A Corpus-Based List of Frequently Used Words in Sesotho

Johannes Sibeko Nelson Mandela University Port Elizabeth, South Africa johannes.sibeko@mandela.ac.za

Abstract

This article describes the development of a list of frequently used words in written Sesotho. The list has been created with the aim of incorporating it into frequency-based text readability metrics. The list was derived using a corpus-based approach. By leveraging three existing Sesotho corpora, frequency lists could be derived, which were subsequently merged and qualitatively analysed and fine-tuned by an experienced speaker of Sesotho. The main challenges in compiling the list included reconciling the spelling variations, the treatment of abbreviations, and the presence of unexpected words in the preliminary lists. The final list comprises 3037 entries and is made publicly available to the research community.

1 Introduction

South African learners struggle with reading comprehension even when reading in their home languages (Pretorius et al., 2020). As a result, they perform poorly on problems involving language (Van der Walt et al., 2008). This is especially pronounced in bilingual and multilingual learners since they develop literacy simultaneously in two languages (Cockcroft, 2016; Wilsenach and Schaefer, 2022). Such learners perform more poorly than monolingual learners who get in-depth exposure to one language (Cockcroft, 2016). This challenge of vocabulary and language demands is increased when South African learners with indigenous languages as their first languages move from the third to the fourth grade of education and where the language of instruction changes from indigenous languages to either English or Afrikaans (Sibanda and Baxen, 2016).

According to Stoffelsma (2019b,a), 78% of South African fourth graders were unsuccessful at extracting meaning from texts. Unfortunately, not being able to extract meaning from texts puts learners at risk of not being able to read as their Orphée de Clercq LT3, Ghent University Ghent, Belgium Orphee.DeClercq@UGent.be

lack of vocabulary affects their ability to read texts with desirable fluency (Pretorius and Stoffelsma, 2017; Stoffelsma, 2019a).

Sadly, learners' inability to read with the expected level of fluency and their inability to extract meaning from texts makes it difficult for teachers and assessors to choose reading passages. Reading interventions to assist learners with less-thandesired reading abilities are needed in most language classes. Unfortunately, teachers may not always be well-trained to teach reading and monitor reading interventions (Pretorius et al., 2020). Furthermore, teachers' levels of command may not always allow for successful interventions (Batinić et al., 2016). For higher pass rates, one might have to resort to using texts expected to be administered to learners with fewer years of schooling. This is particularly unfavourable for research on the development of reading metrics for South African indigenous languages.

Consistent estimations of readability levels are essential in educational contexts where instructors and examiners need to identify and assign texts to readers with specific levels of education. Such consistency in assigning readability estimations prevents instances where learners in higher grades are assigned texts that are easier to read than those that are assigned to learners in lower grades. Without readability metrics, authors, publishers, and readers may not always estimate readability levels accurately or consistently (Humphreys and Humphreys, 2013).

Unfortunately, as far as we are aware, there are no readability metrics for Sesotho. One solution to this may be the development of classical readability formulas. Classical readability metrics use mathematical formulas to estimate the level of education or the grade that a reader needs in order to read a specific text with ease (Gopal et al., 2021). These linear regression formulas are normally based on superficial text properties such as lengths of words, sentences, syllables, and frequency lists. Fortuitously, sentence and word lengths can easily be determined using universal preprocessing tools as they are language-independent. However, determining syllables and frequently used words requires specific language tools. As far as could be ascertained for this article, there are two syllabification systems for Sesotho, see Sibeko and Van Zaanen (2022b). As such, Sesotho syllable information can be extracted from texts. Unfortunately, there is no frequency-based list of the most used words to use in readability studies. Looking at existing readability formulas, one of the most well-known frequency lists is the one developed by Dale and Chall (1948). The strategy to compile this list was to directly present these words in reading to 4th-grade learners, a word made it to the list when it was known by at least 80 per cent of the children (Dale and Chall, 1948; Piu et al., 2020; Glazkova et al., 2021). Ideally, when devising a word list for Sesotho (or other indigenous languages) a similar technique should be used. Unfortunately, such a user-based approach may yield regressive results since the majority of learners in indigenous African language classes have low reading abilities and vocabulary knowledge (Stoffelsma, 2019a,b). In other words, the result of a user-based approach would be a list of reading entry-level words.

This article presents our efforts to create a list of the 3000 most frequently used words in written Sesotho using a corpus analysis approach. We position our work in research on adapting readability metrics (Section 2). We then present our methodology for creating the list (Section 3) and subsequently discuss the results (Section 4). We conclude our article by highlighting the strengths of the list and by formulating a set of suggestions for further improvements (Section 5).

2 Related Work

Digital language resources and human language technologies audits for South African languages indicate that all indigenous languages of South Africa are under-resourced (Grover et al., 2010, 2011; Barnard et al., 2014; Moors et al., 2018a,b). More specifically, Sibeko and Setaka (2022) reviewed the basic language resource kit for Sesotho and concluded that many necessary digital language resources are still lacking.

A few studies have assessed the readability of

Sesotho texts using English readability metrics. For instance, Krige and Reid (2017) manually extracted textual properties used in classical readability metrics to investigate the readability of Sesotho health pamphlets. In another study, Reid et al. (2019) developed a Sesotho health literacy test. Unfortunately, the studies used English metrics without considering the differences between Sesotho and English textual properties. Therefore, the results from these studies may have misrepresented the context of Sesotho's written texts. Furthermore, due to the lack of resources, manual methods for extracting textual properties were used which is highly impractical to be used at a large scale and error-prone.

Multiple studies have explored the adaptation of classical readability metrics to lower-resourced languages. For Norwegian, Jakobsen and Skardal (2007) explored the adaptation of eight classical readability metrics, namely, the Automated Readability Index (ARI) (Smith and Senter, 1967), Coleman-Liau index (CLI) (Coleman and Liau, 1975), Flesch Reading Ease (FRE) (Flesch, 1974), Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Gunning Fog index (GFI) (Gunning, 1969), Lasbarhetsindex (LIX) (Anderson, 1983), Rate index (RIX) (Anderson, 1983), and Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969). Almost all those metrics were originally developed for English, except for LIX and RIX which were developed for Swedish. One of their most important findings was that syllable counts and complex words did not have the same effect on readability levels in Norwegian as they did in the English metrics. In the end, only the Swedish LIX and RIX metrics could successfully be adapted to the lower-resourced Norwegian.

The English FRE and FKGL formulas have also been adapted to Russian (Oborneva, 2006). In this process, the two text characteristics that are included in the formulas, i.e. average sentence and word length, were compared and the corresponding weights adapted (Glazkova et al., 2021). Similarly, the English FRE has been adapted to Dutch, by both Douma (Douma, 1960) and Brouwer (Brouwer, 1963), who each assigned slightly different weights in order to account for the differences in word and sentence length between English and Dutch. The popular FRE formula has also been adapted to French (Kandel and Moles, 1958; Henry, 1975), Czech (Bendová, 2021; Bendová and Cinková, 2021) and Italian (Franchina and Vacca, 1986).

A study by Van Oosten et al. (2010) revealed that the outcomes of classical formulas developed for English, Dutch and Swedish on Dutch data yield strong correlations, which is explained by the formulas' strong reliance on certain languageindependent properties, such as average word or sentence length. However, failed attempts of adapting classical readability metrics using higherresourced languages have also been reported. For instance, Sinha et al. (2012) reports an unsuccessful attempt at adapting the English FRE, GFI, FKGL and SMOG metrics into Hindi and Bangla. They concluded that new metrics which are based on Bangla and Hindi structural properties should be developed as the existing metrics yield out-ofbound results. New textual properties such as Juktaakshars were then introduced in the resulting Hindi metrics. This implies that readability metrics can be adapted when there is comparability such as in the cases of English, Dutch and Swedish, which are all Germanic languages. However, new formulas may need to be developed when language structures are incomparable such as in the case of Hindi and English.

Moreover, certain popular classical formulas also comprise variables based on frequency lists. The most well-known formula in this respect is the Dale-Chall Reading Grade Score (Dale and Chall, 1948). Besides relying on average sentence length, this formula also counts how many words occur in the Dale-Chall word list. This list comprises 3,000 words which are known in reading by at least 80 per cent of fourth-grade children. An updated version of the list was published in 1995 (Chall and Dale, 1995).

The Dale-Chall formula or index is one of the most used metrics in health information (Palotti et al., 2016). Given that the formula was originally developed to assess reading material for children, it is recommended to adjust the formula, and especially the list of frequently used words when using it to measure readability for specific target audiences (Gauthier and Johnson, 2019). On the other hand, the list of common words has been criticised for failing to account for specialised meanings (Yan et al., 2006).

In this work, we wish to explore whether a similar list can be created for Sesotho. Ideally, we would like to also assess the word list with users such as school learners. Even so, before testing can begin, we need an initial list. We are also concerned that testing the list on school learners may be challenging due to the recently implemented gate-keeping procedures aimed at protecting vulnerable participants in the post-pandemic era. Nevertheless, we are exploring other user tests that could make our list applicable to other target audiences, such as adult readers, authors, and publishers.

Currently, there exist at least two lists of frequently used words in Sesotho, namely the Most-CommonWords¹ and the Waston Chen² list. However, a few issues arise when consulting these lists. First of all, both lists have been (machine) translated from English into Sesotho, which means they are less representative of the Sesotho language. Moreover, for some of the translations in the Waston Chen list, Sepedi is being used (such as kgauswi instead of Sesotho haufi 'near'). Because of this translation from English into Sesotho both lists also contain various entries which are not single words but phrases. For instance, the Waston Chen list mentions yuniti eno ya thuto 'that unit of the lesson' for the English 'unit' entry. Similarly, the MostCommonWords list also mentions a sebetsang a 'that work' for the English entry 'active'. The lists also inconsistently interchange between two different orthographies, both the Lesotho Sesotho and the South African Sesotho orthography are being used. Finally, for both lists, it is difficult to find any background information on how the original lists were actually created.

We, therefore, believe that a more structured and purposed development of a frequency list is needed and opt for a corpus-based approach in this paper.

3 Methodology

We collected a total of three corpora which comprise original Sesotho text material, see Table 1 for some corpus statistics.

3.1 Corpus 1: Bible

We extracted text from the Sesotho (Southern Sotho) bible version SSO89 *Bibele* 'Bible'. The Bible texts were downloaded in SQL3lite format from https://www.ph4.org/b4_index. php#google_vignette. All texts were extracted using *bash* scripts. The Bible texts are divided into

¹https://3000mostcommonwords.com

²https://wastonchen.com/6417.html

	# Tokens	# Sentences
Corpus 1	962 916	31 171
Corpus 2	4 614 565	216 854
Corpus 3	2 017 751	85 860
TOTAL	7 595 232	333 885

Table 1: The number of tokens and sentences present in the three Sesotho corpora used for this study.

three sections, namely, (i) bible books containing 66 rows of data, (ii) verses containing 31 171 rows of data and (iii) info containing 10 rows of data. For our corpus, we extracted verse texts, which were subsequently cleaned by removing book numbers, chapter titles and verse information. The texts were tokenized using *ucto* with default settings except for specific settings for displaying each sentence on a new line. In the end, the bible corpus contains around 1 million tokens.

3.2 Corpus 2: Autshumato

The Autshumato machine translation project developed a corpus of translation texts for South African indigenous languages. These texts were manually translated by professional translators from English into the other ten official languages of South Africa, namely, Afrikaans, IsiZulu, IsiXhosa, IsiNdebele, Siswati, Sesotho, Sepedi, Setswana, Xitsonga, and Tshivenda in no particular order of importance. The English-Sesotho texts can be publicly accessed on the South African Centre for Digital Language Resources (SADiLaR) online repository (Mckellar, 2023). The readme file indicates that this corpus may need further cleaning for future uses because it was specifically formatted for training machine translation systems. However, as we were only interested in the words used in the corpus, no further cleaning was necessary. For this article, the corpus was tokenised and sentence segmented using ucto.

In the end, the Autshumato corpus contained around 4.6 million tokens. Unfortunately, the original corpus contained scrambled texts, as such, we could not unequivocally ascertain all text types present in the Sesotho corpus. Even so, McKellar (2022) lists at least four text types, namely magazines, policies, newsletters, translation works and documents crawled from the government (gov.za) domain. We are therefore confident that the corpus comprises different genres.

3.3 Corpus 3: NCHLT

The National Centre for Human Language Technology (NCHLT) project aimed to develop speech and text data to enable HLT development for the 11 official languages of South Africa (Eiselen and Puttkamer, 2014; Badenhorst and De Wet, 2022). The text collection contains data crawled from the South African gov.za domain. Data for each language contains enough training and testing samples for tasks such as language identification (Duvenhage, 2019).

The text corpus contains three sets of data, namely, the source texts, lexica, and corpus ³ (Eiselen and Puttkamer, 2014). Instead of relying on the lexica, we decided to use the actual corpus data to have more control. Both raw and cleaned versions of the corpus are present, we used the cleaned version and again tokenized all text material with *ucto*. Basic settings were used with the sentence segmentation option. In the end, the NCHLT corpus comprises around 2 million tokens.

3.4 Towards a Common Word Frequency List

To derive the frequency list all words were first lowercased. Next, all frequencies were calculated per corpus. In order to have word frequencies which are independent of corpus size, these were normalized to frequencies per million words, which is the preferred standard measure also referred to as relative frequency.

Our primary objective was to end up with a list of 3000 unique words based on the three corpora. To this end, we merged the three lists and made sure to average the relative frequencies of duplicate entries. For example, the entry 'a' appeared in all three lists, with a relative frequency of 54 073.57, 24 853.12 and 27 740.49, respectively. The resulting average relative frequency for this entry is 35 555.73.

After the list was derived automatically using corpus-based frequency measures it was also manually processed by a native speaker of Sesotho with much experience in Sesotho language teaching and writing research in order to end up with a clean list.

4 Results and discussion

As mentioned above the three frequency lists were merged and relative frequencies were calculated for duplicate entries. Afterwards, a qualitative analysis

³The corpus is available from https://hdl.handle.net/ 20.500.12185/336

was carried out on all entries, which is presented next.

4.1 Proper names

The Dale-Chall index considers names of people and organisations as familiar (Barry and Stevenson, 1975). Therefore, they do not need to be included in the list of frequently used words. As a result, we removed the names of people and organisations. People's names included biblical names like *Judase* 'Judas', typical Sesotho names like *Mmalerato* 'Mother of love', and names of public figures like *Madiba* 'the iconic Nelson Mandela'. Organisation names such as the South African Revenue Services abbreviated as SARS were also removed.

4.2 Spelling

According to Chokoe (2020), Sepedi and Setswana do not have rules governing the spelling of loan words. Similarly, loan words in Sesotho also vary in how they are spelt. For instance, the English word 'provinces' is written using four varying spellings in the corpora, that is, *diporofense, diporofensi, diporovense*, and *diprovense*. Inconsistencies like this can be expected when there is flexibility when forming loan words (Kosch, 2013). Although Sesotho words typically do not contain the letter 'v', it is used in loan words that originally contain 'v' letters such as *thelevishene* 'television'.

Taljard (2008) discusses three issues when deciding on the correct spelling. First, it would be easy to take the word that appears the most, however, a large enough corpus representing texts in that domain would be required. Second, the correct candidate word could be chosen based on their best conformity with the target language's standard spelling. In our previous example, this would entail eliminating options with the unusual 'v' letter. Unfortunately, when the language rules governing spelling are superficial, this method of choosing based on conformity is not necessarily the best solution (Taljard, 2008). In the end, all four spelling variations conform to the CV-syllable structure typically preferred in Bantu languages (Ditsele, 2014).

Among others, we noticed a trend of discord in the spelling of (i) *ne* and *ni*, as in the case of *metjhini*, and *metjhine*, (ii) *re* and *ri* as in the example of *rephaboliki* and *riphaboliki*, (iii) *pro* and *poro* as in *porofense* and *profense*, and (iv) *ka* and *kha* as in the case of *kabinete* and *khabinete*. We manually identified instances where spelling varied for one word and decided to retain different spelling variants if they were included in the 3000 most used words. However, these are considered variations and are thus kept in the list as two or more variants of the same entry. In the end, only 30 entries had such varied spellings.

4.3 Plural forms

We treated singular and plural as different entries in the list of frequently used words. For instance, the word dikhemikhale 'chemicals' appears in the lists of frequently used words while the singular form khemikhale 'chemical' does not. Most of the words starting with the letter b and the letter m in the lists are plurals. Given that we want to use our list to identify words that are frequently used, we assumed that the addition of the prefixes to the words changed how the word behaves and therefore should be acknowledged. However, we do acknowledge that the Dale-Chall list counts plurals together with singular forms (Barry and Stevenson, 1975). Identifying all singular and plural forms would result in a very long list where some infrequently used words are falsely identified as frequently used. Furthermore, such analyses would require a trusted lemmatiser. Although lemmatisers have been developed for South African languages (Eiselen and Puttkamer, 2014), the lemmatisers were evaluated on government texts and not on different types of Sesotho texts. As a result, we cannot ascertain their reliability and accuracy in other text genres.

4.4 Abbreviations

A few abbreviations and acronyms were also identified after merging the lists. During the qualitative analysis, it was decided to remove all abbreviations such as *jj* for *jwalojwalo* 'etcetera', *mohl* for *mohla* 'date', and others. Both the abbreviations and their full forms were present in our initial lists of frequently used words. In the end, only the full forms for *jj* and *mohl* were retained in the final list. Even so, we kept both the abbreviation of tv and the full word *thelevishene* 'television'. We decided to keep this abbreviation as it is common in Sesotho. In fact, both the abbreviation and the full word are frequently used Even so, they are retained as one entry with varied spellings (see section4.2 for our treatment of varied spellings).

Acronyms concatenate words into one. For instance, World Health Organisation is counted as one word when abbreviated as WHO (Funk, 1968). For consistency, we removed all instances of abbreviations and acronyms as per Dale-Chall's list which considers abbreviations as unfamiliar. We, however, kept two acronyms, namely, HIV and AIDS as they are globally used abbreviations. The translation of the latter '*eitsi*, was also retained in the final list as it also appeared within the top 3000 most frequently used words before manual editing. Both AIDS and *eitsi* are counted as one entry with varied spellings.

4.5 Unexpected words

A few instances of unexpected words were also identified in the merged list. For instance, letters such as n, d, l, c, r, and b were removed as they do not carry meaning. Instances of non-Sesotho terms were also present on the list. For instance, words such as 'services' and 'language' were identified and removed from the list. Unlike the other unexpected words, the English word 'sister' was not removed from the list. Although there are Sesotho equivalent words for a sister, a matron, a maiden and a nurse, the loan English word 'sister' is more common. In fact, the Sesotho equivalent, *mooki* did not appear in any of the three frequency lists.

4.6 Specific variations

We hope to adapt the English Dale-Chall metric into Sesotho. As a result, we also consider the composition of the Dale-Chall list used in the English metric⁴. The Dale-Chall index uses the formula below to compute estimated grade levels for the Dale-Chall index:

Dale-Chall index =
$$0.0496(\frac{\#words}{\#sentences}) + [11.8(\frac{\#difficultwords}{\#words}) * 0.1579] + 3.6365$$

Difficult words as used in the formula are those that do not appear in the list of frequently used words (Stocker, 1971). As evident from the formula, the identification of difficult words is only one textual property used in determining the readability of texts. Not all words and their variations are listed in the list of frequently used words. However, when computing the scores, specific variations are excluded from this list of difficult words. We focus on two such variations, namely verbs and adverbs.

According to Barry and Stevenson (1975), verbs that end in -s, -ed, -ing, and -ied are not counted as difficult words as they are simply varieties of basic verbs. The Sesotho verb structures do not have an -ing structure. Instead, the continuous tenses are indicated by progressive markers such as '*a*, *ya*, and *ntse*' which are stand-alone words and not suffixed to the verbs. The -ed and -ied structures are indicated by the use of -*wa* and -*uwa* which are suffixed to the verb. For instance, the word *qetwa/qetuwa* 'finished' is derived from *qeta* 'finish'. For the purpose of our list, we disregard these differences and instead count all verb forms as different items.

Finally, English adverbs that end in -ly are not counted as difficult words in the Dale-Chall list (Barry and Stevenson, 1975). Sesotho adverbs take a different type of structure. For instance, the adverb of manner 'lovingly', would be expressed as *ka lerato* as in 'with love'. As a result, distinctions between different adverbs are unimportant in Sesotho and thus for our list.

4.7 Final list

After all these manual steps, we end up with a list of 3000 unique entries. About one third of this list, 993 entries, are words which were frequent in all three corpora. Another third are words frequently appearing in both corpus 2 and 3, namely 975 entries. The final third consists of 723 unique words from corpus 1, 125 unique words from corpus 2, 109 unique words from corpus 3, 49 unique words from corpus 1 and 2 and 63 unique from corpus 1 and 3. Based on these numbers we do notice a difference between corpus 1 (the Bible) and the other two corpora (Autshumato and NCHLT). Further research incorporating this list into actual frequency-based readability metrics and with envisaged end-user will have to corroborate whether the list can be employed for readability purposes.

Please note that in order to allow for spelling variation of certain words (please refer back to Section 4.2 for more information on this) such variations were added to the list for 30 words. As a result, the total number of words included in the list amounts to 3037. The list is made available to the research community at the repository of the Language Resource Management Agency of SADiLaR https://repo.sadilar.gov.

5 Discussion and Conclusion

This article discussed the development of a list of frequently used words in Sesotho using a corpusbased approach. Three different corpora were employed to this end. Although possibly unconven-

⁴*see* https://github.com/words/dale-chall

tional the bible corpus has also been included in this investigation and the results seem to confirm that this corpus behaves somewhat differently. Nevertheless, we believe the inclusion of this corpus is fitting for Sesotho as new orthographies were introduced using bible translations (Makutoane, 2022). For instance, the South African Sesotho orthography was introduced through the 1960 bible translation. Initially, we also wished to incorporate texts from the Sesotho Wikipedia. However, upon closer inspection, it was found that the orthography used in the Wiki pages represents both Sesotho from South Africa and Sesotho from Lesotho depending on who did the translation and/or initial editing of the Wikipedia entry. Such inconsistencies in the orthography may result in misrepresented frequencies. Although some inconsistencies between the orthographies may be semi-automatically corrected, others need a case-by-case analysis.

The aim of designing objective methods for measuring text readability in Sesotho is to enable teachers and assessors, in general, to be able to choose educational texts consistently and fairly. In reality, it is difficult to find texts in an under-researched language such as Sesotho. Finding new texts for learners is not always easy for teachers. In the end, texts are re-used without being re-adjusted to learners' reading abilities. Even so, texts prepared by examiners and textbook compilers at national and provincial levels re-purpose literary texts such as novels and dramas together with magazine and newspaper articles. Some of the texts are even translated from online English texts. The extent to which these texts are suitably adapted to specific learners has not been investigated.

The presence of an empirically developed frequency list such as the one presented in this article allows for taking further steps towards developing classical readability metrics which use frequency lists. Although the texts used for our corpus analysis contain a vast amount of government texts, we believe the final list does not necessarily solely represent government texts, because the meanings of the words are taken out of context. For instance, words such as *leleme* 'tongue/language/gossiper', and moputso 'salary/wage/payment/reward/gain', which both appear on our list, are not only used in government texts. Furthermore, the frequency list from the bible corpus overlapped with more than a third of the frequency lists of the other two corpora. This is an indication that the words are

not necessarily restricted to government texts and that the list may be relevant for use even in other contexts such as educational texts.

In future work, the frequency list will be further validated on a corpus of education texts (Sibeko and Van Zaanen, 2022a), and be incorporated into adaptations of several readability formulas requiring frequency word lists from English to Sesotho. Furthermore, the list will be tested on envisaged end-users such as learners, teachers and publishers for further validation.

Limitations

This creation of a list of frequently used words in Sesotho is limited by the corpora used in this article. As such, the variety of text genres represented in the corpora is limited.

Additionally, although the original English Dale-Chall list was designed with the participation of actual fourth graders, we have relied solely on existing resources. Regrettably, the COVID-19 pandemic national lockdown in South Africa has resulted in stricter restrictions being put in place, making it more challenging to access vulnerable participants such as school learners. As a result, obtaining permission to engage with these participants has become a time-consuming exercise. As a result, we were restricted to text-based resources. We hope to be able to carry out additional validation experiments with the envisaged end-users in future work.

Ethics Statement

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.

Acknowledgements

This article forms part of a doctoral project completed at North-West University. This article was completed under the supervision of Professor Orphée De Clercq during a research stay at Gent University. It is partly funded by the Global Minds Fund at Ghent University.

References

- Jonathan Anderson. 1983. Lix and Rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.
- Jaco Badenhorst and Febe De Wet. 2022. NCHLT auxiliary speech data for ASR technology development in South Africa. *Data in Brief*, 41:107860.
- Etienne Barnard, Marelie H Davel, Charl Van Heerden, Febe De Wet, and Jaco Badenhorst. 2014. The NCHLT corpus of the South African languages. In Proceedings of the 4th International Workshop Spoken Language Technologies for Under-resourced Languages, pages 194–200.
- Jeanne Gardner Barry and Timothy E Stevenson. 1975. Using a computer to calculate the Dale-Chall formula. *Journal of Reading*, 19:218–222.
- Dolores Batinić, Sandra Birzer, and Heike Zinsmeister. 2016. Creating an extensible, levelled study corpus of Russian. In *Proceedings of the 13th Conference* on Natural Language Processing (KONVENS 2016), pages 37–43. Universität Bochum.
- Kiára Bendová. 2021. Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal* of linguistics, 72(2):477–487.
- Kiára Bendová and Silvie Cinková. 2021. Adaptation of classic readability metrics to Czech. In *Proceedings* of the International Conference on Text, Speech, and Dialogue: 24th International Conference, pages 159– 171. Springer.
- RHM Brouwer. 1963. Onderzoek naar de leesmoeilijkheden van nederlands proza (Research into reading difficulty in Dutch prose). *Pedagogische studiën*, 40:454–464.
- Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Sekgaila Chokoe. 2020. Spell it the way you like: The inconsistencies that prevail in the spelling of Northern Sotho loanwords. *South African Journal of African Languages*, 40(1):130–138.
- Kate Cockcroft. 2016. A comparison between verbal working memory and vocabulary in bilingual and monolingual South African school beginners: Implications for bilingual language assessment. *International Journal of Bilingual Education and Bilingualism*, 19(1):74–88.
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Edgar Dale and Jeanne Sternlicht Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

- Thabo Ditsele. 2014. Why not use Sepitori to enrich the vocabularies of Setswana and Sepedi? *Southern African Linguistics and Applied Language Studies*, 32(2):215–228.
- Hessel Douma. 1960. Readability of Dutch farm papers: a discussion and application of readability-formulas. *Wageningen: Afdeling Sociologie en Sociografie van de Landbouwhogeschool*, 17:433–470.
- Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.
- Roald Eiselen and Martin J Puttkamer. 2014. Developing text resources for ten South African languages. In Proceedings of the 9th International Conference on Language Resources and Evaluation, pages 3698– 3703.
- Rudolph Flesch. 1974. *The art of readable writing*, 2nd edition. Harper, New York.
- Valerio Franchina and Roberto Vacca. 1986. Adaptation of flesh readability index on a bilingual text written by the same author both in Italian and English languages. *Linguaggi*, 3:47–49.
- Laverne Carl Funk. 1968. *The reading level of selected magazines as determined by the Dale-Chall readability formula*. Ph.D. thesis, University of Minnesota: Minnesota.
- Martha Gauthier and Nathan Johnson. 2019. Identification and recommendations of readability tests for the evaluation of clinical outcome assessments. *Value in Health*, 22:s828.
- Anna Glazkova, Yury Egorov, and Maksim Glazkov. 2021. A comparative study of feature types for agebased text classification. In Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9, pages 120– 134. Springer.
- Revathi Gopal, Mahendran Maniam, Noor Alhusna Madzlan, Siti Shuhaida binti Shukor, and Kanmani Neelamegam. 2021. Readability formulas: An analysis into reading index of prose forms. *Studies in English Language and Education*, 8(3):972–985.
- Aditi Sharma Grover, Gerhard B van Huyssteen, and Marthinus W Pretorius. 2010. The South African Human Language Technologies audit. In *Proceedings* of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pages 2847– 2850.
- Aditi Sharma Grover, Gerhard Beukes. van Huyssteen, and Marthinus W. Pretorius. 2011. The South African Human Language Technology audit. *Language resources and evaluation*, 45:271–288.
- Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

- Georges Henry. 1975. Comment Mesurer La Lisibilite (How to Measure Readability). Labor, Brussels, Belgium.
- Alexandra H Humphreys and Jere Thomas Humphreys. 2013. Reading difficulty levels of selected articles in the Journal of Research in Music Education and Journal of Historical Research in Music Education. *Music Education Research International*, 6.
- Thomas Jakobsen and Thomas Skardal. 2007. Readability index. Report, Agder University.
- Liliane Kandel and Abraham Moles. 1958. Application de l'indice de Flesch à la langue française (Application of Flesch index to the French language). *Cahiers Etudes De Radio-Télévision*, 19(1958):253–274.
- Peter J 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.
- Inge Kosch. 2013. An analysis of the oxford bilingual school dictionary: Northern Sotho and English (de schryver 2007). *Lexikos*, 23:611–627.
- Daleen Krige and Marianne Reid. 2017. A pilot investigation into the readability of Sesotho health information pamphlets. *Communitas*, 22:113–123.
- 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 G Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.
- Cindy McKellar. 2022. *Autshumato Monolingual Sesotho Corpus*. South African Centre for Digital Language Resources. Available at: https://repo.sadilar.org/handle/20.500.12185/583 Accessed: 28 Jan 2023.
- Cindy Mckellar. 2023. *Autshumato English-Sesotho Parallel Corpora*. Southern African Centre for Digital Language Resources. Available at: https://repo.sadilar.org/handle/20.500.12185/577 [Accessed: 6 Feb. 2023].
- Carmen Moors, Illana Wilken, Karen Calteaux, and Tebogo Gumede. 2018a. Human Language Technology audit 2018: Analysing the development trends in resource availability in all South African languages. In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists, pages 296–304.
- Carmen Moors, Illana Wilken, Tebogo Gumede, and Karen Calteaux. 2018b. Human Language Technology audit 2017/18. Technical report, CSIR Meraka Institute.

- Irina Vladimirovna Oborneva. 2006. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov (Automated estimation of complexity of educational texts on the basis of statistical parameters). Thesis, RAS Institut soderzhaniya i metodov obucheniya [RAS Institute of Content and Teaching Methods].
- Joao Palotti, Lorraine Goeuriot, Guido Zuccon, and Allan Hanbury. 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, pages 965–968.
- Romana Kabir Piu, Kazi Rayed Hossain, Noor Hossain Sabab, and Rakib Bin Mannan Ar Rafi. 2020. *Hatred Message Identification using DistilBERT*. Ph.D. thesis, United International University.
- Elizabeth Pretorius, Nompumelelo Mohohlwane, and Nicholas Spaull. 2020. Investigating the comprehension iceberg: Developing empirical benchmarks for early-grade reading in agglutinating African languages. *South African Journal of Childhood Education*, 10(1):1–14.
- Elizabeth J Pretorius and Lieke Stoffelsma. 2017. How is their word knowledge growing? Exploring grade 3 vocabulary in South African township schools. *South African Journal of Childhood Education*, 7(1):1–13.
- Marianne Reid, Mariatte Neil, and Edgar Janse Van Rensburg-Bonthuyzen. 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.
- Jabulani Sibanda and Jean Baxen. 2016. Determining ESL learners' vocabulary needs from a textbook corpus: Challenges and prospects. *Southern African Linguistics and Applied Language Studies*, 34(1):57– 70.
- Johannes Sibeko and Mmasibidi Setaka. 2022. An overview of Sesotho BLARK content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11.
- Johannes Sibeko and Menno Van Zaanen. 2022a. 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].
- Johannes Sibeko and Menno Van Zaanen. 2022b. 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].
- Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New readability measures

for Bangla and Hindi texts. In *Proceedings of COL-ING 2012*, pages 1141–1151. Association for Computational Linguistics.

- Edgar A Smith and R.J Senter. 1967. *Automated readability index*. University of Cincinnati, Ohio.
- Leonard P Stocker. 1971. Increasing the precision of the Dale-Chall readability formula. *Reading Improvement*, 8(3):87.
- Lieke Stoffelsma. 2019a. English vocabulary exposure in South African township schools: Pitfalls and opportunities. *Reading & Writing-Journal of the Reading Association of South Africa*, 10(1):1–10.
- Lieke Stoffelsma. 2019b. From 'sheep' to 'amphibian': English vocabulary teaching strategies in South African township schools. *South African Journal of Childhood Education*, 9(1):1–10.
- Elsabe Taljard. 2008. Terminology practice in a nonstandardized environment: A case study. In *Proceedings of the XIII EURALEX International Congress* (*Barcelona, 15-19 July 2008*), pages 1073–1080.
- Marthie Van der Walt, Kobus Maree, and Suria Ellis. 2008. A mathematics vocabulary questionnaire for use in the intermediate phase. *South African Journal of Education*, 28(4):489–504.
- 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), pages 775–782. European Language Resources Association (ELRA).
- Carien Wilsenach and Maxine Schaefer. 2022. Development and initial validation of productive vocabulary tests for isiZulu, Siswati and English in South Africa. *Language Testing*, pages 567—592.
- Xin Yan, Dawei Song, and Xue Li. 2006. Conceptbased document readability in domain specific information retrieval. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 540–549.