Developing Resources for Measuring Text Readability in Sesotho

Johannes Sibeko
Department of Linguistics and Applied Linguistics
Nelson Mandela University
Gqeberha, South Africa
johannes.sibeko@mandela.ac.za

Abstract

This article presents a work-in-progress doctoral project that explores measuring text readability in Sesotho, a Bantu language spoken by more than 10 million speakers across Southern Africa. The main project adopts a classical readability formulas approach to text readability analysis. We aim to adapt nine existing readability metrics into Sesotho using English as a higher-resourced helper language. So far, five resources have been developed as part of the study. The rule-based and the TEX-based syllabification systems, the syllable annotated wordlist, and the grade 12 exam reading comprehension and summary writing corpus have been published on the South African Centre for Digital Language Resources' (SADiLaR) online repository. The machine-translated corpus is still under development. This article describes the progress of the PhD project by overviewing the basic digital language resources developed for the project. The metrics under consideration for adaptation into Sesotho are also briefly discussed.

1 Introduction

Automated text readability evaluation has been applied in different application domains such as finding educational materials (Collins-Thompson, 2014). The scholarship of text readability has continued for over a century (Collins-Thompson, 2014; De Clercq and Hoste, 2016). To date, more than 200 metrics have been developed (DuBay, 2004). However, indigenous African languages have been neglected in this area (Sibeko and Van Zaanen, 2021). As a result, matching texts of the right level of readability to readers such as books for learning and teaching in languages where no metrics are available depends on each assessor's intuition. Undoubtedly, intuition-based choices are likely to be flawed, inconsistent and influenced by the content of the text. Unfortunately, although textbooks are the most important teaching material in language teaching discourse, inappropriate textbook use can deskill both language teachers and language learners (Mohammed et al., 2022). Deskilling can be propelled by a mismatch between textbooks and intended readers (Zamanian and Heydari, 2012). Such dissonances can result from incorrect levels of readability which can be expected when texts are chosen based on intuition. To reduce the chances of deskilling the language reader, it is essential to have a system for objectively estimating the readability of reading texts of varied lengths such as textbooks and comprehension texts.

Post-graduate research projects have made significant contributions to the scholarship of text readability. Researchers have used various methods to study text readability, including classical readability metrics (Kondru, 2006; Feng, 2010; Janan, 2011; Bendová, 2021), machine learning (Sjöholm, 2012; Andova, 2017), Natural Language Processing (Dios, 2016), deep learning (Alkaldi, 2022), and eye-tracking (Newbold, 2013). Furthermore, other research has adapted existing readability metrics to lower-resourced languages (Bendová, 2021). Unfortunately, research in lower-resourced languages such as Sesotho is restrained by the lack of training data. As a result, machine learning approaches are not feasible.

This article reports on a work-in-progress doctoral project on measuring text readability in Sesotho using classical readability metrics. A brief overview of writing in Sesotho is presented (Section 2) followed

Johannes Sibeko 2023. Using classical readability formulas to measure text readability in Sesotho. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 120–132. DOI: https://doi.org/10.3384/ecp198012

by a brief synopsis of the classical readability metrics that are adapted in the main project (Section 3). Then the relevance of the project to CLARIN via SADiLaR is briefly highlighted (Section 4) followed by a discussion of the resources produced as part of the project (Section 5). The article is concluded with a discussion of future works (Section 6).

2 Contextualising Sesotho

Sesotho is used by more than ten million speakers in a few countries in Southern Africa including South Africa, Lesotho, and Zimbabwe (Marupi and Charamba, 2022; Sibeko and Setaka, 2022). In fact, it is one of the official languages in Lesotho, Zimbabwe, and South Africa. It is used as a language for learning and teaching in these countries as either a mother tongue, second language, or marginalised language. It is also used for media, political, religious, and other uses. Even so, a recent investigation of the Sesotho Basic Language Resource Kit (BLARK) content has revealed that there is a severe shortage of digital language resources available for Sesotho (Sibeko and Setaka, 2022). As such, Sesotho remains a low-resourced language (Roux and Bosch, 2019; Sibeko and Setaka, 2022). Consequently, automating the process of objectively investigating text readability in Sesotho using classical readability metrics requires the development of a few basic language resources.

In addition to the lack of necessary resources which hinders the development of objective automated metrics for measuring text readability in Sesotho, there are two widely recognised orthographies for Sesotho. The two orthographies are differentiated by the two countries with the most speakers of Sesotho. They are therefore labelled accordingly as the South African Sesotho (SAS) orthography and the Lesotho Sesotho (LS) orthography. The main differences between the two orthographies include the use of w and y in the SAS orthography as opposed to the use of o and e in the LS orthography for representing semivowels. This is exemplified in example 1 below.

(1) *Ke wena le yena. - SAS.*
It's you and her.
'It's you and her.'

Ke oena le eena. - LS.
It's you and her.
'It's you and her.'

The LS orthography also uses l in place of d, and c in place of the digraph tj. The differences are exemplified in example 2 below.

(2) *O dula a tjha. - SAS*
He always is burning.
'He is always burning.'

O lula a cha. - LS
He always is burning.
'He is always burning.'

Although differences such as preferences for certain single letters do not affect the results of the metrics adapted in the main project, the different representations of the semivowels will affect syllable identification. Furthermore, the use of single letters in one orthography and the use of digraphs in another orthography may affect average word lengths. Beyond orthography, there may be region-based vocabulary variations in Sesotho. For instance, see Lemeko (2018) for a discussion of region-based variations of SAS. Nonetheless, these variations are not within the scope of the current project. The research focus is limited to the effects of orthographic conventions.

3 Classical Readability Metrics

The doctoral project described in this article focuses on how text readability could be measured in Sesotho. An automated process for measuring text readability in Sesotho is desired. We believe that classical readability metrics are a good place to start. We identified a total of nine classic readability formulas for adaptation into Sesotho. These metrics are used in the Python 3.2 readability package[1]. All nine metrics have been used in previous research on South African educational texts, for instance, see Sibeko and Van Zaanen (2021). The readability metrics that we hope to include in our web-based platform for measuring Sesotho text readability are briefly described below. In-depth discussions of these metrics are provided elsewhere, for instance, *see* Heydari (2012) and Zamanian and Heydari (2012).

3.1 Syllable-Based Metrics

Four of the readability metrics identified are based on syllable information. The metrics are described below.

3.1.1 Flech-Kincaid Grade Level (FKGL)

The FKGL uses US grades for labelling readability levels (Kincaid et al., 1975). For example, a score of 10 corresponds to the tenth-grade (Toyama et al., 2017). The FKGL metric uses the following formula:

$$\text{FKGL} = 0.39(\frac{\#tokens}{\#sentences}) + 11.8(\frac{\#syllables}{\#tokens}) - 15.59$$

The process for calculating readability follows four steps. First, the total number of words is divided by the total number of sentences and multiplied by the weight given to sentence difficulty, i.e., 0.39. Second, the total number of syllables is divided by the total number of words and multiplied by 11.8. which is the weight given to average word difficulty, that is, the average number of syllables per word. In the third step, the resulting numbers from the first and the second steps are added together. Finally, 15.59 is subtracted from the result of step 3 (Boles et al., 2016).

The FKGL metric was developed for the United States Navy. However, it is suitable for use in multiple contexts including educational contexts (Zhang et al., 2019).

3.1.2 Flesch Reading Ease (FRE)

Flesch's (1948) FRE is calculated using the following formula:

$$\text{FRE} = 206.835 - 1.015(\frac{\#tokens}{\#sentences}) + 84.6(\frac{\#syllables}{\#tokens})$$

The FRE formula outputs scores between zero and 100 (Flesch, 1948). While a text with a score of 100 should be easily readable to a language learner with a fourth-grade education, a text with a score of 0 requires at least a college graduate level for reading with ease.

FRE is one of the most used classical readability formulas. In fact, when combined, the FKGL and FRE can be used for both first and second-language texts (Greenfield, 2004). Both FRE and FKGL are integrated into Microsoft Office (Bendová, 2021). As a result, they can be easily used by anyone who uses Microsoft office products. Furthermore, the FRE metric is the most adapted to other languages (Bendová and Cinková, 2021). For instance, it has been adapted to Italian, French, Spanish, German, Russian, Danish, Bangla, Hindi, and Japanese.

3.1.3 Gunning Fog Index (GFI)

The GFI identifies foggy words which are words comprised of more than two syllables (Zhang et al., 2019; Gunning, 1952; Gunning, 1969; Gunning, 2003). The GFI follows four steps. First, the number of words used per sentence is averaged. Second, the number of foggy words is counted. Third, the percentage of foggy words in the sample is calculated. Finally, the totals are added and multiplied by 0.4 (Eleyan et al., 2020). The following equation is used in the GFI:

[1] https://github.com/andreasvc/readability/

$$\text{GFI} = 0.4[(\frac{\#tokens}{\#sentences}) + 100(\frac{\#complex-words}{\#words})]$$

The readability score generated by the English equation above typically falls within the range of 6 to 20. A score of 6 indicates that the text is suitable for a sixth-grade reading level, while a score of 20 or higher suggests that the text is appropriate for advanced readers, such as those in university postgraduate programs.

3.1.4 Simple Measure of Gobbledygook (SMOG)

When calculating SMOG for long texts, three samples are used, one from the beginning of the text, one from the middle, and one from the end of the text ((Mc Laughlin, 1969; Zhou et al., 2017). Each sample comprises ten sentences. The samples are used to calculate SMOG using the following formula:

$$\text{SMOG} = 3.1291 + 1.043\sqrt{\#polysyllabicwords * (\frac{30}{\#sentences})}$$

Polysyllabic words as indicated in the formula refer to words with more than two syllables (Kasabwala et al., 2012). The SMOG formula also outputs US grade levels.

3.2 Word-Length-Based Metrics

Four of the selected classical readability metrics are based on word lengths. The four metrics are described below.

3.2.1 Lasbarhetsindex (Lix) and Rate Index (Rix)

Lix was originally developed for Swedish (Björnsson, 1983). For access, we use the English version as a point of reference. It is suggested that ten samples comprising ten sentences each be analysed when estimating both Lix and Rix (Anderson, 1983). The Lix and the Rix formulas pay special attention to 'long words,' that is, words that have more than six characters. The Lix formula is presented in table 1.

Lix	**Rix**
Lix = $(\frac{\#words}{\#sentences}) + [\frac{\#longwords}{\#words} * 100]$	Rix = $\frac{\#longwords}{\#sentences}$

Table 1: The Lix and the Rix formulas

The Lix formula outputs numbers that are then converted to grade levels. Anderson (1983) states that fractions can be ignored. This is particularly important in instances where adjusting the scores changes the predicted grade. For instance, when adjusting a Lix score of 47.99 to 48.0, the predicted grade level changes from the 11th to the 12th grade.

The Rix metric is an adaptation of the Lix metric (Courtis, 1987; Anderson, 1983). It considers the ratio of long words to the number of sampled sentences. While shorter texts may be considered as a whole, longer texts may use sentence sampling methods. The Rix metric assigns grade levels through the formula presented in table 1.

3.2.2 Coleman-Liau Index (CLI)

The CLI metric also uses a sampling method (Coleman and Liau, 1975). First, the text is divided into shorter samples of 100 words each. Second, the samples are counted. Third, the number of characters in each word from the samples is calculated. Fourth, the number of characters per word is divided by the number of samples. Fifth, the number of sentences is counted. Sixth, the number of sentences is divided by the number of samples. Finally, the results are applied to the following formula:

$$\text{CLI} = 0.0588(\frac{\#letters}{\#samples}) - 0.296(\frac{\#sentences}{\#samples}) - 15.8$$

According to Coleman and Liau (1975), samples should end with complete sentences. As a result, CLI samples may contain a little less or more than 100 words depending on the last complete sentence sampled.

3.2.3 Automated Readability Index (ARI)

ARI is derived from fractions representing predictions of word and sentence difficulty (Kaur et al., 2018; Smith and Senter, 1967). The process follows a few steps. First, sentence lengths are averaged and multiplied by 0.5. Second, word lengths are averaged and multiplied by 4.7. Third, the totals are combined and 21.43 is deducted. The grade level is assigned through the following formula:

$$\text{ARI} = 4.7(\frac{\#letters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43$$

Letters as used in CLI and ARI, refer to all letters and numbers that build words (Zhang et al., 2019). Thomas et al. (1975) describe it as strokes representing each word.

3.3 Frequency-List-Based Metric

One frequency-list-based metric was identified from Python 3.2's readability package. The metric is described below.

3.3.1 Dale-Chall Index (DCI)

The DCI metric uses a frequency list (Dale and Chall, 1948). The frequency list is based on a list of 3000 words that a grade 4 learner is expected to be familiar with (Stocker, 1971). Difficult words are considered as those that do not appear in the list. Variations include words in plural forms, verbs that end in -s, -ed, -ing, and -ied, adverbs that end in -ly, names of both people and organisations [note that organisation names are counted only two times per 100-word sample], abbreviations, and compound words (Barry and Stevenson, 1975). DCI is computed using the following formula:

$$\text{DCI} = 0.0496(\frac{\#words}{\#sentences}) + [11.8(\frac{\#difficultwords}{\#words}) * 0.1579] + 3.6365$$

It is advisable to use the whole text when texts are too short for sampling. For longer texts, one may sample four sets of 100 words per 2000 words (Barry, 1980; Dale and Chall, 1948).

3.4 Summary

Two important things can be noted from the nine classical readability metrics briefly overviewed in this article. First, the metrics use specific processes for estimating appropriate readability and grade levels. It is important to consider these processes. For instance, this is useful when considering the minimal corpus size necessary for adapting the metrics to Sesotho. Second, the metrics use specific weights that may need to be adapted to Sesotho. For instance, syllable lengths may have minimal effect on the level of readability and therefore need to carry minimal weighting. Additionally, it is not possible to investigate the effect of syllable lengths on Sesotho texts without a system for identifying the syllables. For this reason, we need the resources that are developed in the doctoral project discussed in this article.

Moreover, it is important to consider the expected outputs from the formulas. This is most important for formulas that do not output grade levels, for instance, the FRE, Lix and Rix. For these metrics, we may need to redefine the conversions to suit the context of Sesotho and to reflect South Africa's grade levels. In spite of being used in multiple contexts, the classical readability metrics have been criticised for failing to measure comprehension (Tanprasert and Kauchak, 2021). Furthermore, the use of frequency lists, such as in the DCI list of common words, has been criticised for failing to account for specialised meanings (Yan et al., 2006). Even so, the classical readability metrics remain relevant to our project since our focus is not on meaning or comprehension but on the ease with which the text can be read.

4 Relevance to SADiLaR

This PhD project is conducted at North-West University which hosts the South African Center for Digital Language Resources (SADiLaR). SADiLaR is an observer at the CLARIN European Research Infrastructure Consortium. North-West University functions as a hub of a network of linked nodes for

SADiLaR. There are currently six nodes including four universities and two independent research entities. SADiLaR is a national center supported by the South African Department of Science and Innovation as part of the South African Research Infrastructure Roadmap (Wilken et al., 2018; Roux and Bosch, 2019). It has an enabling function with a focus on all official languages of South Africa (Roux and Ndinga-Koumba-Binza, 2019). It supports research and development in language technologies and language-related studies in the humanities and social sciences. The center impacts three domains, namely, (i) humanities and social sciences, (ii) language technology, and (iii) socio-economic domains. This doctoral project benefits from SADiLaR's humanities and social sciences domain which focuses on building research capacity. For instance, several capacity-building training opportunities were freely provided by SADiLaR. The project also benefits from the language technology domain which focuses on the development of high-level resources and NLP tools for use in applications. For instance, some of the resources discussed in this article resulted from collaborative work with experts from SADiLaR. Furthermore, the project has contributed to the development of digital language resources as part of the main project in adapting classical readability metrics to Sesotho.

5 Developing Resources for Sesotho

This section describes five resources developed for the project. Two syllabification systems are described, followed by three annotated datasets.

5.1 Syllabification Systems

A survey of Sesotho digital language resources listed on SADiLaR's repository web interface indicated the absence of syllabification systems for Sesotho[2] (Sibeko and Setaka, 2022). For this reason, previous assessments of the readability of Sesotho texts using classical readability metrics relied on the manual extraction of textual properties such as syllable information. Sadly, using annotators to manually extract Sesotho syllable information from written texts is laborious (Krige and Reid, 2017). Additionally, reliance on such manual methods for extracting textual properties would not suffice for the envisaged automated tool. For this reason, two syllabification systems were developed. The systems are briefly described below.

As a tonal language, Sesotho carries tone by vowels and nasal consonants (Guma, 1982; Sekere, 2004; Mohasi et al., 2011). According to Guma (1982), nasal consonants, that is, two simple nasal consonants n and m and two complex nasal consonants ŋ and ɲ, and the lateral consonant l can occur as syllables. Furthermore, vowels [ɑ, e, i, o, u, ɪ, ɛ, ɔ, and ʊ] can function as syllables (V). The vowel-only syllables can occur at word initial, word medial, and word-final positions. Nasal consonant-only (C) syllables can also occur in these positions. However, only the complex nasal consonant ŋ can occur at word-final position (Demuth, 2007). Finally, syllables can be composed of consonants and vowels (CV). Table 2 presents the syllable types, subtypes, and examples for each subtype. The syllable boundaries are indicated by the use of dashes (-).

We based our syllabification rules on Guma's (1982) syllable types. For testing the system, we extracted syllabification information from Chitja's (2010) dictionary. The process for extracting syllable information from the dictionary and creating a wordlist is described in section 5.2.1. The wordlist represented all syllabification types presented in Table 2. The rule-based system achieved an accuracy rate of 99.69%. We also experimented with a TeX-based approach. We used the wordlist [see section 5.2.1] for training and testing the machine learning system. This system achieved an accuracy rate of 78.92%. The lower accuracy rate of the TeX-based system is attributed to two unavoidable shortcomings. First, we noticed that there was some human oversight while manually cleaning the training corpus. Second, the TeX-based system cannot handle single-letter syllables at the beginning or end of words. Both systems are publicly available on SADiLaRs's repository (see Sibeko and Van Zaanen (2022a)).

[2] SADiLaR indexes both publicly available, and privately hosted digital language resources. For resources that are not publicly accessible, only metadata is indexed. The contact details of the host and sometimes the creators are then provided in case one needs access to the resource. The repository can be accessed online at `https://repo.sadilar.org/`.

Type	Sub-types	Input	Syllabified	English
V	word-initial vowel	*ama*	*a-ma*	touch
	consecutive vowels	*baena*	*ba-e-na*	brethren
	word-final vowel	*letswai*	*le-tswa-i*	salt
CV	one consonant - one vowel	*panana*	*pa-na-na*	banana
	one consonant - semi-vowel- one vowel	*lwana*	*lwa-na*	fight
	two consonants - one vowel	*tlala*	*tla-la*	hunger
	two consonants - semi-vowel- one vowel	*shwang*	*shwa-ng*	dieing
	three consonants - one vowel	*tlhase*	*tlha-se*	spark
	three consonants - semi-vowel- one vowel	*tshwela*	*tshwe-la*	spit
C	nasal consonant n, m - non-nasal consonant	*ntate*	*n-ta-te*	father
	nasal consonant n, m - nasal consonant	*mme*	*m-me*	mother
	nasal consonant n - complex nasal consonant	*nnyatsa*	*n-nya-tsa*	disrespects me
	complex nasal consonant ŋ - vowel	*ngala*	*nga-la*	abandon
	complex nasal consonant ŋ- non-nasal consonant	*mangmang*	*ma-ng-ma-ng*	so so
	word-ending complex nasal consonant ŋ	*hang*	*ha-ng*	once
	consecutive lateral consonants l	*llela*	*l-le-la*	weep for

Table 2: Syllabification rules and examples.

5.2 Annotated Datasets

5.2.1 Syllabified Wordlist

As part of developing the syllabification systems, we developed a gold-standard syllable information annotated corpus. We extracted dictionary entries and syllable information from *Bukantswe ya Machaba ya Sesotho* 'The international dictionary of Sesotho' (Chitja, 2010). Each dictionary entry contains valuable pieces of information. See, for instance, example 3 below.

(3) ***Diepollo** (di-e-pu-l-law)* /exhumation/ *Ketso ya ho epolla kapa ho ntsha ntho e epetsoeng tlasa mobu. Ketso ya ho ntsha bafu mabitleng. (**bap**. Kepollo).* - LS
'**Diepollo** (di-ep-ul-law) /exhumation/ The act of unearthing things buried under the soil. Acts of digging up deceased people from their tombs. (**comp**. Kepollo).'

In example 3, the dictionary entry indicates the Sesotho word, *Diepollo* 'exhumation', then provides pronunciation information in brackets *(di-e-pu-l-law)*, the English translation 'exhumations', the definition '*Ketso tsa ho epolla kapa ho ntsha ntho tse epetsoeng tlasa mobu. Ketso ya ho ntsha bafu mabitleng.*' Which translates to '[T]he act of unearthing things buried under the soil. Acts of digging up deceased people from their tombs.' Finally, a similar word is provided, that is '*bon. kepollo*' which in this case is the singular form: 'exhumation'. For our project, we extracted the dictionary entries, followed by the pronunciation information as in the example below

(4) *Diepollo (di-e-pu-l-law) - SAS*
Exhumations (ex-hu-ma-tions)

As illustrated in example 4, pronunciation information was not always consistent with orthographic conventions. In instances of such inconsistencies, the wordlist was manually cleaned on a word-for-word basis to ensure consistent orthography. For instance, we altered the pronunciation information at example 4 above. That is, we adjusted the third syllable which illustrated a high tone *o* by using the letter 'u', as in *pu* and the fifth syllable which indicated a lower tone *o* by using the digraph 'aw'. The modified syllables are presented in example 5 below:

(5) *Diepollo* *(di-e-po-l-lo)* - *SAS*
‘Exhumations’

Some pronunciation information included words ending in non-syllabic consonants, others changed the spelling such as in example 4, and others had incorrectly placed syllable boundaries. All of these issues were manually checked and fixed. After manual cleaning and fixing orthographic inconsistencies, we obtained a total of 13 551 words. The cleaned wordlist was also uploaded onto SADiLaR’s repository (see Sibeko and Van Zaanen (2022b)).

5.2.2 Reading Comprehension and Summary Writing Texts

We were granted access to grade twelve exam question papers by the South African National Department of Basic Education (DBE). Grade twelve is the high school exit grade in South Africa. We have since extracted reading comprehension and summary writing texts from the exam question papers. We did this for all eleven official languages of South Africa[3]. The texts in our corpus are split into two categories, that is, the home language (HL) and the first additional language (FAL). Previous research indicated that the English exam texts show consistently lower readability levels for the HL texts as opposed to the FAL texts (Sibeko, 2021; Sibeko and Van Zaanen, 2021). The lengths of texts in the collection vary according to the orthographies such as disjunctive and conjunctive, and text types such as reading comprehension and summary writing. Consistently, the lengths of summary texts are about a third of the reading comprehension texts in all eleven languages. The corpus has been uploaded to SADiLaR’s repository (see Sibeko and Van Zaanen (2022c)). We hope that the differences in text readability and linguistic complexity are uniform throughout the different languages.

5.2.3 Machine-Translated Corpus

Previous studies evaluating the text readability of Sesotho texts, for instance, Krige and Reid (2017) and Reid et al. (2019), assumed that classical readability metrics that are based on syllable information and word-length-based textual properties can be directly used in Sesotho without taking the differences between the superficial textual features of English and Sesotho into consideration. However, it is evident from other studies adapting the weights of these textual features to language-specific conventions that the metrics cannot be applied to new languages without adjustments. The syllabification systems described in this article enable the automatic identification and counting of syllables. This is already different from previous research on Sesotho text readability. However, we still aim to adapt the metrics to the specific context of Sesotho. Such an adaptation is important for accounting for differences in superficial text properties between Sesotho and English. A gold-standard corpus with clear levels of text difficulty is needed to develop an automated readability model (Van Oosten et al., 2010; François and Fairon, 2012). Unfortunately, Sesotho, like other LRLs, does not have corpora readily annotated with levels of difficulty (Filho et al., 2016). The use of translated texts may provide a solution to this lack of levelled texts. For instance, the texts can be easily levelled according to grades in English.

It was observed that when Obonerva’s (2006) readability model that was trained on fiction texts was evaluated on non-fiction texts, exaggerated readability levels were observed (Solovyev et al., 2018). Since we hope that education stakeholders such as teachers, parents, learners, textbook authors, and examiners can use our envisaged automated tool for analysing Sesotho text readability, we are training our models on educational texts. We are relying on texts collected as part of Sibeko and Van Zaanen’s (2022c) corpus described in section 5.2.2 above. We identified texts from grade 12 Sesotho HL and FAL examinations from the collection. As a rule of thumb, we followed Zamanian and Heydari’s (2012) guideline that a text should have at least 200 words for metrics like FRE and FKGL to be applied successfully. As a result, we identified texts of no less than 200 words each. In the end, we could only use longer reading comprehension and summary texts.

For an illustration of the texts collected, table 3 presents the original Sesotho summary writing text from the December 2012 question paper together with the unedited translation from Google translate. In

[3] Since 1994, South Africa recognises eleven official languages, that is, Afrikaans, English, IsiNdebele, IsiXhosa, IsiZulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, and Xitsonga. Efforts are being made to officialise Sign language as the twelfth official language of South Africa.

Ho phedisana le baahisane ba mona	Living with your neighbours
Ho bohlokwa ho ba le dikamano tse ntle le baahisane. Le phela mmoho mme le lokela ho thusana nakong tsa mathata. Leha ho le jwalo, dikamano di ba le diphepetso ha moahisane a rata ditaba. Kopana le moahisane kgafetsa le arolelane mehopolo ya ho ntlafatsa maphelo a lona. Ka ho arolelana tlhahisoleseding, moahisane o tla lemoha hore o sebetsa ka thata ho fumana seo o nang le sona. Ebang a batla ho eketsa ntlo kapa ho reka ho itseng, mo eletse hore a ka fumana hokae dintho ka theko e tlase. Ka tsela ena o tla bona hore o a mo kgathalla ebile o a mo tshehetsa. *Ha le arolelana ditoro le ditabatabelo, nnetefatsa hore dipuo tsa hao ha di mo fe pelaelo ya hore o a ikgantsha, ho seng jwalo o tla batla tsela ya ho o sitisa. Tse ding tsa dintho e be makunutu a hao. Ha a o botsa dipotso ka se itseng, o fane ka karabo e teletsana e arabang dipotso tsa hae. Ha ho na le se sa o kgotsofatseng, bua hantle o ikokobetse, o rarolle qaka ka tsela e ke keng ya baka kgohlano. Ha moahisane a nahana hore o motho wa maemo a hodimo , se etse dintho tse nnetefatsang mohopolo oo wa hae. Dula o ikokobeditse ka dinako tsohle, o se ke wa iketsa betere ho mo feta. Ha moahisane enwa a ka bua leshano ka wena, iphapanye o emele mohla e mong a o tobang mme o nke mohato wa ho bua ka seo.*	It is important to have good relations with your neighbors. You live together and should help each other in times of trouble. However, relationships have challenges when the neighbor likes the news. Meet your neighbor often and share ideas to improve your lives. By sharing information, your neighbor will realize that you are working hard to get what you have. Whether he wants to add to the house or buy something, advise him where to find things at a low price. This way she will see that you care about her and support her. When you share dreams and aspirations, make sure that your words do not give him the suspicion that you are proud, otherwise he will look for a way to distract you. Some of the things should be your secrets. When he asks you questions about something, give a longer answer that answers his questions. If there is something that does not satisfy you, speak clearly and humbly, solve the problem in a way that will not cause conflict. If your neighbor thinks that you are a high-class person, don't do things that confirm that idea. Stay humble at all times, don't pretend to be better than him. If this neighbor lies about you, ignore it and wait for someone else to point it out and take the step to talk about it.

Table 3: Sesotho FAL 2012 summary writing text extracted from the February-March exam together with the corresponding Google translation of the Sesotho FAL 2012 summary writing exam text.

this version of the text, we have removed the sentence markers <utt> that were inserted during Sibeko and van Zaanen's (2022c) tokenization and sentence segmentation process. The text contains examples of figurative language. The Google Translate machine translation in table 3 indicates that at least all the words are successfully translated. Even so, instances of figurative language used in the Sesotho source text were translated out of context and new meanings were created.

The machine translations were post-edited to enable checking whether meaning influenced the readability of texts. Our translation corpus contains the original Sesotho texts, the original machine translations, and the human post-edited versions. The post-editing brief indicated that texts should not be changed unless meaning had been lost. As a result, in the human post-edited versions, machine translations such as liking the news were adapted to meaning-appropriate constructions such as nosy neighbours.

6 Conclusion

This article reported a work-in-progress PhD project. A survey of methods used for measuring text readability in low-resource languages indicated a prevalence of adapting classical readability metrics from high-resourced languages such as English. One of the common methods for adapting classical readability metrics was the use of translated texts between higher-resourced languages as helper languages and lower-resourced languages. Classical readability metrics use shallow textual features such as (i) the number of words, and (ii) the lengths of sentences, both of which can easily be counted, (iii) syllabic information for which we had to develop systems, and (iv) a frequency wordlist. Four resources were created and made available on the SADiLaR repository. One more resource is still under development. The resources include both gold-standard corpora and basic digital language resources such as syllabification systems. Finally, we identified the metrics we hope to adapt to Sesotho while also indicating the textual properties considered by each metric. At this point, we have put together most of the necessary tools for the identification and assessment of surface-level textual properties used in the nine readability metrics chosen. The main aim of the bigger study is to develop a platform for automated measurement of the readability of Sesotho texts. To this end, future works include the development of a list of frequently used words in Sesotho which will then enable the adaptation of the Dale-Chall index to Sesotho. Furthermore, all nine metrics will be adapted and a web-based platform will be developed and made publicly accessible.

Acknowledgements

The doctoral project described in this article is completed at the North-West University in South Africa. It is supervised by Professor Menno van Zaanen of the South African Centre for Digital Language Resources.

References

Alkaldi, W. 2022. Enhancing text readability using deep learning techniques. Université d'Ottawa/University of Ottawa.

Andova, A. 2017. Assessment of text readability using statistical and machine learning approaches. University of Ljubljana.

Anderson, J. 1983. Lix and Rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Barry, J. G. 1980. Computerized readability levels. *IEEE Transactions on Professional Communication*, 23(2): 88–90.

Barry, J. G. and Stevenson, T. E. 1975. Using a computer to calculate the Dale-Chall formula. *Journal of Reading*, 19(3):218–222.

Bendová, K. 2021. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal of linguistics*, 72(2):477–487.

Bendová, K. and Cinková, S. 2021. Adaptation of classic readability metrics to Czech. *Proceedings of the International Conference on Text, Speech, and Dialogue: 24th International Conference*, 159–171.

Björnsson, C.-H. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, 480–497.

Boles, C. D., Liu, Y., and November-Rider, D. 2016. Readability levels of dental patient education brochures. *American Dental Hygienists' Association*, 90(1):28–34.

Chitja, M. 2010. *Phatlamantsoe ya Sesotho ya machaba.* Mazenod Publishers, Maseru, Lesotho.

Coleman, M. and Liau, T.L. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Collins-Thompson, K. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Courtis, J. K. 1987. Fry, SMOG, Lix and Rix: Insinuations about corporate business communications. *The Journal of Business Communication*, 24(2):19–27.

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

De Clercq, O. and Hoste, V. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

Demuth, C. 2007. Sesotho speech acquisition. In McLeod, S. *The international guide to speech acquisition.* 526–538. Thomson Delamar Learning: New York, USA.

Dios, I. G. 2016. Readability assessment and automatic text simplification. The analysis of Basque complex Structures. University of the Basque Country.

DuBay, W. H. 2004. The principles of readability. *Impact Information*, Costa Mesa: Online Submission, 1–76.

Eleyan, D., Othman, A., and Eleyan, A. 2020. Enhancing software comments readability using Flesch Reading Ease score. *Information*, 11(9):430–455.

Feng, L. 2010. Automatic readability assessment. New York: City University of New York.

Filho, J. A. W., Wilkens, R. S., Zilio, L., Idiart, M., and Villavicencio, A. 2016. Crawling by readability level. In *Proceedings of the 12th International Conference on Computational Processing of the Portuguese Language*, Vol 1: 306–318.

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

François, T. and Fairon, C. 2012. An "AI readability" formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Association for Computational Linguistics, 466-477.

Greenfield, J. 2004. Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1):5–24.

Guma, S. M. 1982. *An outline structure of Southern Sotho*. 2nd ed. Shooter and Shuter Publishers: Pietermaritzburg, South Africa.

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

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

Gunning, T. 2003. The role of readability in today's classrooms. *Topics in Language Disorders*, 23(3):175–189.

Hartley, J. 2016. Is time up for the Flesch measure of reading ease? *Scientometrics*, 107(3):1523–1526.

Heydari, P. 2012. The validity of some popular readability formulas. *Mediterranean Journal of Social Sciences*, 3(2):432–423.

Janan, D. 2011. Towards a new model of readability. University of Warwick.

Kasabwala, K., Agarwal, N., Hansberry, D. R., Baredes, S., and Eloy, J. A. 2012. Readability assessment of patient education materials from the American Academy of Otolaryngology—Head and Neck Surgery Foundation. *Otolaryngology–Head and Neck Surgery*, 147(3):466–471.

Kaur, S., Kaur, K., and Kaur, P. 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.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. 1975. Derivation of new readability formulas (Automated Readability index, Fog count and Flesch Reading Ease formula) for navy enlisted personnel. Defense Technical Information Center: *Report*.

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

Kondru, J. 2006. Using part of speech structure of text in the prediction of its readability. The University of Texas at Arlington.

Lemeko, P. A. 2018. Diachronic investigation into current issues in language variation. A case of Sesotho language. Bloemfontein: Central University of Technology, Free State.

Lesotho. 1993. The Constitution of Lesotho. Government Printer: Lesotho.

Marupi, O. and Charamba, E. 2022. Revisiting the effects of – isms in the promotion, development, and revitalisation of indigenous languages in Zimbabwe: The position of Sesotho in Gwanda South, Zimbabwe. *Handbook of Research on Teaching in Multicultural and Multilingual Contexts*, 32–46.

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

Mohammed, L. A., and Aljaberi, M. A., Anmary, A. S. and Abdulkhaleq, M. 2022. Analysing English for Science and Technology reading texts using Flesch Reading Ease online formula: The preparation for academic reading. *International Conference on Emerging Technologies and Intelligent Systems*, 546–561.

Mohasi, L., Mixdorff, E., and Niesler, T. 2011. An acoustic analysis of tone in Sesotho. *ICPhS*, 17–21.

Newbold, N. 2013. New Approaches for Text Readability. United Kingdom: University of Surrey.

Oborneva, I. V. 2006. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov [Automated estimation of the complexity of educational texts on the basis of statistical parameters]. RAS Institut soderzhaniya i metodov obucheniya [RAS Institute of Content and Teaching Methods].

Reid, M., Neil, M., Janse Van Rensburg-Bonthuyzen, E. 2019. Development of a Sesotho health literacy test in a South African context. *African Journal of Primary Health Care and Family Medicine*, 11(1):1–13.

Roux, J. C. and Bosch, S. E. 2019. Preserving and developing indigenous languages in the South African context. *Proceedings of the Language Technologies for All.* European Language Resources Association. 97-–100.

Roux, J. C. and Ndinga-Koumba-Binza, S. African languages and human language technologies. 2019. *The Cambridge Handbook of African Linguistics*, 623–644.

Sekere, N. B. 2004. Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa. University of South Africa, Pretoria.

Sjöholm, J. 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish.

Smith, E. A. and Senter, R. J. *Automated Readability index.* University of Cincinnati, Ohio.

Sibeko, J. 2021. A comparative analysis of the linguistic complexity of Grade 12 English Home Language and English First Additional Language examination papers. *Per Linguam*, 37(2):50–64.

Sibeko, J. and Setaka, M. 2022. An overview of Sesotho BLARK Content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11.

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

Sibeko, J. and Van Zaanen, M. 2022a. Sesotho syllabification systems. *Southern African Centre for Digital Language Resources*. Available at: https://repo.sadilar.org/handle/20.500.12185/555 [Accessed: 3 Jan 2023].

Sibeko, J. and Van Zaanen, M. 2022b. Raw and syllabified wordlist for Sesotho. *Southern African Centre for Digital Language Resources*. Available at: https://repo.sadilar.org/handle/20.500.12185/556 [Accessed: 3 Jan 2023].

Sibeko, J. and Van Zaanen, M. 2022c. Final year high school examination texts of South African Home and First Additional language subjects. *Southern African Centre for Digital Language Resources*. Available at: https://repo.sadilar.org/handle/20.500.12185/568 [Accessed: 29 Dec. 2022].

Solovyev, V., Ivanov, V., and Solnyshkina, M. I. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of intelligent and fuzzy systems*, 5(34):3049–3058.

Stocker, L. P. 1971. Increasing the precision of the Dale-Chall readability formula. *Reading Improvement*, 8(3):87.

Tanprasert, T. and Kauchak, D. Flesch-Kincaid is not a text simplification evaluation metric. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, 1–14.

Thomas, G., Hartley, R. D., and Kincaid, J. P. 1975. Test-retest and inter-analyst reliability of the automated readability index, Flesch Reading Ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.

Toyama, Y., Hiebert, E. H., and Pearson, P. D. 2017. An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, 22(3):139–170.

Van Oosten, P., Tanghe, D., and Hoste, V. 2010. Towards an improved methodology for automated readability prediction. *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, European Language Resources Association (ELRA). 775–782.

Wilken, I., Gumede, T., Moors, C., and Calteaux, K. 2018. Human language technology audit 2018: Design considerations and methodology. *International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, 1–7.

Wong, K. and Levi, J. R. 2017. Readability of pediatric otolaryngology information by children's hospitals and academic institutions. *The Laryngoscope*, 127(4):E138–E144.

Yan, X., Song, D., and Li, X. 2006. Concept-based document readability in domain specific information retrieval. *Proceedings of the 15th ACM international conference on Information and knowledge management*, 540–549.

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

Zhang, Y., Lin, N., and Jiang, S. 2019. A Study on syntactic complexity and text readability of ASEAN English news. *2019 International Conference on Asian Language Processing (IALP)*, 313–318.

Zhou, S., Jeong, H., and Green, P.A 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 6(1):97–111.