

Harnessing Google Translations to Develop a Readability Corpus for Sesotho: An Exploratory Study

Sibeko, Johannes
Nelson Mandela University
johannes.sibeko@mandela.ac.za

Abstract

This article addresses the scarcity of gold-standard annotated corpora for readability assessment in Sesotho, a low-resource language. As a solution, we propose using translated texts to construct a readability-labelled corpus. Specifically, we investigate the feasibility of using Google Translate to translate texts from Sesotho to English and then manually post-editing the texts. We then evaluate the effectiveness of the Google translations by comparing them to the human-post-edited versions. We utilised the Ghent University readability demo to extract the readability levels of both the Google translations and the human-post-edited translations. The translations are then evaluated using three evaluation metrics, namely, BLEU, NIST, and RIBES scores. The translation evaluation results reveal substantial similarities between the machine translations and the corresponding human-post-edited texts. Moreover, the results of the readability assessment and the comparison of text properties demonstrate a high level of consistency between machine translations and human-post-edited texts. These findings suggest that Google Translations show promise in addressing challenges in developing readability-labelled parallel datasets in low-resource languages like Sesotho, highlighting the potential of leveraging machine translation techniques to develop translated corpora for such languages. The evaluation of Google Translations in the context of educational texts in Sesotho and the demonstration of the feasibility and potential of using machine translations for enhancing readability in Sesotho will aid in the quest for developing Sesotho text readability measures.

Keywords Sesotho; Examination Texts; Machine Translation; Text Readability; Google Translate

1 Introduction

The investigation into text readability levels has been a subject of scholarly examination spanning over a century (De Clercq & Hoste 2016, Zamanian & Heydari 2012). However, there remains a notable dearth of research pertaining to text readability in African indigenous languages (Sibeko & De Clercq 2023, Sibeko & Van Zaanen 2021). To put it simply, readability pertains to the ease with which a given text can be read (Dahl 2004).

Text readability measures serve as valuable tools for estimating the level of ease associated with reading particular texts. These estimations hold significant importance, particularly in language learning, wherein language learners rely on appropriately levelled texts to enhance their reading proficiency (Sibeko 2023). Thus, the exploration of readability measures assumes great significance (Collins-Thompson 2014), particularly within the context of African indigenous languages. This is especially relevant in social contexts where texts are scarce and necessitate reuse for varied purposes.

While the available research has primarily concentrated on higher-resourced languages like English, there have been explorations of Sesotho text readability (Krige & Reid 2017, Reid et al. 2019, Sibanda 2019, 2014). However, these studies relied on English text readability measures since there are no known measures for measuring text readability in Sesotho (Sibeko & Van Zaanen 2021). To address the lack of readability measures, one possible solution is to adapt existing traditional readability measures to indigenous South African languages such as Sesotho (Leopeng 2019).

A gold-standard corpus with clear levels of text difficulty is needed to develop an automated readability model (Van Oosten et al. 2010, François & Fairon 2012). In fact, previous studies have emphasised the necessity of gold-standard corpora when adapting text readability measures from high-resource lan-



guages such as English to low-resource languages such as Sesotho (Van Oosten & Hoste 2011, François & Fairon 2012, Curto et al. 2014, 2015).

This article is part of a larger research project focused on developing readability measures for Sesotho. To achieve this goal, we require a gold-standard English-Sesotho parallel corpus that is annotated with text readability levels that can be used to train machine learning models.

In our search for this article, we found a lack of English-Sesotho-aligned corpora labelled for text readability, thus hindering the creation of gold-standard corpora (Sibeko 2023, p. 120). Consequently, a crucial aspect of our larger research project involves the development of a parallel English-Sesotho text readability corpus.

Unfortunately, while human-translated and edited texts would be ideal for creating a parallel corpus, they can be costly (Serhani et al. 2011). Given this constraint, we propose an alternative approach: the utilisation of translated texts to gauge text readability in Sesotho.

Essentially, our suggestion is to leverage machine-translated texts to establish a gold-standard corpus. We, however, acknowledge that the effectiveness of using machine-translated texts to train models for adapting English traditional readability measures to Sesotho remains uncertain. Desirably, one should correlate the English Reading scores to reading scores on known Sesotho documents, however, we do not have access to data of that kind.

Overall, this article explores the feasibility of using Google-translated Sesotho source texts and English target texts in training models for the development of traditional readability measures in Sesotho. To achieve this, the article will:

- examine the similarities between unedited English texts, which are machine-translated from Sesotho, and their human post-edited English counterparts, and
- investigate whether the selected machine-translated texts demonstrate consistent levels of readability when compared to the human

post-edited versions.

1.1 Machine Translation

Machine translation (MT) involves the automated conversion of texts from one language to another. MT systems rely on extensive corpora containing pairs of translations between the source and target languages (Tsai 2019). In this article, we have chosen Google Translate (see translate.google.com), one of the most widely used machine translation service providers (Latief et al. 2020).

Although Google Translate benefits from a vast corpus of texts, it is still prone to errors that are unlikely to be made by human translators (Tsai 2019). Consequently, human translations remain more accurate (Way & Hearne 2011). As such, while Google Translate continues to improve its grammatical accuracy, its translation accuracy, especially for African languages, remains a concern (Patil & Davies 2014, Tsai 2019).

Nevertheless, several studies have demonstrated the utility of Google Translate for translating texts in various research domains, such as health (Patil & Davies 2014), English Foreign Language learning (Tsai 2019), academic language improvement (Groves & Mundt 2015), and English teaching (Medvedev 2016). In this article, we employ Google Translate to translate reading comprehension and summary texts (see (Sibeko & Van Zaanen 2022)) that were extracted from previous examination papers see Sibeko & Van Zaanen (2023) for an overview of the texts.

1.2 Translation evaluation

This article pays attention to three algorithms for comparing the machine translations to the post-edited versions, namely, the Bilingual Evaluation Understudy (BLEU) score (Palotti et al. 2016), the National Institute of Standards and Technology (NIST) score (Doddington 2002), and the Rank-based Intuitive Bilingual Evaluation Score (RIBES) Isozaki et al. (2010).

It is important to note here that these algorithms



| Measure | Formula |
|---------------------------------------|---|
| Flesch-Kincaid Grade Level (FKGL) | $= 0.39(\frac{\#words}{\#sentences}) + 11.8(\frac{\#syllables}{\#words}) - 15.59$ |
| Flesch Reading Ease (FRE) | $= 206.835 - 1.015(\frac{\#words}{\#sentences}) + 84.6(\frac{\#syllables}{\#words})$ |
| Gunning Fog Index (GFI) | $= 0.4[(\frac{\#words}{\#sentences}) + 100(\frac{\#complex-words}{\#words})]$ |
| Simple Measure of Gobbledygook (SMOG) | $= 3.1291 + 1.043\sqrt{\#polysyllabic - words * (\frac{30}{\#sentences})}$ |
| Coleman-Liau Index (CLI) | $= 0.0588(\frac{\#characters}{\#samples}) - 0.296(\frac{\#sentences}{\#samples}) - 15.8$ |
| Automatic Readability Index (ARI) | $= 4.7(\frac{\#letters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43$ |
| Dale-Chall Index (DCI) | $= 0.0496(\frac{\#words}{\#sentences}) + [11.8(\frac{\#difficult-words}{\#words}) * 0.1579] + 3.6365$ |

Table 1: Measures and equations for measuring text readability

were not meant for comparing machine translations to human post-edited translations but rather to compare machine translations to human translations. In this way, they are best suited to comparatively evaluating the precision of different machine translation systems or comparing machine translations to translations produced by professional human translators. For instance, the underlying principle of the BLEU score is that machine translations are expected to contain many n-grams that are similar to the human translations (Lin & Och 2004, Song et al. 2013). Consequently, these similarities are more easily observed in longer texts, such as full documents (Specia et al. 2010). That is, sentence-level comparisons may yield misleading results. Given the focus of this article on document-level comparisons, the metrics are anticipated to produce optimal results.

According to Long et al. (2017) RIBES and BLEU are popular in machine translation evaluation. In fact, the BLEU score has become the standard metric for evaluation since its introduction in the year 2002 (Song et al. 2013, Specia et al. 2010), and continues to be used widely across various languages other than English (Chauhan et al. 2021).

1.3 Text readability

This article examines the application of seven traditional text readability measures: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog index, Simple Measure of Gobbledygook, Automatic Readability index, Lasbarhetsindex, Coleman-Liau index, and Dale-Chall index (Barry & Stevenson 1975, Coleman & Liau 1975, Dale & Chall 1948, Flesch 1948, Gunning 1969, 2003, Kincaid et al. 1975, Mc Laughlin 1969, Smith & Senter 1967). The seven text readability measures and their respective equations are presented in Table 1. Note that polysyllabic words in the SMOG index and complex words in the Gunning Fog index refer to words with more than two syllables (Londoner 1967, Mc Laughlin 1969, Hedman 2008, Christanti et al. 2017).

In the Dale-Chall index, difficult words are defined as words that are not featured in the list of 3000 common words compiled for the measure (MacDiarmid et al. 1998, Nyman et al. 1961, Stocker 1971). The texts used in this article are concise and therefore eliminated the need for sampling and enabled an analysis of complete texts in measures such as Coleman-Liau which requires sampling.



| Year | Tokens | Types | TTR | Sentences | ASL | AWL |
|---------|--------|-------|------|-----------|--------|------|
| 2009 | 592 | 281 | 0.48 | 20.50 | 14.38 | 3.94 |
| 2011 | 1804 | 736 | 0.42 | 19.80 | 18.08 | 3.92 |
| 2012 | 1168 | 543 | 0.47 | 18.50 | 16.08 | 4.06 |
| 2013 | 2022 | 845 | 0.44 | 25.60 | 16.81 | 4.05 |
| 2014 | 1898 | 803 | 0.44 | 22.00 | 18.00 | 4.19 |
| 2015 | 949 | 396 | 0.42 | 17.00 | 19.34 | 4.08 |
| 2016 | 1811 | 834 | 0.46 | 17.50 | 17.27 | 4.33 |
| 2017 | 1232 | 552 | 0.45 | 17.60 | 18.33 | 4.19 |
| 2018 | 1167 | 515 | 0.45 | 15.00 | 19.53 | 4.06 |
| 2019 | 1150 | 535 | 0.47 | 15.50 | 19.45 | 4.15 |
| Total | 13793 | 6040 | | | | |
| Average | 1379.3 | 604 | 0.45 | 18.9 | 17.727 | 4.10 |

Table 2: An overview of Sesotho source texts

| YEAR | Machine Translated unedited texts | | | | | Machine Translated post-edited texts | | | | |
|---------|-----------------------------------|-------|------|-------|------|--------------------------------------|--------|------|-------|------|
| | Tokens | Types | TTR | ASL | AWL | Tokens | Types | TTR | ASL | AWL |
| 2009 | 501 | 277 | 0.55 | 15.59 | 4.39 | 724 | 328 | 0.45 | 16.57 | 4.45 |
| 2011 | 1571 | 766 | 0.49 | 15.59 | 4.39 | 1984 | 978 | 0.49 | 13.92 | 4.36 |
| 2012 | 1022 | 543 | 0.53 | 14.24 | 4.37 | 826 | 482 | 0.58 | 14.49 | 4.46 |
| 2013 | 1805 | 875 | 0.48 | 14.57 | 4.24 | 1100 | 616 | 0.56 | 12.85 | 4.28 |
| 2014 | 1693 | 829 | 0.49 | 16.20 | 4.39 | 1121 | 613 | 0.55 | 15.99 | 4.61 |
| 2015 | 806 | 385 | 0.48 | 16.47 | 4.39 | 845 | 457 | 0.54 | 15.64 | 4.54 |
| 2016 | 1535 | 828 | 0.54 | 14.54 | 4.66 | 1819 | 872 | 0.48 | 15.27 | 4.57 |
| 2017 | 1056 | 537 | 0.51 | 16.05 | 4.43 | 1234 | 595 | 0.48 | 18.73 | 4.30 |
| 2018 | 1002 | 524 | 0.52 | 17.03 | 4.32 | 1255 | 581 | 0.46 | 20.35 | 4.31 |
| 2019 | 1014 | 566 | 0.56 | 17.24 | 4.46 | 1231 | 620 | 0.50 | 15.57 | 4.55 |
| Total | 12005 | 6130 | | | | 12139 | 6142 | | | |
| Average | 1200.50 | 613 | 0.52 | 15.75 | 4.40 | 1213.90 | 614.20 | 0.51 | 15.94 | 4.44 |

Table 3: An overview of translated texts

1.4 Theoretical Underpinning

This article is based on Skopos theory, which views translations as purposeful actions guided by specific intentions (Vermeer 1989, Vermeer & Chesterman 2021, Nord 2018, 2016). According to the theory, translations are considered acceptable only when they align with their intended purposes (Vermeer 1989, Reiss & Vermeer 2014, Tanrikulu 2017, Vermeer & Chesterman 2021). In this approach, translators are bound by the purpose of the translation, limiting their freedom to follow unnecessary impulses (Vermeer 1989, Du 2012, Vermeer & Chesterman 2021).

While Skopos theory may deviate from traditional rules of translation (Koller 1995), it allows for translations that fulfil their intended purposes as the purpose of translation is considered the most important criterion for translators (Reiss & Vermeer 2014), and the usability thereof. Resultantly, assessing the usefulness of translations based on their fulfilment of purpose allows for mechanical translations that closely adhere to the form and formal properties of the source language, sometimes at the expense of meaning (Lu & Fang 2012, Tanrikulu 2017, Suzani & Khoub 2019). Conversely, the meaning conveyed in the source text may be less important than its formal properties (Tanrikulu 2017, Odinye



2019). However, machine translations have been associated with negative perceptions, such as producing senseless texts (Läubli & Orrego-Carmona 2017, Latief et al. 2020).

In this article, machine translations are compared to post-edited versions to investigate whether machine translations can be harnessed to develop a readability corpus for Sesotho. As such, although machine translations have been associated with negative perceptions, such as producing senseless texts (Läubli & Orrego-Carmona 2017, Latief et al. 2020), the machine translations are considered appropriate if they produce the same levels of readability as translations that employ human intervention such as our post-edited dataset.

The rest of this article presents the method in Section 2, the findings in Section 3, and the discussion and conclusions in Section 4.

2 Method and Findings

2.1 The texts

This article utilises the Sesotho reading comprehension and summary writing texts from the corpus of grade twelve examination texts (Sibeko & Van Zanen 2023). These texts are extracted from the first examination paper which focuses on reading comprehension, visual literacy, and language conventions (Department of Basic Education 2011a,b, Van der Walt 2018). The texts have been previously used in studies on text readability (Sibeko & Van Zanen 2021), as well as linguistics complexity (Sibeko 2021).

2.2 Data processing

The data collected were pre-processed through tokenisation and sentence segmentation using *Ucto* (version 0.14). An overview of the Sesotho texts used in this article is presented in Table 2. Note that the disparity in the word count among examination texts across different years is due to the availability of examination papers. That is, certain years provided multiple examination opportunities, making all of the texts accessible for analysis. However, in

some years, only one exam paper could be located, resulting in a limited number of texts for this investigation.

In generating the machine translation data set, the sentence markers, denoted as <utt>, were removed from the texts. Subsequently, each file underwent translation into English using Google Translate¹, which facilitates whole-file translations. A comprehensive summary of the resulting machine translations is provided in Table 3. These translations form the first data set under discussion in this article.

For the post-edited data set, the machine translations underwent a post-editing process to consider cultural nuances and enhance translation quality. The translation brief explicitly guided the post-editing, limiting it to cases where meaning was lost. Consequently, word choices and grammatical constructions that did not affect the meaning were deliberately retained, even if alternative options were available.

By adhering to the editing brief and avoiding unnecessary modifications, the accuracy of the translations remained intact as this approach prioritised readability enhancement over imposing the personal preferences of translators. The summary of the post-edited data set is presented in Table 3. These post-edited versions of the machine translations form the second data set under scrutiny in this article.

Note that the differences in token counts between the machine translations and the human post-edited texts, as presented in Table 3, may be due to various factors, including translationese (Baroni & Bernardini 2006, Toral 2019), over-translation, and a lack of contextual comprehension in the Google translations of idiomatic expressions (Sibeko & Lemeko 2023). As can be observed from Table 3, addressing these translation issues directly influences the final type-token ratios and results in higher type-token ratios for the post-edited texts. This is primarily due to the fact that post-editing introduced idiomatic language (Sibeko & Lemeko 2023) which is missing from the unedited texts.



2.3 Data analysis

Two main steps were followed in the data analysis. The evaluation scores for comparing the data sets were computed using *Rstudio* (version 2023.06.0+421). The BLEU, NIST, and RIBES scores are cross-compared to get a sense of the similarities between the data sets.

The text readability scores were computed using the web-based readability tool developed by the Language and Translation Technology Team at Ghent University². The results obtained through the LT3 readability demo were then extracted and organised into a spreadsheet for further analysis using *RStudio*.

3 Findings

The findings of this investigation are presented in two steps. The evaluation of the translations is presented, followed by the results obtained from the readability assessment.

3.1 Translation evaluation

The comparison of the data sets exhibited a commendable level of similarity. More specifically, as illustrated in Figure 1, the BLEU scores were generally high, ranging from 0.32 to 0.98. It is important to note that the BLEU score measures similarity on a scale from zero to one, where scores closer to zero indicate minimal similarity, while scores closer to one indicate substantial resemblance between the data sets under comparison.

Similar to the BLEU score, the RIBES scores range from zero to one. As can be observed from Figure 1, the RIBES scores are predominantly clustered around zero and 0.3, with a maximum score of 0.94. The mean score of 0.23 indicates a generally low level of similarity between the machine translations and the post-edited versions.

On the other hand, the NIST scores, represent the performance of machine translation systems on a scale from zero to ten. Figure 1 visually demonstrates that the NIST scores predominantly cluster around six and seven, with a minimum of 4.45 and a

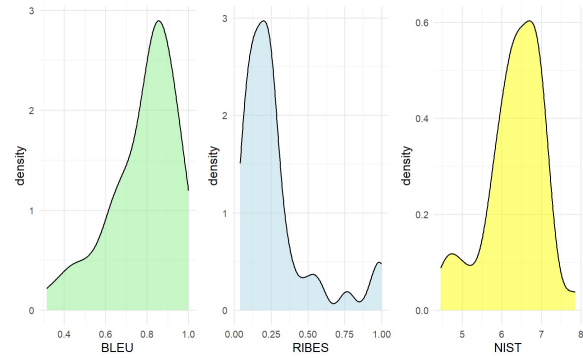


Figure 1: The distribution of machine translation evaluation scores.

maximum of 7.27. The mean score of 6.2 indicates an overall commendable level of quality, signifying that the machine translations exhibit favourable similarity to the post-edited versions.

Furthermore, the Pearson correlation among the three measures was examined, and the results are presented in Table 4. As can be observed in Table 4, the correlation coefficient between BLEU and RIBES is 0.69 ($p < .0001$), indicating a moderate positive correlation. This suggests that there is some degree of similarity between these two measures, although they are not perfectly aligned. Consequently, an increase in the BLEU score is likely to be accompanied by an increase in the RIBES score.

Table 4: Pearson correlations between the different scores for BLEU, RIBES, and NIST.

| Metric | Metric | Coefficient | p -value |
|--------|--------|-------------|------------|
| BLEU | RIBES | 0.69 | <.0001 |
| BLEU | NIST | 0.88 | <.0001 |
| RIBES | NIST | 0.60 | <.0001 |

Additionally, the correlation coefficient between BLEU and NIST is 0.88 ($p < .0001$), indicating a strong positive correlation. In other words, this finding implies a high degree of association between the scores of these two metrics. It is important to note that the correlation coefficient of 0.88 indicates a relatively stronger relationship compared to the correlation between BLEU and RIBES. Finally, the

correlation coefficient between NIST and RIBES is 0.60 ($p < .0001$), indicating a moderate positive correlation.

Overall, the scores for BLEU ($M = 0.78, SD = 0.16$) and NIST ($M = 6.3, SD = 0.72$) demonstrate a high similarity index between the two data sets. Conversely, the scores for RIBES ($M = 0.25, SD = 0.28$) illustrate some level of similarity between the RIBES and the NIST scores.

3.2 Text readability analysis

To get a sense of the distribution of data for each data set, density plots were employed. Similarities are observed in the distribution of data as observed in Figure 2. Additionally, it is clear that the data have similar spreads and consistently peak around the same locations. Additionally, the DCI and ARI measures show comparable directional biases.

Paired t -tests were used to compare the readability values of the two data sets and gain deeper insights into their behaviour. The descriptive outcomes of the t -tests are presented in Table 5, including means and standard deviations in brackets. Additionally, the ASL, which measures the average number of words in each sentence and the AWL which measures the average number of letters in each sentence are included.

Table 5: Results of the t -tests for the machine translated and the post-edited texts indicating the means, standard deviations and p -values.

| Measure | machine translated | | post-edited | | p -value |
|---------|--------------------|------|-------------|------|------------|
| | Mean | SD | Mean | SD | |
| FKGL | 6.86 | 1.60 | 7.11 | 1.63 | .48 |
| FRE | 73.62 | 8.07 | 72.15 | 8.12 | .41 |
| SMOG | 9.29 | 1.42 | 9.54 | 1.39 | .42 |
| GFI | 9.34 | 1.79 | 9.63 | 1.81 | .46 |
| CLI | 8.11 | 1.55 | 8.44 | 1.49 | .33 |
| ARI | 7.05 | 1.94 | 7.39 | 1.93 | .42 |
| DCI | 7.75 | 0.81 | 7.91 | 0.78 | .36 |
| ASL | 15.59 | 2.68 | 15.76 | 2.97 | .78 |
| AWL | 4.39 | 0.25 | 4.45 | 0.24 | .32 |

The post-edited data sets exhibit slightly higher means which are associated with harder-to-read texts, although there is some degree of variability observed within both data sets. Nevertheless, the results indicate that none of the measures displays statistically significant distinctions between the two data sets ($p > .05$). In this way, the findings suggest that there are no real disparities in text readability between the two data sets.

Additionally, there is close proximity in the mean values of average sentence lengths (ASL) ($M_{\text{machine translated}} = 15.59, M_{\text{post-edited}} = 15.76, p = .78$) and average word lengths (AWL) ($M_{\text{machine translated}} = 4.39, M_{\text{post-edited}} = 4.45, p = .32$). In fact, the mean differences between the two data sets are insignificant.

4 Discussion and Conclusions

This article aimed to investigate the feasibility of using texts translated from Sesotho to English using the Google Translate platform in the creation of a text readability corpus for Sesotho that is annotated with levels of text readability. For this, two sets of data were compiled, namely, the machine translations data set and the human-post-edited version of the machine translations data set.

First, the data sets were evaluated for their similarity based on the assumption that the machine-translated data set would only be usable if it demonstrated a certain level of accuracy when we used the post-edited data set as the target standard. For this purpose, we utilised three metrics, namely, the RIBES, BLEU, and NIST scores. The comparison of the three sets of translation evaluation results suggests that the Google Translate-generated translations of the examination texts can be utilised in developing a readability corpus for Sesotho. This finding is based on the results for BLEU and NIST scores which indicate a higher precision. Note that the close correlation between NIST and BLEU is not surprising given that the NIST was based on BLEU scores.

Furthermore, this article aimed to investigate

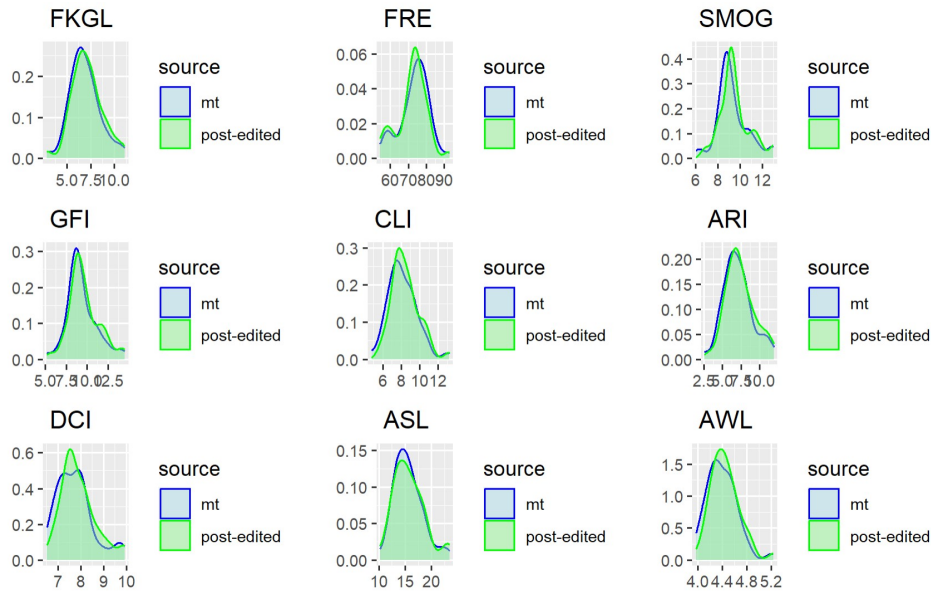


Figure 2: The distribution of readability scores for the machine translated (mt) and post-edited data sets.

whether unedited translations generated by Google Translate would exhibit similar characteristics to human-post-edited versions, thereby justifying the levels of readability that are attached to the texts in the data set proposed for use in developing a readability corpus for Sesotho. To achieve this, the readability of the texts was evaluated using seven traditional readability measures along with sentence and word lengths. The analysis revealed no significant differences in the means of the various measures and text properties.

Based on the findings of this article, that the data sets are highly similar and that they produce texts of similar readability levels, there is justification for utilising translations generated by Google Translate from Sesotho to English in the training of traditional readability models for Sesotho. This conclusion is supported by the Skopos theory, which emphasises that translations should effectively serve their intended purpose. The use of translations generated through Google Translate aligns with the principle of Skopos theory. The comparison conducted between machine-generated translations and human-post-edited translations aimed to assess the similarities between these two translation types.

The results suggest that the machine translations can be deemed valid for the intended purpose of developing a readability corpus.

However, it is essential to note that the discussion in this article is limited, as it only covers the comparison between the post-edited translated texts and the machine translations. Consequently, it does not account for or extensively discuss translation issues which would have provided valuable insights. Furthermore, the section did not investigate whether translations produced by First Additional Language (FAL)³ texts and Home Language (HL) texts exhibited the same level of accuracy or whether the text types contained in the data sets affected the accuracy of the results, as the data sets were considered in their entirety.

Finally, the findings of this article highlight the need for future research. For instance, while this article has demonstrated the consistency of readability levels between machine translations and human-post-edited translations, it has not addressed the development of an aligned English-Sesotho corpus annotated with readability levels. This represents an avenue for further investigation and exploration.

Notes

¹Google Translate is an online machine translation platform accessible at <https://translate.google.com>.

²This tool can be accessed online at the following URL: <https://lt3.ugent.be/readability-demo/>.

³The First Additional Language subject is intended for learners with lower language proficiencies than those in the Home Language classes. Note that previous research such as Sibeko (2021) and Sibeko & Van Zaanen (2021) has clearly shown significant differences in linguistic complexity and text readability levels between these language subjects.

Data Availability

All data used in this study was obtained from publicly available sources. No confidential or sensitive information was used. The study was cleared for ethics by North-West University with the ethics clearance number: NWU-00729-21-A7. The original examination data are available from the South African Centre for Digital Language Resources at <https://repo.sadilar.org/handle/20.500.12185/56>.

Acknowledgements

This article forms part of a PhD study at North-West University under the supervision of Professor Menno van Zaanen from the South African Centre for Digital Language Resources. I, therefore, thank Professor Menno for the extensive feedback he provided before the submission of this article.

References

- Baroni, M. & Bernardini, S. (2006), 'A new approach to the study of translationese: Machine-learning the difference between original and translated text', *Literary and Linguistic Computing* **21**(3), 259–274.
- Barry, J. G. & Stevenson, T. E. (1975), 'Using a computer to calculate the Dale-Chall Formula', *Journal of Reading* **19**, 218–222.
- Chauhan, S., Daniel, P., Mishra, A. & Kumar, A. (2021), 'Adableu: A modified bleu score for mor-

phologically rich languages', *IETE Journal of Research* pp. 1–12.

Christanti, V., Naga, D. S. & Benedicta, C. (2017), 'Measuring reading difficulty using lexile framework and gunning fog index', *Jurnal Teknik dan Ilmu Komputer*.

Coleman, M. & Liao, T. L. (1975), 'A computer readability formula designed for machine scoring', *Journal of Applied Psychology* **60**(2), 283.

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.

Curto, P., Mamede, N. & Baptista, J. (2014), 'Automatic readability classifier for european portuguese', *System* **5**(1), 309–324.

Curto, P., Mamede, N. & Baptista, J. (2015), 'Automatic text difficulty classifier: Assisting the selection of adequate reading materials for european portuguese teaching', in M. Helfert, M. T. Restivo, S. Zvacek & J. Uhomobhi, eds, 'Proceedings of the 7th International Conference on Computer Supported Education (CSEDU-2015)', Vol. 1, Science and Technology Publications, pp. 36–44.

Dahl, O. (2004), *The growth and maintenance of linguistic complexity (Studies in Language Companion Series)*, Vol. 71, John Benjamins Publishing, Amsterdam.

Dale, E. & Chall, J. S. (1948), 'A formula for predicting readability: Instructions', *Educational research bulletin* pp. 37–54.

De Clercq, O. & 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.

Department of Basic Education (2011a), *English First Additional Language, Further Education and Training Phase Grades 10–12: Curriculum*



- and Assessment Policy Statement (CAPS), Government printing works, Pretoria. Policy Document.
- Department of Basic Education (2011b), *English Home Language, Further Education and Training Phase Grades 10–12: Curriculum and Assessment Policy Statement (CAPS)*, Government printing works, Pretoria. Policy Document.
- Doddington, G. (2002), Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in ‘Proceedings of the second international conference on Human Language Technology Research’, pp. 138–145.
- Du, X. (2012), ‘A brief introduction of Skopos theory’, *Theory and Practice in Language Studies* 2(10), 2189–2193.
- Flesch, R. (1948), ‘A new readability yardstick’, *Journal of applied psychology* 32(3), 221–233.
- François, T. & Fairon, C. (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’, Association for Computational Linguistics, pp. 466–477.
- Groves, M. & Mundt, K. (2015), ‘Friend or foe? Google Translate in language for academic purposes’, *English for Specific Purposes* 37, 112–121.
- 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.
- Hedman, A. S. (2008), ‘Using the SMOG formula to revise a health-related document’, *American Journal of Health Education* 39(1), 61–64.
- Isozaki, H., Hirao, T., Duh, K., Sudoh, K. & Tsukada, H. (2010), Automatic evaluation of translation quality for distant language pairs, in ‘Proceedings of the 2010 conference on empirical methods in natural language processing’, pp. 944–952.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. (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.
- Koller, W. (1995), ‘The concept of equivalence and the object of translation studies’, *Target. International Journal of Translation Studies* 7(2), 191–222.
- Krige, D. & Reid, M. (2017), ‘A pilot investigation into the readability of Sesotho health information pamphlets’, *Communitas* 22, 113–123.
- Latief, M. R. A., Saleh, N. J. & Pammu, A. (2020), ‘The effectiveness of machine translation to improve the system of translating language on cultural context’, *IOP Conference Series: Earth and Environmental Science* 575(1), 1–9.
- Läubli, S. & Orrego-Carmona, D. (2017), When Google Translate is better than some human colleagues, those people are no longer colleagues, in ‘Translating and the Computer’, AsLing, London, Zurich Open Repository and Archive, pp. 59–69.
- Leopeng, M. T. (2019), Translations of informed consent documents for clinical trials in South Africa: Are they readable?, Thesis, University of Cape Town.
- Lin, C.-Y. & Och, F. J. (2004), Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in ‘Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)’, pp. 605–612.
- Londoner, C. A. (1967), *A readability analysis of randomly selected Basic Education and Vocational Education curriculum materials used at the Atterbury Job Corps Center as measured by the Gunning Fog Index*, ERIC, United States.



- Long, Z., Utsuro, T., Mitsuhashi, T. & Yamamoto, M. (2017), 'Translation of patent sentences with a large vocabulary of technical terms using neural machine translation', *arXiv preprint arXiv:1704.04521*.
- Lu, W. & Fang, H. (2012), 'Reconsidering Peter Newmark's Theory on literal translation', *Theory & Practice in Language Studies* 2(4).
- MacDiarmid, S. A., Goodson, T. C., Holmes, T. M., Martin, P. R. & Doyle, R. B. (1998), 'An assessment of the comprehension of the American Urological Association Symptom Index', *The Journal of Urology* 159(3), 873–874.
- Mc Laughlin, G. H. (1969), 'SMOG grading-a new readability formula', *Journal of reading* 12(8), 639–646.
- Medvedev, G. (2016), 'Google translate in teaching English', *Journal of teaching English for specific and academic purposes* 4(1), 181–193.
- Nord, C. (2016), 'Function+ loyalty: Theology meets skopos', *Open Theology* 2(1).
- Nord, C. (2018), *Translating as a purposeful activity: Functionalist approaches explained*, Routledge.
- Nyman, P., Kearsley, B. E. & Powers, R. D. (1961), 'An attempt to shorten the word list with the Dale-Chall readability formula', *Educational Research Bulletin* 40, 150–152.
- Odiinye, S. I. (2019), 'Translation: The problem of meaning and equivalence', *Journal of African Studies and Sustainable Development* 2(2).
- Palotti, J., Goeuriot, L., Zuccon, G. & Hanbury, A. (2016), Ranking health web pages with relevance and understandability, in 'Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval', pp. 965–968.
- Patil, S. & Davies, P. (2014), 'Use of Google Translate in medical communication: evaluation of accuracy', *Bmj* 349.
- Reid, M., Nel, 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 & family medicine* 11(1), 1–13.
- Reiss, K. & Vermeer, H. J. (2014), *Towards a general theory of translational action: Skopos theory explained*, Routledge.
- Serhani, M. A., Jaffar, A., Campbell, P. & Atif, Y. (2011), 'Enterprise web services-enabled translation framework', *Information Systems and e-Business Management* 9, 497–517.
- Sibanda, L. (2014), 'The readability of two grade 4 natural sciences textbooks for South African schools', *South African Journal of Childhood Education* 4, 154–175.
- Sibanda, L. (2019), A case study of the readability of two Grade 4 Natural Sciences textbooks currently used in South African schools, MA thesis, Rhodes University.
- 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, A Journal for Language Teaching* 37(2), 50–64.
- Sibeko, J. (2023), Using classical readability formulas to measure text readability in sesotho, in T. Erjavec & M. Eskevich, eds, 'Selected papers from the CLARIN Annual Conference 2022', Electronic Conference Proceedings, pp. 120–132.
- Sibeko, J. & De Clercq, O. (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)', Association for Computational Linguistics, Dubrovnik, Croatia, pp. 32–41.
- Sibeko, J. & Lemeko, P. A. (2023), Developing a text readability corpus using Google Translate for low-resourced South African indigenous languages. Presented at the Sol Plaatje University International Language Conference. 6 –



- 8 September 2023. Kimberley, Northern Cape, South Africa. [Unpublished].
- Sibeko, J. & 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. & Van Zaanen, M. (2022), *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 Mar. 2023].
- Sibeko, J. & Van Zaanen, M. (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.
- Smith, E. A. & Senter, R. (1967), *Automated readability index*, University of Cincinnati.
- Song, X., Cohn, T. & Specia, L. (2013), 'BLEU deconstructed: Designing a better MT evaluation metric.', *International Journal of Computational Linguistics and Applications* 4(2), 29–44.
- Specia, L., Raj, D. & Turchi, M. (2010), 'Machine translation evaluation versus quality estimation', *Machine translation* 24, 39–50.
- Stocker, L. P. (1971), 'Increasing the precision of the Dale-Chall readability formula', *Reading Improvement* 8(3), 87.
URL: <https://eric.ed.gov/?id=EJ052421>
- Suzani, S. M. & Khoub, V. M. A. (2019), 'A study of Nida's (1964) types of equivalence in three English translations of letters 40 to 60 of Nahj-al-balagheh', *Specialty Journal of Language Studies and Literature* 3(4), 161–179.
- Tanrikulu, L. (2017), 'Translation theories', *Neuşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi* 7(1), 97–108.
- Toral, A. (2019), Post-editeese: An exacerbated translationese, in 'Proceedings of Machine Translation Summit XVII: Research Track', pp. 273–281.
- Tsai, S.-C. (2019), 'Using Google Translate in EFL drafts: a preliminary investigation', *Computer Assisted Language Learning* 32(5-6), 510–526.
- Van der Walt, C. (2018), 'The category language structures and conventions in the CAPS for English First Additional Language: A critical analysis', *Journal for Language Teaching* 52(1), 170–200.
- Van Oosten, P. & Hoste, V. (2011), Readability annotation: Replacing the expert by the crowd, in 'Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications', pp. 120–129.
- Van Oosten, P., Tanghe, D. & Hoste, V. (2010), Towards an improved methodology for automated readability prediction, in 'Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)', European Language Resources Association (ELRA), pp. 775–782.
- Vermeer, H. J. (1989), Skopos and commission in translational action, in L. Venuti, ed., 'The translation studies reader', Routledge, London and New York, pp. 221–232.
- Vermeer, H. J. & Chesterman, A. (2021), Skopos and commission in translational action, in 'The translation studies reader', 4th edn, Routledge, London and New York, pp. 219–230.
- Way, A. & Hearne, M. (2011), 'On the role of translations in state-of-the-art statistical machine translation', *Language and Linguistics Compass* 5(5), 227–248.
- Zamanian, M. & Heydari, P. (2012), 'Readability of texts: State of the art', *Theory and Practice in Language Studies* 2(1), 43–53.

