# Compiling a List of Frequently Used Setswana Words for Developing Readability Measures

**Johannes Sibeko**
Nelson Mandela University
University Way, Summerstrand, Port Elizabeth, 6019, South Africa
johanness@mandela.ac.za

**Abstract**

This paper addresses the pressing need for improved readability assessment in Setswana through the creation of a list of frequently used words in Setswana. The end goal is to integrate this list into the adaptation of traditional readability measures in Setswana, such as the Dale-Chall index, which relies on frequently used words. Our initial list is developed using corpus-based methods utilising frequency lists obtained from five sets of corpora. It is then refined using manual methods. The analysis section delves into the challenges encountered during the development of the final list, encompassing issues like the inclusion of non-Setswana words, proper names, unexpected terms, and spelling variations. The decision-making process is clarified, highlighting crucial choices such as the retention of contemporary terms and the acceptance of diverse spelling variations. These decisions reflect a nuanced balance between linguistic authenticity and readability. This paper contributes to the discourse on text readability in indigenous Southern African languages. Moreover, it establishes a foundation for tailored literacy initiatives and serves as a starting point for adapting traditional frequency-list-based readability measures to Setswana.

**Keywords:** Setswana, Frequently used words, Indigenous language, Readability, Low-resourced

## 1. Introduction

There is consensus that words that are frequently encountered in reading become easier to read (Chen and Meurers, 2016; Rello et al., 2013). This connection between word exposure and ease of reading extends to improved word familiarity and subsequent knowledge (Chen and Meurers, 2016). Conversely, the adverse impact on reading fluency is evident when readers are confronted with unfamiliar words or grammatical structures (Newbold and Gillam, 2010).

Therefore, it becomes imperative to delve into the frequencies of words for the development of readability measures. This awareness of word frequencies can serve as a valuable tool to assess and manipulate levels of text readability.

Understanding text readability is important in South Africa, where literacy levels among language learners are consistently low across various languages, both in home and additional language classes. This concern of low literacy skills is particularly emphasised among language learners such as those in Setswana classes who demonstrate greater proficiency in oral skills than in reading (Lekgoko and Winskel, 2008). According to Mophosho et al. (2019), focusing on enhancing reading proficiency, especially among Setswana learners, is crucial.

While research has been conducted on reading ability in Setswana, such as the work by Pule and Theledi (2023), which delves into challenges in reading proficiency and underscores the influence of prosodic features on Setswana comprehension, and the study by Probert (2019), which advocates for targeted research on reading skills in African languages, pinpointing syllables as crucial units for connected reading in isiXhosa and Setswana, there remains a noticeable lack of knowledge regarding strategies for acquiring reading proficiency in African languages when compared to resource-rich languages like English.

In this paper, we use corpora to develop a list of frequently used words in Setswana. The primary aim of developing this list is to facilitate the adaptation of the Dale-Chall readability index for Setswana.

The rest of this paper provides background to frequency-based readability measures in Section 2, it then discusses the need for measuring text readability in the South African context in Section 3, followed by the method for data collection and analysis in Section 4, the findings that outline problems and solutions in Section 5, a discussion of the findings and the implication of the list of frequently used words in Section 6, and the conclusion with recommendations.

## 2. Background

Setswana, alternatively referred to as 'Tswana,' 'Chuana,' or 'Sechuana,' is a Bantu language (Bennett et al., 2016). It forms part of the Sotho-Tswana language group with Sesotho and Sepedi. In South Africa, the Sotho-Tswana language group has over 16 million primary speakers (Fraser, 2023).



*The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL)*, pages 37–44
25 May, 2024. 

Setswana constitutes 8.3% (5.15 million speakers), alongside Sepedi (10%, or 6.2 million speakers) and Sesotho (7.8%, or 4.84 million speakers). Although the majority of Setswana speakers are from South Africa, Setswana is also an official language in Botswana, a recognised national language in Zimbabwe, and a marginalised spoken language in Namibia (Otlogetswe, 2001).

Despite the prevalence of Setswana, we are not aware of prior efforts to develop readability measures in the language.

In a review of readability measures, DuBay (2004) notes that over 200 measures have been developed for English, reflecting the extensive scholarship on text readability spanning over two centuries in high-resource languages (Collins-Thompson, 2014; De Clercq and Hoste, 2016; DuBay, 2004).

We adhere to the definition of text readability proposed by Bailin and Grafstein (2001), who define it as the ease with which a text can be read. Our focus on readability does not extend to text comprehension, understandability, extra-textual properties or reader characteristics.

Our search for readability measures for African indigenous languages revealed at least three readability measures for Afrikaans. A comprehensive examination of the three Afrikaans readability measures is presented in (McDermid Heyns, 2007). In essence, the three readability measures for Afrikaans drew inspiration from the English Flesch-Reading Ease measure.

Furthermore, recent developments indicate initiatives to formulate nine readability measures for Sesotho (Sibeko, 2023; Sibeko and Van Zaanen, 2022). These measures encompass four syllable-information-based metrics, four word-length-based metrics, and one frequency list-based metric. For this purpose, Sibeko and De Clercq (2023) crafted a list of frequently used words in Sesotho, intended for incorporation into the development of the Dale-Chall index for Sesotho. Similarly, the efforts in this paper are geared towards the development of a frequency list for inclusion in the Dale-Chall index for Setswana.

Setswana serves various functions including education. Written texts constitute a significant component of communication in Setswana. Consequently, access to written information in Setswana holds paramount importance. Regrettably, despite the inclusion of Setswana in educational curricula from basic to tertiary levels in Southern Africa, a portion of the language users lack formal education, while others possess only limited educational attainment. As a result, the absence of readability measures for these languages poses a significant challenge, especially when readers encounter difficulties extracting information from written communications.

## 3. Frequency-list-based Measures

A widely accepted hypothesis among readability scholars posits that the readability of a text can be quantified using specific formulas. One prominent category of these formulas includes frequency-list-based readability measures, which operate on the principle that frequently encountered words are easier to recognise, making them easier to read than less common words in texts (Brysbaert et al., 2011). This approach assesses word difficulty by counting infrequently used or challenging words (Gopal et al., 2021). Therefore, the foundation of traditional readability measures, which gauge word familiarity, lies in having a comprehensive list of frequently used words.

To illustrate this principle in practice, George Spache developed the Spache Readability Formula (Spache, 1953). This formula relies on a compilation of familiar words tailored to learners in specific grades. Texts are then segmented into 100-word sections to ascertain the number of unfamiliar words not included in the grade-specific word list (Spache, 1953; Smith, 2016). A higher average of unfamiliar words correlates with harder-to-read texts.

Similarly, Dale and Chall, in their Dale-Chall Index (Dale and Chall, 1948), employ a list of words familiar to and comprehensible by Grade 4 learners. The average of these words is computed, and a higher prevalence of unfamiliar words, absent from the designated list, corresponds to texts that are harder to read.

In this paper, we rely on general corpora and not texts that are tailored for language learners. Even so, our list can serve as a foundation for the development of a frequency-list-based readability measure specifically designed for Setswana.

## 4. Methodology

We collected five corpora to construct a frequency wordlist, aiming to encompass various genres by gathering texts from different sources. The preparation of each corpus for analysis involved lowercasing using *bash* and tokenisation with *ucto*, including the specific requirement for sentence segmentation. The corresponding sentence information is detailed in Table 1. Below are brief overviews of the five corpora.

### 4.1. Corpus 1: NCHLT

The objective of the National Centre for Human Language Technology (NCHLT) project was to generate speech and text data to support the development of Human Language Technologies (HLTs) for the 11 official written languages of South Africa

| File | Sentences | Lines | Marks | Numbers | Tokens | Words | Types |
|---|---|---|---|---|---|---|---|
| NCHLT | 58 443 | 58 520 | 147 058 | 41 351 | 1 400 737 | 1 249 980 | 38 864 |
| Autshumato | 104 976 | 103 425 | 266 388 | 82 062 | 2 887 117 | 2 596 847 | 53 810 |
| PuoBERTa | 67 388 | 67 071 | 143 214 | 25 745 | 2 396 525 | 2 248 475 | 41 037 |
| Wikipedia | 48 541 | 47 718 | 106 459 | 21 379 | 1 179 331 | 1 063 236 | 42 157 |
| Bible | 37 526 | 30 891 | 106 386 | 4 995 | 958 692 | 834 748 | 18 765 |

Table 1: Summary of Text Properties

(Eiselen and Puttkamer, 2014; Badenhorst and De Wet, 2022). The text collection[1] consists of translated data acquired from the South African Government domain, with ample training and testing samples for language identification tasks in each language (Duvenhage, 2019).

The original dataset includes source texts, lexica, and the corpus (Eiselen and Puttkamer, 2014). We utilised the cleaned corpus data (approximately 1 249 980 words) and not the raw or source files.

### 4.2. Corpus 2: Autshumato

The Autshumato Machine Translation project[2] developed a translation text corpus for South African indigenous languages. The texts were manually and professionally translated from English into the other ten official written languages of South Africa. The English-Setswana texts are publicly accessible on the South African Centre for Digital Language Resources (SADiLaR) online repository (Mckellar, 2023).

The Autshumato English-Setswana parallel corpora consist of three distinct sets. The Set 1 collection comprises data that has been translated from English into Setswana by professional translators. This set encompasses a total of 324 342 Setswana words. The Set 2 collection contains data sourced as translated file pairs from reliable translators, with a total of 1 099 509 Setswana words. Lastly, the Set 3 collection comprises data crawled from various government websites, containing a total of 1 172 172 Setswana words.

Ultimately, the Autshumato corpus comprised approximately 2 596 847 words. McKellar (2022) outlines at least four text types from the dataset, including magazines, policies, newsletters, and translation works, in addition to documents obtained from the `gov.za` domain.

### 4.3. Corpus 3: PuoBERTa

We also collected the PuoBerta corpus (Marivate and Wagner, 2023). The PuoBERTa corpus functions as a News Categorisation dataset (Marivate et al., 2023). Its primary objective is to facilitate the development of monolingual resources for Setswana, encompassing tasks such as part-of-speech (POS) tagging, named entity recognition (NER), and mainly, news categorisation. The dataset was derived from online news articles accessible that were provided by the Botswana Government.

The Berta corpus comprises three data files: the development set (230 373 words), the training set (1 806 813 words), and the test set (226 614 words). We amalgamate the texts to compile a corpus of 2 248 475 words pre-processing.

### 4.4. Corpus 4: Wikipedia

Our Wikipedia corpus is sourced from Leipzig-Corpora-Collection (2020), offering three downloadable corpora. The first corpus, Leipzig-Corpora-Collection (2017), involves texts crawled from general Wikipedia, totalling 660 041 words. The second corpus, from 2018, comprises 232 210 words collected in Botswana. The third corpus, from 2020, consists of 229 987 words from South Africa. Both the 2020 and 2018 files include 10 000 sentences each. In total, our Wikipedia corpus encompasses 1 063 236 words.

### 4.5. Corpus 5: Bible

We make use of bible texts sourced from the MyBible project which is a non-profit religious initiative that offers its resources freely at `https://mybible.zone/en/`. This project and website provide Bible translations in various languages, including all the written languages of South Africa. The site provides two Setswana Bible versions including *Beibele e e boitshepo*, the 1907 version that uses the initial and founding orthography of Setswana and *Beibele*, the 1970 version that employs the refined orthography of Setswana. For our paper, we use the 1970 version.

The Bible texts were acquired in SQL3lite format from `https://www.ph4.org/b4_index.php#google_vignette`. All text extraction procedures were executed using *bash* scripts. The Bible texts are categorised into three sections: (i) Bible books with 66 rows of data, (ii) verses with 31,170 rows of data, and (iii) info with 10 rows

[1]Access the NCHLT corpus at `https://repo.sadilar.org/handle/20.500.12185/343`

[2]Access the Autshumato corpus at `https://repo.sadilar.org/handle/20.500.12185/404`

of data. Specifically for our corpus, we extracted verse texts, which then underwent cleaning processes involving the removal of book numbers, chapter titles, and verse information.

Religious texts have been successfully employed for corpus development in previous studies. For instance, Agic and Vulic (2019) utilised parallel articles from the Jehovah's Witness website. Similarly, Marivate et al. (2020) employ Bible texts from both Sepedi and Setswana for a news topic classification task.

### 4.6. A common frequency list

We generated different frequency lists for the five sets of corpora by calculating frequencies for each. To achieve independence from corpus size, we employed relative frequencies (Brysbaert et al., 2011; Leech et al., 2014; Van Heuven et al., 2014), normalising the frequency lists to occurrences per million tokens. We aimed to extract the most frequent 3 500 words from each set of data. Some of the data sets contained more words on the same level of frequency and thus resulted in longer lists than the intended 3 500 words. The first step resulted in a total of 17 683 words.

Our primary objective was to end up with a list of 3 000 unique words based on the five corpora. To accomplish this, we merged the five lists and ensured the average relative frequencies of duplicate entries. For instance, the entry *'go'* appeared in all five lists with relative frequencies of 28 315.13, 52 336.45, 52 303.89, 61 184.58, and 58 209.73, respectively. The resulting average frequency for this entry is 50 470.156 per million words. We then identified the top 3 000 most frequently used words for the final list.

## 5. List Analysis

The initial list of 3 000 words was generated automatically using corpus-based frequency measures and later refined through manual processing. The final compilation comprises 3 006 entries, including 2 992 unique entries and 14 instances of varied spellings. The subsequent section provides a detailed account of the curation process involved in finalising the list.

### 5.1. Non-Setswana Words

We identified a total of 60 instances of non-Setswana words from our initial list. Examples included terms like '*superintendent*,' '*of*,' and '*society*." These instances were excluded due to their lack of Setswana origin and absence of normalised or naturalised Setswana orthography. Nevertheless, contemporary terms such as '*corona*" and '*covid*,' which also deviate from Setswana's naturalised orthography, were retained on the list. This decision was based on the recognition that these terms are more commonly used in the indigenous languages of South Africa than their translated counterparts.

Furthermore, considering linguistic conventions in South Africa, where certain terms like month names, for example, '*June*,' are typically written in English, we have retained these names in the list. However, it is worth noting that not all months are included in the list, as we aim to maintain fidelity to the corpus under analysis. Nonetheless, there are also instances of months in Setswana, such as '*Motsheganong*'.

We also chose to include the entry '*eish*' in our current list. While acknowledging its primary association with a magazine, we opted to retain it due to its additional usage as a borrowed exclamation. This term appears in three of our source corpora, where in the NCHLT corpus, it pertains specifically to the '*Eish*' magazine, and in the Wikipedia and Autshumato corpora, where it is employed both as an exclamation and in reference to the magazine.

### 5.2. Abbreviations and Acronyms

The initial list included abbreviated words. For example, words such as '*Mopofof*' - representing '*Mopofofesa*' as in professor, '*Moh*' - standing for '*Mohumagadi*' as in Miss, and '*jj*' - for etc. were identified. These abbreviations were retained although full versions for '*Moh*' and '*jj*' were excluded from the list to maintain fidelity to the list.

We also noted that there were instances of unfamiliar abbreviations, such as the ambiguous '*rbn*.' A closer examination revealed that this abbreviation originated from the Autshumato collection, where '*rbn*' referred to a specific company. Consequently, we decided to remove this particular entry from our list.

Secondly, the initial list included acronyms, such as '*SARS*' representing the South African Revenue Service. Note that these entries were anticipated since some texts were sourced from government websites. Among the expected acronyms were '*SAPS*,' denoting the South African Police Service, and '*SASSA*,' an acronym for the South African Social Security Agency, the current distributor of welfare grants in South Africa. Despite this anticipation, we made the decision to eliminate these entries from the list. The rationale behind this choice is twofold. Firstly, these acronyms deviate from the typical Setswana words as they are not normalized into Setswana. Additionally, they demonstrate a specific inclination towards a domain, which further justifies their exclusion.

Even so, we opted to retain globally recognised acronyms such as '*HIV*' (human immune virus),

as they are typical in Setswana texts beyond our current corpus. Interestingly, the Sesotho list of frequently used words (Sibeko and De Clercq, 2023) contains both *HIV* and *aids* while our list only contains *HIV*.

### 5.3. Proper Names

Our initial list contained at least 80 proper names most of which were biblical names such as '*Gileate, Hesekia, Abesalomo, Jerobeame, Nebukatenesare*' and others. These biblical names were naturalised into Setswana and used expected Setswana orthography.

The list also contained names of African icons such as *Mandela*, as well as names of places such as '*Francistown, Gauteng, Zimbabwe*', and '*Vaal*'. Similar to Sibeko and De Clercq (2023), we removed all instances of proper names. According to Dale and Chall (1948), proper names are automatically deemed familiar and need not be included in the frequency list.

### 5.4. Multifaceted Meanings

There were instances of words where the meaning was unclear and not immediately discernible without context. For example, the entry '*time*' could be interpreted in English to refer to the passage of time, a specific point in time, or planning. In Sotho-Tswana languages, it can also be used to signify switching off. Similarly, the entry '*rate*' may mean to evaluate or assess in English, but it typically carries the meaning of love in Setswana.

Despite the potential for ambiguity, these words are retained in the final list. This decision acknowledges the diverse meanings they hold across different linguistic contexts. The assumption is that readers will interpret these words in Setswana rather than in English when reading Setswana texts. It is important to note, however, that this ambiguity may pose challenges in the context of multilingual texts where the reader will have to rely on context to aid in identifying the correct language and expected pronunciation when reading.

### 5.5. Unexpected Words

The preliminary list included unexpected entries. Firstly, there were instances of non-word entries, including numerical values. All such instances were removed from the list as we are interested only in frequently used words.

Secondly, we observed the presence of isolated letters such as '*p*', '*d*', '*g*', '*s*', '*f*', '*i*' and others. We systematically removed all instances of isolated consonants from the list because individual consonants do not qualify as valid Setswana words.

Furthermore, even though certain Setswana vowels can constitute words in a vowel-only context (for instance, '*a,' 'e*' and '*o*'), it was noted that the vowel '*i*' does not serve as a standalone word. Consequently, we excluded this particular entry from our list. Nonetheless, a more thorough analysis revealed that the letter '*i*' was predominantly used as a page number reference in the source documents.

### 5.6. Spelling and Orthography

There is a general consensus that Sotho-Tswana languages lack specific rules for governing the orthography of loanwords (Chokoe, 2020). This absence of clear regulations manifested in our preliminary list, leading to diverse spellings for the word '*Afrika*.' We identified at least four spelling variations, including '*Africa*,' '*Aferika*,' '*Aforika*,' and '*Afrika*,' with the '*Afrika*' spelling exhibiting a higher frequency. These varied spellings are also associated with related terms such as '*Afrikaborwa*,' '*Afrikan*,' '*Pan-Afrikan*,' '*MoAfrikan*,' '*MaAfrikan*,' and others. Likewise, additional spelling variations yield similar words, as seen with '*MoAforikaborwa*' and '*MaAforikaborwa*,' both present in the list. Like Sibeko and De Clercq (2023), we retained all spelling variations as long as they were part of the initial list of frequently used words.

We were surprised to encounter a misspelled word, namely '*bosetphaba*', which, upon contextual analysis, was identified as originating from the NCHLT corpus. The correct form is '*bosetšhaba*', meaning '*national*'. Recognising it as a typographical error, we have excluded this entry from our current list.

Furthermore, we observed the inclusion of dash-compounded words like '*ba-na-le*' and '*bokone-bophirima*'. To maintain a focus on individual words, we have opted to remove compound entries from our list.

## 6. Discussion and Conclusion

This paper contributes to the development of readability measures for lower-resourced indigenous languages of Southern Africa by developing a list of frequently used words for Setswana. As detailed in the introduction, the scholarly focus on high-resource languages has left indigenous Southern African languages, including Setswana, understudied in the realm of text readability.

Our research aims to improve the applicability of readability measures to the Sotho-Tswana language group. We draw inspiration from the ongoing Sesotho readability project (Sibeko, 2023), which introduces readability measures based on word length, syllable information, and frequency lists. However, before our work, the transferability

of frequency-list-based measures to Setswana encountered difficulties because there was no curated list specifically tailored for integration into readability measures.

Inspired by established readability measures such as the Dale-Chall Readability Index currently in development for Sesotho, our goal is to adapt and extend these measures to address the unique linguistic context of Setswana. The Dale-Chall Index, known for relying on a list of familiar words, aligns seamlessly with our objective of enhancing readability by prioritising frequently used Setswana words.

### 6.1. Challenges and Decision-Making

The choice to preserve diverse spelling variations aligns with recommendations in the literature (Sibeko and De Clercq, 2023), highlighting the significance of inclusivity in representing frequently used words. This strategy results in a comprehensive frequency list that acknowledges the linguistic richness and variations present in Setswana. While this approach may introduce a potential mismatch between word use frequency and their inclusion in our list, considering that words may be spelled differently in various corpora, it overlooks the aspect of familiarity for readers encountering different word forms. Consequently, we opted to include words only if they were part of our original shortlist, maintaining fidelity to the actual appearances of words in the list.

Additionally, as illustrated in Section 5, the manual cleaning process revealed non-Setswana words, abbreviations, proper names, and unexpected terms within the Setswana corpus. These occurrences presented challenges during the compilation of the frequency list. Consequently, specific measures were implemented to either retain or exclude these words from the list.

### 6.2. Implications

The literacy challenges faced by school language learners, particularly those in the Sotho-Tswana language group, underscore the pressing need for tailored readability measures. Unfortunately, Setswana is not well-explored in Natural Language Processing (Marivate et al., 2023). Our goal is to help address this gap by focusing on readability studies specifically in the context of Setswana.

The presence of our list of frequently used words not only offers valuable insights into reading proficiency but also serves as a foundation for the development of Setswana-tailored readability measures that rely on lists of frequently used words. These measures will enable educators and policymakers to make informed decisions, providing targeted strategies to enhance reading proficiency among Setswana learners.

Curriculum developers, assessors, and teachers can leverage our list to guide their language teaching decisions and to select desirable reading materials for both instruction and assessment.

### 6.3. Limitations

Note that traditional readability measures are criticised for many shortcomings. For instance, according to Crossley et al. (2021), these measures commonly rely on estimates for measuring lexical and syntactic features, while neglecting semantic features and discourse structures, text cohesion and style elements. Furthermore, they are limited in reading criteria and are susceptible to age group and domain specificity. The current paper does not address these shortcomings. Instead, it focuses on the development of a frequency list that can be used in the development of a frequency-based readability measure based on the Dale-Chall index.

The findings presented in this paper exhibit certain limitations. Notably, unlike the Spache Readability Formula examples (Spache, 1953) and the Dale-Chall Index instances (Dale and Chall, 1948), our approach involves compiling a list of frequently used words in the language as observed from limited corpora rather than tailoring it for specific readers in a particular grade level.

While our research was constrained by the absence of originally written texts in Setswana designated for educational purposes and the resulting unavailability of educational corpora, we recommend that future research explores the development of grade-level lists. This refinement could enhance the applicability and precision of readability measures for Setswana, aligning them more closely with the educational context and readership levels targeted in language-related studies.

### 6.4. Future Directions

Building on our current work, we envision several avenues for future research that will contribute to the ongoing development of Setswana readability measures and broader linguistic studies. Comparative studies with other Sotho-Tswana languages, such as Sepedi and Sesotho, will identify shared linguistic patterns and assess the generalisability of common word lists and readability measures across these languages.

Additionally, extending the analysis of a frequency list such as the one proposed in this paper to include a diverse range of text types, including educational materials, news articles, and literary works, will capture the breadth of Setswana language usage and ensure the applicability of readability measures across contexts. Nonetheless, the

exploration of reading proficiency methodologies in African languages, as advocated by Probert (2019), remains imperative.

## 7. Bibliographical References


Željko Agic and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.

Jaco Badenhorst and Febe De Wet. 2022. NCHLT auxiliary speech data for ASR technology development in South Africa. *Data in Brief*, 41:107860.

Alan Bailin and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301.

William. G Bennett, Maxine Diemer, Justine Kerford, Tracy Probert, and Tsholofelo Wesi. 2016. Setswana (South African). *Journal of the International Phonetic Association*, 46:235–246.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental psychology*.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th workshop on innovative use of nlp for building educational applications*, pages 84–94.

Sekgaila Chokoe. 2020. Spell it the way you like: The inconsistencies that prevail in the spelling of Northern Sotho loanwords. *South African Journal of African Languages*, 40(1):130–138.

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

Scott A Crossley, Aron Heintz, Joon Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The commonlit ease of readability (clear) corpus. In *EDM*.

Edgar Dale and Jeanne Sternlicht Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq and Véronique Hoste. 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.

William H DuBay. 2004. The principles of readability. Technical report, Impact Information, Costa Mesa.

Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.

Roald Eiselen and Martin Puttkamer. 2014. Developing text resources for ten South African languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*, pages 3698–3703, Paris. European Language Resources Association (ELRA).

Luke Fraser. 2023. These are the most spoken languages in South Africa. Technical report, BusinessTech. Accessed: 11 Jan 2024.

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.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Olemme Lekgoko and Heather Winskel. 2008. Learning to read Setswana and English: Cross-language transference of letter knowledge, phonological awareness and word reading skills. *Perspectives in Education*, 26(4):57–73.

Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. 2023. Puoberta: Training and evaluation of a curated language model for setswana. In *Artificial Intelligence Research. SACAIR 2023. Communications in Computer and Information Science*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, Abiodun Modupe, and South Africa CSIR. 2020. Pedi. *arXiv preprint arXiv:2004.13842*.

Jacques McDermid Heyns. 2007. Readability statistics for Afrikaans. In *LSSA/SAALT/SAALA Joint Annual Conference, North-West University, Potchefstroom, South Africa*.

Cindy McKellar. 2022. *Autshumato Monolingual Sesotho Corpus*. ONLINE. 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[Lastmodified:15Dec.2022].

Munyane Mophosho, Lesedi L Sebole, and Katijah Khoza-Shangase. 2019. The reading comprehension of grade 5 Setswana-speaking learners in rural schools in South Africa: Does home language matter? *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 35(3):59–73.

Neil Newbold and Lee Gillam. 2010. Populating a framework for readability analysis: Word frequency = word difficulty.

Thapelo Otlogetswe. 2001. The BNC design as a model for a setswana language corpus. *extraction*, page 1.

Tracy N Probert. 2019. A comparison of the early reading strategies of isiXhosa and Setswana first language learners. *South African Journal of Childhood Education*, 9(1):1–12.

Violet Maphefo Sefolaro Pule and Kgomotso Theledi. 2023. The impact of the presence of prosodic features (tone markings) on comprehending Setswana words in reading. *African Journal of Inter/Multidisciplinary Studies*, 5(1):1–12.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*, pages 203–219. Springer.

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

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

Johannes Sibeko and Menno Van Zaanen. 2022. Developing a text readability system for Sesotho based on classical readability metrics. In *Digital Humanities Conference: Responding to Asian diversity, Book of Abstracts*, volume 2022, pages 571–572. Short Paper. Available at: https://dh-abstracts.library.virginia.edu/works?keywords=11133 Accessed: 15 Apr. 2023.

Terry Smith. 2016. The problems with current readability methods and formulas: missing that usability design. In *2016 IEEE International Professional Communication Conference (IPCC)*, pages 1–4. IEEE.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

## 8. Language Resource References

Leipzig-Corpora-Collection. 2017. *Tswana community corpus based on material from 2017*. University of Leipzig. PID https://corpora.uni-leipzig.de?corpusId=tsn_community_2017. Accessed: 2024-01-20.

Leipzig-Corpora-Collection. 2020. *Download Corpora Tswana*. University of Leipzig. PID https://wortschatz.uni-leipzig.de/en/download/Tswana#tsn_wikipedia_2021. Accessed: 2024-01-20.

Marivate, Vukosi and Wagner, Valencia. 2023. *Daily News - Dikgang Categorised News Corpus*. Data Science for Social Impact Group. PID https://github.com/dsfsi/PuoBERTa.