procedure relied on a single expert reviewer and did not include a second independent rater, so no inter-rater agreement was calculated.

This systematic methodology allowed us to overcome the initial data scarcity by creating a bespoke, thematically robust Sepedi dataset. This dataset, while modest in size, is grounded in the empirical findings of our EDA and explicitly tailored for fine-tuning LLMs for rabies-related health communication, forming the foundation for the experiments detailed in Section 4.

4 Modelling & Experimentation

This section details the experimental methodology, including the selection of large language models, the fine-tuning process, the evaluation strategy, and the results of our comparative analysis.

4.1 Model Selection

Two models were selected for this study: OpenAI's `gpt-4o-2024-08-06` and Google's `gemini-1.5-flash-001`. These models were chosen for both practical and strategic reasons. Both support parameter adaptation via accessible APIs and offer mature Python SDKs, which facilitate integration, experimentation, and inference orchestration. Their underlying transformer architectures and token efficiency also made them suitable candidates for our low-resource context. The primary goal was to compare two leading models under both base and fine-tuning conditions to assess their ability to generate accurate Sepedi-language responses.

4.2 Fine-Tuning Methodology

The fine-tuning stage used the validated Sepedi question–answer dataset described in Section 3.3. Each of the 60 reviewed pairs was converted into the provider-specific JSONL structure required by the OpenAI and Google fine-tuning APIs (OpenAI, 2024; Google DeepMind, 2024). Fine-tuning jobs were then initiated via the respective SDKs, and the adapted GPT-4o and Gemini-1.5 Flash models were exposed through their inference endpoints. This subsection, therefore, focuses on how the curated dataset was consumed by the models and deployed for inference, rather than re-describing the data creation and verification pipeline already presented in Section 3.3.

4.3 Evaluation Strategy

Standard automated NLP metrics, such as BLEU and ROUGE, are known to be unreliable in morphologically rich, low-resource languages like Sepedi, as they do not capture culturally inappropriate phrasing or medically unsafe guidance. For this reason, the evaluation was designed around human judgment rather than relying solely on text-similarity scores.

A held-out evaluation set of questions was created that did not appear in the fine-tuning data but followed the same six core public health themes defined in Section 3 (epidemiology and surveillance, prevention and control, virology and diagnosis, clinical presentation, geographical risk areas, and public awareness and education). Each question in this held-out set was posed to multiple model variants, and all model responses were anonymised and randomised before scoring. This blind setup ensured the reviewer could not see which system generated which answer, reducing bias toward any specific vendor model.

The reviewer was instructed to assess each answer for factual accuracy, clarity, and cultural appropriateness in Sepedi, with specific attention to safety-relevant content, including advice on exposure, reporting, and treatment timing. This procedure treated the task as health communication, rather than generic translation, and aimed to approximate real-world usefulness for a Sepedi-speaking end-user.

A structured set of held-out evaluation questions, distinct from the training set but aligned with the same six public health themes, was developed. Each question was posed to four model configurations:

- Base GPT-4o with Context
- Fine-tuned GPT-4o
- Base Gemini-1.5 Flash with Context
- Fine-tuned Gemini-1.5 Flash

The "with Context" configurations refer to a simple prompt engineering strategy in which each query was prepended with a static, manually crafted summary of key rabies information. This approach ensured that the base models had access to relevant domain knowledge without the complexity of a whole Retrieval-Augmented Generation (RAG) pipeline. All model outputs were anonymised and their order randomised to facilitate a blind evaluation by our Sepedi-speaking expert, thereby minimising potential bias.

4.4 Evaluation Rubric

Model responses were scored against a five-dimension rubric, detailed in Table 4. The criteria, fluency, factual accuracy, relevance, clarity, and cultural appropriateness, were rated on a five-point scale from 1 (Poor) to 5 (Excellent). This structured framework enabled a detailed, qualitative analysis of output quality, capturing nuances that automated metrics would miss.

The rubric operationalised "quality" as it matters in practice for health guidance: Can a Sepedi speaker understand the answer? Is the advice factually safe? Is the tone locally appropriate rather than culturally alien or condescending?

Each criterion was scored on a five-point scale. This allowed us to compare model variants on clinically and socially relevant dimensions in high-stakes public health messaging, rather than only on text overlap with a reference.

4.5 Results and Discussion

The human evaluation results, summarised in Figure 4, reveal essential differences in model performance. Contrary to our initial hypothesis, the base GPT-4o model achieved the highest scores across all five criteria, marginally outperforming its own fine-tuned variant. This suggests that for a model with exceptionally strong pre-existing multilingual capabilities, fine-tuning on a small dataset provides limited, if any, additional benefit.

In contrast, fine-tuning yielded a marked improvement for the Gemini model. The fine-tuned version demonstrated significantly better factual accuracy and relevance compared to its base counterpart, which scored lowest across all metrics. This highlights the value of domain-specific adaptation for models with weaker baseline performance in a given low-resource language.

These findings challenge the assumption that fine-tuning is universally superior. Instead, they indicate that adaptation effectiveness is highly dependent on both the strength of the base model and the scale of the available training data. It is important to note that a single expert reviewer conducted this evaluation. While this ensured consistency, it also introduces subjectivity; the results should therefore be interpreted as indicative of performance within this pilot study.

5 Conclusion

This pilot study investigated the feasibility of using multilingual LLMs to deliver critical public health information on rabies to Sepedi-speaking communities. By focusing on a low-resource language, this work addressed a significant gap in the accessibility of digital health resources in South Africa. It explored a pathway toward more linguistically

Table 4: Scoring rubric used for human evaluation of model-generated Sepedi responses.

Criterion	1 (Poor)	2 (Fair)	3 (Good)	4 (Very Good)	5 (Excellent)
Fluency / Grammar	Unnatural, many errors	Frequent grammar or phrasing issues	Understandable but some awkward parts	Smooth with minor issues	Native-like, fluent and natural
Factual Accuracy	Incorrect or misleading	Several factual errors	Mostly correct but minor inaccuracies	Almost entirely accurate	Completely accurate
Relevance to Prompt	Off-topic or unrelated	Partially addresses the question	Mostly relevant	Very relevant	Directly and completely answers the prompt
Clarity / Understandability	Confusing or incoherent	Difficult to follow	Understandable with effort	Clear and mostly easy to follow	Very clear and easily understood
Cultural & Contextual Appropriateness	Offensive or inappropriate	Culturally out of touch or awkward	Mostly appropriate	Suitable and respectful	Culturally aligned, respectful, and sensitive

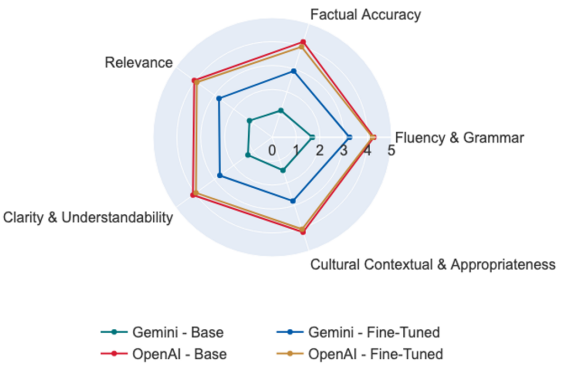


Figure 4: Radar plot comparing human evaluation scores across five criteria. While fine-tuning improved Gemini's performance, the base OpenAI model outperformed all other configurations, including its fine-tuned version. Scores reflect the assessment of a single Sepedi-speaking expert reviewer.

equitable AI.

The methodology followed a multi-stage process, beginning with an exploratory analysis of authoritative materials to identify core public health themes. This analysis informed the creation of a curated dataset of 60 Sepedi question–answer pairs, which was subsequently used to fine-tune and evaluate OpenAI's GPT-4o and Google's Gemini-1.5 Flash. Using a human-centred rubric that assessed linguistic fluency, cultural appropriateness, and factual accuracy, the study compared the performance of base and fine-tuned model variants.

The evaluation yielded a key, and somewhat counterintuitive finding: the base GPT-4o model outperformed all other configurations, including its own fine-tuned version. This result suggests that for models with powerful foundational multilingual capabilities, the benefits of fine-tuning on a small, specialised dataset may be marginal. In contrast, fine-tuning significantly improved the Gemini model's performance, confirming the value of task-specific adaptation for LLMs that are less capable at baseline in low-resource languages.

This asymmetry has direct practical consequences. It suggests that in a low-resource public health setting, a competent multilingual model may already produce usable Sepedi guidance when paired with careful prompt conditioning and contextual priming, even without task-specific fine-tuning. By contrast, a weaker baseline model can still be made viable through fine-tuning on a small but verified dataset. In practice, the question is not simply "should we fine-tune," but "which model type is being deployed." This distinction is crucial for any real-world rollout where both labelled data and expert review capacity are limited.

In summary, this research provides an early but tangible contribution towards equitable AI for health communication. Its primary output is not just a set of evaluated models, but a foundational workflow, from thematic data curation to human-centred evaluation, that can be adapted for other under-resourced languages and health domains. Future efforts should prioritise the development of larger, community-validated datasets and robust, infrastructure-aware deployment strategies to en-

hance the relevance, reliability, and public impact of these technologies.

A key barrier to deployment is the lack of infrastructure. The current approach assumes online access to inference endpoints, which is often not guaranteed in the rural settings where rabies risk is highest. Any real-world rollout will require offline-capable or low-bandwidth channels (e.g., SMS or USSD), along with human oversight for urgent escalations. Without this, technically strong models will not translate into practical access.

The workflow developed here is intentionally modular. Thematic extraction from authoritative sources, controlled generation of QA material, native-speaker correction for linguistic and cultural fit, and human-centred blind evaluation can, in principle, be repeated for other diseases and other South African languages. This is important for scalability, as it suggests that public health agencies and language communities can replicate the process for priorities such as maternal health or tuberculosis awareness without having to rebuild the entire methodology from scratch.

6 Limitations

While this pilot study demonstrates the potential of LLMs for Sepedi health communication, its findings are subject to several limitations that define a clear agenda for future work.

First, the primary constraint is the small, synthetic dataset of 60 question-answer pairs used for fine-tuning. Although curated for thematic relevance and linguistic accuracy, its limited size and synthetic origin may introduce translation artefacts and fail to capture the diversity of authentic, community-specific queries fully. This data scarcity likely explains why fine-tuning yielded only marginal improvements for a robust base model like GPT-4o, suggesting that the small-scale adaptation did not significantly enhance its strong foundational capabilities.

Second, the evaluation framework, while qualitatively rich, relied on a single expert reviewer. This approach, though consistent, introduces potential subjectivity and limits the generalisability of our findings. A more robust assessment would require a broader panel of reviewers and incorporate structured metrics for detecting factual inaccuracies or hallucinations, which are critical for safe deployment in a public health context.

Third, the project's scope was intentionally narrow, focusing on a single disease (rabies) and a single language (Sepedi). The direct applicability of these models and findings to other health domains or low-resource languages is not guaranteed without further adaptation and validation. Finally, the web-based prototype assumes consistent internet access, which remains a significant barrier in many of the rural communities this work aims to serve. By focusing on a low-resource language, this work addressed a substantial gap in the accessibility of digital health resources in South Africa. It explored a pathway toward more linguistically equitable AI.

However, practical deployment remains constrained by the availability of infrastructure. The current approach assumes access to internet-connected inference endpoints, which is not guaranteed in many rural communities with the highest risk of rabies exposure. This creates a gap between technical feasibility and actual delivery. Any real-world rollout would require offline-capable or low-bandwidth channels, such as SMS or USSD, or lightweight on-device responders, as well as a mechanism for human oversight in cases where medical escalation is time-sensitive.

The validated Sepedi rabies QA dataset described in this study will be made publicly available under an open licence and can be found at this [GitHub link. This is intended to support reproducibility and independent evaluation.

These limitations highlight several key directions for future research:

- **Dataset expansion through community participation:** Future work should prioritise the creation of larger and more diverse datasets to address the current data bottleneck. This can be achieved through participatory methods, including engagement with community health workers and local speakers to collect authentic questions and co-create culturally resonant answers.

- **Robust, multi-reviewer evaluation:** To improve reliability, future evaluation will adopt a multi-reviewer framework involving a panel of native speakers. Protocols for systematic hallucination detection and factual verification against trusted medical sources will complement this.

- **Accessible deployment models:** Infrastructure constraints are a practical barrier. A key

priority is to explore lightweight models suitable for offline or low-bandwidth deployment. This includes developing interfaces accessible via SMS or USSD to ensure equitable access for communities with limited internet connectivity.

- **Scaling to new domains and languages:** The established workflow was designed for reuse. Future projects will adapt this framework to other critical public health topics (e.g., maternal health, non-communicable diseases) and other under-resourced Southern African languages.

By addressing these areas, this work can be extended toward technically robust, culturally aligned, and practically accessible AI tools for public health communication across the African continent.

Acknowledgements

The authors gratefully acknowledge the support of the ABSA Chair of Data Science and the Data Science for Social Impact (DSFSI) Lab at the University of Pretoria. This work was supported by UK International Development and the International Development Research Centre (IDRC), Ottawa, Canada, under the AI4D Africa Program. DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI, and Meta.

We also extend our thanks for the baseline knowledge support provided by the National Institute of Communicable Diseases of the National Health Laboratory Services, South African National Department of Health.

References

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, and Minh Dinh. 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655.

Clément Christophe, Praveen K. Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco A. F. Pimentel, and Shadab Khan. 2024. Med42: Evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. *Preprint*, arXiv:2404.14779.

Crina Cotoc and Stephen Notaro. 2022. Race, zoonoses and animal-assisted interventions in paediatric cancer. *International Journal of Environmental Research and Public Health*, 19(13):7772.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Mack Makgatho, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner. 2021. Training cross-lingual embeddings for setswana and sepedi. *arXiv preprint arXiv:2111.06230*.

Unarine Makungo, Lehlohonolo Kumalo, Jacqueline Weyer, Tumiso Malatji, Lebitsi Ranoto, and Veerle Msimang. 2025. Epidemiological trends of animal bites and human rabies cases in limpopo, south africa, 2011–2023: A retrospective review. *Public Health Bulletin South Africa*, 6(3). Polokwane Mankweng Research Ethics Committee Ref: PMREC 30 October UL 2024/A.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2004.13842*.

Akira Murakami, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. ‹what is this corpus about?›: using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277.

Jacob Murel and Eda Kavlakoglu. 2024. What is topic modeling?

National Institute for Communicable Diseases. 2025. Rabies (disease index).

Ting Fang Tan, Kabilan Elangovan, Liyuan Jin, Jie Yao, Yong Li, Joshua Lim, Stanley Poh, Wei Yan Ng, Daniel Lim, Yuhe Ke, Nan Liu, and Daniel Shu Wei Ting. 2024. Fine-tuning large language model (LLM) artificial intelligence chatbots in ophthalmology and LLM-based evaluation using GPT-4. *Preprint*, arXiv:2402.10083.

World Health Organization. 2024. Rabies.

Traditional Readability Approaches to Sesotho and isiZulu: A (First) Overview

Johannes Sibeko
Nelson Mandela University
Gqeberha, South Africa
johanness@mandela.ac.za

Mthuli Buthelezi
University of Kwa-Zulu Natal
Mazisi Kunene Road, eThekwini, South Africa
mthulibuthelezi08@gmail.com

Abstract

This paper presents a conceptual overview of traditional readability metrics adapted for two South African Indigenous languages, isiZulu and Sesotho, which differ orthographically with conjunctive and disjunctive writing systems, respectively. Both languages are low-resource, lacking extensive corpora, lexicons, and pre-trained models necessary for automatic readability assessment. By critically examining these adaptations, we highlight the challenges of applying English-based metrics to morphologically complex African languages and emphasise the need for language-specific digital resources that reflect local linguistic structures. Our work aligns with ongoing efforts to develop and enhance language resources for under-resourced African Indigenous languages, thereby supporting their evolving presence and accessibility in the digital age, including contexts shaped by large language models.

1 Introduction

Reading proficiency remains a persistent challenge in Southern Africa. Although recent scholarship has examined common contributing factors such as inadequate teacher training and various learning difficulties, there remains a notable gap in the literature concerning the kinds of texts prescribed to learners and how these texts shape reading development (Sibeko, 2024a). That is, text readability, the ease with which texts can be read (Dahl, 2004, 2009) has received limited attention, especially in low-resource Indigenous languages such as those of Southern Africa.

This paper investigates existing research on readability in Sesotho and isiZulu, which remain significantly under-resourced in the digital language domain (Roux and Bosch, 2019). That is, in terms of Natural Language Processing (NLP), both Sesotho and isiZulu are considered resource-scarce languages (Mahloane and Trausan-Matu, 2015; Moors et al., 2018; Wills et al., 2020), particularly when evaluated against the availability of pre-trained models, basic language resources, corpora, and training data necessary for automating tasks (Marivate et al., 2020), such as text readability analysis (Cucu et al., 2014; Magueresse et al., 2020; Bansal et al., 2021; Zupon et al., 2021).

Before readability measures were developed for specific Southern African languages, studies in the South African context relied largely on traditional English-language formulas (see Grabar et al. (2014), Joubert and Githinji (2014), Sibanda (2014), Buthelezi (2017a), Leopeng (2019), De Wet (2021)). These formulas are calibrated to the U.S. education system, limiting their relevance for local contexts. Bargate (2012) attempted to align these metrics with South African grade levels, but only in English, as no comparable tools existed for Indigenous languages at the time. The emergence of readability measures for languages like Sesotho and isiZulu opens new possibilities for contextually appropriate readability assessment.

The adaptation of readability measures for low-resource languages rests on the premise that such measures must differ from those developed for high-resource languages. However, there is limited comparative research on how these metrics function across different Indigenous languages in Southern Africa. This paper addresses that gap by examining the application of traditional readability measures in Sesotho and isiZulu, asking whether readability can be assessed in the same way across these typologically and orthographically distinct languages.

We focus on Sesotho and isiZulu because, to our knowledge, they are the only South African Bantu languages with existing readability metrics. They also represent orthographically and structurally distinct languages, Sesotho being disjunctive and isiZulu conjunctive. These features pose challenges

for applying English-based readability formulas and underscore the need for language-specific adaptations suited to African linguistic contexts.

The remainder of this paper is structured as follows: Section 2 reviews related literature to contextualise our discussion. Section 3 examines the text properties of Sesotho that influence the applicability of traditional readability measures, while Section 4 considers comparable properties in isiZulu. Section 5 of this paper elaborates on the surface-level features of both isiZulu and Sesotho. Section 6 synthesises these insights, highlighting key differences in the development of readability measures across the two languages. Finally, Section 7 concludes the paper by summarising the findings of our comparative analysis.

2 Related Literature

Generally, readability can be approached from either a relative or an absolute perspective. The relative approach, as seen in Bunch et al. (2014) and Collins-Thompson (2014), frames readability in terms of reading and text difficulty, with comprehension varying by reader. In this view, readability depends on how well a specific reader understands a given text. On the other hand, the absolute approach quantifies linguistic features to estimate readability independently of individual readers. Generally, modern objective measures combine traditional readability formulas with more nuanced textual characteristics (Daelemans et al., 2017).

Previously, researchers have used various methods to investigate text readability, including traditional readability measures (Kondru, 2006; Feng, 2010; Janan, 2011; Bendová, 2021), machine learning (Sjöholm, 2012; Andova, 2017), Natural Language Processing (Gonzalez-Dios, 2016), deep learning (Alkaldi, 2022), and eye-tracking (Newbold, 2013). While we specify them in this article, we acknowledge that most of these methods overlap as they use some form of machine learning. Overall, these research projects have made significant contributions to the scholarship of text readability. As far as we are aware, for this article, there is a paucity in the employment of these methods to develop readability measures for the Indigenous official languages of South Africa. Thus, we discuss the traditional readability measures for Sesotho and isiZulu that are adapted from the English readability measures.

We are aware of readability measures developed for Afrikaans (see van Rooyen (1986), Jansen et al. (2017) and Du Plessis and Vos (2022)). Note that the Afrikaans VivA-Leesbaarheidsindeks' accessible at `https://viva-afrikaans.org/afdelings/tegnologie` presents possibly the only online or web-based platform for measuring text readability in Afrikaans or any other non-English official language of South Africa. The portal includes Combrink's (1992), which is described in detail by McDermid's (2007) and Jansen et al.'s (2017). It also includes the Flesch Reading Ease and Flesh-Kincaid Grade Level formulas for English.

In their study on the readability of medical texts translated into isiXhosa, Afrikaans, and English, Leopeng (2019) employed the Læsbarhetsindex (LIX) formula, operating under the assumption—shared by Naveed (2024, 1750)—that the formula's lack of language-specific constants renders it applicable to both English and foreign languages, such as our Indigenous languages. This assumption, however, risks overlooking the unique structural and orthographic features of individual languages. That is, the perceived simplicity and ease of application of the LIX formula do not guarantee its validity across typologically diverse languages. Indeed, the generalisability of traditional and more advanced readability measures must be empirically substantiated rather than taken for granted. For instance, the significance of tailoring methodological approaches to the linguistic and resource conditions of specific languages is clearly illustrated in the work of Grabar et al. (2014). Their cross-lingual investigation into the comprehension of medical terminology in French and isiXhosa highlights a stark contrast in methodological sophistication. That is, while French benefited from more intricate and resource-rich approaches, isiXhosa was limited to more basic methods due to constraints in linguistic resources.

3 Readability and Disjunctive Writing System

Despite a wide usage, a recent assessment of the Sesotho Basic Language Resource Kit (BLARK) revealed a severe shortage of digital language resources, confirming its classification as a low-resource language (Roux and Bosch, 2019; Sibeko and Setaka, 2022). Unfortunately, this lack of foundational language resources restricts the use of

more advanced readability measures that go beyond surface-level features.

While the development of Sesotho readability metrics is detailed elsewhere (see Sibeko (2024b) and Sibeko and van Zaanen (2024a)), the discussion therein is not contextualised in the discussion of the writing system of Sesotho or other languages in Southern Africa.

Sesotho employs a primarily disjunctive orthography, in which words and morphemes are written separately. This has important implications for readability, as word segmentation affects both visual parsing and syntactic interpretation. A brief overview of the Sesotho writing system is necessary to highlight how text complexity, word length, and sentence structure influence readability measures.

Sesotho is written in two widely recognised orthographies, which are generally associated with the countries in which they are predominantly used. These have been referred to as the Lesothan Sesotho (LS) orthography and the South African Sesotho (SAS) orthography (Mohasi and Mashao, 2005; Motjope-Mokhali et al., 2020). Although there is limited literature on Zimbabwean Sesotho, studies have identified at least minor differences between the Lesothan and South African variants in terms of orthography (Eldredge, 2002; Makutoane, 2022), vocabulary (Motjope-Mokhali et al., 2020), and syntax (Sekere, 2004). In addition, Lesotho and South Africa each maintain distinct spelling conventions for Sesotho (Matlosa, 2017; Bardill and Cobbe, 2019).

Additionally, Sibeko and van Zaanen (2024a) pay special attention to the semi-vowels: (w and y), the lateral consonant: (l), the glide: (g), and as used in the SAS orthography (Demuth, 2007; Nkolola-Wakumelol et al., 2012; Anghelescu, 2016; Chokoe, 2020). The SAS orthography was preferred in the development of the readability measures for Sesotho. Thus, these orthographical differences are important to note in the use of Sibeko and van Zaanen's (2024a) readability measures.

Understanding the disjunctive nature of Sesotho, along with the variations between Lesothan and South African orthographies, is essential for accurate readability assessment. That is, word segmentation, orthographic conventions, and the treatment of semi-vowels and glides directly influence text complexity and parsing. Consequently, readability measures for Sesotho must account for these orthographic features to provide linguistically informed and reliable evaluations.

4 Readability and Conjunctive Writing System

Like Sesotho, isiZulu is an agglutinative language. However, unlike Sesotho's disjunctive writing system, isiZulu employs a conjunctive system, in which short morphemes attach to a stem to form single orthographic words (Land, 2016). As a result, isiZulu words are often long and morphologically complex; one written word in isiZulu can correspond to multiple words in English. This agglutinative structure and conjunctive orthography have important implications for reading and text readability, as processing each morphologically complex word requires more cognitive effort than reading simpler, disjunctive text.

Languages such as isiZulu have largely been excluded from the digital revolution due to the absence of extensive corpora, comprehensive lexicons, and supporting software tools (Buthelezi, 2025). This digital scarcity means that few tools, including readability formula tools, have been developed for isiZulu. Consequently, assessing the readability of isiZulu texts remains a significant challenge, since word and sentence segmentation differ fundamentally from languages for which conventional metrics were created. Complex internal structure of words in isiZulu and related Bantu languages poses major difficulties for building digital tools such as readability formulas (Prinsloo and Schryver, 2002). Despite its large speaker base, isiZulu has relatively scarce scholarly and technological support – confirming its status as a low-resource language.

IsiZulu uses a Latin alphabet writing convention and has a close grapheme-phoneme relationship. In addition, it has five vowels a, e, i, o, u, and two non-phonemic vowels ɛ, and ɔ(Buthelezi, 2024). Further, IsiZulu has 49 distinct sounds (Simelane et al., 2024).

IsiZulu exhibits a relatively constrained set of permissible letter combinations compared to languages with more diverse orthographic patterns. The language's syllable structure limitations result in frequent repetition of short letter sequences throughout texts (Land, 2016).

Research consistently demonstrates that isiZulu texts are read more slowly than comparable texts in languages with different orthographic and morphological characteristics like Sesotho. For instance,

eye-tracking studies reveal that skilled isiZulu readers exhibit longer fixation durations, shorter saccade lengths, and more frequent regressions compared to readers of languages with less complex morphological systems (Land, 2016). Moreover, isiZulu readers appear to employ different comprehension strategies than readers of more analytic languages. The morphological complexity of isiZulu words necessitates small-grain processing approaches where readers must attend to individual morphemes rather than processing entire words as units (Land, 2016). This processing approach affects reading speed and may influence comprehension accuracy, thereby directly shaping text readability.

The conjunctive writing system and agglutinative morphology of isiZulu result in long, morphologically complex words that demand fine-grained processing at the morpheme level. This structural complexity slows reading, increases cognitive load, and affects comprehension, as shown by eye-tracking studies (Land, 2016). Consequently, any readability measure for isiZulu must account for its conjunctive orthography and internal word structure to provide meaningful and linguistically informed evaluations of text difficulty.

5 Surface-level Features

Surface-level features provide a foundational measure of text complexity. For low-resource languages like Sesotho and isiZulu, they offer a reliable starting point for readability assessment, enable comparison across disjunctive and conjunctive writing systems, and support the adaptation of formulas to capture meaningful differences in text difficulty. We focus on at least four main features, including syllable, word, sentence, and tone features.

5.1 Syllable information

In early reading instruction for languages such as Sesotho and isiZulu in South African schools, learners are typically introduced to syllabic phonics, where they learn to decode sound–symbol correspondences through patterns such as a, e, i, o, u and ba, be, bi, bo, bu. This syllable-based decoding approach aligns closely with the consonant-vowel (CV) syllabic structure of these languages. Although not yet empirically tested, our initial assumption is that while syllable structures may aid early reading instruction because they are teachable and recognisable, they may have a limited impact on reading fluency and the computational assessment of text readability in Sesotho. While both Sesotho and isiZulu share similar syllable structures such as vowel-only, consonant-vowel - allowing up to five onset consonants and one vowel per syllable (Land, 2015), their treatment in computational readability analysis differs significantly. Sesotho benefits from a rule-based syllabification system that enables the automatic extraction of 16 distinct syllable types and possible syllable breaks, facilitating its integration into automated traditional readability measures (Sibeko and van Zaanen, 2024b). In contrast, isiZulu, despite its predominantly (V), (CV), and syllabic nasal (N) syllable structures (Buthelezi, 2017b), lacks an automated syllabification tool, attesting to the low-resource concern we raised earlier in this article. This is particularly limiting given that isiZulu words often contain four or more syllables (Buthelezi, 2017a), a feature that, by English readability standards, would indicate higher complexity. Because of this restriction - the lack of a syllabification system - the automation of readability assessment is limited.

5.2 Word information

Another typical characteristic of traditional readability measures is their focus on word lengths. In Sesotho, the lengths of words can be affected by the differences in orthographies. That is, SAS orthography uses longer words through the use of digraphs in places where the LS orthography uses single letters. For instance, the SAS orthography uses the digraph (tj) where the LS orthography uses (c).

While variations such as preferences for specific single letters (e.g., l and d) do not influence the outcomes of the measurements, the use of single letters in one orthography compared to digraphs in another (e.g., c, š vs tj, sh) may impact linguistic properties such as average word length.

IsiZulu's agglutinative morphology represents perhaps the most significant surface-level feature affecting readability. Single words can contain multiple prefixes, stems, and suffixes that each contribute distinct grammatical or semantic information (Keet and Khumalo, 2017). This morphological richness allows for the expression of complex ideas within individual words but creates challenges for reading comprehension and text processing. The agglutinative nature of isiZulu means that

readers must parse multiple morphemes within single orthographic words to extract complete meaning. This parsing process requires different cognitive strategies than those employed in languages with more analytic word structures. Readers must identify morpheme boundaries, recognise individual morpheme meanings, and integrate these components to derive overall word meaning (Land, 2015).

5.3 Sentence length

While it is generally accepted that shorter sentences improve readability (Duvenhage et al., 2017), there are no widely agreed-upon optimal sentence lengths for Sesotho and isiZulu. Like isiZulu, Sesotho is an agglutinative language, meaning that it constructs meaning through sequences of morphemes. However, Sesotho uses a disjunctive orthography, in contrast to isiZulu's conjunctive system. This orthographic difference significantly affects how sentence length is measured. In isiZulu, grammatical elements such as subject, tense, and object are often combined into a single word, making sentences appear shorter in word count while masking underlying complexity. In Sesotho, these elements are written separately, inflating word counts and making sentences appear longer. As a result, applying English-based readability measures, many of which rely on word counts to assess sentence length, can lead to distorted assessments in both languages. While such measures may be more easily applied to Sesotho, they still fail to capture the full grammatical complexity inherent in the language.

5.4 Tone Marking

Sesotho carries meaning by use of tone (Raborife et al., 2015). As a result, the same words, as used in examples 1 and 2 for Lesothan Sesotho and example 3 for South African Sesotho below, can carry different meanings based on the tone used during reading. Examples 1 and 2 illustrate how tone can be carried through vowels in the Lesothan Sesotho.

(1) *Ke bóna batho.* – LS
'I see people.'

(2) *Ké boná batho.* – LS
'They are the people.'

(3) *Ke bona batho.* – SAS
'I see people.' or 'They are the people.'

In Lesothan Sesotho, diacritics are more commonly used to indicate tone, which is phonemically distinctive and changes meaning.

According to Dickens (1978) and Matlosa (2017), seven phonological vowels are recognised across the Sotho language group: ɑ, e, i, o, u, ɛ, and ɔ. However, in writing, the South African Sesotho (SAS) orthography reduces these to five morphological vowels: a, e, i, o, and u.

In addition to tonal distinctions carried by vowels, Sesotho also uses four consonants to mark tone: m, n, ŋ, and ɲ. These tonal cues are typically inferred from context rather than explicitly marked in writing.

Similar to Sesotho, the absence of tone marking in written isiZulu represents a significant challenge for text readability, as tonal variation can change meaning even when the orthographic form remains identical. Examples 4 and 5 illustrate how unmarked tone in writing can be represented.

(4) *Le nkomo ingahlatshwa.* – Land (2015, 165)
'This cow must not be slaughtered.'

(5) *Le nkomo ingahlatshwa*
'This cow can be slaughtered.'

Although both sentences are spelled the same, their meanings differ entirely depending on tone. While the ambiguity may not affect reading fluency, this ambiguity requires readers to rely on contextual and inferential processing to determine the intended interpretation. Consequently, unmarked tone increases the cognitive load during reading, making decoding slower and more effortful. While traditional readability measures such as those adapted to Indigenous languages do not account for processing and comprehension difficulties, we acknowledge that they have an influence that should be studied in more depth for the languages in question.

6 Discussion

6.1 Sesotho Readability Measures

Recent efforts to adapt traditional English readability formulas for Sesotho have resulted in a set of language-specific metrics calibrated to the linguistic features of the language (Sibeko and van Zaanen, 2024a). These measures listed in Table 1 include adaptations of well-known formulas such as the Flesch–Kincaid Grade Level (FKGL), Flesch Reading Ease (FRE), Gunning Fog Index (GFI),

Table 1: Readability measures and corresponding adapted Sesotho formulas (Sibeko and van Zaanen, 2024a).

Measure	Formula
$FKGL_{Sesotho}$	$=-14.08905 + 0.43405(\frac{\#words}{\#sentences}) + 5.86314(\frac{\#syllables}{\#words})$
$FRE_{Sesotho}$	$= 209.3286 - 1.7930(\frac{\#words}{\#sentences}) - 46.6548(\frac{\#syllables}{\#words})$
$SMOG_{Sesotho}$	$= 0.28788 + 0.68741(\sqrt{\#polysyllabic - words * (\frac{30}{\#sentences})})$
$GFI(1)_{Sesotho}$	$= -4.30942 + 0.28610(\frac{\#words}{\#sentences}) + (\frac{\#complex-words}{\#words})$
$GFI(2)_{Sesotho}$	$= -1.77916 + 0.40861((\frac{\#words}{\#sentences}) + 30.9982(\frac{\#complex-words}{\#words}))$
$CLI_{Sesotho}$	$= -3.683470 + 0.038782(\frac{\#letters}{\#samples} * 100) - 0.727659(\frac{\#sentences}{\#samples} * 100)$
$ARI_{Sesotho}$	$= -13.66031 + 2.87106(\frac{\#letters}{\#words}) + 0.49323(\frac{\#words}{\#sentences})$
$LIX_{Sesotho}$	$= 0.46038 + 1.14736(\frac{\#words}{\#sentences}) + 0.60841(\frac{\#long-words}{\#words} * 100))$
$RIX_{Sesotho}$	$= 0.02180 + 0.76883(\frac{\#long-words}{\#sentences})$
$DCI_{Sesotho}$	$= 4.66547 + 0.14199(\frac{\#words}{\#sentences}) + 0.03264(\frac{\#difficult-words}{\#words} * 100)$

Simple Measure of Gobbledygook (SMOG), Coleman–Liau Index (CLI), Automated Readability Index (ARI), LIX, RIX, and Dale–Chall Index (DCI). Each formula has been re-calibrated using corpus-based data from Sesotho texts and evaluated in relation to linguistic units such as syllables, polysyllabic words, and morphologically complex words.

The adapted formulas maintain the basic structure of their English counterparts but apply coefficients derived from statistical analysis of Sesotho data. For instance, $FKGL_{Sesotho}$ and $FRE_{Sesotho}$ retain the core logic of sentence and syllable length influencing grade level or ease, but the weights reflect sentence structure and word formation patterns more typical of Sesotho. Similarly, $GFI_{Sesotho}$ and $SMOG_{Sesotho}$ capture syntactic and morphological complexity by focusing on sentence length and the presence of polysyllabic or complex words—units that require tailored identification strategies in agglutinative languages.

Notably, $CLI_{Sesotho}$ and $ARI_{Sesotho}$ incorporate letter and sample-based metrics, useful in typologically disjunctive orthographies like Sesotho, where word segmentation more closely aligns with English. $LIX_{Sesotho}$ and $RIX_{Sesotho}$ focus on long words and sentence structure, while $DCI_{Sesotho}$ introduces a focus on "difficult words" based on frequency or familiarity, providing a more lexically sensitive measure.

Together, these adapted formulas provide a starting point for standardised readability assessments in Sesotho.

6.2 The IsiZulu Text Readability Calculator (ITRC)

The IsiZulu Text Readability Calculator (ITRC) is a corpus-based software programme that estimates text readability based on word frequency. Its algorithm calculates a readability index by summing the frequency values of words in the input text. The ITRC has a built-in corpus. In gauging readability, words in the input text are compared to the ones in the built-in corpus. Thus, the value of each word in the input text is summed and the result is divided by the number of words in the text. Words that occur more frequently in the built-in corpus are assumed to be easier to read, while rarer words signal higher difficulty.

The ITRC also provides both upper and lower readability thresholds and categorises texts into ‘beginner’, ‘intermediate’, or ‘advanced’ levels. Users can upload isiZulu text samples from their device and compute the readability score by clicking the ‘calculate readability index’ button. The ITRC interface displays a score intended to support human judgment of readability.

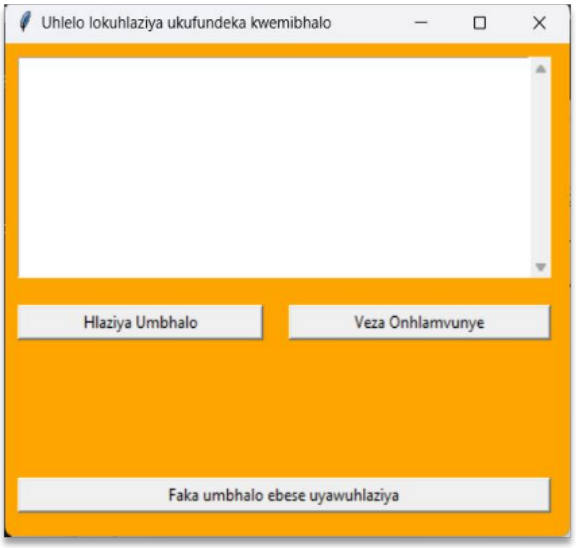


Figure 1: IsiZulu text readability tool

6.2.1 *Uhlelo lokuhlaziya ukufundeka kwemibhalo* (The programme to analyse the readability of texts)

This programme determines readability by calculating the average word length in a given isiZulu text. Texts with an average word length above five characters are classified as difficult to read, while those with five characters or fewer are considered easier. The value 5 sits at the statistical mean and perceptual midpoint for isiZulu word length (Buthelezi, 2017a), hence its selection. Although word length is a simple metric, it serves as an initial proxy for textual difficulty, particularly in agglutinative languages like isiZulu.

Like the ITRC, this tool is designed to assist rather than replace human judgment That is, it supports semi-automated readability measurement.

According to (Buthelezi, 2025), to extend this tool to other African languages, its algorithm would need to be adapted to reflect the average word length of the target language. While average word length is a generalisable metric, language-specific readability thresholds should be established through various means such as corpus analysis.

6.2.2 The GUI of *uhlelo lokuhlaziya ukufundeka kwemibhalo*

The Graphical User Interface (GUI) of *uhlelo lokuhlaziya ukufundeka kwemibhalo*, presented in Figure 1, is designed for ease of use. At its centre is a scrollable text box where users can type or paste text. Users may also upload a file using the '*faka umbhalo ebese uyawuhlaziya*' (upload and analyse text) button, which displays the content in the text box. Below the input area are three main buttons:

- 'hlaziya umbhalo' (analyse text),
- 'veza onhlamvunye' (show n-grams); and
- result and n-gram labels.

The programme represents a significant step toward computational readability assessment in isiZulu. While its word-length metric offers a basic objective measure, readability remains multifaceted and is best assessed through models that capture the unique linguistic and structural properties of each African language. In agglutinative and conjunctively written languages like isiZulu, word length alone cannot adequately reflect the cognitive effort required for reading.

7 Conclusion

This paper has offered a conceptual overview of readability measures for isiZulu and Sesotho, two structurally distinct African indigenous languages. While readability plays a critical role in education, language accessibility, and digital resource development, it is often assessed using metrics designed for Indo-European languages. We argue that such metrics are not easily transferable to agglutinating, low-resourced languages like isiZulu and Sesotho.

By comparing the application of existing readability tools in these languages, we have highlighted the importance of language-specific approaches that reflect typological and orthographic differences. By examining how readability is approached in isiZulu and Sesotho, this paper contributes to broader conversations on the digital representation of linguistic structures in low-resourced African languages. Our analysis underscores the need for language resources that are attuned to local linguistic realities—particularly as large language models increasingly shape digital language technologies. In this way, the paper speaks directly to the workshop's theme of language resources in the age of large language models.

Although this paper does not propose new tools, it provides a foundation for future research and development. In the case of isiZulu, future enhancements—such as refining classification thresholds,

incorporating syllable-based metrics, and expanding on morphological and word density (Buthelezi, 2025)—could improve readability assessments and facilitate cross-language adaptation. For Sesotho, ongoing research involving school teachers in human evaluation projects will help refine existing measures through practical classroom engagement.

Together, these efforts point to the growing importance of context-sensitive approaches in the development of readability tools for African languages—tools that must be grounded in linguistic realities rather than imposed from external models.

Limitations

This article presents a conceptual overview rather than an empirical evaluation. While we reflect on existing tools and theoretical frameworks, we do not present new experimental data or corpus-based analyses. As such, our observations about the performance and applicability of readability measures remain indicative and would benefit from further empirical validation.

Second, the article does not include end-user evaluation. Readability is ultimately experienced by readers, yet this study does not incorporate their perspectives directly. As noted above, however, ongoing research on Sesotho involves human evaluation with school teachers, which will inform the refinement of existing measures.

Third, although we focus on isiZulu and Sesotho—two widely spoken South African languages—the comparative scope is limited. Our findings may not generalise to other Indigenous African languages, particularly those with different orthographic or morphological characteristics.

Finally, we have not yet explored advanced computational approaches such as deep learning or psycholinguistic modelling, which are increasingly applied in high-resource contexts. We hope that as more foundational language resources are developed for Sesotho and isiZulu, future research will be able to experiment with these methods to enhance the accuracy and relevance of readability assessments in African languages.

References

Wejdan Alkaldi. 2022. *Enhancing text readability using deep learning techniques*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Andrejaana Andova. 2017. *Assessment of text readability using statistical and machine learning approaches*. Ph.D. thesis, University of Ljubljana.

Andrei Anghelescu. 2016. Distinctive transitional probabilities across words and morphemes in Sesotho. *University of British Columbia Working Papers in Linguistics*, 3(1):1–17.

Rachit Bansal, Himanshu Choudhary, Ravneet Punia, Niko Schenk, Jacob L Dahl, and E Pagè-Perron. 2021. How low is too low? a computational perspective on extremely low-resource languages. *arXiv:2105.14515*, preprint. https://aclanthology.org/2021.acl-srw.5.

John E Bardill and James H Cobbe. 2019. *Lesotho: dilemmas of dependence in Southern Africa*. Routledge, Cape Town, SA.

Karen Bargate. 2012. The readability of managerial accounting and financial management textbooks. *Meditari Accountancy Research*, 20(1):4–20.

Kiára Bendová. 2021. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal of linguistics*, 72(2):477–487.

George C Bunch, Aída Walqui, and P David Pearson. 2014. Complex text and new common standards in the united states: Pedagogical implications for English learners. *Tesol Quarterly*, 48(3):533–559.

Mthuli Buthelezi. 2017a. A corpus analysis of readability and the degree of complexity of isiZulu texts. *Uploaded on Research Gate*, pages 1–22. Available at: https://doi.org/10.13140/RG.2.2.22908.74888.

Mthuli Buthelezi. 2017b. Isizulu and its readability levels: Investigating the applicability of selected readability formulae to isizulu texts. Master's thesis, University of KwaZulu-Natal, Durban.

Mthuli Buthelezi. 2024. A description of articulated isizulu and chinese vowels for pedagogical purposes. *Journal of African Language and Culture Studies*, 6:18–67.

Mthuli Buthelezi. 2025. *IsiZulu and algorithms: A language-agnostic approach to developing software for African languages*. Uluntu Algorithms, eThekwini.

Sekgaila Chokoe. 2020. Spell it the way you like: The inconsistencies that prevail in the spelling of Northern Sotho loanwords. *South African Journal of African Languages*, 40(1):130–138.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Johannes Gert Hendrik Combrink. 1992. *Hoe om'n verslag te skryf*. Tafelberg.

Horia Cucu, Andi Buzo, Laurent Besacier, and Corneliu Burileanu. 2014. SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian. *Speech Communication*, 56:195–212.

Walter Daelemans, Orphée De Clercq, and Véronique Hoste. 2017. Stylene: An environment for stylometry and readability research for Dutch. In *CLARIN in the Low Countries*, pages 195–210. Ubiquity Press, London.

Osten Dahl. 2004. *The growth and maintenance of linguistic complexity (Studies in Language Companion Series)*, volume 71. John Benjamins Publishing, Amsterdam.

Osten Dahl. 2009. Testing the assumption of complexity invariance: the case of elfdalian and swedish. In G Sampson, D Gil, and P Trudgill, editors, *Language Complexity as an Evolving Variable*, page 50–63. Oxford: Oxford University Press.

Anneliese De Wet. 2021. *The development of a contextually appropriate measure of individual recovery for mental health service users in a South African context*. Thesis, Stellenbosch University.

Katherine Demuth. 2007. Sesotho speech acquisition. In S McLeod, editor, *The international guide to speech acquisition. Thomson Delmar Learning*, pages 526–538. Thomson Delamar Learning, New York, USA.

Patrick Dickens. 1978. A preliminary report on Kgalagadi vowels. *African Studies*, 37(1):99–106.

Carmen Du Plessis and Elize Vos. 2022. Selfgerigte tekskeuse vir afrikaans huistaal in die intermediêre fase met behulp van 'n leesbaarheidsindeks. *Journal for Language Teaching*, 56(2).

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017. Improved text language identification for the South African languages. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 214–218. IEEE.

Elizabeth A Eldredge. 2002. *A South African kingdom: The pursuit of security in nineteenth-century Lesotho*, volume 78. Cambridge University Press, Camridge.

Lijun Feng. 2010. *Automatic readability assessment*. New York: City University of New York.

Itziar Gonzalez-Dios. 2016. *Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures*. Phd thesis, University of the Basque country.

Natalia Grabar, Izak Van Zyl, Retha De la Harpe, and Thierry Hamon. 2014. The comprehension of medical words - cross-lingual experiments in french and xhosa. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 334–342.

Dahlia Janan. 2011. *Towards a new model of readability*. Ph.D. thesis, University of Warwick.

Carel Jansen, Rose Richards, and Liezl Van Zyl. 2017. Evaluating four readability formulas for Afrikaans. *Stellenbosch Papers in Linguistics Plus*, 53:149–166.

Karin Joubert and Esther Githinji. 2014. Quality and readability of information pamphlets on hearing and paediatric hearing loss in the gauteng province, South Africa. *International journal of pediatric otorhinolaryngology*, 78(2):354–358.

C Maria Keet and Langa Khumalo. 2017. Grammar rules for the isizulu complex verb. *Southern African linguistics and applied language studies*, 35(2):183–200.

Jagadeesh Kondru. 2006. *Using part of speech structure of text in the prediction of its readability*. Ph.D. thesis, The University of Texas at Arlington.

Sandra Land. 2015. Reading and the orthography of isiZulu. *South African Journal of African Languages*, 35(2):163–175.

Sandra Land. 2016. Automaticity in reading isizulu. *Reading & Writing-Journal of the Reading Association of South Africa*, 7(1):1–13.

Makiti Thelma Leopeng. 2019. *Translations of informed consent documents for clinical trials in South Africa: Are they readable?* Thesis, University of Cape Town.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *ArXiv*, abs/2006.07264. https://api.semanticscholar.org/CorpusID:219636137.

Malefu Justina Mahloane and Stefan Trausan-Matu. 2015. Metaphor annotation in Sesotho text corpus: Towards the representation of resource-scarce languages in NLP. In *The 20th International Conference on Control Systems and Computer Science*, pages 405–410, New York. IEEE.

Tshokolo J Makutoane. 2022. 'The people divided by a common language': The orthography of Sesotho in Lesotho, South Africa, and the implications for Bible translation. *HTS Teologiese Studies/Theological Studies*, 78(1):9.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*.

Litšepiso Matlosa. 2017. Sesotho orthography called into question: the case of some Sesotho personal names. *Nomina Africana: Journal of African Onomastics*, 31(1):51–58.

Heyns J McDermid. 2007. Readability statistics for afrikaans. Paper read at the LSSA/SAALT/SAALA Joint Annual Conference, North-West University, Potchefstroom, 4 July 2007.

Lehlohonolo Mohasi and Daniel Mashao. 2005. Phonetization for text-to-speech synthesis in Sesotho. In *The sixteenth annual symposium of the Pattern Recognition Association of South Africa*, pages 121–122, Langebaan, South Africa. Citeseer.

Carmen Moors, Illana Wilken, Tebogo Gumede, and Karen Calteaux. 2018. Human Language Technology audit 2017/18. Available at: https://sadilar.org/index.php/en/2-general/284-health-resources [Accessed: 14 May. 2023].

Tankiso Lucia Motjope-Mokhali, Ingeborg M Kosch, and Munzhedzi James Mafela. 2020. Sethantso sa Sesotho and Sesuto-English dictionary: A comparative analysis of their designs and entries. *Lexikos*, 30:1–17.

Muhammad Shumail Naveed. 2024. Readability of wikipedia pages on covid-19. *Universal Access in the Information Society*, pages 1–12.

Neil Newbold. 2013. *New Approaches for Text Readability*. Thesis, University of Surrey.

Mildred Nkolola-Wakumelol, Liketso Rantsoz, and Keneiloe Matlhaku. 2012. Syllabification of consonants in Sesotho and Setswana. In H S Nginga-Koumba-Binza and S Bosch, editors, *Language Science and Language technology in Africa: Festschrift for Justus C. Roux*, pages 10–13. Sun Express, Stellenbosch, South Africa.

Daniele Prinsloo and Gilles-Maurice De Schryver. 2002. Towards an 11 x 11 array for the degree of conjunctivism/disjunctivism of the south african languages. *Nordic Journal of African Studies*, 11(2):249–265.

Mpho Raborife, Sigrid Ewert, and Sabine Zerbian. 2015. Improving a tone labeling algorithm for sesotho. *Language Resources and Evaluation*, 49(1):19–50.

Justus C Roux and Sonja E Bosch. 2019. Preserving and developing indigenous languages in the South African context. In *Proceedings of the Language Technologies for All (LT4All), Paris, France*, pages 97—100, Paris. European Language Resources Association.

Ntaoleng Belina Sekere. 2004. *Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa*. Thesis, University of South Africa.

Lucy Sibanda. 2014. The readability of two grade 4 natural sciences textbooks for South African schools. *South African Journal of Childhood Education*, 4:154–175.

Johannes Sibeko. 2024a. Exploring the readability of sesotho grade 12 examination texts using english readability metrics. *Southern African Linguistics and Applied Language Studies*, 42(sup1):S271–S284.

Johannes Sibeko. 2024b. *Measuring Text Readability in Sesotho*. Ph.D. thesis, North-West University, Potchefstroom, South Africa.

Johannes Sibeko and Mmasibidi Setaka. 2022. Sesotho BLARK content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11.

Johannes Sibeko and Menno van Zaanen. 2024a. Adapting nine traditional text readability measures into sesotho. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 66–76, Torino, Italia. ELRA and ICCL.

Johannes Sibeko and Menno van Zaanen. 2024b. Developing and testing syllabification systems for south african sesotho. *Language Resources and Evaluation*, pages 1–16.

Fikile Winnie Simelane, Jaclyn de Klerk, and Elizabeth Henning. 2024. Challenges of teaching isizulu reading in the early grades. *Southern African Linguistics and Applied Language Studies*, 42(sup1):S35–S52.

Johan Sjöholm. 2012. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. Thesis, Linköpings Universitet.

Rien van Rooyen. 1986. Eerste afrikaanse leesbaarheidsformules. *Communication*, 12(1):59–69.

Simone Wills, Pieter Uys, Charl Johannes Van Heerden, and Etienne Barnard. 2020. Language modeling for speech analytics in under-resourced languages. In *Interspeech 2020*, pages 4941–4945.

Andrew Zupon, Evan Crew, and Sandy Ritchie. 2021. Text normalization for low-resource languages of Africa. *preprint*, arXiv:2103.15845.

An Ideational Analysis and Integration of African Folktale in Science, Technology, and Education

Malete Elias Nyefolo
Department of African Languages, University of the Free State,
P.O. Box 339, Bloemfontein, South Africa
Tel. 051 401 2178, Fax. 051 401 2332
maleteen@ufs.ac.za

Abstract

Folktales are literary forms that reveal the soul of any society; they express its wishes, desires, hopes, and beliefs about the world. They have fictional characters and situations, mostly oral traditions, before they were written down. According to Cynthia McDaniel (1993), folktales can be used in all disciplines to convey knowledge and communicate ideas; they serve as an inherent vehicle for intergenerational communication that prepares and assigns roles and responsibilities to different generations in their communities. They are more pedagogic devices and less literary pieces. They cultivate universal values such as compassion, generosity, and honesty while disapproving of attributes such as cruelty, greed, and dishonesty. To illustrate McDaniel's claims, this paper will firstly use the ideational metafunctional framework found in Systemic Functional Linguistics, which expresses the clausal experiences and content from a grammatical perspective, coupled with syntagmatic analysis, which describes the text (folktale) in chronological order as reported by the storyteller. Secondly, the presentation will use a textual metafunctional framework that fulfills the thematic function of the clause, coupled with the paradigmatic analysis where the folkloristic text's patterns are regrouped more analytically to reveal the text's latent content, or theme. The Voyant Tool, a web-based text reading and analysis environment designed to facilitate the analysis of various text formats, was used to extract and analyze data from a Sesotho folktale to illustrate how folktales may be integrated with technology for research and educational purposes. This paper employed a descriptive research design that incorporates qualitative (content analysis) and quantitative (statistical analysis) methodologies to analyze and interpret the story. It is observed, through the Voyant tool, that the story is built out of 191 Sesotho word formations, and through the ideational analysis, that the storyteller employed more material process types than mental process types, and lastly, with the textual interpretation, indicating the value of oral literature in our daily lives as well as the significant role folktales may play in interpreting sociopolitical events in contemporary communities.

Keywords: Ideational metafunction; Syntagmatic analysis; Paradigmatic analysis; Sociopolitical interpretation, Sesotho folktales, Voyant Tool

1 Introduction

Rosenberg (1997) describes a folktale as a story that, in its plot, is pure fiction and has no location in either time or space. It is a symbolic way of presenting the different means by which human beings cope with the world in which they live. It concerns people, their loyalty, or common folk, and or animals that speak and act like people. They focus on the behaviour of the individual and are not usually concerned with the human being's relationship to divinity. They are literary forms that reveal the soul of any society; they express their wishes, desires, hopes, and beliefs about the world. They are often ancient, feature fictional characters, and were mostly passed down through oral traditions before being written down. (Finnegan:1970). In terms of their significance, McDaniel (1993) suggests that folktales can be utilized in all disciplines to convey knowledge and communicate ideas, helping us achieve social maturity, academic success, and moral compliance. Dorji Penjore (2005) further states that Folktales serve as an inherent vehicle for intergenerational communication that prepares and assigns roles and responsibilities to different generations in their communities. They are more pedagogic devices and cultivate universal values such as compassion, generosity, and honesty while disapproving of attributes such as cruelty, greed, and dishonesty. Moleleki (1988) continues to say

that folktales serve to keep our culture and belief systems as they portray the existence of another world yonder, that there is life after death, and there is someone superior to man. Education is the process by which society deliberately and actively transmits its accumulated knowledge, skills, and values from one generation to another. In the African context, traditional informal education was well conceived and pursued collectively for the good of the community. (Wiysahnyuy, LF, and Valentine, NB (2023)). It is important to integrate folktales into our curriculum as they form part of narrative prose and can analysed as such. They can be used to stimulate creative writing, nurture oral presentations, and assist in developing critical analysis of life situations. They help us to visit the primitive worlds, understand spiritual worlds, and cope with the present. In them, we get our identity, social cohesion, and moral upliftment; they promote the 'Botho' principles in general. Mweti (1999) proposed that the content of African folktales can be used to support pupil well-being in five ways: to build bridges between cultures through the shared tradition of storytelling; to engage pupils' emotions; to help pupils to recognize the shared experiences and problems faced by people from different cultural backgrounds; to serve a cathartic effect by allowing the safe expression of emotions and to allow pupils to confront fears and solve problems "at one step removed"; and to help pupils to make sense of their worlds. The application of technology in folktale studies has opened new avenues for research, preservation, and analysis of African oral traditions. Digital tools such as text-analysis software (AntConc, NVivo, Voyant Tools) enable scholars to identify recurring motifs, linguistic features, and evaluative expressions within folktale narratives, enhancing both linguistic and cultural interpretation. Corpus building allows for the systematic digitization and organization of folktales across languages and regions, facilitating comparative studies on themes, narrative structures, and moral values. Furthermore, digital mapping techniques—using platforms like StoryMapJS or ArcGIS StoryMaps—support the visualization of folktale variants, illustrating their diffusion and adaptation among different African communities. Advances in machine learning and natural language processing further enable the automated classification and thematic analysis of folktales, contributing to a deeper understanding of African worldviews and identity formation. Collectively, these technological approaches not only preserve Indigenous Knowledge Systems (IKS) but also promote accessibility, multilingual scholarship, and innovative pedagogical integration of folktales in education. Several scholars of African Languages, just to mention a few, have researched the structure and significance of folklore in general: Nakin (2017) analysed one Sesotho Folktale called 'Ngwana ya Kgwedi Sefubeng', using a deconstruction approach and concluded that binary oppositions can be used to analyse Sesotho folktales. Masowa (2024) explored the relevance of folktales in the current generation, where he uses the folktale 'Tselane le Dimo, using the Functionalist approach, and demonstrated that folktales are relevant to the current societies. Phindane (2019) explored Propp's functional approach in Sesotho folktales and observed that Sesotho folktales comply with 7 functions of Propp's 32 functions. The application of the Voyant tool, the metafunctional approach to Sesotho folktales, has not been explored adequately, if there are any attempts. The main purpose of this study is three-fold: First, to apply Systemic Functional Linguistics (SFL) to examine one Sesotho folktale, an integration into science, where the focus is on the ideational metafunction that expresses the clausal experiences. Secondly, to explore the thematic function of the folktales, an integration of folktales into education, where the folktale is used for the socio-political interpretation, and thirdly, to analyse the folktale using the Voyant Tool, the digital text analysis software to identify linguistic features, and evaluative expressions within the folktale narrative. This paper employed a descriptive research design that incorporates both qualitative (content analysis) and quantitative (statistical analysis) methods to analyse and interpret the folktale. Data from the folktale was collected through the online Voyant tool, a semi-automated data analysis, which assisted with word formation categories and word count, as well as identifying the distribution of the folktale characters across the story. In terms of ideational analysis, the frequency lists were produced to categorize clauses into process types to determine animal characters in the folktale.

2 Theoretical Framework

Hoang (2021:08) opines that the functions of a language are approached from different perspectives, such as the ethnographical perspective, psychological perspective, communicative perspective, and

educational perspective. All these models, according to Hoang (2021:08), are constructed on a conceptual framework in non-linguistic terms, looking at language from outside and using non-linguistic terms to interpret different ways that people use language. As a result, the approach equates function with *use*. The concept of function is synonymous with that of the concept of use. Halliday and Hasan (1989) proposed a theory that will identify functions at a high level of abstraction so that they can be recognized as essential to all uses of language. This theory allowed the entire linguistic social process to be viewed as integral to the system of language and explained the nature of its internal structure by relating it to its social use. It is for this reason that Halliday (1973,1975) incorporated the social dimension into his linguistic theory, connecting children's function of language to that of adults' general functions of language. The Proto-Language of adults has two levels only: the *content* (*semantic* and *lexicogrammar*) and the *expression* (*phonology* and *phonetics*). In addition to the four levels, Halliday added context (which constitutes *field, tenor,* and *mode*) (Hoang, 2021). It is within this context that Halliday's three metafunctions emerged, namely the Ideational metafunction, the Interpersonal metafunction, and the Textual metafunction.

Systemic Functional Linguistics (SFL henceforth) is a linguistic theory within discourse analysis that considers the contextual dimensions of language. Sabao (2012) observes SFL as a theory that views language as a social semiosis, a systemic resource for expressing meaning in context. In terms of data, it does not address how language is processed or represented within the human brain, but rather examines the discourses produced (whether spoken or written) and the contexts in which these texts are created.

2.1 The ideational function of a language

Halliday (1961) defines the ideational or experiential metafunction of language as an instrument of thought or conceptualization, or representation of the experiential or real world, including the inner world of consciousness. In using the clause to represent or conceptualize the real world, the principal grammatical element identified by Halliday in the ideational metafunction of language is Transitivity. This term stands for the verbal group in the clause, and Halliday defines it as the overall resource for constructing experience. It means the kind of activity expressed by the sentence participants and the way the participants. Bakuuro (2017:213) defines language as a means of reflecting on things representing the individual's real outside and inner worlds. It is a means of acting on things employing a symbolic system. Language is also used for interactivity between or among people. Clause as a representation has three distinct components. The participant(s) (subject(s) and objects(s)), the process (verb(s)), and the circumstances (adjunct(s) and adverb(s)). It is these identifiable components that outline the human experience, conceptualized or represented. Sabao (p. 49) describes it as the content function of the language. It is realized intransitively (i.e., the relationships established between: (a)processes involved in a clause, (b) the participants implicated. (c)circumstances encoded in a clause. It serves to represent situations and events in the world. It is concerned with explaining experience. It is in the social function that the author/speaker (text producers) embodies in a language their experience – what is going on, who is doing what to whom, where, when, why, and how.

3 The collection of Data through the Voyant Tool

Below is the Sesotho version of the folktale, 'Phokojwe' (Jackal), which was submitted to the Voyant Tool website for analysis. The story text is displayed through a corpus table of terms, which indicates terms in the corpus, the frequency of the term in the corpus, and trends, a graph that shows the distribution of relative frequencies in the corpus. The story is captured by five parts of the screenshots with highlighted terms, in this case, the names of the animal characters, followed by a graph below.

3.1 Part 1.

PHOKOJWE
E ne ere e le diphoofolo kaofela, li hloka metsi moo di ka nwang teng, tsa fumana sedibanyana se eso ka se ba se fatwa. Yaba di re: Ha re fateng kaofela, re tle re tsebe ho nwa metsi a mangata. Yaba phokojwe e hana ho fata. Jwale yare ha di qeta ho fata, yaba di re: Ha ho lebelwe. Na ho tla lebela mang, phokojwe a tle a se ke a nwa, hoba o ile a hana ho fata? Yaba di re: Ha ho lebele hlolo. Yaba Phokojwe o ikela thabeng.
Yaba di a tloha moo, sedibeng. Ha di se di tlohile, Phokojwe a tla. Yaba o re ho hlolo: He hlolo e, he hlolo e! Dumela. Yaba hlolo o re: E. Yaba Phokojwe o a tla; o fihla a ntsha mokotlana, ha a fihla ho hlolo mona sedibeng. Yaba o kenya letsoho ka mokotlaneng, yaba o ntsha dinotshi; yaba o re ho hlolo: O a bona, nna ha nke be ke nyorwe; ke ja ntho e monate. Yaba o a ja; yaba hlolo o re: Ako mphe, mokanaka. Yaba o mofa ha nyenyane. Yaba o re: Kgele ke ntho e monate! Yaba o re: Ako mphe haholo, mokanaka. Yaba phokojwe o re: Tjhe, ekare ke tla o fa haholo, ka o tlama matsoho, ka a isa ka morao, wa qethoha ka seetse, ke tle ke tsebe ho o tshollela ka hanong. Jwale a qethoha. Eitse ha a qethohile, phokojwe a ya sedibeng a nwa metsi a neng a lebetswe ke hlolo. Jwale ha a qeta ho nwa, a itsamaela, a ikela thabeng.

3.2 Part 2.

Jwale tse ngata tsa fihla, tsa re: Hlolo, o entse jwang? Hlolo: Ha se phokojwe, ha se elwa thabeng! O ile a ntlama matsoho, a re o tla mpha ntho e monate, athe o a nthetsa hore a nwe metsi. Yaba di re: Hlolo o sethato ha o tlohetse phokojwe a nwa metsi, phokojwe a hanne ho fata! Jwale tsa re: Ho ya lebela mang ya bohlale! Yaba mmutlanyana o re: Ke nna ya tla lebela. Yaba mmutlanyana o a lebela jwale. Yaba di a tsamaya. Yaba phokojwe o a tla, ha di tsamaile, yaba o re: He mmutlanyana e, he mmutlanyana e! dumela. Mmutlanyana a re: E. A re: He ntsubise kwae eo. Yaba mmutlanyana o re: Ha e yo. Jwale yaba phokojwe o a tla, a a fihla a dula fatshe pela mmutlanyana. Yaba o ntsha mokotlana, yaba o kenya letsoho kahare, yaba o ntsha dinotshi, yaba o a ja, o re: Mm! Yaba o re: Kgele, ke ja ntho e monate, mmutlanyana! Yaba mmutlanyana o re: Na ke eng? Yaba phokojwe o re: Ke kolobisa diqhoqhoqhwana. Yaba o re: Nna ha nke be ke nyorwe ha ke ja ntho ena, mmutlanyana. Yaba o re: Kea kgolwa, lona mmutlanyana,

3.3 Part 3.

Na ke eng? Yaba phokojwe o re: Ke kolobisa diqhoqhoqhwana. Yaba o re: Nna ha nke be ke nyorwe ha ke ja ntho ena, mmutlanyana. Yaba o re: Kea kgolwa, lona mmutlanyana, le bolawa ke lenyora. Yaba mmutlanyana o re: Ako nkutlwise hle, mokanaka. Yaba phokojwe o mo utlwisa hanyenyane. Yaba o re: Tjhe bo! Mmutlanyana ekare ha o tla utlwe monate, ka o tlama matsoho, ka a isa kamorao, wa qethoha ka seetse, ke tle ke tsebe ho o tshollela ka hanong. Yaba mmutlanyana o re: A ko nketse jwalo mokanaka. Yaba o a motlama, a mo isa matsoho kamorao. Yaba jwale phokojwe o ikela ka sedibeng, ha a se a mo tlamile, o ya nwa metsi. Jwale ha a se a nwele, a ikela thabeng. Jwale tsa fihla tse ding diphoofolo tse ngata. Tsa re: Mmutlanyana o entse jwang. Re ne re itse o bohlale, wa re wena o tseba ho lebela; wa re wena ha o lebela phokojwe a ke ke a nwa metsi! Jwale metsi a kae? Jwale ha re nyoruwe hakale, re tla nwa kae? Yaba mmutlanyana o re: Ke phokojwe, o ile a tla le ntho e monate, yaba o re o mpha yona; yaba o re, ekare ha a tla mpha haholo, a ntlama matsoho a isa kamorao. Yaba di re: Na jwale ho tla lebela mang? Yaba

3.4 Part 4.

nkwe o re: Ha ho lebele kgudu. Yaba kgudu o a lebela. Yaba dia tsamaya di a aloha. Yaba phokojwe o a tla, a fumana ho lebetswe kgudu. Yaba o re: He, he kgudu e! Yaba kgudu o a thola. Yaba o boetse o a pheta, o re: He, he, kgudu e! Yaba kgudu o a thola. Yaba phokojwe o re: Ho lebetswe sethato kajeno, ke tla fihle ke mo rahe ka leoto, ke nwe metsi. Yaba o fihla ho kgudu; yaba o re: Kgudu! Yaba kgudu o a thola. Yaba o sututsa kgudu, hore a tloha pela sediba a nwe. Yaba o se a inamela sedibeng, phokojwe. Eitse hoja a re o a nwa, kgudu a mo tshwara ka leoto. Yaba phokojwe o re: Itjhi, Itjhi wa nthoba! Yaba kgudu o a motiisa. Phokojwe a ba a ntsha mokotlana, a ba a re o sa nkgisa kgudu dinkong. Jwale yaba kgudu o tadima hosele, o fapana le mokotlana wa hae. O itse ka re o mo nea ona a re, "ke wa hao", kgudu a hana a mo tiisa.

3.5 Part 5.

The graph above shows the distribution and frequency of the participants (animal characters). The blue colour represents the main character, *Phokojwe* (Jackal), and his participation is distributed across the storyline. The purple colour represents the *hlolo* (rabbit), and his participation starts from the beginning and ends at point 5. In terms of the storyline, the rabbit was disqualified from guarding the fountain and was replaced by the *Mmutlanyana* (hare). The hare is represented by the brownish colour, and his participation starts at point 5 and ends at point 8. In terms of the storyline, the hare was also disqualified to be replaced by the *kgudu* (tortoise), whose participation starts from point 9 and ends at point 10, the same as the Jackal. The graph above confirms that in terms of the spread of the blue line, the jackal should be the main character, and the others are the supporting characters in the story.

Diphoofolo tsa ba tsa fihla. Eitse ha di fihla, a pshemola ho kgudu a baleha. Yaba di fihla di re: E, ha se moo, kgudu, o mohale; kajeno re tla tseba ho nwa metsi, ka hore o ile wa tshwara phokojwe, a se ke a nwa metsi! E.Jacottet (1909:16-18)

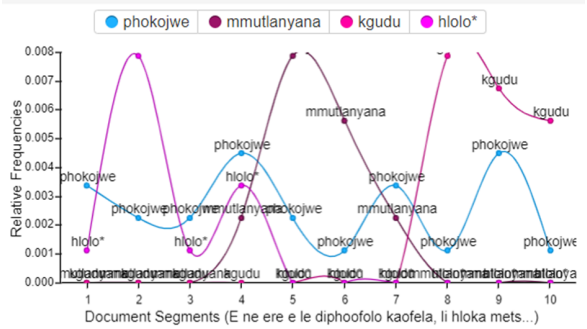


Table 1 illustrates the graph representation of the main characters in the folktale. The word Phokojwe appears 25 times in the story, which is interpreted as the villain who represents someone who puts others into temptation. The storyteller is considerate of the fact that the jackal is the main character, by giving it the title of 'The Jackal'. The jackal is followed by the tortoise 18 times, even though it appeared in the last segment of the graph. The tortoise is the protagonist who stands for honesty and courage. The rabbit and the hare appear 12 and 16 times, respectively. They are supporting characters, and they stand for weaknesses in the story. The name, Diphoofolo (animals), appears only 3 times and represents the community. They are the community that is starved of service delivery by both the rabbit and hare due to their weakness when they succumbed to bribery in the story. The leopard (nkwe) appears only once in the story and represents the tribal chief or leader. It is the leopard that instructed the tortoise to guard the fountain, and it succeeded. The table and the graph further give the roles played by these fictional characters in the story, and the socio-political interpretation of the whole story reflects real-world experiences. These interpretations will be described further under the textual metafunction analysis.

4 Ideational Analysis of the Folktale: Phokojwe

As indicated earlier, ideational metafunction serves to represent situations and events in the world, and it is concerned with explaining experience. Clause as a representation has three distinct components. The participant(s) (subject(s) and objects(s)), the process (verb(s)), and the circumstances (adjunct(s) and adverb(s)). The ideational analysis of the story

Table 1: Participants with high frequency, important roles in the story, and social interpretations.

	Participants	Appearances	Analysis	Interpretation
1	Phokojwe (Jackal)	25	Main Character - Villain	Represents temptation
2	Kgudu (Tortoise)	18	Protagonist - Hero	Honesty
3	Mmutlanyana (Hare)	16	Supporting Participant	Represents weakness
4	Hlolo (Rock Rabbit)	12	Supporting Participant	Represents consensus
5	Diphoofolo (Animals)	3	Community	Service delivery
6	Nkwe (Leopard)	1	Leadership	Secondment

will be analysed in terms of the verb frequency, which, according to the ideational analysis, will be referred to as process types.

Kapau and Banda (2019:291) classify the grammar of the clause into six process types. Dom (2014:12) also opines that six process types are not to be considered separate and distinct categories but as neighbouring regions in a continuous space. There are three basic processes: material, mental, and relational, with the processes that can be described as a mix of two of the basic ones situated between them - behavioural, existential, and verbal. It should also be noted that this is a prototype network: some processes can be categorized under one process type. (p12), In a functional approach to language, where the meaning of linguistic elements is central, the participants involved in different types of processes logically receive different names according to the roles they play in the goings-on and the different properties each process ascribes to them. Some of these participants are displayed under the names of the process types: Material process type the subject takes the name of 'Actor / Agent'; Behavioural process type take name of 'behaver', menta process type takes 'Sensor', verbal process type takes the name of 'Sayer', relational take the name of 'Carrier' and 'identifier' while existential process type take the name of 'Existent;. The process types are listed below:

a) *material process type*: [MaP] - the process of doing and happening in the physical world – they have two inherent participants- actor and goal. According to Dom (2014:12), these are events of 'doing and happening', and always involve a central participant called the Actor, which is "the source of the energy bringing about the change" (H&M 2014: 224). The Actor can be the *sole participant* featured in the process, in which case the clause is analysed in traditional frameworks as an *intransitive clause*.

b) *mental process type*: [MeP] - encoding the meaning of feeling, thinking, and perception – they have inherent participants of sensor and phenomenon.

c) relational process type: [ReP] - the process of being and having – realized by intensive verbs, possessive verbs, and circumstantial relations (e.g., is at).

d) *Verbal Process*: [VeP] - The process of saying and telling.

e) *Behavioural process*: [BeP] - The process of physiological and psychological behaviour.

f) *Existential process*: [ExP] - They are realised in ReP and MeP.

In this story, the analysis of process types will be conducted according to the main characters of the story, namely *Phokojwe* (jackal), *Kgudu* (tortoise), *Mmutlanyana* (hare), *Hlolo* (rock rabbit), *Diphoofolo* (animals), and *Nkwe* (leopard).

5 The process types on Phokojwe (Jackal)

The statistics in Table 2, above, indicate that the jackal had more action and more interaction with the other animals. As is the case in the story, his name has the highest frequency; he communicated more (45%) with the rabbit, the hare, and the tortoise when he was trying to bribe them into allowing him to drink the water. The behavioral aspect is reflected only when the jackal refuses to cooperate with other animals to dig the fountain, and this occurred in only 4.8% of cases.

6 The process type on Kgudu (tortoise)

The statistics in Table 3 reflect what transpired between the jackal and the tortoise. There has been more action than dialogue. On several occasions, the Jackal tried to create a conversation to lure the tortoise into a bribe, and the tortoise resorted to

Table 2: Frequency of process types of the participant, Jackal.

Process Types	Count	Percentage	Analysis
MaP	29	46.7%	More action and interaction
MeP	1	1.6%	Fewer emotions
VeP	28	45%	More dialogue
BeP	3	4.8%	Less behavioural indication
ReP	1	1.6%	Less relational demonstration
Total	62	99.774%	

Table 3: Frequency of process types of the participant, Tortoise.

Process Types	Count	Percentage	Analysis
MaP	17	99%	More action and interaction
MeP	0	0%	No emotions
VeP	0	0%	No dialogue
BeP	1	1%	Less behavioural indication
ReP	0	0%	No relational indication
Total	18	100%	

silence, avoidance, and honesty as its strategies to defeat the Jackal. The 1% behavioral process type is demonstrated by the tortoise when it turned away from the jackal's trick of tasting the honey.

7 The process types on Hlolo (Rabbit)

Table 4 indicates that there is more dialogue (66,6%) between the Jackal and the Rabbit than the action (25%). The Jackal spent some time convincing the rabbit to allow him to drink the water from the fountain. Though a small percentage of (8,3%) on Mental process types, the tasting of honey by the rabbit did the trick, and the Jackal succeeded in defeating the rabbit.

8 The process type on Mmutlanyana (Hare)

The same applies to the Hare. The statistics in Table 5 show that there was somehow a balance between Verbal Process type (VeP), representing dialogue (50%), and Material Process type (MaP), representing action (43,75). The Mental process type (6,25%) is only represented by the tasting of the honey and the quenching of thirst by the hare, and the Jackal succeeded in defeating the Hare as well.

9 The Process types of the animals in the story.

The analysis in Table 6 shows that there are more clauses with material verbs (51,2%) in this folktale, followed by Verbal Processes (39,3%), confirming that folktales are popular stories to tell a story to amuse. Clauses with Mental process types and clauses with Behavioral process types constitute 3,25% each. The relational process types are almost non-existent. The interpretation of these statistics is that the storyteller places more emphasis on action, demonstration, and telling, providing less information about mental aspects such as thinking, smell, sight, and mostly focusing on taste that relates to eating honey and drinking water. The rabbit and the hare do not use their brains; hence, they easily fall into Phokojwe's tricks.

10 The thematic interpretation of the folktale: Pedagogical Significance

The central message of this story is very simple: *"Ha ho tume di melala, le Nketjwane le yena ke motho."* (Heroes always come from humble beginnings). If one goes further to dig deeper into the story, using the paradigmatic approach, one gets a latent message from this folktale: We get the story behind the story:

In terms of Systemic Functional Linguistics, the textual metafunction of the clause plays both linguistic and social roles in the use of the language. The theme is the starting point of the message, as the message starts off from there, whilst the rheme is the rest of the message. The textual analysis goes hand in hand with the paradigmatic structural analysis, which describes the pattern that underlies the

Table 4: Frequency of process types of the participant, Rabbit.

Process Types	Count	Percentage	Analysis
MaP	3	25%	Action and interaction
MeP	1	8,3%	Fewer emotions
VeP	8	66,6%	More dialogue
BeP	0	0%	Less behavioural indication
ReP	0	0%	Less relational demonstration
Total	12	99,96%	

Table 5: Frequency of process types of the participant, Hare.

Process Types	Count	Percentage	Analysis
MaP	14	43,75%	Action and interaction
MeP	2	6,25%	Fewer emotions
VeP	16	50%	More dialogue
BeP	0	0%	Less behavioural indication
ReP	0	0%	Less relational demonstration
Total	32	100%	

Table 6: Frequency of process types of all participants.

Process Types	Jackal	Tortoise	Rabbit	Hare	Total	%
MaP	29	17	3	14	63	51,2
MeP	1	0	1	2	4	3,25
VeP	28	0	8	16	52	39,3
BeP	3	1	0	0	4	3,25
ReP	1		0	0	0	0,81
Total	62	18	12	32	124/123	100

Folkloristic text not as a sequential structure, but as elements regarded in a more analytical scheme. It manifests latent content, and it is in this latent content that this paper is more interested, and how folktales within the context of social and political environments can be used as pedagogic devices. If this is the case, Folktales as part of our Indigenous Knowledge System can help us achieve our aspirations, attitudes, and values through development. Some African Values and Principles are depicted by the Folktale Phokojwe:The story of 'The Jackal' carries a cultural and socio-political message and educational lessons towards life.

10.1 i) Interactive Forum: Pitso

Boon (2007:104) describes umhlangano/ pitso as a community gathering, a place where its participants can have a deep discussion. It is a mechanism that drives personal accountability, a forum in which opinions can be aired, where decisions that can affect values are made and a place where rules of consensus are made. In the story of 'The Jackal', the animals were living together when the drought attacked the land. The need for water became a challenge because no one can live without water. The animals had to apply their minds to how to find water and preserve it for the future. They gathered to have a discussion, and they all agreed to dig a fountain, but amongst them was the jackal who refused to work with others.

10.2 ii) Democracy and Consensus

Boon (2007:82-3) first describes democracy as a phenomenon that implies the majority rule, fairness, respect, tolerance, a classlessness that recognizes the basic equality before the law of all the people. On the other hand, the concept of consensus is a collective and inclusive decision-making process, which goes beyond simple democracy, where a majority is no longer acceptable. Consensus decision-making is only possible where there is trust, where the group is trying to do what is right, not trying to promote a particular selfish point. Consensus decision-making is reliant on the feelings of the

group, while intellect is only part of the group, and subsequently, it is in tune with the traditional culture of Africa. After the building of the fountain, a meeting was then called again to map a way forward as to how to guard the fountain and how to deny those who refused to work access to water. A consensus was then reached that the Rabbit would have to guard the fountain while other animals were out grazing.

10.3 iii) Traditional Philosophy of Botho

This folktale also depicts the concept of Ubuntu, or Botho in Sesotho, where animals worked together to build a fountain or a well. Ubuntu is a term that refers to the collective interdependence and solidarity of communities. It embraces hospitality, caring about others, and being willing to go an extra mile for the sake of another. In Africa, the family shared values and equality. [Boon,2007:25]. Botho is about morality, humaneness, compassion, care, understanding, and empathy. In the words of Mbigi (2005: xvi), botho is a concept based on communal fellowship – on the notion that our personal survival and salvation lie in our shared destiny with others. Our religion is not about intellectual facts but a spiritual experience based on faith in God, with our ancestor spirits as facilitating intercessors. In this folktale, animals share the communal chores, communal frustrations, but also depend on the individuals for the common survival. If animals can work together, surely human beings can work together as well.

10.4 iv) Communication

The story tells us that Africans considered communication very important, which is why animals speak in their stories. In the story of the jackal, the author has given these animals an important gift: Language. They can call meetings and resolve issues through dialogue. As animals, their status has been raised to that of human beings; they can use language to manage others, to establish relationships, and even to communicate new ideas to others. Language is a speech of our parents, siblings, friends and the community, it is the code we use to communicate in the most powerful and intimate experiences of our lives, a central part of our personality, an expression and a mirror of what we are and wish to be, and any scorn for the language of other is a scorn for those who use it, and as such another form of social discrimination.

10.5 v) Environmental Issues

The idea of environmental sensitivity is depicted by the tale, where animals discover that they can store water by digging a fountain or making a well, and that it is necessary to protect it by employing guards or security. Today in South Africa, we have big dams, which are also guarded. It is evident that we not only have to care for our natural resources, but also for our human resources, such as the guards. They cared about their Natural Resources like water; they believed in one of the Ubuntu principles, which is communalism- working together to achieve, they believed in the principle of consensus, the principle of volunteerism, and respected authority and looked with scorn at bribery. You may recall that we said folktales reveal the soul of any society; they express its wishes, desires, hopes, and beliefs about the world.

10.6 vi) Needs Hierarchy

Boon (2007:64) refers to Maslow's hierarchy of needs, where he believes that within every human being there exists a hierarchy of five needs: i) Physiological needs, which include satisfying hunger, thirst, shelter, sex, and other physical bodily needs. The 2nd important need is safety, which includes security and protection from physical and emotional harm, followed by social needs, which entail affection, belonging, acceptance, and friendship. The 4th in terms of significance is the esteem, which refers to self-respect, recognition, and attention, and lastly, self-actualization, which refers to the drive to achieve one's potential and self-fulfillment. In this folktale, the need for water was survival, and all the animals paid attention to it. However, the jackal, due to self-actualization need, decides on a narrow and selfish decision. We then find what we call moral economy, where a person is viewed and identified from within a certain cultural context by a thumb rule. You are either nominated or marginalized by consensus. The jackal was never accepted for his behaviour, but the tortoise was applauded. Democratic principles have been a norm even in African Culture, where people are nominated by consensus.

10.7 vii) Volunteerism

In our communities, volunteerism is also one of the African social principles, and in our story, this concept is captured very well. The hare, when the animals were in a dilemma, volunteered to guard

the fountain. In Africa, cultural people will volunteer to do the work to help their community reach a certain goal, and others will volunteer to work in so-called "letsema" (collective work) and help on different occasions. In the folktale, we see the hare volunteer to protect the fountain, even though it also failed due to vulnerability to bribery.

10.8 viii) Bribery

It also discourages people from accepting bribes, as this will cause failure in performing their duties as expected. The hare and rabbit were so weak, had low moral values, and were only selfish to have accepted the bribes from the traitor, like the jackal. Accepting bribes leads to failures. Other things that lead to failure are of lack of knowledge and a lack of skills. Again, it is to lose focus and control. The results of failure to deliver what is expected cause painful things:

10.9 x) Interactive Leadership

In our story, the issue of leadership authority is raised: when everybody was in despair, when every animal was scratching its head, trying to find out who would be the next to take responsibility, the tiger appointed the tortoise. In African culture, the hide of the tiger symbolizes chieftaincy or royalty. It is also normal for the king or a leader to appoint someone with good qualities to serve their nation. As chiefs are custodians of culture, they have a vision and mission for their followers. The tiger succeeded in performing the duty he was entrusted with. A chief is not an autocrat, and must rely on councillors representing the people to assist him, must be guided by consensus and if not, people will ignore his decision or his 'law', people must always be strongly represented, and the entire community (adult) should attend court or a 'hearing', The people have a responsibility to each other and collectively, ensure that the law and values of the community are upheld. Boon (2007:114) states that there is a degree of leadership in every person; all leaders are responsible for nurturing, stimulating, and awakening the leader that exists within every human being. Boon (2007:114) further outlines four broad groups of leaders according to 'the progression – Leadership Model: i) Spectators: This section makes up the greatest sector of humanity: They are observers and critics, have a high concern for themselves, and are very critical of the change agents. They never expose themselves and avoid vulnerability at all costs. They are not leaders. They represent the greatest challenge to the rest of the group. The challenge of leadership is to grow people away from spectorship, to stimulate maturity, and nurture the individual accountability within them. ii) Players: It is the start of the real active leadership. They are generally positive, and they accept accountability. iii) Captains: Through motivation, players are motivated to become captains. Captains display compelling physical characteristics. They a highly concerned for others and extremely sensitive to the difficulties experienced by lower people. They accept accountability. iv) Coaches: They are visionary leaders. They are impassionate, charismatic, and they share their dreams. They are mature and well-balanced.

11 Conclusion

The main purpose of the paper was to analyse the Sesotho folktale through the lens of Systemic Functional linguistics, using the Voyant tool as the data collection tool, which further assisted in the identification of Sesotho word categories, analysing the frequency appearances of the main characters of the story, to explore the thematic function of the folktales, an integration of folktales into education, where the folktales are used for the socio-political interpretation. The paper presents observations that the story of 'The Jackal' carries a socio-political message and education towards life. The African folktales can be integrated into science through the application of Systemic Functional linguistics to analyse them, be brought into broader discourse through the application of Voyant Tool as a technological tool of text analysis, in our case, the African folktales. This paper wishes to conclude that Halliday's metafunction of language as the theoretical framework suggests that SFL linguistics has the potential applicability in text analysis.

12 References

Isidore, O. (1992). *African Oral Literature.* University Press.

Guma, S. M. (1967). *The form, content, and structural analysis, significance, and interpretation of Sesotho folktale: Phokojwe.*

Moleleki, M. A. (1988). *A study of some aspects of K.P.D. Maphalla's poetry.* Master's thesis, University of Cape Town.

Mike, B. (2007). *The African Way.* Zebra Press, Cape Town.

Mokoena, L. S. (2011). *Structural analysis, significance and interpretation of Sesotho folktale: Phokojwe.*

Philip, A. (1998). *African Myths and Legends.* Belitha Press, UK.

Rosenberg, D. (1997). *Folklore, Myths, and Legends.* NTC, USA.

Russell, H. K. (1993). *Foundations in Southern African Oral Literature.* Witwatersrand University Press, USA.

Sabao, C. (2013). *The reporter voice and objectivity in cross-linguistic reporting of controversial news in Zimbabwean newspapers: An appraisal approach.* Doctoral dissertation, Stellenbosch University.

Bakuuro, J. (2017). African folktales as repositories of indigenous knowledge and values. *Journal of African Cultural Studies, 29*(3), 215–229.

Daniel, C. (1993). Folklore and education in Africa: An ideological exploration. *African Studies Review, 36*(2), 67–82.

Dom, S. (2014). African oral literature and the preservation of indigenous knowledge systems. *Journal of African Studies and Development, 6*(4), 75–83.

Dorji Penjore. (2005). *Folktales and Education: The Role of Bhutanese Folktales in Value Transmission.* Thimphu: Centre for Bhutan Studies.

Finnegan, R. (1970). *Oral literature in Africa.* Oxford University Press.

Guma, S. M. (1969). *The form, content, and technique of traditional literature in Southern Sotho.* Oxford University Press.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word, 17*(2), 241–292.

Jacobs, G. M., & Cleveland, P. (1999). *Language, culture, and education: Integrating perspectives in teaching.* Routledge. [details may vary]

Kapau, L., & Banda, F. (2019). Integrating indigenous knowledge systems into science education in Africa: A discourse analysis. *Southern African Linguistics and Applied Language Studies, 37*(2), 103–117.

Masowa, L. (2024). Folktales as tools for science and technology learning in African classrooms. *African Journal of Educational Research and Development, 14*(1), 55–72.

Mweti, C. (1999). The use of stories and their power in the secondary curriculum among the Akamba of Kenya. University of Wisconsin-Madison, Madison, WI, United States.

Nakin, J. B. (2017). The use of indigenous stories in teaching natural sciences: An African pedagogical model. *Perspectives in Education, 35*(3), 45–59.

Phindane, P. (2019). An analysis of the Sesotho folktale *Kgubetswana Le Talane* using the binary opposition approach. *South African Journal for Folklore Studies, 29*(2), 1–12.

Pirkova-Jakobson, R. (2006). *Language, meaning, and cultural narrative: A semiotic perspective on folklore.* Cambridge Scholars Publishing.

Sabao, C. (2012). Language and ideology in African folktales: A critical discourse analysis perspective. *Journal of Language and Communication, 9*(2), 112–128.

Wiysahnyuy, L. F., & Valentine, N. B. (2023). Folktales as indigenous pedagogic tools for educating school children: A mixed methods study among the Nso of Cameroon.

Multimodal Classification System for Hausa using LLMs and Vision Transformers

Ali Mijiyawa[1] **Fatiha Sadat**[1]

[1]Université du Québec à Montréal (UQAM), Montreal, QC, Canada

mijiyawa.ali@courrier.uqam.ca | sadat.fatiha@uqam.ca

Abstract

This paper presents a classification-based Visual Question Answering (VQA) system for the Hausa language, integrating Large Language Models (LLMs) and vision transformers. By fine-tuning LLMs on monolingual Hausa text and fusing their representations with those of state-of-the-art vision encoders, our system predicts answers from a fixed vocabulary. Experiments conducted on the HaVQA dataset, under offline text–image augmentation regimes, tailored to the specificity of Hausa as a low-resource language, show that this augmentation strategy yields the best performance over the baseline, achieving 35.85% accuracy, 35.89% WuPalmer similarity, and 15.32% F1-score.

1 Introduction

The task of VQA can be approached through three paradigms: *open-ended classification*, selecting an answer from a fixed vocabulary; *multiple-choice (MCQ)*, choosing from given options; and *generative*, producing free-form text responses, sometimes with rationales (Dua et al., 2020). This study adopts a classification-based approach, framing VQA as a closed-vocabulary, multi-class task where the model selects the correct answer from a predefined label set given an image and a question.

Transformer-based language models such as BERT (Devlin et al., 2018) have enabled significant progress in VQA for high-resource languages. However, many African languages, including Hausa, remain underrepresented due to the lack of large annotated multimodal corpora (Kumar et al., 2022; Hedderich et al., 2021), which hinders the development of VQA systems capable of capturing their linguistic and cultural specificities. To the best of our knowledge, no previous study has explored the fine-tuning of Large Language Models (LLMs) combined with vision transformers for classification-based VQA in African low-resource contexts. Furthermore, no prior work has examined data augmentation techniques specifically tailored to VQA systems in African languages.

In this study, we aim to explore the non-generative potential of LLMs in this setting, focusing on the Hausa language. We present a classification-based Hausa VQA system combining fine-tuned LLMs with state-of-the-art vision transformers.

Using the HaVQA dataset (Parida et al., 2023b), we evaluate training paradigms to assess the impact of text and image data augmentation, focusing on offline augmentation adapted to Hausa, where data are expanded before training via text rewriting and image perturbations, and compare this to a baseline without augmentation.

Our main contributions are twofold: (i) We conduct a comprehensive multimodal benchmark of nine LLMs and four vision transformers (36 model variants in total) within a unified fine-tuning framework for Hausa VQA; and (ii) We propose a low-resource data augmentation and multimodal enhancement framework, combining text and image transformations tailored to Hausa linguistic and cultural characteristics. This expands the HaVQA dataset (Parida et al., 2023a) into HaVQA_aug[1] and yields measurable improvements in classification accuracy, Wu-Palmer similarity, and F1-score across models.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 2.4 provides background on the Hausa language; Section 3 formalizes the VQA task and presents the proposed methodology;

Section 4 presents our experiments and evaluations; Section 5 discusses the results; and Section 6 concludes the paper with future directions.

[1]https://github.com/Alimiji/LLM_QRV_Hausa_HaVQA_aug

2 Related Work

2.1 VQA for African Languages

Recent advances in VQA leverage transformer-based models such as BERT (Devlin et al., 2018) and Vision Transformer (Dosovitskiy et al., 2020) to integrate visual and textual information (Tan and Bansal, 2019; Lu et al., 2019). In Africa, only three VQA datasets exist: HaVQA for Hausa using non-large models (Parida et al., 2023a), CVQA covering multiple African languages except Hausa with large multimodal models (Romero et al., 2024), and SwahiliVQA with 10,000 images and 41,448 Q&A pairs, achieving 38.38% accuracy with non-large models (MBWANA and Long Hoang, 2025). Data augmentation improves dataset diversity and model generalization (Chen et al., 2022; Yang et al., 2022). Challenges remain, including limited performance of models like GPT-4o (Olufemi et al., 2025) and cultural biases highlighted by CulturalVQA and WorldCuisines (Nayak et al., 2024; Winata et al.).

2.2 LLMs for African Languages

Multilingual LLMs such as mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020) often underperform on African languages due to their scarcity in pretraining corpora (Hedderich et al., 2021; Blasi et al., 2022). Adaptive fine-tuning (e.g., MAFT (Alabi et al., 2022)) and from-scratch models (e.g., AfriBERTa (Ogueji et al., 2021), AfroLM (Doe et al., 2023)) improve alignment with African languages as well as text classification and question answering (Yu et al., 2025). For Hausa, challenges include limited datasets, dialectal variation, and suboptimal tokenization (Muhammad et al., 2025). Community resources like HausaNLP, AfroBench (Ojo et al., 2023), and IrokoBench (Adelani et al., 2025) support NLP development but highlight persistent performance gaps. Integrating these language-specific models with vision transformers and culturally-aware data augmentation is key for effective Hausa VQA.

2.3 Data Augmentation

Recent advances in data augmentation enhance both text and images. Text techniques include synonym replacement, EDA (Wei and Zou, 2019), back-translation (Sennrich et al., 2016), and LLM-based paraphrasing (Ding et al., 2024), while image methods use geometric and photometric transformations, MixUp, and CutOut (Shorten and Khoshgoftaar, 2019a). In VQA, multimodal augmentation combines visual perturbations with question reformulation, QA generation, and adversarial examples (Chen et al., 2022). Such strategies, integrated with adapted LLMs, are crucial for improving classification-based VQA in low-resource African languages like Hausa.

2.4 Hausa Language

Hausa, a Chadic language of the Afroasiatic family, is spoken by over 200 million people across West and Central Africa (Muhammad et al., 2025). It functions as a regional lingua franca and has a rich literary tradition in both *boko* (Latin) and *ajami* (Arabic-derived) scripts. Despite its cultural significance, Hausa remains low-resource in NLP, facing limited datasets, suboptimal tokenization, and dialectal variation (Muhammad et al., 2025). Recent work has advanced text classification, sentiment analysis, machine translation, NER, and POS tagging, while speech datasets like Common Voice and multimodal resources such as HaVQA support VQA research. Pretrained models like AfriBERTa and the HausaNLP catalog further enhance accessibility and development in these tasks.

3 Proposed Methodology

The Hausa VQA task adapts visual question answering to Hausa, aiming to build models that understand Hausa questions about images and provide accurate answers by combining language comprehension with visual reasoning (Parida et al., 2023a; Antol et al., 2015). This work promotes inclusive AI for speakers of low-resource languages (Nekoto et al., 2020; Hedderich et al., 2021), with datasets like HaVQA (Parida et al., 2023a) providing benchmarks for multilingual and multimodal research.

3.1 System Architecture

Our proposed system consists of a text encoder (LLM) that converts a Hausa question q into a dense vector, an image encoder (ViT) producing patch-level visual embeddings, a fusion layer combining both modalities via concatenation or cross-modal attention, and a classification head mapping the fused representation to a global Hausa answer vocabulary $\mathcal{A}$. Given a question–image pair (q, I), the model predicts a unique label $a' \in \mathcal{A}$. The answer space $\mathcal{A}$ comprises 2,991 gold-standard Hausa labels, ensuring each input pair maps to exactly one label.

We adopt a classification-based VQA approach for Hausa, following HaVQA (Parida et al., 2023b).

This simplifies modeling and enables direct supervision with a fixed label set, avoiding the challenges of free-form answer generation in low-resource languages.

3.2 Offline vs. No Augmentation

As shown in Figure 1, the baseline (no augmentation) model is fine-tuned solely on the original HaVQA dataset (Parida et al., 2023a), serving as a performance reference. In contrast, offline augmentation (Figure 2) expands the dataset by duplicating images with geometric transformations (e.g., random flips and rotations) and paraphrasing English question–answer pairs using synonym replacement (Wei and Zou, 2019), which are then translated into Hausa (Parida et al., 2023a). These augmented pairs are linked to the transformed images, creating a larger, fixed dataset. Comparing these two settings enables systematic evaluation of data augmentation's impact on classification accuracy, semantic alignment, and robustness in low-resource scenarios.

4 Experiments and Evaluations

4.1 HaVQA dataset and evaluation setup

We use the HaVQA dataset (Parida et al., 2023a), which contains 6,022 English–Hausa question–answer pairs spanning 2,991 unique Hausa labels and 1,555 images from Visual Genome (Krishna et al., 2017). Due to Hausa's low-resource nature, we merged the original development and test sets into a single evaluation set (Kann et al., 2019) to maximize training data. This practice, common in low-resource NLP, increases training robustness while maintaining comprehensive evaluation.

4.2 Training details (Hyperparameters)

All models were trained for 10–20 epochs using HuggingFace's Trainer (final runs: 17 epochs) with a fixed seed of 12345. Training and evaluation used a batch size of 32 per device. We employed the AdamW with $learning\_rate = 1 \times 10^{-5}$ and $weight\_decay = 1 \times 10^{-4}$. Precision was set to bfloat16, and evaluation, logging, and checkpointing occurred every 100 steps, retaining the last three checkpoints. The best model was selected based on WUPS. Data loading used 8 workers and kept all columns. Training was conducted on a single NVIDIA A100 or L4 GPU.

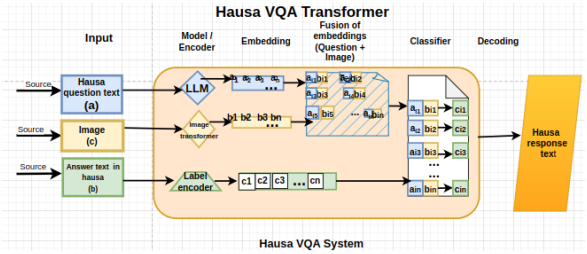


Figure 1: Baseline system: Hausa VQA model combining a large language model (LLM), inspired by Parida et al. (2023b).

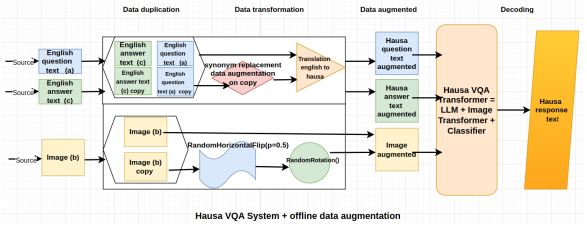


Figure 2: Offline augmentation: Hausa VQA model trained on pre-augmented data before multimodal fusion (Zhang et al., 2015; Wei and Zou, 2019; Parida et al., 2023b).

4.3 Metrics meaning

We evaluate Hausa VQA using Accuracy, F1 Score (Pedregosa et al., 2011), and WuPalmer (WUP) Similarity (Wu and Palmer, 1994). Accuracy measures the proportion of correct predictions, while F1 Score balances precision and recall to account for class imbalance. WUP similarity, computed via wup_similarity in NLTK WordNet[2], assesses semantic relatedness between predicted and reference answers. Together, these metrics provide a comprehensive evaluation, summarized in Table 1.

4.4 Hausa VQA task models training

Tables 2 and 3 list the nine LLMs and four vision transformers evaluated in our Hausa VQA study. Pairing each LLM with every vision encoder yields 36 model variants, all trained as multiclass classifiers over a fixed set of Hausa answer labels, without sequence generation. Each question–image pair from HaVQA (Parida et al., 2023a) is encoded by the LLM and the vision transformer, with fused embeddings fed to a classifier. We compare performance using offline augmentation on the expanded dataset versus the baseline with no augmentation.

4.4.1 No Augmentation training

In this setup, the model is trained on the original HaVQA split: 4,816 training, 1,204 test pairs and covering 1,555 images in total. Each training in-

[2] https://www.nltk.org/api/nltk.corpus.reader.wordnet.html

Metric	Objective	Interpretation	Value Range
Accuracy	Exact match with the reference answer	Answer is entirely correct for high value	0–1
F1 Score	Harmonic mean of precision and recall	Answer is both precise and complete for high value	0–1
WUP Similarity	Semantic similarity via WordNet hierarchy	Answer is semantically close to the reference for high value	0–1

Table 1: Comparison of VQA evaluation metrics. Sources: (Antol et al., 2015; Rajpurkar et al., 2016; Malinowski and Fritz, 2014).

LLMs	Pretrained on Hausa?	Fine-tuned on Hausa?	Parameters
mt0-base	Yes	No	580M
mt0-large	Yes	No	1.2B
afriberta_large	Yes	No	126M
afro-xlmr-large	Yes	No	560M
gemini	Yes	No	770M
bloomz560	No	No	560M
bloomz1b7	No	No	1.7B
llama-3.2-1B	No	No	1.23B
deepseek-R1-1.5B	No	No	1.5B

Table 2: Information on Hausa pretraining, fine-tuning, and the number of parameters of the different LLMs used for the Hausa VQA system.

stance is a triplet $(q,\ I,\ a)$: a Hausa question q, its associated image I, and the gold-standard answer a. The question q is tokenised and embedded using a large language model (LLM) encoder, while the image I is patchified and processed by a vision transformer to obtain visual embeddings. The answer space $\mathcal{A}$ is defined as a fixed set of label vectors, each encoding a valid Hausa answer. Textual and visual embeddings are fused, typically via cross-modal attention, into a unified representation. A classification head then projects this fused vector onto the label space, and the top-ranked label is decoded into Hausa text as the final prediction. This baseline serves as the reference configuration to assess the impact of data augmentation strategies.

4.4.2 Offline Augmentation training

To construct the augmented dataset, each English question–answer–image triplet $(q_{en},\ i_{en},\ a_{en})$ from the HaVQA corpus (Parida et al., 2023a) is duplicated: the original instance is preserved, while its copy is transformed. For the textual components, synonym replacement (Wei and Zou, 2019) is applied exclusively to the duplicated English questions q_{en} by using the WordNet module of the NLTK library. This produces a parallel English question–answer set containing both the original questions and their paraphrased variants, each paired with the same gold-standard answers. The resulting English dataset is subsequently translated into Hausa (Parida et al., 2023a) using the Translator module from the googletrans library. For the visual components, only the duplicated images undergo geometric transformations such as random horizontal flips and slight rotations (Shorten and Khoshgoftaar, 2019b).

By combining the original triplets with their augmented counterparts, we obtain the augmented dataset HaVQA_aug[3], whose size is approximately twice that of the original HaVQA corpus. Each training instance is represented as $(q_{ha\text{-}aug},\ i_{ha\text{-}aug},\ a_{ha\text{-}aug})$, where $q_{ha\text{-}aug}$ denotes the paraphrased and translated Hausa question, $i_{ha\text{-}aug}$ the transformed image, and $a_{ha\text{-}aug}$ the corresponding Hausa answer.

The training architecture mirrors the baseline setup: the question is encoded by an LLM, the augmented image is processed by a vision transformer, and the fused multimodal representation is classified into a fixed set of Hausa answer categories. The offline approach enriches both text and image modalities prior to training.

Image Encoder	Parameters
vit-base-patch16-224-in21k	86.4M
clip-vit-base-patch32	149M
mae-base	86M
deit-base-patch16-224	86M

Table 3: Number of parameters of the different image encoders used for the Hausa VQA system.

Dataset	Train	Test	Total Images
HaVQA_aug	9,625	2,407	3,110

Table 4: Details of the HaVQA dataset's offline-augmented partitions for training and evaluation.

5 Results and Discussion

The best Accuracy and WuPalmer scores achieved by each LLM under these regimes are presented in Tables 5 and 6.

Under the no-augmentation baseline, llama-3.2-1B + clip-vit-base-patch32 achieves 19.68% Accuracy and WUP, with the highest F1 among all pairs (Table 5). This remains below the 30.86% WUP reported for DeiT-Base-P-224 + BERT-Base-Hausa (Parida et al., 2023a). Notably, llama-3.2-1B performs robustly across all image encoders despite lacking Hausa-specific pretraining. With offline augmentation, most LLM–image pairs show larger gains, particularly for Hausa-pretrained LLMs. Gemini + vit-base-patch16-224-in21k reaches Wu-

[3] https://github.com/Alimiji/LLM_QRV_Hausa_HaVQA_aug

LLMs	Image Encoder	WuPalmer	Accuracy	F1
mt0-base	deit-base-patch16-224	15.45	15.45	0.35
mt0-large	clip-vit-base-patch32	15.61	15.87	1.03
afriberta_large	vit-base-patch16-224-n21k	18.27	18.42	1.05
afro-xlm-large-7.6b	vit-base-patch16-224-n21k	15.24	15.32	0.49
gemini	vit-base-patch16-224-n21k	35.89	35.85	1.86
bloomz560	deit-base-patch16-224	15.40	15.42	0.34
bloomz1b7	deit-base-patch32	17.62	17.64	0.89
deepseek-R1-1.5B	deit-base-patch32	19.86	19.84	1.73
llama-3.2-1B	deit-base-patch32	18.46	18.36	1.36

Table 5: Best Accuracy, WuPalmer, and F1 score per LLM on HaVQA — *baseline configuration*.

LLMs	Image Encoder	WuPalmer	Accuracy	F1
mt0-base	mae-base	32.19	32.16	9.89
mt0-large	mae-base	35.30	35.27	13.45
afriberta_large	clip-vit-base-patch32	32.74	32.70	7.84
afro-xlmr-large-76L	clip-vit-base-patch32	28.45	28.42	4.47
gemini	vit-base-patch16-224-in21k	35.89	35.85	15.32
bloomz560	clip-vit-base-patch32	18.01	17.95	1.24
bloomz1b7	vit-base-patch16-224-in21k	18.36	18.36	0.63
deepseek-R1-1.5B	clip-vit-base-patch32	21.70	21.69	3.42
llama-3.2-1B	clip-vit-base-patch32	17.99	17.99	3.11
gemini	deit-base-patch16-224	33.52	33.49	12.79

Table 6: Best Accuracy, WuPalmer, and F1 score per LLM on HaVQA — *offline* augmentation.

Palmer 35.89%, Accuracy 35.85%, and F1 15.32% (Table 6). mT0-base/large with MAE-Base also benefits substantially (WuPalmer 32.19/35.30%, Accuracy 32.16/35.27%, F1 9.89/13.45%). Non-Hausa-pretrained models such as bloomz and llama-3.2-1B gain modestly; e.g., llama-3.2-1B + clip-vit-base-patch32 shows F1 3.11%, similar to inline augmentation, indicating that its baseline robustness persists but improves little under offline augmentation. Overall, offline augmentation is most effective for Hausa-pretrained backbones.

5.1 Systems Analysis and Limitations

Despite promising results, our Hausa VQA system has several limitations. First, HaVQA is limited in size and diversity, and automatically augmented Hausa texts can introduce noise.bSecond, the system handles only static images, lacking richer multimodal inputs (e.g., video, spatial context), and dialectal or orthographic variation in Hausa is underrepresented, limiting generalization. Model interpretability is also limited, as large models remain opaque without integrated explainability mechanisms.

Our analysis shows that offline augmentation benefits Hausa-pretrained LLMs (mT0, AfriBERTa, Afro-XLM-R, Gemini) considerably, while non-Hausa models (BloomZ, LLaMA, DeepSeek) gain modestly (Table 6). This suggests improvements stem not merely from increased data volume, but from linguistically compatible paraphrases exploited by Hausa-aware tokenizers. However, none of our models is fine-tuned on Hausa (Table 2),

Image Encoder	Text Encoder	WuPalmer
BEiT-L-P224	Bert-Hausa	27.76
ViT-B-P224	Bert-Hausa	28.91
ViT-L-P224	Bert-Hausa	29.67
DeiT-B-P224	**Bert-Hausa**	**30.86**

Table 7: WuPalmer scores of *Bert-Hausa* with four visual encoders on HaVQA(Parida et al., 2023b).

so offline augmentation may act as a proxy for adaptation, potentially overstating gains. Future work includes controlled experiments to separate volume and diversity effects and stricter filtering of synthetic paraphrases.

6 Conclusion and Future Work

This research presents a classification-based Hausa VQA system, combining fine-tuned LLMs with state-of-the-art vision transformers. Experiments on HaVQA show that offline multimodal augmentation, tailored to Hausa linguistic and cultural features, substantially improves performance, achieving 35.85% Accuracy, 35.89% WuPalmer, and 15.32% F1—exceeding prior benchmarks by over 5%. These results highlight the value of language-specific pretraining and multimodal enrichment in low-resource VQA. Future work includes extending HaVQA into a multilingual MT-VQA benchmark, improving interpretability via cross-modal attention analysis, and generalizing the framework to other African languages. We also plan to enhance deployment through model compression, knowledge distillation, and multimodal extensions to speech and video. We will also emphasize ethical alignment, cultural sensitivity, and bias mitigation to foster inclusive and equitable multimodal AI for underrepresented languages.

References

David Ifeoluwa Adelani and 1 others. 2025. IrokoBench: A new benchmark for african languages in the age of large language models. https://aclanthology.org/2025.naacl-long.139/.

Jesujoba Olabode Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Findings of ACL*, pages 1–17.

Stanislaw Antol, Arjun Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering.

In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *European Conference on Computer Vision (ECCV)*.

Alexis Conneau, Karthik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Felipe Guzmän, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705.

John Doe, Alice Smith, and Rahman Mohammed. 2023. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. Beyond answering: Towards multi-step reasoning in machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8849–8863.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*, volume 123, pages 32–73. Springer.

Raghavendra Kumar, Michael Hedderich, Bonaventure F Dossou, Chris C Emezue, Krešimir Šojat, Heike Adel, and Iryna Gurevych. 2022. Towards data and benchmarking for african languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3554–3573.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*, 27.

Francis Stephen MBWANA and Dang Long Hoang. 2025. Swahilivqa: A dataset for visual question answering in swahili language. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. HausaNLP: Current status, challenges and future directions for Hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 176–191, Vienna, Austria. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Wilhelmina Nekoto, Vukosi Marivate, Taiwo Fagbohungbe, and et al. 2020. Participatory research for low-resourced machine translation: A case study

in african languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–11.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Victor Tolulope Olufemi, Oreoluwa Boluwatife Babatunde, Emmanuel Bolarinwa, and Kausar Yetunde Moshood. 2025. Challenging multimodal LLMs with african standardized exams: A document VQA evaluation. In *CVPR 2025 Workshop Vision Language Models For All*.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen H. Muhammad, Aneesh Bose, Guneet S. Kohli, Ibrahim S. Ahmad, Ketan Kotwal, Sayan D. Sarkar, Ondrej Bojar, and Habeebah A. Kakudi. 2023a. Havqa: A dataset for visual question answering and multimodal research in hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023b. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Connor Shorten and Taghi M. Khoshgoftaar. 2019a. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Connor Shorten and Taghi M. Khoshgoftaar. 2019b. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 8 others. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. ACL '94, page 133–138, USA. Association for Computational Linguistics.

Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun othersZhuang, and 1 others. 2025. IN-JONGO: A multicultural intent detection and slot-filling dataset for 16 African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 649–657.