Adapting Nine Traditional Text Readability Measures into Sesotho

Johannes Sibeko, Menno van Zaanen
Nelson Mandela University, South African Centre for Digital Language Resources
University way, Summerstrand, Port Elizabeth, Internal Box 340, Private bag X6001, Potchefstroom
johanness@mandela.ac.za, menno.vanzaanen@nwu.ac.za

Abstract

This article discusses the adaptation of traditional English readability measures into Sesotho, a Southern African indigenous low-resource language. We employ the use of a translated readability corpus to extract textual features from the Sesotho texts and readability levels from the English translations. We look at the correlation between the different features to ensure that non-competing features are used in the readability metrics. Next, through linear regression analyses, we examine the impact of the text features from the Sesotho texts on the overall readability levels (which are gauged from the English translations). Starting from the structure of the traditional English readability measures, linear regression models identify coefficients and intercepts for the different variables considered in the readability formulas for Sesotho. In the end, we propose ten readability formulas for Sesotho (one more than the initial nine; we provide two formulas based on the structure of the Gunning Fog index). We also introduce intercepts for the Gunning Fog index, the Läsbarhets index and the Readability index (which do not have intercepts in the English variants) in the Sesotho formulas.

Keywords: Text Readability, Sesotho, Low-resource language

1. Introduction

The reports from the Progress in International Reading Literacy Study (PIRLS) show consistent subpar performance among learners reading in South African indigenous languages (Roux et al., 2021). In the PIRLS standards, learners who perform below the 400-point benchmark, struggle to extract fundamental information from the text, making it challenging for them to respond to even the simplest questions. Regrettably, at least 81% of learners in the South African indigenous languages have been performing below the 400-point benchmark (Roux et al., 2021). As a result, such performance hinders the achievement of inclusive and equitable quality education in essentially all high school subjects as learners cannot access information from written sources. Steps need to be taken to address this literacy challenge as highlighted by the fourth of United Nations' (UN) seventeen Sustainable Development Goals, which focuses on the importance of ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all.

A possible solution to low literacy levels is to make sure children learn to read properly, which can only be attained through practising reading (van Bergen et al., 2018). In other words, learners need to read in order for their reading skills to improve. One way of igniting the desire to read is providing learners with both opportunities to select texts and reading time (Rasheed, 2023). According to Rasheed (2023), learners who have the autonomy to choose their reading materials tend to perform better than those who are assigned texts. However, it is essential to note that a poor choice of reading materials can hinder the development of reading skills when the texts are not well-matched to the reader's level of proficiency (Mohammed et al., 2023). Keeping this in mind, it becomes evident that education stakeholders require a tool to assess the readability of texts to enable the identification of texts that align with the reader's reading ability level. The development of readability measures for the different indigenous languages of South Africa will allow for objective measurements of text readability.

Note that the indigenous languages of South Africa are low-resourced. As such, the choice of approaches to the exploration of text readability is somewhat limited. Here, we propose the use of traditional readability measures that focus on shallow text properties (Van Oosten et al., 2010; Zamanian and Heydari, 2012).

Despite a longstanding research interest in readability assessment, traditional readability measures have not been tailored for South African indigenous languages (Leopeng, 2019). The lack of text readability measures for South African indigenous languages so far has led to the use of (unmodified) English readability measures for readability analyses in indigenous South African languages such as isiZulu (Land, 2015), isiXhosa (Carel, 2019; Leopeng, 2019), and Sesotho (Krige and Reid, 2017; Reid et al., 2019). Recently, Sibeko (2023) reports attempts to develop text readability measures for Sesotho. Their article focuses on the basic language resources for Sesotho required to develop the readability measures. However, they

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 66–76

do not tackle the actual development of readability measures.

In this article, we focus on the development of readability measures for Sesotho and not all twelve official languages of South Africa. Even though a similar approach may be applied to the other languages as well, sign language, one of the twelve official languages, may require a different approach. Overall, we address the research question:

> How can traditional readability measures be effectively modified and adapted to suit the specific characteristics of Sesotho?

To answer this question, we adapt traditional readability measures to Sesotho using English as a high-resource helper language for the low-resource Sesotho. The underlying assumption is that texts that are easy to read in Sesotho will also be easy to read when translated into English and difficult Sesotho text will be translated into difficult English texts. First, the background of this investigation is presented in Section 2, then the methodology is described in Section 3, followed by the evaluation in Section 4. Finally, we present our discussion and conclusions in Section 5.

2. Background

2.1. An overview of Sesotho

Sesotho is a language spoken in Southern Africa. It is one of the two official languages in Lesotho (Government of Lesotho, 1993), one of the twelve official languages in South Africa (Republic of South Africa, 2023), and one of the marginalised official languages in Zimbabwe (Parliament of Zimbabwe, 2021). Furthermore, Sesotho is spoken in Zambia, Namibia, and Botswana. At least more than ten million people use Sesotho on a daily basis. It is used and taught in both basic and higher education sectors.

Sesotho has at least six recognised dialects, namely, the Sekwena, Sekgolokwe, Serotse, Setlokwa, Sephuthi, and Setaung (Kula and Marten, 2008; Mohasi and Mashao, 2005; Nhlapo, 2021). Of these dialects, Sekwena was promoted and has thus become the standard of writing in Sesotho (Nakin, 2009; Sekere, 2004). Moreover, there are at least two officially recognised orthographies for Sesotho, namely, the South African and the Lesothan orthographies (Makutoane, 2022; Setaka, 2018; Setaka and Prinsloo, 2020; Sibeko, 2022). The research described in this article is based on texts that are written using the South African Sesotho orthography.

2.2. Traditional Readability Measures

In this article, we explore nine traditional English readability measures for adaptation to Sesotho. These measures include the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Flesch-Reading Ease (FRE) (Flesch, 1948, 1974), Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969; Zhou et al., 2017), and Gunning Fog Index (GFI) (Gunning, 1952, 1969), which rely on syllable-related information, as well as the Coleman-Liau index (CLI) (Coleman and Liau, 1975), Automatic Readability index (ARI) (Kaur et al., 2018; Smith and Senter, 1967), Readability index (RIX), and Läsbarhets index (LIX) (Björnsson, 1968; Björnsson, 1983) measures which are based on word-length information. Finally, we also explore the Dale-Chall index (Dale and Chall, 1948) which draws from a list of commonly used words. The formulas of each of these measures, as well as the type of output, are presented in Table 1.

The general approach of the syllable-based measures is to consider the number of syllables in each word and process the results in measure-specific ways. The FKGL and the FRE process syllable information by evaluating the number of syllables per word while the SMOG and the GFI measures exclude "simple" words with two or fewer syllables, thereby focusing only on words with three or more syllables.

Given that the number of syllables in long words is language-dependent, we suspected that the English requirement of 3+ syllables may not be indicative of long words as measured by the number of syllables per word in Sesotho. For instance, in a similar study, Kusec et al. (2002) adjusted the minimum syllables counted from the English helper language to the low-resource language, Croatian. They compared the top 100 frequently used words in English and Croatian to determine the differences between syllable counts in the two languages in order to determine the number of syllables that are typical in Croatian long words. In the end, they adjusted the requirement for polysyllabic words to 4+ syllables. We consider both 3+ and 4+ syllable long words in our experiments.

In addition to syllable information, word length and sentence lengths are also common features used in the measures as is evident in Table 1. Orthographic word length, that is, the lengths of words as measured by the number of letters per word (Ziegler et al., 2001), has been a topic of interest in language studies, with research indicating variations across languages and over time. For instance, Bochkarev et al. (2015) investigate the evolution of word lengths in English and Russian as observed through e-libraries, Google Books, and Google Ngram Viewer. Their findings indicate an increase in the average length of words in both

Measure	Formula	Output
FKGL	$= 0.39(\frac{\#tokens}{\#sentences}) + 11.8(\frac{\#syllables}{\#tokens}) - 15.59$	grade
FRE	$= 206.835 - 1.015(\frac{\#tokens}{\#sentences}) + 84.6(\frac{\#syllables}{\#tokens})$	level
SMOG	$= 3.1291 + 1.043\sqrt{\#polysyllabicwords * (\frac{30}{\#sentences})}$	grade
GFI	$= 0.4[(\frac{\#tokens}{\#sentences}) + 100(\frac{\#complex-words}{\#words})]$	grade
CLI	$= 0.0588(\frac{\#letters}{\#samples}) - 0.296(\frac{\#sentences}{\#samples}) - 15.8$	grade
ARI	$= 4.7(\frac{\#letters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43$	grade
RIX	$= \frac{\#longwords}{\#sentences}$	grade
LIX	$= (\frac{\#words}{\#sentences}) + [\frac{\#longwords}{\#words} * 100]$	grade
DCI	$= 0.0496(\frac{\#words}{\#sentences}) + (\frac{\#difficultwords}{\#words} * 0.1579) + 3.6365$	grade

Table 1: Selected classical readability measures (Flesch-Kincaid Grade Level (FKGL), Flesch-Reading Ease (FRE), Simple Measure of Gobbledygook (SMOG), Gunning Fog Index (GFI), Coleman-Liau index (CLI), Automatic Readability index (ARI), Readability index (RIX), Läsbarhets index (LIX), Dale-Chall index (DCI)), corresponding formulas, and type of output.

languages, with English increasing from 4.4 letters per word in the year 1700 to 4.6 in the year 2000. Additionally, they note that these numbers were reported differently in other studies where the average length of words in English was 5.1 letters per word while that of Russian was slightly higher at 5.28 letters per word (Bochkarev et al., 2015). According to Hefer (2013) words in Sesotho are on average almost a full character shorter than in English. Conversely, Loukatou (2019) indicates an average word length of 4.24 for Sesotho and a lower average of 3.02 letters per word for English in their over-segmentation corpus.

3. Methodology

According to De Clercq et al. (2014), there are at least three steps to describe when developing readability measures. Those are (i) the development of a readability corpus, (ii) describing a methodology, and (iii) undertaking the prediction tasks (François and Fairon, 2012; Collins-Thompson, 2014). We structure the discussion of our methodology for adapting the traditional readability measures into Sesotho using these three steps below.

3.1. Step 1: A readability corpus

Within the context of indigenous languages of South Africa, including Sesotho, the unavailability of readily annotated corpora with readability levels highlights the need to develop new corpora or repurpose existing corpora to train readability measures. In this context, Sibeko and Van Zaanen (2021) suggest the use of examination texts for the creation of readability corpora for South African indigenous languages.

For our study, we employ Sibeko's (2024) readability corpus of Sesotho-English translations. This corpus includes document-level parallel translations of 80 Sesotho reading comprehension and summary writing texts sourced from the grade 12 examination corpus (Sibeko and Van Zaanen, 2023). For texts produced after 2011, the English translations are essentially back translations as the texts were originally translated from English to Sesotho for exam purposes. Note that the Sesotho exam texts indicate that the original source is in English, but they do not indicate exactly where the English texts can be found (hence the back translation process is applied).

The corpus comprises 13,793 words, consisting of 6,040 types, with an average sentence length of 17.73 words in Sesotho. Additionally, the English translations include 12,005 words with 6,130 types, featuring an average sentence length of 15.75 words. The examination texts span from the year 2009 to 2019.

3.2. Step 2: A methodology

The overall methodology consists of three steps. First, we extract relevant text features from Sesotho texts. Second, we use the English translations that correspond to the Sesotho texts to determine readability levels for the texts using traditional readability measures. With this approach, we follow El-Haj and Rayson (2016) who illustrate that the readability of texts in a higher-resourced language can be utilized as a benchmark for the estimation of the readability of texts in a low-resource language. Similarly, we

align the distribution of readability levels in Sesotho with those observed in English translations. Third, we use linear regression models to determine the impact of the text features from the Sesotho texts on the overall scores of the different readability measures computed on the English translations.

To provide some additional insight into the impact of the different text features, we examine the text characteristics employed in traditional readability measures. The following brief discussion outlines some of the text features considered in this article.

3.2.1. Word lengths by letters

There are two main concerns with average word lengths in Sesotho. On the one hand, as an agglutinative language, words may be expected to be relatively long in Sesotho (Blanchard, 2011). On the other hand, monosyllabic words which may comprise between one and four letters (and especially single-letter words) may result in shorter averages for Sesotho texts (Messerschmidt et al., 2003). Furthermore, overall text word length by the letters may be affected by the use of subject concords in Sesotho. Within our dataset, English words exhibit an average of 4.34 letters per word, while Sesotho words demonstrate an average of 4.07 letters per word. Nonetheless, given that the average word length in Sesotho is relatively similar to that of English, we follow the English guideline for the LIX and RIX measures and thus consider words with more than six letters as long words.

3.2.2. Word length by syllables

Polysyllabic words refer to words with more than one syllable. However, the traditional measures used in this research, particularly the SMOG and the GFI measures consider only words with three or more syllables as polysyllabic, foggy, and complex. Within our data set, the English words exhibit an average of 1.26 syllables per word while the Sesotho texts demonstrate 2.0 syllables per word. Sesotho words tend to have more syllables than English words. As such, although we define polysyllabic words (as used in the different metrics) as words with three or more syllables, we also investigate the possibility of increasing the minimum syllables in polysyllabic words to words with four syllables.

3.2.3. Common words

The DCI measure is based on the assumption that there are words that are commonly used and should therefore be easy to read. According to this method, words that do not appear on the list of frequently used words are considered difficult. For our experiment, we use the list of common Sesotho words compiled by Sibeko and De Clercq (2023). We need to use this list with caution, however, since it was not derived from educational texts. Unfortunately, we are not aware of any other word lists available for use in this context.

3.2.4. Samples

Some formulas, like the DCI and CLI measures, require sampling of small amounts of text. As the texts in these experiments are relatively short, we forgo the sampling steps. In this way, for instance, the number of sentences in the CLI formula refers to all sentences in the text instead of a small set of sampled sentences. As can be observed in Table 1, the CLI formula focuses only on word lengths as counted in letters, and sentences in the whole text (for both the English and Sesotho formulas).

3.3. Step 3: Prediction tasks

3.3.1. Correlations

Before we develop text readability measures for Sesotho, we first investigate the interrelationships among the different textual features that underpin the readability measures. This exploration provides insights into the nature of Sesotho text features.

The exploration of the interrelationships between the text features used in the traditional readability formulas was computed using the Pearson correlation measure. The outcomes of these correlations are presented in Table 2. Note that the Labels V1-16 are used to represent the features in columns 1 to 16. Notably, all correlations are significant with $p < .05$.

The examination of Sesotho text features through correlation analysis reveals interesting findings. For example, perfect alignments are uncovered between word and syllable counts, as well as between syllable and letter counts. This suggests a consistent and predictable relationship between these features in that more syllables will result in longer words. Furthermore, strong positive correlations emerge, highlighting the association between syllables per word and the frequency of polysyllabic words, while negative correlations indicate that an increase in letter counts per word may result in fewer sentences, words, and long words.

We also investigated the correlations between syllable-based formulas and syllable-related text features. The findings in Table 3 reveal weak negative correlations between the number of syllables and the scores of the FKGL, the GFI, and the SMOG index. This observation suggests that syllable counts in Sesotho align with those in English, indicating that texts with higher syllable counts are likely to be more challenging to read.

As mentioned earlier, we also considered modifying the criteria for defining polysyllabic words

Feature	Label	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
number of sentences	V1	1.00	.87	.91	.72	.86	.86	.82	.84	.54	-.34	-.29	.29	-.48	.58	-.34
number of words	V2	.87	1.00	.93	.92	.98	1.00	.97	.99	.10	-.24	-.24	.24	-.07	.43	-.21
number of difficult_words	V3	.91	.93	1.00	.84	.94	.93	.91	.92	.28	-.20	-.18	.17	-.24	.71	-.21
number of long_words	V4	.72	.92	.84	1.00	.91	.95	.97	.96	-.05	.11	.10	-.10	.08	.38	.17
number of types	V5	.86	.98	.94	.91	1.00	.98	.96	.98	.11	-.20	-.20	.19	-.09	.50	-.18
number of syllables	V6	.86	1.00	.93	.95	.98	1.00	.99	1.00	.08	-.15	-.15	.15	-.05	.44	-.13
number of polysyllables	V7	.82	.97	.91	.97	.96	.99	1.00	.99	.05	-.04	-.03	.03	-.03	.45	-.02
number of letters	V8	.84	.99	.92	.96	.98	1.00	.99	1.00	.06	-.12	-.13	.13	-.03	.43	-.10
ratio of sentences per word	V9	.54	.10	.28	-.05	.11	.08	.05	.06	1.00	-.24	-.16	.16	-.95	.48	-.32
ratio of letters per word	V10	-.34	-.24	.20	.11	-.20	-.15	-.04	-.12	-.24	1.00	.93	-.93	.21	-.01	.89
ratio of syllables per word	V11	-.29	-.24	.18	.10	-.20	-.15	-.03	-.13	-.16	.93	1.00	-1.00	.14	.05	.88
ratio of words per syllable	V12	.29	.24	.17	-.10	.19	.15	.03	.13	.16	-.93	-1.00	1.00	-.13	-.05	-.89
ratio of words per sentence	V13	-.48	-.07	.24	.08	-.09	-.05	-.03	-.03	-.95	.21	.14	-.13	1.00	-.44	.31
ratio of difficult_words per word	V14	.58	.43	.71	.38	.50	.44	.45	.43	.48	-.01	.05	-.05	-.44	1.00	-.08
ratio of long_words per word	V15	-.34	-.21	.21	.17	-.18	-.13	-.02	-.10	-.32	.89	.88	-.89	.31	-.08	1.00

Table 2: The correlation of text features used in the readability measures computed from the Sesotho texts.

by exploring a potential increase from three to a minimum of four syllables (4+ syllables). Our findings reveal that maintaining a minimum of three syllables consistently demonstrates stronger correlations with readability scores compared to a minimum of four syllables. Consequently, for Sesotho, we also consider only 3+ syllables as in the original English formulas.

3.3.2. Linear Regression Models

Finally, we create linear regression models using the '`lm`' linear regression model function in R to determine the coefficients of the different textual features using our Sesotho training data and the readability levels computed on the English texts. The structures of the Sesotho linear regression models mimic that of the English readability measures. In this way, we try to ensure that the readability values computed using a particular readability measure are used to create a Sesotho readability measure that uses a similar structure and the same textual features as the English measure.

We then created linear regression models for the different measures. The formulas are presented in Table 4. Our proposed readability formulas for Sesotho maintain a degree of structural preservation for the DCI, CLI, SMOG, FRE, and FKGL formulas. Note that a more simplified version of the CLI formula would use the actual counts and not percentages and result in $CLI_{Sesotho} = -3.683470 + 3.8782(\frac{\#letters}{\#words}) - 72.7569(\frac{\#sentences}{\#words})$.

When comparing the weights of the Sesotho formulas with those of the English formulas, we observe several things. First, there is a reduction in the coefficients of syllables per word within the Sesotho formulas concerning the English ones. For example, this manifests as a heightened and negative weighting for syllables per word within the Sesotho FRE formula.

Second, we propose two structures for the GFI formula. Both versions introduce an intercept for the formula, involving a deduction of 0.177916, which is different from the original formulation. The first proposed formula, $GFI(1)_{Sesotho}$, follows the structure of the original English formula more closely although an intercept is added. The second formula, $GFI(1)_{Sesotho}$ introduces a coefficient to the percentage of complex words, thereby deviating from the original structure.

The English LIX and RIX, do not include weights. To align the readability values that were acquired through the application of English readability measures on the translated examination texts, with the text features observed in the Sesotho texts, it was necessary to introduce weighting factors. This adjustment ensured a more accurate correspondence between the readability values and the adapted for-

	Label	F1	F2	F3	F4	F5	F6	F7	F8
KFGL	F1	1.00	-.96	.93	.87	-.04	.01	.33	.16
FRE	F2	-.96	1.00	-.86	-.85	.08	.01	-.45	-.16
GFI	F3	.93	-.86	1.00	.97	-.02	.03	.29	.18
SMOG	F4	.87	-.85	.97	1.00	-.03	.03	.36	.19
syllables	F5	-.04	.08	-.02	-.03	1.00	.99	-.15	.90
3+ syllables	F6	.01	.01	.03	.03	.99	1.00	.00	.92
%3+ syllables	F7	.33	-.45	.29	.36	-.15	.00	1.00	.14
4+syllables	F8	.16	-.16	.18	.19	.90	.92	.14	1.00

Table 3: The correlation of syllable-based measures and syllable information computed on the Sesotho texts.

Measure	Formula
$FKGL_{Sesotho}$	$= -14.08905 + 0.43405(\frac{\#words}{\#sentences}) + 5.86314(\frac{\#syllables}{\#words})$
$FRE_{Sesotho}$	$= 209.3286 - 1.7930(\frac{\#words}{\#sentences}) - 46.6548(\frac{\#syllables}{\#words})$
$SMOG_{Sesotho}$	$= 0.28788 + 0.68741(\sqrt{\#polysyllabic - words * (\frac{30}{\#sentences})})$
$GFI(1)_{Sesotho}$	$= -4.30942 + 0.28610(\frac{\#words}{\#sentences}) + (\frac{\#complex-words}{\#words})$
$GFI(2)_{Sesotho}$	$= -1.77916 + 0.40861((\frac{\#words}{\#sentences}) + 30.9982(\frac{\#complex-words}{\#words}))$
$CLI_{Sesotho}$	$= -3.683470 + 0.038782(\frac{\#letters}{\#samples} * 100) - 0.727659(\frac{\#sentences}{\#samples} * 100)$
$ARI_{Sesotho}$	$= -13.66031 + 2.87106(\frac{\#letters}{\#words}) + 0.49323(\frac{\#words}{\#sentences})$
$LIX_{Sesotho}$	$= 0.46038 + 1.14736(\frac{\#words}{\#sentences}) + 0.60841(\frac{\#long-words}{\#words} * 100))$
$RIX_{Sesotho}$	$= 0.02180 + 0.76883(\frac{\#long-words}{\#sentences})$
$DCI_{Sesotho}$	$= 4.66547 + 0.14199(\frac{\#words}{\#sentences}) + 0.03264(\frac{\#difficult-words}{\#words} * 100)$

Table 4: Readability measures and corresponding adapted Sesotho formulas.

mulas. For the LIX, the impact of words per sentence is accorded weight, while the percentage of long words remains unaltered. However, in the RIX formula, we ascribe weight to the fraction of long words per sentence. Note that we also introduce intercepts for both the $LIX_{Sesotho}$ and $RIX_{Sesotho}$.

Moreover, a noteworthy decrease in the weight attributed to sentences per word[1] is evident in the Sesotho version of the CLI, when contrasting with its English counterpart. Similarly, the intercept of the $CLI_{Sesotho}$ is appreciably lower compared to the English variant. Similarly, a contrast is discernible in the intercept of the $ARI_{Sesotho}$ formula. Despite the consistent coefficient of words per sentence, the Sesotho ARI entails a reduced weighting of letters per word.

Finally, the coefficient of difficult words appears somewhat lower in the Sesotho CLI formula, as opposed to the English formula. Conversely, the Sesotho CLI formula bestows a higher coefficient for words per sentence.

4. Evaluation

The linear regression summary output provides five statistics to assess the performance of each model and the significance of their coefficients. We consider the Adjusted R-squared, F-statistic, and residual standard error. The outcome of the evaluations is presented in Table 5.

[1]The ratio of the number of sentences to the number of words

	FKGL	FRE	SMOG	GFI1	GFI2	CLI	ARI	LIX	RIX	DCI
F-statistic	293.3	118.3	113.0	124.7	109.1	117.4	433.4	144.0	269.5	23.3
Adjusted R^2	.881	.748	.586	.610	.732	.746	.916	.784	.773	.361
Residual std. error	0.647	4.806	0.873	1.166	0.966	0.848	0.626	2.758	0.441	0.631

Table 5: Evaluations of the adapted linear regression models for Sesotho. Note that the p-values are all significant at $p < .001$.

First, the F-statistic is an indicator of the comprehensive validity of the models. It highlights its statistical significance across all models, as evidenced by the observed p-values ($p < .001$). This affirmation attests to the composite contribution of the predictor attributes in elucidating the variation in text readability, thereby proving that results are highly unlikely to be the result of random chance.

Second, the Adjusted R-squared metric indicates how much the independent variables describe the variance of the data. Our analysis reveals higher values particularly for the $ARI_{Sesotho}$, signifying that the variables of letters per word and sentences per word account for approximately 91% of the predictive capacity associated with ARI scores. However, contrasting outcomes are observed for the SMOG formula, where the number of polysyllabic words accounts for only 59.16% of the overall predictive influence. This variation highlights the varying degrees of contribution made by predictor features across the formulated models.

Finally, the lower residual standard errors describe the standard deviation of the residuals, where lower values indicate better results.

5. Discussion and conclusions

The underlying rationale of this research is the absence of an objective method for identifying the readability levels of texts in Sesotho. We postulate that the already low literacy levels in the indigenous languages of South Africa, including Sesotho, could potentially be worsened by the inappropriate selection of textual materials, especially given the limited pool of texts available in the indigenous languages. We expect that being able to gauge the extent of text readability in the language will assist both learners and teachers in the identification of correctly levelled reading materials. In turn, the use of an objective readability assessment framework will improve access to quality (reading education and hence general) education in Sesotho. However, we acknowledge that the limitations previously ascribed to traditional readability measures remain applicable, even in the context of our proposed adaptations.

Given that Sesotho, like most other indigenous languages in Southern Africa, is a low-resource language, no suitable readability labelled corpora exist. To resolve this issue, educational texts originally written in Sesotho were translated into English. The English translations then formed the source of the readability assessment as traditional English readability measures can be applied. Given the extracted textual features from the Sesotho text, combined with the English readability values, linear regression models can be created. The structure of the linear regression models (e.g., the textual features and how they fit together in the formula) is taken from the corresponding English metrics that were used to compute the readability values.

Among the readability formulas adapted within this article, six depend on the sentence length variable. The adapted Sesotho formulas consistently ascribe greater weight to sentence length in comparison to the original English formulas we adapted to Sesotho. Note that, despite this emphasis, no strong correlations emerge between sentence length and the other variables considered in the sentence length-focused formulas. Nonetheless, the CLI formula's coefficient analysis highlights the impact of a strong negative correlation between sentences per word and words per sentence. This observation accentuates that while sentence length is ascribed higher coefficients in numerous Sesotho formulas, the sentences per word variable receive substantially lower weight in the $CLI_{Sesotho}$ formula, thus underscoring the prominence of the sentence length feature in determining Sesotho text readability. It is, however, also important to note that these features are inverse of each other. As such, they are expected to affect readability levels in contrasting ways.

Furthermore, an examination of the correlation between sentence length and word length in terms of syllable counts reveals a modest negative association. This suggests that as sentence length extends, syllables per word exhibit a slight reduction. In essence, an increase in sentence length corresponds to a marginal decrease in both syllables and letters per word, due to the prevailing negative correlation with letters per word. This observation accentuates sentence length as a dominant predictor of Sesotho text readability.

To the best of our knowledge, the findings of this article present the first formulas for an indigenous language of South Africa and the first for the Sotho-Tswana language group in Southern Africa.

Although our models are trained on educational texts, specifically reading comprehension and summary writing texts, the availability of standardised and objective readability formulas provides a solid starting point for employing machine learning approaches to measure text readability in Sesotho. Furthermore, the methods outlined in this article can be employed in the development of readability measures for other low-resource languages.

5.1. Limitations

Our approach in this article is limited by the reliance on written texts and the existing readability measures that were not originally developed for Sesotho. First, we make the assumption that the translated texts have similar readability. The texts are automatically translated and manually corrected to ensure the most similar texts in English compared to the Sesotho texts. A possible solution would be to develop a text collection that is specifically targeted to readability measures based on Sesotho texts (only). Given that not many texts are publicly available in Sesotho, this will remain a challenge.

Second, to ensure the practical usability of the metrics, an empirical examination involving human participants should be undertaken. Such an evaluation would involve selecting and grading texts using our proposed formulas, and subsequently administering these texts to learners within the grade levels indicated by our formulas. This approach is crucial for evaluating the societal impact of our formulas. However, it also presents a challenge in the need for well-defined criteria to distinguish success or failure in the reading tests that would be administered to participants for the evaluation of the readability levels suggested by our formulas for Sesotho.

Finally, to properly measure the impact of the readability metrics through the effectiveness of the selection of suitable texts, criteria for identifying "success" in reading Sesotho texts will need to be developed. This step is important in the context of Sesotho (and the other South African indigenous languages), in particular given the prevailing challenges that South African learners generally encounter in reading.

5.2. Future studies

The findings discussed in this section reveal a number of avenues for possible future studies. First, a further investigation into the used features may provide more suitable metrics for Sesotho. For instance, the optimal number of letters per word that correspond to long Sesotho words remains an intriguing avenue. This entails scrutinising correlations between differing letter counts per word and the resultant scores to ascertain the highest positively correlated number of counts to the readability levels for defining long words for Sesotho. In this article, we use the English definition for long words.

Second, the utilisation of existing grade levels, albeit untested within the South African education context, underscores an avenue for future research. Future inquiries should investigate the applicability of the FRE, CLI, LIX, and other indicators to South African grade levels to refine the contextual relevance of these measures. Currently, we rely on the existing adaptation of readability scores for the FRE, LIX, and RIX measures within the South African context based on the works of Bargate (2012), and Leopeng (2019). Perhaps future works can consider recalibrating such scores to the South African grades through human-based evaluation methods.

Finally, the metrics are not developed in isolation. The practical use of the metrics will need to be investigated in a proper educational context. Do the metrics indeed allow for the identification of suitable texts for a learner? Can we rely on teachers to evaluate this or do we need other evaluation methodologies? Of course, additionally, we will need to investigate how readers experience the readability metric results. Ultimately, we hope that this research help in improving the reading skills of learners in South Africa which we hope to see in future PIRLS results.

6. Bibliographical References

Karen Bargate. 2012. The readability of managerial accounting and financial management textbooks. *Meditari Accountancy Research*, 20(1):4–20.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Bokförlaget Liber, Stockholm, Sweden.

Carl-Hugo Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.

Daniel Blanchard. 2011. Unsupervised word segmentation: An investigation of sub-word features. Online on github. https://dan-blanchard.github.io/papers/proposal.pdf.

Vladimir V Bochkarev, Anna V Shevlyakova, and Valery D Solovyev. 2015. The average word length dynamics as an indicator of cultural changes in society. *Social Evolution and History*, 14(2):153–175.

David Carel. 2019. 4 ways to stay safe online-developing a text difficulty indicator for isiXhosa early grades. In *Policy Commons*. Research on Socio-Economic Policy. Available at: https://policycommons.net/artifacts/2105074/

4-ways-to-stay-safe-online/2860372/.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

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.

Edward Dale and J S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Mahmoud El-Haj and Paul Rayson. 2016. Osman: A novel Arabic readability metric. *Procedia Computer Science*, 142:38–49.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Rudolph Flesch. 1974. *The art of readable writing*, 2nd edition. Harper, New York.

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

Government of Lesotho. 1993. *The Constitution of Lesotho*. Government Printer, Maseru.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill: New York.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Esté Hefer. 2013. Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English. *Southern African Linguistics and Applied Language Studies*, 31(3):359–373.

Sukhpuneet Kaur, Kulwant Kaur, and Parminder Kaur. 2018. The influence of text statistics and readability indices on measuring university websites. *International Journal of Advanced Research in Computer Science*, 9(1):403–414.

Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease formula) for navy enlisted personnel. Report, Defense Technical Information Center.

Daleen Krige and Marianne Reid. 2017. A pilot investigation into the readability of Sesotho health information pamphlets. *Communitas*, 22:113–123.

Nancy C Kula and Lutz Marten. 2008. Central, east and southern african languages. In Peter Austin, editor, *One Thousand Languages*, pages 86–111. Ivy Press/University of California Press.

Sanja Kusec, Miroslav Mastilica, Gordana Pavlekovic, and Luka Kovacic. 2002. Readability of patient information on diabetes on the Croatian Web sites. In *Health Data in the Information Society, Netherlands*, pages 128–132. IOS Press, Amsterdam.

Sandra Land. 2015. Reading and the orthography of isiZulu. *South African Journal of African Languages*, 35(2):163–175.

Makiti Thelma Leopeng. 2019. *Translations of informed consent documents for clinical trials in South Africa: Are they readable?* Thesis, University of Cape Town.

Georgia Loukatou. 2019. From phonemes to morphemes: Relating linguistic complexity to unsupervised word over-segmentation. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, Florence, Italy*, New Jersey, USA. Association of Computational Linguistics.

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.

Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Hans Messerschmidt, JJE Messerschmidt, and DP Thulo. 2003. A human-assisted computer generated LA-grammar for simple sentences in Southern Sotho. *Southern African Linguistics and Applied Language Studies*, 21(1–2):41–47.

Lubna Ali Mohammed, Musheer Abdulwahid Aljaberi, Antony Sheela Anmary, and Mohammed Abdulkhaleq. 2023. Analysing English for science and technology reading texts using Flesch Reading Ease online formula: the preparation

for academic reading. In *International Conference on Emerging Technologies and Intelligent Systems*, pages 546–561, New York. Springer.

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.

Rosalia Moroesi Nakin. 2009. *An examination of language planning and policy in the Eastern Cape with specific reference to Sesotho: A socioliguistic study*. Ph.D. thesis, Nelson Mandela Metropolitan University, South Africa.

Moselane Andrew Nhlapo. 2021. *Historical perspectives on the development of Sesotho linguistics with reference to syntactic categories*. Ph.D. thesis, University of the Free State, South Africa.

Parliament of Zimbabwe. 2021. *The Constitution of Zimbabwe*. Veritas, Harare.

Michelle Rasheed. 2023. Kindling a desire to read: A review of three young adult novels. *South Carolina Association for Middle Level Education Journal*, 2(1):109–113.

Marianne Reid, Mariette Nel, and Ega Janse Van Rensburg-Bonthuyzen. 2019. Development of a Sesotho health literacy test in a South African context. *African journal of primary health care & family medicine*, 11(1):1–13.

Republic of South Africa. 2023. *Constitution eighteenth amendment bill*. Department of Justice and Correctional Services, Pretoria.

Karen Roux, S van Staden, and M Tshele. 2021. *Progress in International Reading Literacy Study 2021: South African Preliminary Highlights Report*. Department of Basic Education, Pretoria, South Africa.

Ntaoleng Belina Sekere. 2004. *Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa*. Thesis, University of South Africa.

Mmasibidi Setaka. 2018. *Corpus-based Lexicography for Sesotho*. Ph.D. thesis, University of Pretoria, South Africa.

Mmasibidi Setaka and Danie J Prinsloo. 2020. A critical evaluation of three sesotho dictionaries. *Lexikos*, 30:445–469.

Johannes Sibeko. 2022. Tshebediso ya melao kabong ya dinoko tsa Sesotho. *Southern African Linguistics and Applied Language Studies*, 40(4):494–506.

Johannes Sibeko. 2023. Using classical readability formulas to measure text readability in Sesotho. In Tomaž Erjavec and Maria Eskevich, editors, *Selected papers from the CLARIN Annual Conference 2022*, volume 198, pages 120—132. Linköping Electronic Conference Proceedings, Prague, Czechia.

Johannes Sibeko. 2024. Harnessing google translations to develop a readability corpus for sesotho: An exploratory study. *Journal of the Digital Humanities Association of Southern Africa*, 5:1–12.

Johannes Sibeko and Orphée De Clercq. 2023. A corpus-based list of frequently used words in Sesotho. In *Proceedings of the Fourth workshop on Resources for African Indigenous Language (RAIL 2023), Dubrovnik, Croatia*, pages 32–41, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

Johannes Sibeko and Menno Van Zaanen. 2021. An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11.

Johannes Sibeko and Menno Van Zaanen. 2023. A data set of final year high school examination texts of South African home and first additional language subjects. *Journal of Open Humanities Data*, 9(9):1–6.

Edgar A Smith and R.J Senter. 1967. *Automated readability index*. Clearing house for Federal Scientific and Technical information, Cincinnati, Ohio.

Elsje van Bergen, Margaret J Snowling, Eveline L de Zeeuw, Catharina EM van Beijsterveldt, Conor V Dolan, and Dorret I Boomsma. 2018. Why do children read more? the influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, 59(11):1205–1214.

Phillip Van Oosten, Dries Tanghe, and Véronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta*, pages 775–782, Paris. European Language Resources Association (ELRA).

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.

Shixiang Zhou, Heejin Jeong, and Paul A Green. 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 6:97–111.

Johannes C Ziegler, Conrad Perry, Arthur M Jacobs, and Mario Braun. 2001. Identical words are read differently in different languages. *Psychological science*, 12(5):379–384.