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.

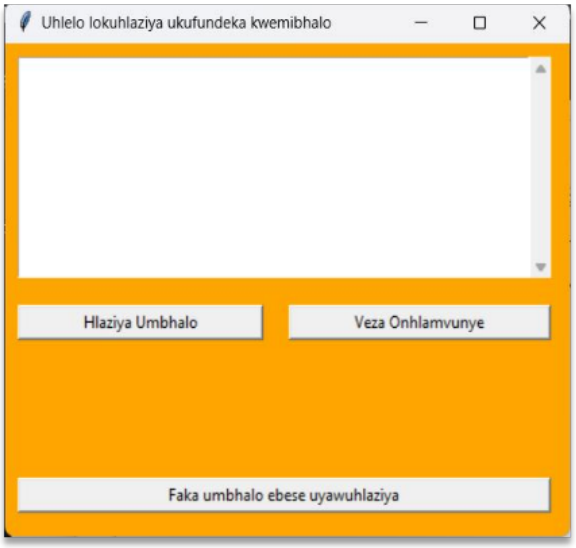

Figure 1: IsiZulu text readability tool

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.

Acknowledgements

This research was supported in part by funding from the National Research Foundation (NRF) of South Africa under the Thuthuka Post PhD Track (Grant Number *TTK240411213481*). The views and conclusions expressed in this article are those of the author and do not reflect those of the NRF or Nelson Mandela University.

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.