

Measuring Text Readability in Sesotho

J. Sibeko

 orcid.org/0000-0003-3586-7491

Thesis accepted in fulfilment for the degree *Doctor of Philosophy
in Linguistics and Literary Theory* at the North-West University

Promoter: Prof M van Zaanen

Graduation: May 2024

Authorship Declaration

I, Johannes Sibeko, declare that:

- This thesis uses an article format.
- The content of this thesis is my unaided work except for co-authored pieces herein.
- This thesis has not been submitted elsewhere for any degree or non-degree purposes.
- All the resources that I have used or quoted have been indicated and acknowledged through complete reference.
- I also confirm that I have submitted the Turnitin report for this dissertation to my supervisor.

Date: 06 February 2024

Acknowledgements

This research was funded in part by the:

- (i) The National Research Foundation (NRF) South Africa (Grant number: 128875).
- (ii) Global Minds Fund Short Research Stays program (GMF.SRS.2022.0001.01).
- (iii) Nelson Mandela University through the Teaching Relief Grant.
- (iv) North-West University through Postgraduate Funding

I extend my deepest appreciation to:

- Prof Menno van Zaanen for his unwavering support throughout this journey. Your mentorship and commitment ensured a successful outcome, and I hope to have the privilege of being under your guidance for my second PhD.
- Prof Orphee de Clerq, your impressive multitasking is truly inspiring.
- Prof Naydene de Lange, thank you for consistently reminding me that I can do it.
- Dr Jacqui Lück, your guidance has been akin to that of an academic aunt.
- Prof Anna Susanna Coetzee-van Rooy, I apologize for not completing our PhD project. Thank you for introducing me to Prof Menno; I do not regret heeding your advice.

Dedications

I dedicate this doctoral thesis to my mother in heaven, Betty Nomsebenzi Hadebe. Your memory has been a guiding light throughout my academic journey, and I know that you would be proud of me today. Emma Angela Nomvula Hadebe, you were a constant source of support, and I am grateful for your unwavering love and encouragement, may you continue resting in peace.

Mme Nomhlahlo Nomangoma Elsie Hadebe and Bhuti Nkomo Jantjie Hadebe in heaven, you have always been there for me, and I love you all. Bhuti Vusimuzi Hadebe, I finally finished the PhD, I wish you were to witness this milestone.

To my siblings, Ntonto Nosibununyana Adeline Hadebe-Nkuna, Lizbeth Lee Lizzy Hadebe-Khanya, and Papsita Doctor Ball ‘The Round One’ Hadebe, thank you for your love and support. Your belief in me kept me going during the most challenging times.

My Pontjonono, Seun Moholi Motjhotjhonono Sean Elias Mpetjekana, you have been my rock (except that one time you took my laptop when I needed to work). To my children, Nomhle, Nompumelelo, Justice, Zakhele, Khanyisile, Katleho, Nkosana, Musa, Lilly, Nomusa, and Nhlakanipho, this is for you, my babies.

To my family, Ndlaleni, Nokuhlupheka, Deliwe, Nompoti, Maphelile, and my late family, Zodwa, and Khale, thank you for being part of this journey.

All the people mentioned in this dedication have tremendously impacted my life, and I am grateful. Finally, **to God**, I owe everything. You have been the foundation of my life, and without you, none of this would have been possible. Thank you for your endless grace and love.

I continue to sing “Ha le mpotsa tshepo ya ka, ke tla re ke Jeso” because I know that you are my hope at all times. I also sing Ungakhohlwa izifungo by the Holy Cross Choir because I need you to always remember the promises you made to my Mom who prayed for my success while she lived.

English Abstract

The main objective of the thesis is to investigate and propose objective text readability measures for Sesotho. This investigation serves as a case study for the development of readability measures for low-resource languages.

This thesis assumes that Sesotho texts can be ranked according to their levels of text readability. To measure text readability objectively, we would like to use computational analyses for this. Unfortunately, Sesotho is a low-resource language, lacking sufficient basic language resources necessary for many computational natural language processing methods.

We first survey existing readability measures for both high-resource languages, like English, and approaches to develop such measures for low-resource languages. Based on these findings, we identify existing resources needed to develop the measures for Sesotho.

As such, various sets of data are collected, including a dictionary lexicon for extracting syllable information, examination texts in all official languages of South Africa from 2008 to 2020, Sesotho Bible translations, Autshumato machine translation texts, and the National Centre for Human Language Technology machine translation texts. In addition, we develop several language resources, such as a syllable annotated wordlist, a corpus-based list of common words in Sesotho, Google Translate machine translations of grade twelve Sesotho examination texts to English, and both pattern-based and rule-based syllabification systems.

To create readability measures for Sesotho, we extract text features from Sesotho high school examination texts. The readability values of these texts are computed from their machine-translated English version using existing English readability measures. This allows us to create linear regression models that fit the features against the readability values which leads to readability measures specifically for Sesotho. The structure of the linear regression models for the Sesotho readability measures follows that of the English readability measures. This results in a collection of nine readability measures for Sesotho.

Overall, our findings indicate that Sesotho text can be ranked according to levels of readability. Equally important, our results demonstrate that traditional (English) formulas can be adapted to low-resource languages like Sesotho. However, additional efforts may

be needed to prepare resources for such projects. We provide a practically useful approach by transitioning from an initial absence of a readability corpus, common vocabulary, and syllable identification to attaining a finalised product.

To the best of our knowledge, the findings of this thesis present the first formulas for an indigenous language of South Africa and the first for the Sotho-Tswana language group in Southern Africa. Although our models are trained on educational texts, specifically reading comprehension and summary writing texts, the availability of standardised and objective readability formulas provides a solid starting point for employing machine learning approaches to measure text readability in Sesotho. Furthermore, the methods outlined in the thesis can be employed in the development of readability measures for other low-resource languages.

Keywords: Sesotho; Readability; Readability measures; Traditional readability measures; Text readability; Corpora; Sesotho linguistics; Reading ability; BLARK; Basic Language Resource Kits; Grade 12; Exam texts; Syllables, Syllabification.

Sesotho Abstract

Thesis ena e nka hore dingolwa tsa Sesotho di ka hlophiswa ho ya ka maemo a tsona a paleho ya dingolwa. Ho lekanya ho balwa ha mongolo ka nepo, re lakatsa ho sebedisa tlhahlobo ya khomphutha bakeng sa sena. Ka bomadimabe, Sesotho ke puo e nang le disebediswa tse mmalwa, mme se ya haellwa ke disebediswa tsa puo bakeng sa mekgwa e mengata ya khomphutha e sebedisang puo ya tlhaho.

Sepheo se ka sehlohong sa thesis ena ke ho fuputsa le ho sisinya mekgwa ya paleho ya mongolo wa Sesotho. Patlisiso ena e sebetse jwalo ka mohlala bakeng sa ntshetsopele ya mehato ya ho balwa bakeng sa dipuo tse nang le disebediswa tse tlaase.

Re qala ka ho hlahlola mekgwa ya paleho e sebediswang dipuong tse nang le disebediswa tse ngata, jwalo ka Senyesemane, le mekgwa ya ho theha mekgwa ya paleho bakeng sa dipuo tse hlokanang mekgwa ya paleho. Ho ipapisitse le diphuputso tsena, re hlwaya mehlodi e teng le e hlokahalang ho ntshetsapele mekgwa ya Sesotho.

Ka hona, re bokella dihlopha tse fapaneng tsa disebediswa tsa mongolo, ho kenyeletsa lethathama la mantswe a sebediswang ho hlwaya dinoko tsa Sesotho, ditemana tsa tlhahlobo dipuong tsohle tsa molao tsa Afrika Borwa ho tloha ka selemo sa 2008 ho ya ho sa 2020, diphetolelo tsa Bible tsa Sesotho, ditemana tsa phetolelo ya motjhini wa Autshumato, le Setsi sa Sechaba sa Puo ya Batho le ditemana tse fetoletsweng ka motjhini wa theknoloji. Ho feta moo, re theha disebediswa tse mmalwa tsa dipuo, jwalo ka lethathamo la mantswe a hlwailweng dinoko, lethathamo la mantswe a tlwaelehileng a Sesotho, diphetolelo tsa motjhini wa Google Translate tsa ditemana tsa tlhahlobo tsa Sesotho tsa Kereiti ya 12 ho ya Senyesemaneng, le disisitimi tse pedi tsa kabodinoko tse thehilweng ho mokgwa-kekiso le tse thehilweng melaong.

Ho theha methati ya paleho bakeng sa Sesotho, re sebedisa ditemana tsa mongolo ho tswa ho dihlahlobo tsa Sesotho tsa kereiti ya 12 ho hlwaya dibopeho tsa mongolo. Maemo a ho balwa a ditemana tsena a balwa ho tswa phetolelong ya tsona ya Senyesemane e fetoletsweng ka motjhini ho sebediswa mehato e teng ya paleho ya Senyesemane. Sena se re utullela bokgoni ba ho rala dimotlolo tsa kgutlisetso tse ithehileng hodima tsamaisano ya dikarolo tsa mongolo le maemo a paleho ho rala mekgwa ya paleho ya Sesotho.

Sebopeho sa mekgwa ya dimotlolo tsa kgutlisetso tsa paleho ya Sesotho e latela ya mekgwa

ya paleho ya Senyesemane. Sena se fella ka pokello ya mekgwa e robong ya ho bala ya Sesotho.

Ho hlahile dintlha tse pedi diphuputsong tsa rona. Re hlokometse hore boima bo abetsweng dinoko ka lentse le leng le le leng ka hara difomula tsa rona tse rerilweng di dula di le tlase ho feta tse ngotsweng ka mekgwa wa Senyesemane, haholo-holo fomuleng ya Flesch Reading Ease. Ho feta moo, re eleletswe paterone e fapaneng, moo dibopeho tsa bolelele ba dipolelo ho ya ka mantswe polelong ka e fihletseng boima bo hodimo ho difomula tseo re di sisinyang.

Ka maswabi, re kopane le kgaello ya dingodilweng tse fanang ka tlhahlobo ya papiso ya dibopeho tsa mantswe dipakeng tsa Senyesemane le Sesotho, hammoho le tlhahlobo ya bolelele ba polelo ka pakeng tsa dipuo tsena tse pedi. Ka lebaka leo, ha re kgone ho tokafatsa kapa ho hlalosa sesosa sa bo tlaase ba boleng bo tsamaisanang le dinoko lentse le leng le bolelele ba dipolelo. Leha ho le jwalo, ho bohlokwa ho ela hloko hore tlhahisoleseding ka dinoko le bolelele ba dipolelo di na le tshusumetso e kgolo mekgweng eo re e kgothallatsang bakeng sa Sesotho. Ho feta moo, dikarolo tse sebedisitsweng ho difomula tse fapaneng di bontsha matla a phahameng a ho aba paleho ya mongolo.

Ka kakaretso, diphuputso tsa rona di netefatsa menahano ya rona ya hore mongolo wa Sesotho o ka hlophiswa ho latela maemo a paleho. Habohlokwa ka ho tshwanang, diphetho tsa rona di bontsha hore diforomo tsa moetlo tsa Senyesemane di ka fetolelwa dipuong tse hlokanang disebediswa, jwalo ka Sesotho. Leha ho le jwalo, ho ka nna ha hloka hloka boitekoko bo eketsehileng ho lokisetsa disebediswa bakeng sa merero e jwalo. Re fana ka mekgwa o molemo ka ho tloha ho bosikong ba disebediswa le pokello ya mengolo, lethathama la mantswe a tlwaelehileng, le mekgwa ya nokofatso ya ho fumana sehlahiswa se seng se fedile.

Ho ya ka tsebo ya rona, diphuputso tsa thesis ena di hlalisa mekgwa ya pele ya paleho ya puo tsa matswalla a Afrika Borwa le ya pele bakeng sa sehlopha sa puo ya Sotho-Tswana ya Afrika Borwa. Le hoja mehlala ya rona e rupetswe ka ditemana tsa thuto, haholo-holo dingolwa tsa ho bala ka kutlwisiso le ho ngola dikgutsufatso, ho ba teng ha difomula tse leka-lekaneng le tse abang ka nepo ho fana ka qaleho e tiileng ya ho sebedisa mekgwa ya maemo a hodimo ya ho ithuta ka motjhini ho metha paleho ya Sesotho. Ho feta moo, mekgwa e hlalositse ho thesis ena e ka sebediswa ho ntshetsopele ya mekgwa ya ho bala dipuo tse ding tse nang le disebediswa tse tlaase.

Mantsweshloho: Sesotho; Paleho; Metjha ya paleho; Diforomo tsa moetlo; Paleho ya mongolo; Pokellomengolo; Thutopuo; Bokgoni ba ho bala; Disebediswa tsa Motheo tsa Dipuo; Sehlopha sa 12; Ditemana tsa tlhahlobo; Dinoko; Nokofatso.

Table of abbreviations

The following abbreviations are used in the text.

ANOVA: Analysis of Variance

ARI: Automatic Readability index

ASL: Average Sentence Length

AWL: Average Word length

BASH: Bourne Again SHell

BLARK: Basic Language Resource Kits

CAPS: Curriculum and Assessment Policy Statements

CLI: Coleman-Liau Index

DCI: Dale-Chall Index

FAL: First Additional Language

FKGL: Flesch-Kincaid Grade Level

FRE: Flesch Reading Ease

GFI: Gunning Fog Index

HL: Home Language

HRL: High-Resourced Language

LIX: Läsbarhetsindex

LRL: Low-Resourced Language

LS: Lesotho Sesotho

MT: Machine Translation

NCHLT: National Centre for Human Language Technology

NLP: Natural Language Processing

RIX: Readability Index

SA: South Africa

SAS: South African Sesotho

ST: Source Text

SMOG: Simple Measure of Gobbledygook

T_EX: *Technikí Setarísmatos* (Technique of typesetting)

TT: Target Text

TTR: Type-Token Ratio

Related publications

The following journal articles are published and are based on work in this thesis:

- Sibeko, J. and 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(1):9. <https://doi.org/10.5334/johd.108>
- Sibeko, J. 2022. Exploring linguistic complexity of English First Additional Language Grade 12 and Grade 11 exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 4(02):1–11. <https://doi.org/10.55492/dhasa.v4i02.4039>
- Sibeko, J. and Setaka, M. 2022. An overview of Sesotho BLARK Content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11. <https://doi.org/10.55492/dhasa.v4i01.4440>
- Sibeko, J. 2021. A comparative analysis of the linguistic complexity of Grade 12 English Home Language and English First Additional Language examination papers. *Per Linguam*, 37(2):50–64. https://hdl.handle.net/10520/ejc-perling_v37_n2_a5
- Sibeko, J. and Van Zaanen, M. 2021. An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11. <https://doi.org/10.55492/dhasa.v3i01.3864>

The following journal article is accepted for publication:

- Sibeko, J. and Van Zaanen, M. 2023. Developing and testing Sesotho syllabification systems. *Accepted with minor changes by Journal of European Language Resources Association*, pending final review.

The following conference papers are delivered and are based on work in this thesis:

- Sibeko, J. 2023. Using classical readability formulas to measure text readability in Sesotho. *Selected papers from the CLARIN Annual Conference 2022*. In Tomaž Erjavec and Maria Eskevich, eds. pages 120–132. *Linköping Electronic Conference Proceedings*. <https://doi.org/10.3384/ecp198012>.
- Sibeko, J. and De Clercq, O. 2023. A Corpus-Based List of Frequently Used Words in Sesotho. In *Proceedings of the Fourth Workshop for Resources for African Indigenous Language*, May 6, pages 32–41, Dubrovnik, Croatia. Association of Computational Linguistics, New Brunswick, New Jersey, USA.

- Sibeko, J. and Setaka, M. 2023. Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Language (RAIL 2023)*, May 6, pages 76–85, Dubrovnik, Croatia. Association for Computational Linguistics, New Brunswick, New Jersey, USA.
- Sibeko, J. and Van Zaanen, M. 2022. Developing a text readability system for Sesotho based on classical readability metrics. *Digital Humanities 2022: Responding to Asian diversity*, July, pages 571–572, Tokyo, Japan. Short Paper.
- Sibeko, J. 2022. Using Classical Readability Formulas to Measure Text Readability in Sesotho. *CLARIN Annual Conference Proceedings*. PhD Student papers, 10–12 October 2022, Prague, Czechia.

The following basic language resources are developed and made publicly available as part of this thesis. Note that the syllabification resource contains two types of syllabification systems:

- Sibeko, J. and Van Zaanen, M. 2022. Sesotho syllabification systems. *South African Centre for Digital Language Resources*.
Available at: <https://repo.sadilar.org/handle/20.500.12185/555>.
- Sibeko, J. and Van Zaanen, M. 2022. Raw and syllabified word list for Sesotho. *South African Centre for Digital Language Resources*.
Available at: <https://repo.sadilar.org/handle/20.500.12185/556>.
- Sibeko, J. and Van Zaanen, M. 2022. Final year high school examination texts of South African Home and First Additional language subjects. *South African Centre for Digital Language Resources*.
Available at: <https://repo.sadilar.org/handle/20.500.12185/568>.

The following journal papers are submitted:

- Sibeko, J. 2023. Measuring the readability of Sesotho grade 12 exam texts. *Submitted*.
- Sibeko, J. and Van Zaanen, M. 2023. Measuring text readability in lower-resourced languages: A systematic literature review. *Submitted*.

Contents

Authorship declarations	i
Acknowledgements	ii
Dedications	iii
English abstract	iv
Sesotho abstract	vi
Abbreviations	viii
Related publications	x
1 Introduction	1
1.1 Introduction	2
1.2 Background	3
1.3 Problem statement	4
1.4 Research questions	5
1.5 Scope of the thesis	6
1.6 Relevance of the thesis	6
1.7 Thesis structure	8
1.8 Academic integrity	10
1.9 Summary	10
2 Methodology, reviews, and resource survey	11
2.1 Introduction	12

2.2	Background on Sesotho	13
2.2.1	Speakers of Sesotho	13
2.2.2	The history of writing in Sesotho	15
2.2.3	Sesotho orthography	15
2.2.4	Lesotho versus South African orthographies	16
2.2.5	Summary	17
2.3	Measuring text readability in lower-resourced languages: A systemic literature review	18
2.3.1	Introduction	18
2.3.2	Method	19
2.3.3	Results	24
2.3.4	Discussion and conclusions	35
2.4	Overview of English readability measures	36
2.4.1	Introduction	36
2.4.2	Syllable-based measures	36
2.4.3	Word-length-based measures	39
2.4.4	Frequency-list-based measure	41
2.4.5	Example use of readability formulas	42
2.4.6	Summary	42
2.5	An overview of resources and tools available to Sesotho	43
2.5.1	Introduction	43
2.5.2	Literature review	44
2.5.3	Methodology	46
2.5.4	Findings	47
2.5.5	Discussion and conclusions	53
2.6	Summary	54

3	Developing digital language resources	56
3.1	Introduction	57
3.2	Corpus 1: Final year reading comprehension and summary writing examination texts	59
3.2.1	Introduction	59
3.2.2	Methods	61
3.2.3	Data description	61
3.2.4	Summary	63
3.3	Corpus 2: A list of common words in Sesotho	63
3.3.1	Introduction	64
3.3.2	Background	65
3.3.3	Methodology	67
3.3.4	Results and discussion	69
3.3.5	Conclusions	72
3.4	Corpus 3: Machine translation texts	73
3.4.1	Introduction	73
3.4.2	Method	74
3.4.3	Results	78
3.4.4	Discussion and conclusions	80
3.5	Developing Sesotho syllabification systems	82
3.5.1	Introduction	82
3.5.2	Methodology	83
3.5.3	Results	90
3.5.4	Discussion and conclusions	90
3.6	Summary	93

4	Adapting readability measures to Sesotho	94
4.1	Introduction	95
4.2	Using traditional readability measures to evaluate English grade twelve examination texts	97
4.2.1	Introduction	97
4.2.2	Background	98
4.2.3	Methodology	99
4.2.4	Results	100
4.2.5	Discussion and conclusions	104
4.3	Assessing Sesotho examination text readability with English traditional measures	106
4.3.1	Introduction	106
4.3.2	Methodology	107
4.3.3	Results	108
4.3.4	Discussion and conclusions	111
4.4	Adapting traditional readability measures into Sesotho	112
4.4.1	Introduction	112
4.4.2	Background	113
4.4.3	Methodology	114
4.4.4	Evaluation	121
4.4.5	Discussion and conclusions	122
4.5	Summary	124
5	General discussion	125
5.1	Introduction	126
5.2	Research assumptions	126
5.3	Methodology overview	127

5.4	Comparison to previous research	127
5.4.1	Literature review summary	128
5.4.2	Comparison to previous research	128
5.4.3	Contribution to the field	130
5.5	Strengths and weaknesses	130
5.5.1	Critical evaluation of research design	130
5.5.2	Methodological strengths and weaknesses	131
5.5.3	Data analysis strengths and weaknesses	131
5.6	Future directions	132
5.6.1	Unanswered questions	132
5.6.2	Future research directions	133
5.7	Summary	134
6	Summary and conclusions	135
6.1	Introduction	136
6.2	Summary of findings	136
6.3	Implications of research	140
6.4	Summary	142
	References	143
A	Common Sesotho words	181
B	Ethics approval letters of study	198
C	Ghent University invitation	203
D	Permissions to use co-authored publications	206

List of Figures

2.1	Overview of countries in Southern Africa in which Sesotho has official language status.	14
2.2	PRISMA graph depicting the workflow used in this research.	20
2.3	Publications on text readability measures in the sample published over time.	23
2.4	Spread of sources in the sample according to source type.	24
2.5	Distribution of textual feature groups used in studies.	31
2.6	Overview of the distribution of classifiers commonly used.	33
2.7	The readability statistics are readily available on Microsoft Word.	38
3.1	The LT3 readability end-user interface.	77
3.2	The distribution of machine translation evaluation scores.	78
3.3	The distribution of readability scores for the machine translated (mt) and post-edited data sets.	80
4.1	Data distributions box plots for all of the measures, separated by the HL and FAL texts.	100
4.2	Data distribution density plots for all of the measures, separated by the English HL and FAL texts.	103
4.3	Data distribution Q-Q plots for all measures.	104
4.4	The distribution of readability levels depicted in box plots.	109
6.1	The integration of language resources and their relevance to the study. . . .	137

List of Tables

2.1	Search terms used in “Publish or Perish” using the Google Scholar database and the corresponding results obtained.	21
2.2	Overview of corpora sizes by the number of documents considered in the different studies.	29
2.3	Overview of the number of textual features considered in the different studies.	30
2.4	Example application of readability measures and their outputs.	42
2.5	The results of our search queries.	46
2.6	Translated word lists available in Sesotho.	48
3.1	The distribution of texts per language and subject level.	60
3.2	Token and type count per language and subject level for the full exam texts.	62
3.3	Token and type count per language and subject level for the extracted reading comprehension and summary writing texts.	62
3.4	The number of tokens and sentences present in the three Sesotho corpora used for this section.	67
3.5	An overview of Sesotho source texts, machine translations, and human-post-edited machine translations.	76
3.6	Examples of Sesotho source text (ST) and English target text (TT) sentences.	77
3.7	Pearson correlations between the different scores for BLEU, RIBES, and NIST.	79
3.8	Results of the <i>t</i> -tests for the machine translated and the post-edited texts indicating the means, standard deviations and <i>p</i> -values.	81
3.9	Rules of syllabification examples.	87

3.10	Total number of words, number of erroneously syllabified words, and accuracy for both pattern-based and rule-based systems for each of the folds as well as their means and standard deviations.	90
3.11	The different types of errors in the rule-based system output. Here WL indicates the wrong location, MB indicates a missing boundary, NE indicates a non-existent word, and LW indicates a loan word.	91
4.1	Mean and standard deviation (SD) scores for English examination text properties used in the measures.	101
4.2	Means and standard deviations (SD) of the different readability measures for the English HL and FAL examination texts.	102
4.3	Correlation between the different readability formulas. All correlations are significant ($p < .0001$).	102
4.4	ANOVA results. <i>Difference</i> indicates the estimated difference between the English HL and FAL values. <i>Error</i> indicates the standard error. $F(1,23)$ provides the results of the F -test with the corresponding p -values between the texts.	105
4.5	Mean scores and standard deviations (SD) for the properties of Sesotho texts used in the measures.	108
4.6	Means and standard deviations (SD) of the different readability measures for the Sesotho HL and FAL examination texts.	109
4.7	Correlation between the different readability formulas. All correlations are significant ($p < .0001$).	110
4.8	ANOVA results. <i>Difference</i> indicates the estimated difference between the Sesotho HL and FAL values. <i>Error</i> indicates the standard error. $F(1,75)$ provides the results of the F -test with the corresponding p -values.	111
4.9	Text features used in the traditional readability formulas adapted in this thesis.	116
4.10	The correlation of text features used in the readability measures computed from the Sesotho texts.	118
4.11	The correlation of syllable-based measures and syllable information computed on the Sesotho texts.	120
4.12	Readability measures and corresponding adapted Sesotho formulas.	121

4.13	Evaluations of the adapted linear regression models for Sesotho. Note that the p -values are all significant at $p < .001$	122
------	---	-----

Chapter 1

Introduction

This chapter serves as an introduction to the core topic of the thesis, which is the absence of a standardised objective yardstick for determining text readability levels in Sesotho (Southern Sotho). It aims to establish the significance of this thesis by providing background information that underscores the existing gap in the field. Moreover, the chapter addresses the research questions focusing on the need for an objective method for analysing text readability in Sesotho. To achieve this, the scope and relevance of the thesis are discussed. An overview of the thesis structure is provided.

1.1 Introduction

Text readability studies are premised on the assumption that easy-to-read and hard-to-read texts can be predicted (Janan, 2011). Although many definitions for text readability have been proposed, it is generally accepted that text readability is what “makes some texts easier to read than others” (DuBay, 2004).

Typically, we are inclined to rely on competent language users’ estimations of text readability. The scholarship of text readability encompasses both manual and automated procedures for analysing text readability levels (Van Oosten and Hoste, 2011). Even so, the accuracy, consistency, and comparability of language users’ manual assessments of text readability may be questionable for general use as their process of choosing texts is mainly subjective and intuitive (De Clercq and Hoste, 2016). First, the rating point of view and the rating standards vary depending on the evaluators. Second, the same piece of text could be given different ratings by the same evaluator if blindly re-analysed. Third, the evaluator’s judgement could be affected by subjectivity influenced by factors such as authors’ writing styles.

One of the main purposes of readability information is such that it can be used to match intended readers to texts of the desired levels of ease or difficulty (Meyer, 2003; Benjamin, 2012; Janan and Wray, 2012; Wray and Dahlia, 2013; Janan and Wray, 2014). Unarguably, subjectivity in choosing texts by following intuition poses some concerns. In the end, intuition-based approaches often result in a mismatch between the text and the intended reader (Zamanian and Heydari, 2012).

One solution to choosing texts and matching texts to readers is the development of a standardised objective yardstick for measuring text readability which can be used in choosing texts. In other words, having a standardised method for evaluating text readability will allow for better matching of texts to readers.

The field of linguistics has seen significant growth of interest in the scholarship of linguistic complexity (Miestamo, 2017). One of the focal areas has been matching texts to both learners and general language users (Van Oosten and Hoste, 2011; De Clercq et al., 2014). Such scholarship of text readability has continued for over a century (Collins-Thompson, 2014; De Clercq and Hoste, 2016). To date, more than 200 readability measures have been developed (DuBay, 2004).

The process of automatically labelling texts with levels of readability ease has been explored for an extensive number of years in high-resource languages (HRLs) (Gonzalez-Garduno and Sogaard, 2017; Antunes and Lopes, 2019b; Benzahra and Yvon, 2019) with skewed attention to some European languages (Tsvetkov, 2017; Hedderich et al., 2020). Even so, it remains a relevant and challenging topic (Collins-Thompson, 2014), especially for low-resourced languages (LRLs) (Zupon et al., 2021), such as indigenous African

languages which have been neglected in both linguistic complexity and text readability research.

We adopt Lignos et al.’s (2022) definition of LRLs, which recognises disparities in available digital language resources. In other words, we categorise languages as either a HRL or a LRL based on a comparative sorting strategy. In this approach, the status of a language is considered with respect to the language with which it is compared. Hence, the term ‘low’ refers to a language that may be deemed lower-resourced when compared to a language with more resources for language technology development. However, the same ‘low’ resourced language may be regarded as a HRL when compared to a language with even fewer resources. For instance, Arabic is classified as a LRL when compared to English (Zouhair, 2021), and a HRL when compared to Hindi (Tsvetkov, 2017).

We sort languages based on the adequacy of pre-trained models (Bansal et al., 2021), basic language resources, large enough corpora, and training data for performing Natural Language Processing (NLP) tasks with a focus on automatic text readability analysis (Cucu et al., 2014; Magueresse et al., 2020; Bansal et al., 2021; Zupon et al., 2021). Specifically, we consider the low-resourced Sesotho (Southern Sotho) and compare it to the high-resourced English. Note that Sesotho is a resource-scarce language (Mahloane and Trausan-Matu, 2015; Moors et al., 2018c; Wills et al., 2020) that remains deficient in digital language resources (Roux and Bosch, 2019).

1.2 Background

The concept of linguistic complexity is complicated (McWhorter, 2012) and difficult to define and measure (Comrie, 1992; Mikulecky, 2001; Ansaldo and Nordhoff, 2009). As such, multiple conceptualisations need to be considered when explaining it (Rescher, 1998; Sinnemäki, 2011). Nonetheless, we limit our discussion to the two conceptualisations of linguistic complexity, namely, the relative and the absolute approaches (Miestamo, 2006).

The relative approach to linguistic complexity seeks to answer a simple question of “to whom a language is or is not complex,” (Kusters, 2003). As a result, it is subjective and is often used interchangeably with text difficulty (Miestamo, 2008; Bulté and Housen, 2012). That is, it focuses on non-linguistic features such as the reader (Miestamo, 2006). Dahl (2004) recommends using ‘difficulty’ for the relative linguistic complexity.

On the other hand, absolute complexity is often used interchangeably with grammatical complexity (Berendes et al., 2018), which refers to the complexity of syntax and morphology of the text in question. That is, the absolute approach to linguistic complexity is objective as it focuses on the linguistic features and their relationship to each other (Miestamo, 2006, 2008, 2017).

Similar to linguistic complexity, text readability can also be viewed from a relative or

an absolute perspective. For instance, Bunch et al. (2014) and Collins-Thompson (2014) view text readability through a reading and text difficulty. From this perspective, text readability is measured through comprehension and is affected by the user of the text in that the extent to which a specific reader comprehends the text will affect the assigned level of text readability.

In the absolute approach, features of a language are quantified and the overall readability of a text is estimated. Modern objective measures for assessing text readability incorporate some features of the readability formulas and the more intricate characteristics of the text (Daelemans et al., 2017).

In this thesis, we explore text readability from the technical perspective which uses traditional formulas. Some traditional ways of using quantifiable linguistic characteristics are by use of syllable-based readability measures and word-length-based measures. These formulas do not take readers' special characteristics into account and are thus related to absolute measures of text readability. According to Bailin and Grafstein (2001), these formulas are characterised by quantification and ease of understandability.

We concentrate on nine traditional measures of readability that were developed for English, specifically the Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), and the Gunning Fog Index (GFI). These measures are based on syllable information. Additionally, we consider the Coleman-Liau Index (CLI), the Läsbarhetsindex (LIX), the Readability Index (RIX), and the Automatic Readability Index (ARI), which centre around word lengths. Furthermore, we examine the Dale-Chall Index (DCI), which utilizes a predetermined list of commonly used words.

1.3 Problem statement

Text readability evaluation has been applied to different domains, including assessing appropriate educational materials such as textbooks (Collins-Thompson, 2014). Unfortunately, textbooks from different publishers and authors may present texts of varied levels of difficulty for the same school grades. In this way, the absence of text readability measures poses a problem, as the use of textbooks can lead to deskilling of both language teachers and learners when textbooks are not effectively matched to their intended readers (Mohammed et al., 2023). For this reason, the development and use of appropriate readability measures become crucial in addressing the issue of readability in educational contexts.

Measuring the readability of texts depends on the measurement of language-specific properties. As such, readability measures are language dependent. Until now, prevailing literature on readability has focused on traditional readability measures devised for English.

Regrettably, research on South African indigenous languages (in general, but more specif-

ically for readability purposes) continues to be marginalised. The dearth of resources required for the development of readability measures tailored to these languages compounds the obstacles in conducting automatic text readability analysis. It is currently unclear whether similar approaches to readability measurement are optimally suited for gauging readability in South African indigenous languages (Leopeng, 2019).

To this point, studies on text readability in the South African contexts have limited themselves to using English traditional text readability formulas, see, for instance, Grabar et al. (2014); Joubert and Githinji (2014); Sibanda (2014); Buthelezi (2017); Leopeng (2019); De Wet (2021). The scores from these formulas are interpreted based on the United States schooling system. In Bargate’s (2012) research, the adaptations of the English readability scores are introduced as a method of contextualising them within the framework of South African school grades. Unfortunately, such contextualisation is limited in that the scores are specifically for English measures since indigenous languages do not have readability measures.

To address the lack of readability measures in South African indigenous languages, Leopeng (2019) suggests that the current readability measures should be adapted to these languages. Against this background, this thesis investigates how text readability can be measured for Sesotho. We skew our focus on how existing traditional measures can be adapted to Sesotho.

1.4 Research questions

This thesis rests upon three underlying assumptions: (i) Sesotho texts can be ranked according to different levels of readability, (ii) the adaptation of traditional readability measures can enable the assessment of Sesotho texts’ readability, and (iii) an automated process will standardise and simplify the process of ranking Sesotho texts according to their different levels of readability.

Against this background, the primary inquiry of this thesis revolves around the following overarching problem statement:

PS: How might one measure text readability in Sesotho?

To fully investigate and accurately resolve the problem statement above, the following research questions will be explored:

RQ 1: What methods, techniques, and approaches are used in measuring text readability in low-resourced languages?

RQ 2: What language resources are needed for measuring text readability in Sesotho?

RQ 3: What can we learn from existing traditional text readability measures?

RQ 4: How can existing English traditional readability measures be modified and adapted to Sesotho?

1.5 Scope of the thesis

This thesis aims to explore and develop readability measures for Sesotho, a South African indigenous LRL. This thesis will focus on the adaptation of traditional readability measures for Sesotho. The development of a method for analysing text readability in Sesotho will undoubtedly add to the development of the language. While this thesis focuses on Sesotho, related languages such as Sepedi and Setswana (Prah, 2007), may benefit from the results of this thesis as well.

The choice for developing measures for analysing text readability in Sesotho is based on (i) the expertise of the researcher on linguistic properties of Sesotho, (ii) the fact that Sesotho is one of the most used home languages of South Africa, and (iii) Sesotho was one of the first Southern African languages to be transcribed into writing (Motjope-Mokhali et al., 2020).

The research methodology involves a comprehensive review of existing literature on text readability analysis methodologies for LRLs, a survey of available resources for enabling the assessment of readability in Sesotho, data collection, creation of corpora, as well as the development of readability measures for Sesotho. However, the thesis will not present the creation of completely new readability measures, but will focus on adapting existing traditional measures to suit the linguistic characteristics of Sesotho.

We believe that having objective measurements of readability will be highly useful in educational contexts, where varying levels of text difficulty in reading texts can lead to either boosting experiences or deskilling of language teachers and learners. However, it is important to note that the thesis does not purport to align the readability measures with the educational standards of South African grade levels. Such endeavours have already been expounded upon in other scholarly works, as exemplified by the work of Bargate (2012, p. 11), Leopeng (2019), and Mankayi et al. (2023, p. 10). Additionally, the scope will not extend to the implementation of the readability measures on a large scale or the evaluation of its impact in educational settings. It will, however, provide a strong foundation for future studies to explore such practical applications.

1.6 Relevance of the thesis

Previously, researchers have used various methods to investigating text readability, including traditional readability measures (Kondru, 2006; Feng, 2010; Janan, 2011; Bendová, 2021), machine learning (Sjöholm, 2012; Andova, 2017), Natural Language Processing

(Gonzalez-Dios, 2016), deep learning (Alkaldi, 2022), and eye-tracking (Newbold, 2013). These research projects have made significant contributions to the scholarship of text readability.

This thesis follows former research that focused on the adaptation of existing readability measures to LRL using a machine learning approach, such as Bendová (2021). Unfortunately, readability research that utilises machine learning methods in LRLs, such as Sesotho, is restrained by the lack of training data. For this reason, the development of training data is vital for enabling machine learning approaches to the development of text readability measures in LRLs including Sesotho.

There is evidence of interest in text readability analyses in the South African context. For instance, Leopeng (2019) argued that the LIX formula is the best for measuring text readability in any language and consequently applied it to English, Afrikaans, and isiXhosa. Meanwhile, De Wet (2021) argued that the Fry Readability graph and the FRE formulas are the best and subsequently applied them in testing her suggested tool for measuring mental health recovery texts. Earlier, Joubert and Githinji (2014) contended that SMOG is the best measure in that it provides the best analyses of readability levels and subsequently applied it in their assessment of health pamphlets.

Nonetheless, previous research on text readability in South African indigenous languages like Sesotho (Krige and Reid, 2017; Reid et al., 2019), isiXhosa (Carel, 2019; Leopeng, 2019), and isiZulu (Land, 2015) have relied on the use of English readability measures that are not tailored for the indigenous languages. In fact, as far as could be ascertained for this thesis, none of the existing studies on text readability in the South African context has attempted to develop measures for text readability in the indigenous¹ languages. Thus far, there have been at least four adaptations of the FRE to the European-rooted Afrikaans². An overview of these formulas is provided in Jansen et al. (2017).

Like languages with a European origin, such as English and Afrikaans, in Southern Africa, texts written in Sesotho span a diverse range including, for instance, dictionaries, academic books for primary, secondary and tertiary education, magazines, newspapers, literature for leisure, and religious books (Webb et al., 2010). This is because Sesotho is used for day-to-day functions in both urban and rural areas (Moleleki, 1999).

¹We use the word indigenous cautiously in this instance pending the ongoing debates on the status of Afrikaans as an indigenous African language.

²These four formulas for Afrikaans include two formulas developed by Van Rooyen (1986), one presented in an unpublished conference paper presentation by McDermid Heyns (2007), and one presented in a poster presentation by McKellar (2008).

1.7 Thesis structure

The current chapter introduces and contextualises the thesis, highlights the need for an investigation of this nature and provides a general outline of the whole thesis. A brief discussion of the approach, chapter outlines, and ethics-related issues are also presented.

The rest of the thesis is organised according to the research questions posed in Section 1.4. In total, the thesis consists of six main chapters. The structure of each chapter and its contents are briefly outlined below.

Chapter 2 addresses the first research question on the methods, techniques, and approaches to text readability in LRLs. To achieve this, a brief background on Sesotho, the history of writing in Sesotho, a general overview of traditional readability measures, and reviews of methods and resources are provided. Typically, literature reviews highlight research outcomes, research methods, theories, practises and applications (Cooper, 1988). Furthermore, literature reviews introduce the terminology that is used in the thesis while also contextualising the thesis and building a bibliography (Rowley and Slack, 2004). In this thesis, the literature reviews also provide information on the research methods, practices and applications.

First, we conduct a systemic literature survey to investigate the methods, measures, resources and approaches to the evaluation of text readability in LRLs. The aim of this survey is to provide a generalised view of approaches to assessing text readability in LRLs. Second, following the generalised overview of approaches to text readability assessment in LRLs, attention is paid to Sesotho as an under-resourced language (Moors et al., 2018c; Wills et al., 2020).

The second survey, amongst other things, uncovers already existing tools that could aid the analyses of text readability in Sesotho and highlights the need for the resources that are subsequently developed to enable appropriate adaptation of existing readability measures to Sesotho. Although there are many ways of typifying Sesotho, in this thesis, Sesotho is viewed from the perspective of recent trends in the scholarship of text readability that classifies languages as either resourced or low-resourced (see for instance Islam et al. (2012) and Wang et al. (2019)).

Chapter 3 addresses the second research question on the needed language resources for measuring text readability in Sesotho. It does this by discussing the development of resources necessary for the adaptation of readability measures into Sesotho. Overall, we delve into the creation and curation of two corpora, two lexicons, and two modules, which subsequently form the foundation for the construction of readability metrics tailored to Sesotho.

Chapter 4 addresses two research questions. First, we consider what can be learned from existing English readability measures by evaluating the use of traditional readability

measures on South African school texts. In this thesis, we use nine traditional readability measures focusing on varied text properties. We pay special attention to high school exit grade twelve examination reading comprehension texts. The different articles forming the basis of the chapter are briefly outlined below:

The first article investigates the differences in the readability of texts at the home and the first additional language levels of the English grade twelve comprehension texts. The second article used the rule-based syllabification system, the list of frequently used words in Sesotho, and the exam text corpora to evaluate the readability of texts in Sesotho using English traditional readability measures. The aim of the studies is to ascertain whether the levels of the readability of the examination texts are differentiated and whether they would be justifiably used in the development of readability measures for Sesotho.

Chapter 4 also addresses the question of how one might go about measuring text readability in Sesotho using the available resources (including both the pre-existing resources and those that were developed as part of this thesis). In this chapter, the nine traditional readability measures are adapted to Sesotho and new adapted formulas are presented.

Prominent recent tools, such as the Coh-Metrix³ (McNamara et al., 2014), the LCA⁴ (Ai and Lu, 2010), the L2SCA⁵ (Lu and Ai, 2015), and the LT3⁶, have successfully utilised the web as a platform to expand access for linguistic complexity and text readability assessment. Building on this trajectory, our readability measures are poised to similarly encourage widespread use. By facilitating effortless accessibility and delivering rapid results, our measures aim to cater to an array of users, including educators, evaluators, learners, education stakeholders, and the general public. To ensure ease of access of the developed Sesotho measures, we are planning to make our readability measures available as an online tool.

Chapter 5 provides a discussion of the results and findings of this thesis. It also compares the research in the current thesis to previous research and explains the specific contribution of this thesis to the field of measuring text readability in LRLs. Furthermore, the chapter also highlights the strengths and weaknesses of this thesis and presents proposed future research.

Chapter 6 concludes the thesis by summarising the thesis findings. It emphasises the implications of the thesis findings to practical applications. Furthermore, it discusses the limitations of the research. Additionally, the chapter provides an answer to how one might go about measuring text readability in Sesotho.

³Available at: <http://tool.cohmetrix.com/>.

⁴Available at: <https://aihaiyang.com/software/lca/>.

⁵Available at: <https://aihaiyang.com/software/l2sca/>.

⁶Available at: <https://lt3.ugent.be/readability-demo/>.

1.8 Academic integrity

This thesis is cleared for ethics by the North-West University Language Matters Research Ethics Committee with ethical clearance number (NWU-00729-21-A7). See Appendix A for the clearance certificate. No participants are involved in this thesis. Instead, the research in this thesis uses publicly accessible question papers, corpora, applications, and an open-access dictionary corpora.

For all the co-authored sections, I am the first and main author and have been responsible and involved in the full process, including conceptualisation, methodology, data collection, data pre-processing, data analysis, writing, and revising. Professor Menno van Zaanen, as my promoter, has aided me in the conceptualisation and has consistently reviewed my writing. For the wide range of methodologies and analyses used within this thesis, I received advice from experts in the field. These experts and hence co-authors have supported me with methodology (for instance in Chapters 3 and 4, Professor Orpheé de Clercq of Ghent University, and Miss Mmasibidi Setaka of the South African Centre for Digital Language Resources).

1.9 Summary

Unlike HRLs, such as some European languages, which benefit from a plethora of text readability measures (Benjamin, 2012; De Clercq et al., 2014), indigenous African languages, including South African indigenous languages, remain understudied in the context of text readability.

This thesis is motivated by the lack of readability measures specifically developed for South African indigenous languages. As such, this research seeks to investigate ways to adapt readability measures to the context of Sesotho, one of the indigenous languages of South Africa.

Throughout this chapter, we provide an introduction to the thesis by outlining the background, identifying the research gap, and presenting the research problem, questions, and scope of the thesis. Our main objective is to offer insights into the methods used in assessing readability in LRLs, explore the existing and needed resources for developing readability measures in Sesotho, and leverage findings from English readability measures to apply them effectively to Sesotho texts.

This research is significant as it contributes to the broader field of language teaching, education, and linguistic research by addressing the scarcity of readability measures tailored to indigenous African languages. By focusing on Sesotho, our findings are likely to have implications for other LRLs in the Southern African region.

Chapter 2

Methodology, reviews, and resource survey

This chapter provides the background of Sesotho as a written language, the strategies used in the measurement of text readability in LRLs, available digital language resources available to Sesotho, and the nine traditional readability measures explored in this thesis.

2.1 Introduction

This chapter builds on the foundation established in the previous chapter, by addressing the first research question. That is, the chapter provides a detailed analysis of the different methods and techniques available for measuring text readability in LRLs by drawing on case studies and examples from various languages beyond Sesotho.

To structure the discussion, the chapter is divided into four sections. First, the background of writing and the orthographies of Sesotho is provided in Section 2.2. Here we also provide context through a short description of information on the speakers of Sesotho.

Second, we provide a systematic review of the literature on text readability for LRLs and automated methods of text readability analyses in Section 2.3. The goal of this review is to survey methodologies employed in analysing text readability in LRLs. Our literature review is centred on 71 publications on readability measures in such languages. We discuss machine learning and cross-lingual approaches to measuring text readability. We also explore the adaptation of English measures into measures for LRL. The findings indicate that text readability measures depend on some form of gold-standard corpora. In particular, educational texts with school grades and other pre-annotated texts are preferred for such gold-standard corpora. Unfortunately, corpora sizes are not prescribed. Similarly, the number of and the types of textual features considered in different studies are inconsistent and range between three and 406 features. Additionally, although Support Vector Machine (SVM) classifiers are frequently used, we suggest exploring a combination of classifiers to identify one that yields the highest predictive accuracy.

Third, a brief outline of the nine English traditional readability measures investigated in this thesis is provided in Section 2.4. These are well-known measures that have been developed for English specifically. In a later chapter, we will use the structure of these measures as the basis for the development of readability measures for Sesotho.

Finally, we provide an overview of digital language resources available for Sesotho in Section 2.5. The South African Center for Digital Language Resources (SADiLaR) repository is used as a reference as it is the official host of various language resources for South African languages. A total of 77 resources are identified. Among other things, we observed that the majority of applications and modules available for Sesotho are in fact general resources aimed at all eleven¹ official South African languages. The available resources indicate an inclination to the development of entry-level, basic language resources and an absence of middle and higher resources with functionalities such as semantic analyses for written resources and prosody prediction for spoken resources. The investigation is hindered by the dearth of resource-specific evaluations and related research and is further challenged by the absence of some of the resources in the repository.

¹In July 2023, South African sign language has been recognized as the twelfth official language in South Africa.

Through this chapter, we aim to uncover frequently used methods in measuring text readability in LRLs. Furthermore, we aim to contribute to a better understanding of the unique opportunities for and challenges facing under-resourced languages such as Sesotho. The discussion of available resources for Sesotho and the discussion of traditional readability measures highlight some of the challenges in LRLs such as the lack of corpora and basic language resources for extracting text properties.

Beyond this thesis, we hope that these reviews will facilitate further research and development of resources in text readability measures, ultimately leading to improved access to information for people in under-resourced language communities such as those speaking Sesotho. The chapter concludes with a summary of the main findings and contributions, as well as recommendations for future research in this area.

2.2 Background on Sesotho

This section briefly contextualises Sesotho by discussing the speakers, the history of writing and the different orthographies of Sesotho.

2.2.1 Speakers of Sesotho

Sesotho is used by more than ten million speakers in a few countries in Southern Africa (Marupi and Charamba, 2022; Sibeko and Setaka, 2022). In fact, it is one of the official languages in Lesotho, Zimbabwe, and South Africa (Lesotho, 1993; Republic of South Africa, 1996; Ndlovu, 2011, 2013; Kadenge and Mugari, 2015; Wissing and Roux, 2017). The geographic location of the three countries where Sesotho has official status is presented in Figure 2.1. The 2021 census of South Africa revealed Sesotho as the seventh most common home language with 3,849,563 speakers out of the 50,961,443 people reached by the census data collection team (Lehohla, 2011). Furthermore, a total of 99.7% of Lesotho's population speaks Sesotho as a home language (Kamwangamalu, 2013). Unfortunately, we could not locate the language statistics of Sesotho users in Zimbabwe where Sesotho is one of the marginalised languages (Ndhlovu, 2012; Nyika, 2012; Marupi and Charamba, 2022). Nonetheless, the number of Sesotho users is seemingly high. It is hoped then that this research will impact a relatively large group of marginalised language users.

Linguistically, Sesotho forms part of the bigger Sotho group composing Southern Sotho (Sesotho), Northern Sotho (Sepedi), and Western Sotho (Setswana) (Makoni, 2003; Van Heerden et al., 2010; Nkolola-Wakumelol et al., 2012; Riep, 2013; Mojela, 2016). Sesotho is used as a language for learning and teaching as either a mother tongue, second language, or marginalised language. It is also used for media, political, religious, and other uses. Even so, a recent investigation of the Sesotho Basic Language Resource Kit (BLARK) content has revealed a severe shortage of digital language resources available for Sesotho

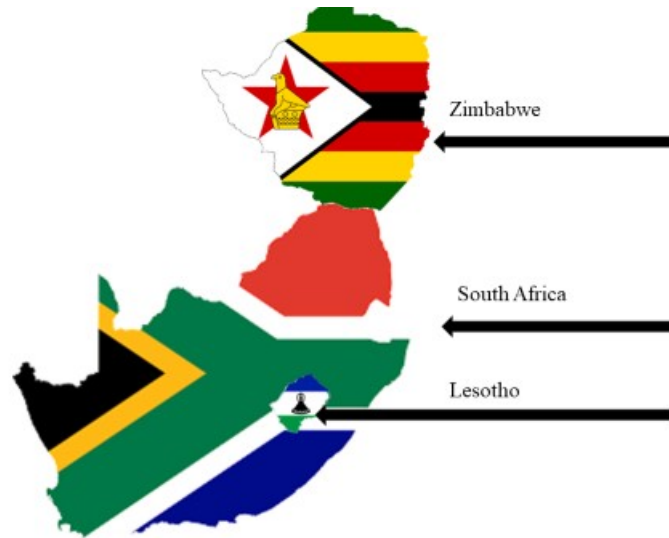


Figure 2.1: Overview of countries in Southern Africa in which Sesotho has official language status.

(Sibeko and Setaka, 2022). As such, Sesotho remains an LRL (Roux and Bosch, 2019; Sibeko and Setaka, 2022). Unfortunately, automating the process of objectively investigating text readability in Sesotho using traditional readability measures is hindered by the lack of necessary primary language resources needed in more advanced traditional readability measures that consider more intricate text properties above the level of word and sentence complexities as measured by the number of letters.

In addition to the lack of necessary basic language resources which hinders the development of objective automated measures for measuring text readability in Sesotho, there are two widely recognised orthographies for Sesotho. The two orthographies are differentiated by the two countries of their respective origins. They have therefore been labelled accordingly as the Lesotho Sesotho (LS) orthography and the South African Sesotho (SAS) orthography (Mohasi and Mashao, 2005; Motjope-Mokhali et al., 2020). Although we are unaware of the literature on Zimbabwean Sesotho, there are at least minor differences in the orthography and vocabularies of Lesotho Sesotho and South African Sesotho (Motjope-Mokhali et al., 2020) as well as in the syntax (Sekere, 2004) of the two Sesotho variants. Furthermore, South Africa and Lesotho each have their own rules for spelling words in Sesotho (Matlosa, 2017; Bardill and Cobbe, 2019). The main differences between the two orthographies are discussed elsewhere (for instance, see Eldredge (2002), Makutoane (2022), and Sibeko (2022b)).

2.2.2 The history of writing in Sesotho

The earliest known Sesotho writing dates back to the 19th century when European Christian missionaries, such as Thomas Arbousset and Eugène Casalis, translated the Bible into Sesotho (formerly spelt Sesuto). These religious translations were instrumental in the spread of Christianity in the region and the development of the Sesotho language especially as a written language.

During the pre-colonial period, Sesotho was used primarily for oral communication, and the language was not widely written. However, in the 20th century, Sesotho began to be used more widely in written form, particularly in literature and journalism. In Lesotho, the establishment of the Morija Printing Works in 1833, which published Sesotho books, newspapers, and pamphlets, played a significant role in developing Sesotho's written literature.

According to Selepe (1997), literature in Sesotho developed over three periods including (i) the 1900s predominated by Lesotho authors focusing on religious and educational texts, (ii) the 1930s when the Basotho authors from South Africa joined the writing industry, and (iii) the post-1960s where issues other than religion, education, and the economy were written about. According to Zulu (2003), this focus on ordinary topics is clear in prose texts between 1990 and 2002. Sesotho literature gained prominence in the mid-20th century, particularly with the emergence of writers such as Thomas Mofolo, who wrote the first novel in Sesotho, "Chaka", in 1925. Other notable Sesotho writers include K. P. Dinoko, who wrote poetry and plays, and S. E. K. Mqhaya, who adapted and translated Shakespeare's works into Sesotho.

2.2.3 Sesotho orthography

Historically, South African Sesotho orthography used 22 letters from the 26 letters of the Latin alphabet: (a, b, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, w, and y) (in both upper and lower case). Recently, however, as can be observed from the Sesotho International Dictionary (Chitja, 2010), the remaining letters (c, v, x, z) are now also used in both LS and SAS.

As a tonal language, Sesotho carries tone by vowels and nasal consonants (Guma, 1982; Sekere, 2004; Mohasi et al., 2011). Seven phonological vowels (a, e, i, o, u, ε, and ɔ) (Dickens, 1978; Matlosa, 2017) are accepted across the Sotho language group (Barnard and Wissing, 2008; Matlosa, 2017). However, SAS orthography does not explicitly indicate all seven phonological vowels (Anghelescu, 2016). To this end, we limit our vowel discussions to morphological vowels (a, e, i, o, and u) as these are indicated in writing (note that the aim of this thesis is to evaluate the readability of written texts and not spoken or multimedia texts). In addition to the vowels, special attention is also paid to the semi-

vowels: (w and y), the lateral consonant: (l), the glide: (g), and the four nasal consonants: (m, n, ŋ, and ɲ) as used in the SAS orthography (Demuth, 2007; Nkolola-Wakumelol et al., 2012; Anghelescu, 2016; Chokoe, 2020).

2.2.4 Lesotho versus South African orthographies

Some of the differences between the LS and the SAS orthographies affect the evaluation of text readability using traditional readability measures. One such difference affects the syllable counts. For instance, the use of (e) in the LS orthography for representing the semivowel (y) and the use of (o) for representing the semivowel (w) results in exaggerated syllable counts when Lesotho orthography is used. See the illustration in Example 2.1².

(2.1) *Ke wena le yena. (Six syllables) — SAS.*
 It's you and her.
 'It's you and her.'

Ke oena le eena. (Eight syllables) — LS.
 It's you and her.
 'It's you and her.'

The lengths of words can also be affected by the differences in orthographies. That is, SAS orthography uses longer words through the use of digraphs in places where the LS orthography uses single letters. For instance, the SAS orthography uses the digraph (tj) where the LS orthography also uses (c). The differences are exemplified in Examples 2.2 and 2.3.

(2.2) *Se a tjha. (Seven letters) — SAS*
 It is burning.
 'It is burning.'

Se a cha. (Six letters) — LS
 It is burning.
 'It is burning.'

In Example 2.3, the SAS orthography uses the (sh) digraph while the LS orthography uses the single accented (š) letter. Note that the South African Sesotho orthography does not contain any words with accents.

(2.3) *Tshepa tshepiso. (14 letters) — SAS*
 Trust promise.
 'Trust the promise.'

²The discussion of syllable extraction is presented in Chapter 3.

Tšepa tšepiso. (12 letters) — LS
 Trust promise.
 ‘Trust the promise.’

Some of the differences in orthography do not affect the readability of texts. For instance, the LS orthography uses (l) in place of (d). In this way, neither word length nor syllable counts are affected. The differences are exemplified in Example 2.4.

(2.4) *O dula mona. — SAS*
 He/she stays here.
 ‘He stays here.’

O lula mona. — LS
 He/she stays here.
 ‘He stays here.’

Although differences such as preferences for certain single letters such as (l) and (d) do not affect the results of the measures, the different representations of the semivowels will affect syllable information. Furthermore, the use of single letters in one orthography and the use of digraphs in another orthography may affect properties like average word lengths.

Beyond orthography, there are also region-based vocabulary variations in Sesotho. For instance, see Lemeko (2018) for a discussion of region-based variations of SAS. Nonetheless, these variations are not within the scope of the current project. The research focus is limited to the effects of orthographic conventions. Differences between SAS and LS orthographies are discussed in more detail in studies such as Demuth (2007) and Matlosa (2017).

2.2.5 Summary

It is clear that the differences between the orthographies of the South African Sesotho and the Lesotho Sesotho necessitate a decision to adopt either one of the orthographies or accommodate both, but they cannot be used interchangeably, especially for the purposes outlined in this thesis. Therefore, the preferable choice in this context is to utilize the South African orthography given our intended use of South African reading comprehension and summary writing exam texts.

The brief history of writing in Sesotho may be responsible for the lack of freely accessible textual data in Sesotho. Unfortunately, even the copyrighted textual data is inaccessible and the publishing houses remain unresponsive to requests to use the texts for research.

2.3 Measuring text readability in lower-resourced languages: A systemic literature review

This section is adapted from: Sibeko, J. and van Zaanen, M., 2023. Measuring text readability in lower-resourced languages: A systemic literature review. *Manuscript under consideration for publication, submitted 20 July 2023*.

2.3.1 Introduction

This section sets out to answer the following research question: *How can text readability measures be developed for low-resource languages?* To answer this question, we first look at definitions of text readability and LRLs. Next, we present a systematic literature review that surveys methods that are used in developing text readability measures for LRLs.

There is an adequacy of surveys on text readability from different perspectives, including features for readability assessment (Feng et al., 2010), the contribution of artificial intelligence (François, 2015), recent developments (Benjamin, 2012), current and future trends (Collins-Thompson, 2014), the United Kingdom educational context (Wray and Janan, 2013), English Foreign Language perspectives (Greenfield, 2004; Zamanian and Heydari, 2012), and fluctuations of the use of ‘readability’ in academic literature (Solnyshkina et al., 2020). However, to our knowledge, no state-of-the-art overview of text readability from the perspective of LRLs exists.

2.3.1.1 Defining text readability

Authors often define text readability in their own ways, which allows them to provide definitions that suit their intentions (Van Oosten et al., 2010). As a result, text readability is defined and conceptualised in various ways (Benjamin, 2012). Unfortunately, most definitions tend to be vague, which makes readability difficult to research (O’Brien, 2010; Phani et al., 2019). Even so, it is generally accepted that readability refers to “what makes some texts easier to read than others” (DuBay, 2004).

Text readability perspectives can be typified into cognitive and traditional approaches (Zifar and Namaziandost, 2019). Cognitive approaches encompass the scholarship of text readability that depends on readers’ characteristics. In this sense, text readability is associated with text complexity (Pires et al., 2017; Frantz et al., 2015; Amendum et al., 2018), reading difficulty (Collins-Thompson and Callan, 2005; Heilman et al., 2008; Newbold and Gillam, 2010; Collins-Thompson, 2014; Prasanna and Vaishnavh, 2014; Soh, 2019), and understandability (Stenner and Stone, 2004; Buse and Weimer, 2008; Islam et al., 2012; Razon and Barnden, 2015; Paterson and Hartmann, 2016; Plavén-Sigray et al., 2017; Soh, 2019). In this manner, the reader, the text and the author are seen as contributing to the

subjectivity of text readability (Newbold, 2013). Additionally, legibility may also be considered in this perspective of text readability as it affects cognition (Smith, 2016; Battisti et al., 2020).

Traditional approaches to readability, in contrast, deal with analyses using quantitative parameters, such as word and sentence complexity, with disregard for reader qualities (Tongtep et al., 2014; Solnyshkina et al., 2018a). In this context, quantitative parameters are combined into traditional text readability measures to measure text readability (Hiebert and Mesmer, 2013). Traditional readability measures are mainly used to assess the extent to which reading materials are suitable for readers at a certain grade or level (Woodsend and Lapata, 2011; Balakrishna and Meurers, 2014; Soh, 2019). Readability measures are essentially mathematical formulas used to analyse readability obtained through, for instance, regression analyses (Greenfield, 2004; Heilman et al., 2008; Mandic et al., 2012) and empirical correlations (Crossley et al., 2007; Lissón, 2017). In these methods, text properties such as word length, syllable count (Si and Callan, 2001), and sentence length (Stahl, 2003; Antunes and Lopes, 2019b), the ratio of simple and difficult terms, together with constant weights (Van Oosten et al., 2010), form the basis for measuring text readability (Azpiazu and Pera, 2020). By focusing on quantitative parameters, these readability measures ignore the semantic content of the text (Curto et al., 2014). Several traditional readability measures are also referred to as classical as they have been around for a while and are considered “standard”. For instance, these classical traditional measures have served as the basis for adaptation into LRLs. See Section 2.4 for an overview of traditional readability measures.

In this section, text readability is conceptualised from the quantitative parameters point of view with a special focus on traditional text readability measures. Furthermore, texts are weighed through comparison with other texts and not authors, readers or legibility (Amendum et al., 2018).

2.3.2 Method

To identify suitable studies for analysis, we follow the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) protocols³ (Moher et al., 2015; Shamseer et al., 2015; Moher et al., 2016). According to Page et al. (2021), the PRISMA statement was intended to enable systematic reviewers to provide clear and transparent explanations regarding the purpose of the review, the actions taken by the authors, and the discoveries made during the process. In this way, a PRISMA literature review employs explicit and systematic methods to collate all relevant evidence that fits pre-specified eligibility criteria, aiming to answer a specific research question while minimizing bias in the identification, selection, synthesis, and summary of studies, ultimately providing reliable findings from which conclusions can be drawn and decisions made (Moher et al., 2015). Although

³See <https://www.prisma-statement.org/>.

primarily intended for health-related research (Moher et al., 2009), the PRISMA method has undergone many adjustments for different purposes (Page et al., 2021). As such, it has been employed in other subjects such as natural language processing (Alahyari et al., 2022; Valera and Luštrek, 2022), hate speech (Jahan and Oussalah, 2023), and school bullying (Xie et al., 2023). The PRISMA process involves multiple steps such as identification, screening, and eligibility determination, leading to a comprehensive examination of a collection of studies.

For our purpose, we adapt the flow diagram for systematic reviews which include searches of databases and registers⁴. First, we do not consult any registers. Second, we report the results of the eligibility screening from the initial screening stage. Figure 2.2 presents the flow chart of the research process and the totals for the searches following an adaptation of the PRISMA model (Moher et al., 2009).

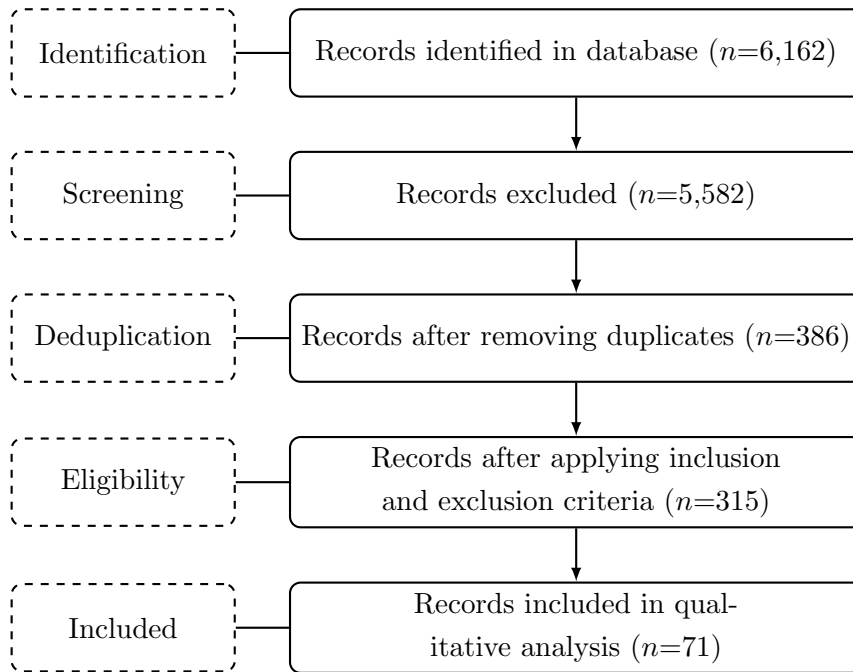


Figure 2.2: PRISMA graph depicting the workflow used in this research.

The remainder of this section describes the information sources used in this section, followed by a description of the exclusion and eligibility criteria and the data extraction process.

⁴See <https://www.prisma-statement.org/PRISMAStatement/FlowDiagram> for the PRISMA flow diagram.

2.3.2.1 Information sources

Using the Microsoft Windows desktop version of the “Publish or Perish” application⁵ (Harzing, 2007), the Google Scholar database⁶ was searched. This application returns a maximum of 1,000 results for each search query⁷. We performed searches using eleven search term variations to identify publications published between the years 2002 and 2022. A combined total of 6,162 results were returned from the eleven search queries. The results for each search query are presented in Table 2.1.

Table 2.1: Search terms used in “Publish or Perish” using the Google Scholar database and the corresponding results obtained.

Search terms	# of results
text + readability + measures	980
“text readability” + “metrics”	820
“text readability” + “lower-resource”	121
“text readability” + “lower-resource” + “cross-lingual”	28
“text readability” + “african” + “cross-lingual”	10
“text readability” + “cross-lingual”	1,000
“text readability” + “marginalised”	27
“text” + “marginalised” + “readability”	982
“text” + “readability” + “metric”	980
“readability” + “less” + “resourced”	989
“readability” + “less-resourced” and “readability” + “less resourced” ^a	225
Total	6,162

^aBoth searches resulted in exactly the same publications.

2.3.2.2 Eligibility and exclusion criteria

From the sources that were identified using the Google Scholar searches described above, we performed a manual selection process. In particular, we were interested in publications on:

- (i) the development of text readability measures in languages other than English,
- (ii) adaptation of English measures into LRLs, and
- (iii) cross-lingual approaches.

As such, we included all publications on these topics.

⁵Available at: <https://harzing.com/resources/publish-or-perish>.

⁶See <https://scholar.google.com/>.

⁷We enabled the system option to ignore citations.

Our comparative sorting strategy recognises English as an upper-level HRL (Zamanian and Heydari, 2012; Tsvetkov, 2017). To this end, we exclude all publications focusing on English (only) text readability. Furthermore, although we consider publications focusing on the adaptation of English measures to LRLs, we exclude studies that customise the Coh-Metrix, since it is primarily used for HRLs which can process text features such as cohesivity, narrativity and semantic relations (Graesser et al., 2004; De Almeida and Aluísio, 2009; McNamara et al., 2010b; Scarton and Alusio, 2010; Graesser et al., 2011; McNamara et al., 2011; McNamara and Graesser, 2012; Graesser et al., 2014; McNamara et al., 2014). Methods that require language-specific analyses such as those employed in the Coh-Metrix approach fall outside the scope of this research as LRLs typically do not have suitable analysis methods available.

Within the identified studies, we recognised two themes closely related to text readability, namely, summarisation and simplification. First, we reviewed sources on text summarisation, which investigates textual features inclusive of grammar, redundancy, clarity, structure, and cohesion (Pitler et al., 2010; Vadlapudi and Katragadda, 2010a,b; Lloret and Palomar, 2012; Kryściński et al., 2018; Verma and Om, 2019; Verma and Verma, 2020; El-Kassas et al., 2021). Unfortunately, none of the studies on text summarisation described the development of measures for measuring text readability. As a result, all sources on text summarisation were excluded.

Second, we evaluated sources on text simplification, which refers to the process of reducing text complexity (Balakrishna and Meurers, 2014; Al-Thanyyan and Azmi, 2021) and thereby improving text readability and intelligibility without affecting its content and meaning (Wubben et al., 2012; Banerjee et al., 2013; Shardlow, 2014; Sai et al., 2019). In this way, automatic text simplification relies on readability measures for identifying and evaluating targets for simplification (Sowmya and Detmar, 2014; Lasecki et al., 2015; Brunato et al., 2016). Although text simplification is one of the two main applications of text readability measures (Newbold and Gillam, 2010), simplification is beyond the scope of this review. As a result, studies on text simplification that do not specifically address the creation of text readability measures in LRLs were excluded. In the end, only three studies on text simplification were retained in our discussion.

Unfortunately, due to access restrictions, textbooks were excluded from this review. Subsequently, book reviews were also excluded. In summary, we follow the following exclusion criteria:

- (i) development of measures for English (only) text readability,
- (ii) application methods where no measure is developed,
- (iii) unavailable full-texts,
- (iv) hardcover sources,

- (v) texts in languages other than English,
- (vi) textbooks, and
- (vii) book reviews.

After conducting the eleven separate search queries, the identified sources were processed using Endnote X9. The relevance of each result was assessed by reviewing the titles, specifically excluding those that did not mention text readability. From the initial pool of 6,162 results, a subset of 580 studies was exported to a merged Endnote library for further analysis. Additionally, utilizing Endnote X9's duplicate removal functionality, 194 duplicate sources were eliminated. This left a remaining total of 386 sources that were chosen for comprehensive full-text analysis.

2.3.2.3 Data extraction

Following the screening process based on the inclusion and exclusion criteria, all 386 studies were read in detail and the relevant studies were selected manually. This led to a total of 71 sources that were identified for further qualitative analysis. Figure 2.3 presents the total publications on text readability publications within our sample grouped per year since 2002.

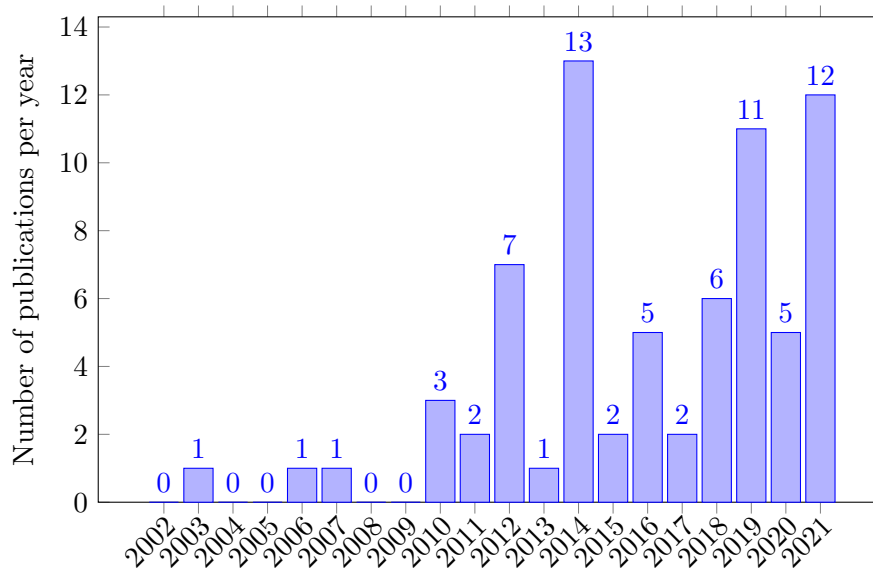


Figure 2.3: Publications on text readability measures in the sample published over time.

The final selection of sources comprised a total of 38 conference papers, 28 journal articles, one report, and four theses (consisting of one Bachelor, one Master of Arts, and two Doctoral theses). The final selection of studies per type of source is presented in Figure 2.4. These sources form the basis of our discussion. They are investigated thoroughly for methods used in measuring text readability in LRLs.

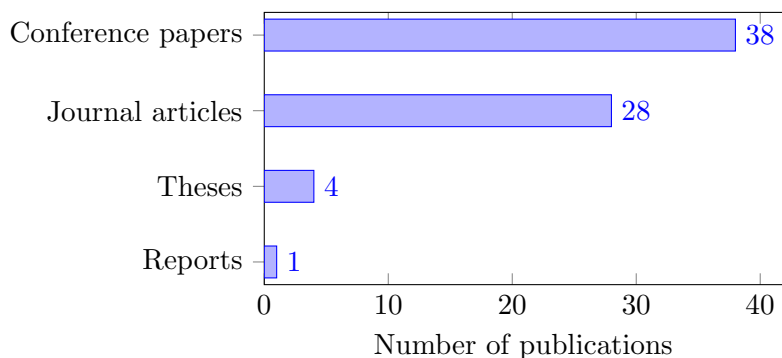


Figure 2.4: Spread of sources in the sample according to source type.

2.3.3 Results

The development of traditional text readability measures relies on the use of machine learning approaches (François and Fairon, 2012; Collins-Thompson, 2014). These approaches involve three stages:

- (i) developing or identifying a gold-standard corpus as training and testing material,
- (ii) identifying features to compute from the text, and
- (iii) training the machine learning system to predict gold label standards from the computed features.

In this section, we discuss the realisation of these steps for LRLs, beginning with findings on corpora, followed by a discussion on the textual features used. Next, machine learning approaches (including cross-lingual approaches and the adaptation of existing measures into LRLs) will be considered.

2.3.3.1 Use of corpora

Given that traditional readability approaches rely on quantitative parameters extracted from text, these approaches require example corpora. Properties from the texts in these corpora are extracted and form the input to the machine learning approaches. As such, the choice of suitable corpora is extremely important. We will first investigate the used text types and gold-standard corpora followed by an overview of the sizes of the used corpora.

2.3.3.1.1 Text types Different corpora used in the context of readability for LRLs comprise texts from various genres, including educational, non-fiction, fiction, and jour-

nalistic texts. Section 2.3.3.1.2 contains a discussion on the use of educational texts. Non-fiction texts that have been used encompass topics such as politics, health, culture, and official documents (Van Oosten et al., 2010; Battisti et al., 2020). Fictional texts include blogs, opinion articles (Van Oosten et al., 2010; Sinha et al., 2014), children’s and adult literature, novels, and short stories (Sinha and Basu, 2016; Phani et al., 2019). Journalistic texts comprise popular science (Scarton et al., 2010; Van Oosten et al., 2010), magazines (Falkenjack and Jönsson, 2014; Narasinh, 2019; Škvorc et al., 2019), and newspapers (Scarton and Alusio, 2010; Van Oosten et al., 2010; Dell’Orletta et al., 2011; Falkenjack and Mühlenbock, 2012; Falkenjack et al., 2013; Dell’Orletta et al., 2014a; Sinha et al., 2014; Sinha and Basu, 2016; Al-Khalil et al., 2020).

2.3.3.1.2 Gold-standard training corpus To use (supervised) machine learning approaches to develop readability measures, it is essential to use a gold-standard corpus that has clearly annotated levels of text difficulty (Van Oosten et al., 2010; François and Fairon, 2012). For example, Curto et al. (2014, 2015) used a corpus that was previously annotated with text readability levels using phases of the Framework for Teaching Portuguese Abroad (QuaREPE). However, LRLs may not have readily available corpora that are annotated with difficulty levels (Filho et al., 2016; Xia et al., 2019). Without a gold-standard corpus, machine learning-based text readability analyses cannot be conducted. In the following sections, we discuss five potential solutions to this challenge.

School texts Textbooks are often aimed at learners in specific grades. For this reason, textbook texts with accompanying grades are favourable for the development of readability measures. Different subject textbook texts are used, for instance language (Cavalli-Sforza et al., 2018; Soh, 2019; Lin et al., 2021; Sarimanah et al., 2021), reading comprehension (François and Fairon, 2012), social studies and life curriculum (Chen et al., 2011), and language instruction (Curto et al., 2014; Forsyth, 2014; Curto et al., 2015).

Even so, texts from school textbooks are to be used with caution. For instance, the Russian Readability Corpus comprises textbook texts and additional measures are implemented to ensure that they are representative (e.g., social sciences and history textbook texts) (Ivanov et al., 2018; Solnyshkina et al., 2018b; Solovyev et al., 2018, 2019).

Among other considerations are the issue of text types and expected textual features. For instance, when the model of Osborneva (2006), which was based on fiction texts, is tested on the moderated model of Solovyev et al. (2018), which was based on non-fiction textbook texts, it shows exaggerated readability levels. As a result, texts from different school subjects are used to account for a more general academic corpus such as the one used in Antunes and Lopes (2019b) and Soh (2019).

Similarly, corpora should contain texts from multiple authors to improve generalisability. For instance, texts from a small number of authors (or publishers) may be (too) consistent

in writing style (Solovyev et al., 2018, 2019). Unfortunately, this consistency will affect the readability assessment when texts from different authors or publishers are analysed. Therefore, using a range of authors and publishers is preferred to ensure that the style and expected textual properties vary (Balakrishna, 2015). Furthermore, texts that lead to considerable bias in the corpus can be excluded from the data set. For instance, Solnyshkina et al. (2017) exclude one of their texts that scores twice as low as other texts.

Typically, texts are collected from different school grades as they are assumed to represent different levels of readability (Qiu et al., 2018). When different grades are considered, it is observed that texts in lower grades are generally shorter than texts in higher grades (Islam and Rahman, 2014; Sun et al., 2014; Filho et al., 2016). In this way, controlling for text length is not always possible. Even so, 1000-word samples can be deemed as representative of academic texts (Biber, 1990; Ivanov et al., 2018).

Corpora with known readability levels Corpora with explicitly known differences in readability can be used. For instance, we consider the La Repubblica and the Due Parole newspaper corpora, which are commonly used in the development of Italian readability measures (Dell’Orletta et al., 2011; Sjöholm, 2012; Dell’Orletta et al., 2014b). The La Repubblica corpus contains normal everyday newspaper texts while the Due Parole comprises simplified newspaper texts. In this way, easier and harder-to-read texts can be identified based on the corpus of origin. Even so, when harder-to-read sentences are removed, from the La Repubblica corpus, higher predictive precision is observed (Dell’Orletta et al., 2014b). Although the texts from the two corpora are not aligned, they are comparable in genre and different on the basis of being easier and harder to read.

Translated corpora Two types of translated texts are identified in the literature. First, intra-lingual parallel texts containing original and simplified texts are used. For instance, Battisti et al. (2020) create corpora containing German texts at two levels, namely, everyday German and simplified German texts.

Second, inter-lingual parallel texts containing translated texts from different languages are used. For instance, El-Haj and Rayson (2016) use English and Arabic texts. The use of translated and parallel texts enables the exploration of linguistic differences between languages so that the correct linguistic properties affecting the specific language for which a readability measure is developed can be established. For example, using translated and parallel texts, Antunes and Lopes (2019a) discover that English formulas using syllable counts and complex words need adjustment before conversion into Portuguese measures. Translated texts are then aligned either at the sentence level as in the case of Cardon and Grabar (2020) or at paragraph level akin to El-Haj and Rayson (2016). Additionally, Battisti et al. (2020) demonstrate the use of back-translation in developing more synthetic texts.

Crowd and expert sourcing Crowds and language experts are also used in developing gold-standard corpora. The literature does not provide guidelines for study population sizes (Amendum et al., 2018). Even so, this method can be used in sorting texts on broad categories such as easy or hard (Van Oosten et al., 2010). In fact, annotators can give objective information about the text. In their study, Laposhina et al. (2018) collect texts from the Cambridge International Education (CIE) resources and textbooks with identified target levels. However, they discover that clear Common European Framework of Reference (CEFR) levels are not always indicated and that difficulty levels sometimes indicate the authors’ subjective evaluations. As a remedy, they employ crowdsourcing for highlighting difficult words and annotating three texts with three CEFR labels, namely, A1, B1, and B2. They found that difficult texts as indicated by their measures correlate positively with crowd-highlighted difficult words (Laposhina et al., 2018). In this way, textual properties that make texts easy to read can be highlighted by the crowd. For instance, crowdsourced insights indicate that sentence and word lengths are non-predictive of the ease of readability in Bangla (Sinha and Basu, 2016) and Hindi (Sinha et al., 2014; Venugopal et al., 2021).

When using crowdsourcing, a few considerations should be heeded. First, non-expert participants should be given some training even if it is just definitions of core concepts (Mohammadi and Khasteh, 2020). Second, differences caused by varied educational backgrounds can be minimised by targeting participants with similar academic backgrounds (Das and Roychoudhury, 2006) or similar age groups (Phani et al., 2019). Alternatively, language experts can be used. Studies using expert participants generally accommodate very few experts. For instance, Nahar et al. (2019) use only one participant. Similar to non-expert participants, language experts are also given test texts to evaluate their grading of texts before they continue to develop a gold-standard annotated corpus (Dell’Orletta et al., 2014b).

Web crawling Lastly, web crawling techniques are used in collecting gold-standard training data. Easy-to-read texts, such as simplified texts from newspaper websites, can be collected using this method. Additionally, Filho et al. (2016) propose the use of a web crawling readability filter for collecting texts and generating classified corpora. Two main issues impede the successful implementation of this method for generating gold-standard corpora. First, this method is unusable in the absence of already existing readability measures. Second, the lack of presence in web spaces renders this method far-fetched for many LRLs.

Corpora sizes All studies reviewed in this section rely on some form of a corpus, but the corpus sizes used vary considerably. Here, we consider corpus size measured based on tokens, sentences, and the number of texts per corpus.

Regarding corpus size by tokens, two studies use fewer than twenty thousand tokens

(Dell’Orletta et al., 2011; Venugopal et al., 2021), four studies use between 20,000 and 100,000 tokens (Forsyth, 2014; Curto et al., 2015; Saddiki et al., 2015; Nassiri et al., 2022b), six studies comprised between half a million and a million tokens (Filho et al., 2016; Ivanov et al., 2018; Solovyev et al., 2018, 2019; Bengoetxea and Gonzalez-Dios, 2021; Khallaf and Sharoff, 2021), and three studies use more than a million tokens (Solnyshkina et al., 2018b; Al-Khalil et al., 2020; Cardon and Grabar, 2020). Notably, Bendová (2021) uses more than forty-three million tokens for Czech as a pivotal translation language and more than a hundred million combined tokens for Czech, English, Russian, French, and Italian.

In terms of corpus size by the number of sentences, approaches range from 1,000 sentences (Naderi et al., 2019) to almost four million sentences for Czech as a pivotal translation language, with over nine million combined sentences for Czech, English, Russian, French, and Italian (Bendová, 2021). Furthermore, several studies use more than a thousand sentences (El-Haj and Rayson, 2016; Filho et al., 2016; Ivanov et al., 2018; Battisti et al., 2020; Chakraborty et al., 2021; Khallaf and Sharoff, 2021; Nassiri et al., 2022b).

Finally, we consider corpus size measured by the number of texts in each corpus. The largest number of documents per corpus in the literature is 14,581 texts (Škvorec et al., 2019). Table 2.2 presents the spread of texts per corpus, indicating that there are no prescribed guidelines for corpus sizes.

In conclusion, the use of corpora is essential for developing readability measures using machine learning-based approaches. Although there are no set guidelines for corpus sizes, the range and variety of corpus sizes that are used demonstrate the flexibility and the adaptability of machine learning approaches in assessing text readability in LRLs.

2.3.3.1.3 Textual features In order to explicate information obtained from the texts under consideration, specific textual properties or features are extracted. The decision on the level of readability is then decided upon by considering the textual features (only). The choice of which (and how many) textual features to use may have a major impact on the performance of the readability measure.

The total number of textual features considered in the different studies ranges from three (Sinha et al., 2012) to 406 features (François and Fairon, 2012). Interestingly, although François and Fairon (2012) investigate 406 features, they conclude that four lexico-syntactic features yield the same results as the many feature models. The total number of features used in different studies is presented in Table 2.3.

Seven major textual feature groups can be identified from the different studies. The total number of studies considering each of the major text feature groups is presented in Figure 2.5.

Surface features, also called raw features (Dell’Orletta et al., 2014a), refer to textual fea-

Table 2.2: Overview of corpora sizes by the number of documents considered in the different studies.

Texts	Studies stating corpus size
1–100	Das and Roychoudhury (2006); Scarton et al. (2010); Sinha et al. (2012); Islam and Rahman (2014); Sinha et al. (2014); Filho et al. (2016); Solnyshkina et al. (2018b); Narasinh (2019); Phani et al. (2019); Soh (2019); Solovyev et al. (2019); Mohammadi and Khasteh (2020); Al Shaye and Al Dhefeiri (2021); Sarimanah et al. (2021)
101–500	Dell’Orletta et al. (2011); Islam et al. (2012); Curto et al. (2014); Forsyth (2014); Gonzalez-Dios et al. (2014); Curto et al. (2015); Saddiki et al. (2015); Gonzalez-Dios (2016); Cavalli-Sforza et al. (2018)
501–1,000	Chen et al. (2011); Dell’Orletta et al. (2011); François and Fairon (2012); Dell’Orletta et al. (2014b); Falkenjack and Jönsson (2014); Sun et al. (2014); Qiu et al. (2018); Battisti et al. (2020); Nassiri et al. (2022b)
1 001–2,000	Falkenjack and Mühlenbock (2012); François and Fairon (2012); Falkenjack et al. (2013)
2 001–10,000	Azpiazu and Pera (2020); Battisti et al. (2020)
10 001–	Škvorc et al. (2019)

tures used in traditional readability formulas, such as word and sentence lengths (Dell’Orletta et al., 2011; Falkenjack et al., 2013; Gonzalez-Dios et al., 2014; Gonzalez-Dios, 2016). They include features that can be extracted after tokenisation by counting words and characters (Falkenjack and Jönsson, 2014). Among others, syllable, word, and sentence lengths are considered. *Lexical* features require lemmatisation (Falkenjack et al., 2013). Optionally, frequency and uncommon words lists and dictionaries are used for lexical feature extraction in measures for languages such as Slovene (Škvorc et al., 2018, 2019), Bengali (Chakraborty et al., 2021), Latin (Lin et al., 2021), Italian (Dell’Orletta et al., 2011, 2014a; Bolli et al., 2017), Chinese (Xu et al., 2019), Russian (Solovyev et al., 2018, 2019), French (François and Fairon, 2012), and Arabic (Forsyth, 2014).

Lexical changes have been shown to be important in text simplification (Drndarevic and Saggion, 2012). For instance, uncommon words are replaced with common words from the frequency levels in the Real Academia Española corpus⁸ (Drndarevic and Saggion, 2012; Bott and Saggion, 2014; Štajner et al., 2015).

Part of Speech (POS) features provide information on the syntactic roles of words in a sentence. POS-based analyses depend on the availability of a POS-tagger (Forsyth, 2014). We differentiate between three POS-tagged feature groups. The first group is concerned

⁸See <https://www.rae.es/>.

Table 2.3: Overview of the number of textual features considered in the different studies.

Features	Studies stating the number of textual features
1–10	Sinha et al. (2012); Sinha and Basu (2016) Soh (2019); Phani et al. (2019); Battisti et al. (2020)
11–50	François and Fairon (2012); Karpov et al. (2014); Sinha et al. (2014); Curto et al. (2015); Castilhos (2016); Ivanov et al. (2018); Škvorc et al. (2019); Soh (2019); Al Shaye and Al Dhefeiri (2021); Venugopal et al. (2021)
51–100	Qiu et al. (2018); Naderi et al. (2019); Nassiri et al. (2022b)
101–150	Jiang et al. (2014); Laposhina et al. (2018)
151–	François and Fairon (2012); Forsyth (2014); Bengoetxea and Gonzalez-Dios (2021)

with *POS-based* features that do not rely on syntactic parsing but depend on specific POS. The second group relies on *syntactic* features and relies on syntactic parsers (Falkenjack et al., 2013; Falkenjack and Jönsson, 2014; Forsyth, 2014). It includes the ability to consider dependency trees and therefore analyses sentence complexity (Azpiazu and Pera, 2020; Hansen et al., 2021). The third POS-based group of features, *morphosyntactic* features, focuses on content words and their ratio to the total tokens in the considered text (Dell’Orletta et al., 2011, 2014b; Gonzalez-Dios, 2016).

The information transfer features are grouped into two, namely, *semantic*, and *entropy* features. They go beyond tokens and text structures. As such, it is not surprising that entropy and semantic features are considered in relatively few studies since most studies focus on lexical and surface-level features, therefore less dependence on higher-level analysis is expected. Semantic features can encompass investigation into personalisation levels, lexical cohesion, and conceptual density (François and Fairon, 2012). Unfortunately, since very few semantic features are generally considered, the predictive accuracy of semantic features is relatively low (François and Fairon, 2012; Azpiazu and Pera, 2020; Bengoetxea and Gonzalez-Dios, 2021). Similar to conceptual density, entropy is borrowed from information theory, as such, lower entropy levels are associated with easy-to-read texts (Islam and Rahman, 2014; Sun et al., 2014; Soh, 2019).

Lastly, *morphological* features analyse how words are formed (Azpiazu and Pera, 2020). In order to analyse morphological features to predict inflectional and derivational morphologies, a morphological processor and a lemmatizer are needed (Forsyth, 2014; Škvorc et al., 2018; Naderi et al., 2019; Škvorc et al., 2019). Morphological features yield higher prediction accuracy when combined with other features (Gonzalez-Dios et al., 2014; Gonzalez-Dios, 2016).

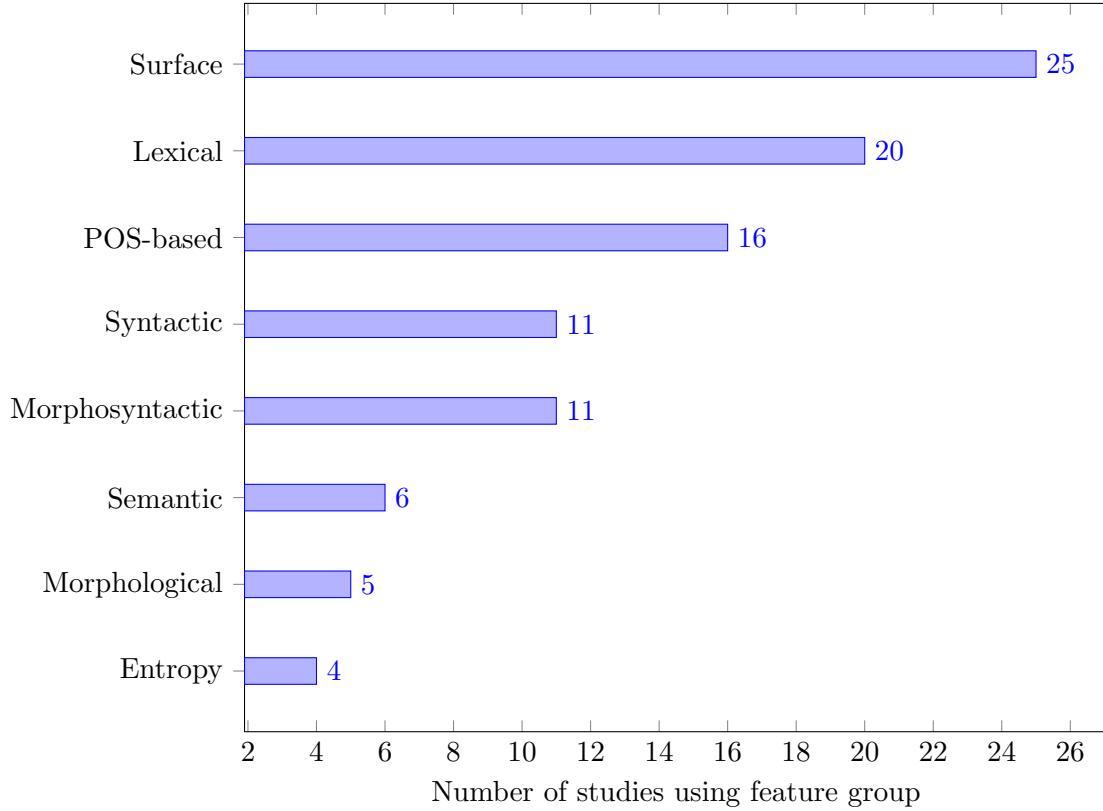


Figure 2.5: Distribution of textual feature groups used in studies.

2.3.3.1.4 Frequently used methods This section focuses on the commonly used methods employed in the development of readability measures for LRLs. It begins by discussing cross-lingual approaches and subsequently delves into machine-learning approaches, including the adaptation of traditional readability measures for LRLs.

Cross-lingual approaches Cross-lingual readability assessment which enables the rating of text readability in different languages is a relatively novel area of research (Azpiazu and Pera, 2020). Despite its nascent status, promising outcomes have already been documented. We discuss two distinct approaches to cross-lingual readability assessment.

The first approach involves identifying similar features and using these selected features to assess readability, irrespective of the language under examination. This approach accommodates the use of less comparable data sets. For instance, Bengoetxea and Gonzalez-Dios (2021) employed separate corpora in English, Basque, and Spanish to create a cross-lingual readability measure. Specifically, the researchers utilised 400 Basque texts, which were divided into 200 easy and 200 difficult texts. Their Spanish corpus contained 100 texts split into 50 easy and 50 difficult texts. Moreover, the 567 English texts were categorised into three proficiency levels: 189 elementary texts, 189 intermediate texts, and 189 advanced

texts. In the end, the developed MultiAzterTest for English-Spanish-Basque texts showcased a high level of prediction accuracy (Bengoetxea and Gonzalez-Dios, 2021). The intersection of feature selection across the three languages demonstrated lower accuracy for English (76.53%) and the highest accuracy for Basque (92%).

The second approach in cross-lingual readability assessment involves using an LRL as the base language and a higher-resource language as a helper language (Azpiazu and Pera, 2020). The helper language provides additional training texts that complement the base language. For instance, in their study, Azpiazu and Pera (2020) employ a sampling method that is worth noting. They randomly select 5,388 documents from various languages, including English, Spanish, Basque, Italian, French, and Catalan. To ensure comparability, they limit the maximum number of documents for each language to 449, ensuring an equal number of training documents for each language, considering that Basque only has 449 documents available in the document collection. Furthermore, they select texts that facilitate a comparison between simple and complex texts. Note that the effectiveness of the base and helper relationship improves when the languages involved are typologically related (Azpiazu and Pera, 2020). It is important to note that the model developed in their study is specifically designed for categorising texts as either easy or complex. However, the approach holds promise if the parameters that determine text readability in a given language can be accurately identified (Azpiazu and Pera, 2020).

Machine learning approaches When the values of textual features change, this may have an effect on the measured readability levels. The changed values could lead to either higher or lower measured readability levels (Gunning, 2003). Correlation tests are run to investigate such relationships. The literature typically uses either the Spearman Correlation (El-Haj and Rayson, 2016; Laposhina et al., 2018; Azpiazu and Pera, 2020) or the Pearson’s Product Moment Correlation Coefficient (Das and Roychoudhury, 2006; Dell’Orletta et al., 2011; Jiang et al., 2014; Laposhina et al., 2018).

Correlation results, unfortunately, are restricted to revealing relationships between individual textual features and the overall readability of the texts. Even so, the preferred end goal for readability assessment is an indication of specific levels or grade targets (Solovyev et al., 2019). As such, these can be attained using regression or classification models (Sinha and Basu, 2016; Mohammadi and Khasteh, 2020). While real numbers can be obtained using regression models, classifiers indicate a categorical classification. Commonly used regression and classification models are discussed below.

Linear regression models are used for identifying the influence of features on readability (Ivanov et al., 2018). This allows, for instance, for the adjustment of coefficients from the English readability measures to specific languages (Solovyev et al., 2018). Linear regression models produce few errors compared to other regression models (Sinha and Basu, 2016) and overall perform almost as well as random forest regression (Naderi et al., 2019). Different variations of linear regression models can be identified from our sampled

studies, including, basic linear regression models (Curto et al., 2014; Sun et al., 2014; Amendum et al., 2018; Laposhina et al., 2018; Antunes and Lopes, 2019a; Narasinh, 2019; Solovyev et al., 2019; Mohammadi and Khasteh, 2020; Lin et al., 2021), linear regression with ridge regression (Ivanov et al., 2018; Laposhina et al., 2018; Solovyev et al., 2019), linear regression with lasso regularisation (Solovyev et al., 2019), linear regression with ElasticNet regularisation (Solovyev et al., 2019), and logistic regression (François and Fairon, 2012).

Classification approaches tend to provide more accurate results than regression models (Aluisio et al., 2010). Although multiple classifiers are used in different studies, we summarise common classifiers that are used in at least five studies. Common classifiers are presented in Figure 2.6.

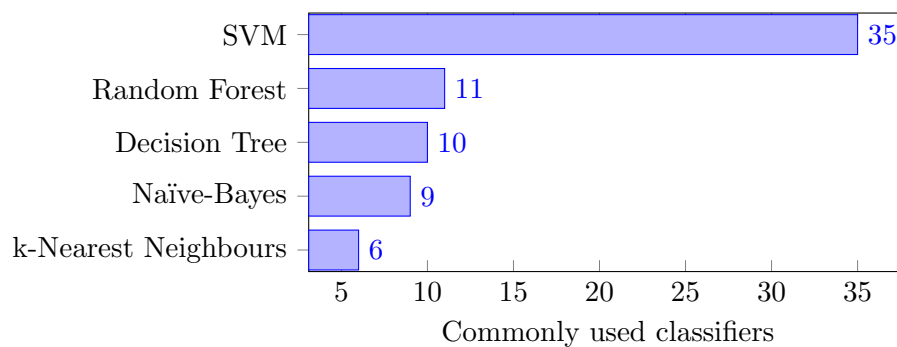


Figure 2.6: Overview of the distribution of classifiers commonly used.

According to Mohammadi and Khasteh (2020), Support Vector Machines (SVM) are the most commonly employed classifiers in text classification tasks. As demonstrated in Figure 2.6, this observation remains consistent even in our own investigation. However, other studies have also identified Random Forest as one of the frequently utilised classifiers in text readability research (Solovyev et al., 2019; Azpiazu and Pera, 2020). In our analysis, as presented in Figure 2.6, we find that Random Forest is the second most frequently used classifier among the studies we survey.

It is worth emphasizing that achieving better classification performance is more likely when large-scale training data is available, particularly when features are thoroughly studied (Tanaka-Ishii et al., 2010). This suggests that having a substantial amount of training data and a comprehensive understanding of the features can greatly enhance the performance of classification models.

Adapting English readability formulas

In the studies investigated, we identify the adaptation of eight traditional readability measures into LRLs. These include syllable information-based measures such as FRE, FKGL, GFI, and SMOG, and word-length-based measures such as ARI, CLI, LIX, and RIX (see Section 2.4 for an overview of the measures).

We identified two main methods for adapting existing traditional readability measures for LRLs. Two main methods can be identified, namely, the monolingual-based adaptation, and the translated or parallel-text methods.

First, Jakobsen and Skardal (2007) explore the eight above-mentioned traditional readability measures for adaptation into Norwegian. After comparing superficial text features using web-crawled texts, they discovered that syllables and complex words do not have the same effect on readability levels in Norwegian as they do in the original measures. In the end, only the LIX and RIX measures are adapted to Norwegian. Similarly, New Russian readability measures draw inspiration from the established English measures, FRE and FKGL, as evidenced in prior works by Solnyshkina et al. (2018b) and Solovyev et al. (2018). To develop these measures, they leverage the Russian Readability Corpus (Solovyev et al., 2018) and the Russian Frequency Dictionary (Lyashevskaya and Sharov, 2009) as primary resources. Their models are based on average word length and average sentence length and are assessed through the mean squared error and the coefficient of determination. Amongst others, they find that corpus size is important.

Second, the FRE measure is also adapted into Czech (Bendová, 2021). This is done using the InterCorp parallel corpus (Čermák and Rosen, 2012; Rosen, 2016) for Czech as a pivot language for languages including English, Russian, French and Italian. These four languages have language-specific versions of the FRE formula, that is English (Flesch, 1974), Russian (Oborneva, 2006), French (Kandel and Moles, 1958; Henry, 1975), and Italian (Franchina and Vacca, 1986). Initial correlation tests indicate that the Italian and French models deviate greatly from the Czech, English, and French models. As a result, they are excluded from further investigation. The English scores indicate a closer relationship between the Czech expected outcomes and the English measure outcomes. As a result, a new regression model for analysing readability in Czech is formulated based on the root mean squared error (RMSE) between the English and the Czech FRE readability scores. In a similar study, the FRE, FKGL, ARI, and CLI measures are adapted into Czech using the InterCorp and CzEng 2.0 corpora (Bendová and Cinková, 2021). A non-linear optimize.curve fit algorithm is used to determine specific Czech parameters. The end products are four new Czech traditional readability formulas. Note that, there is no justification for the resultant difference between the coefficient in the FRE formula for Czech ($\text{FRE} = 206.935$) in Bendová and Cinková (2021) and the coefficient ($\text{FRE} = 206.835$) in Bendová (2021) although the RMSE for both is constant at 4.639.

Unfortunately, successful adaptation of HRL readability measures is not always guaranteed. For instance, Sinha et al. (2012) reports an unsuccessful attempt at adapting the FRE, GFI, FKGL, and SMOG measures into Hindi and Bangla. They conclude that new measures that are based on Bangla and Hindi structural properties should be developed as the existing measures yield out-of-bound results (Sinha et al., 2012). New textual properties such as Jukta-Akshars, that is Asian language characters formed by combining two or more consonant letters into a single graphical unit, are then introduced in Hindi measures.

2.3.4 Discussion and conclusions

This systematic review aimed to identify methods that are used in developing text readability measures for LRLs. Within the reviewed text readability publications, two major approaches were identified, namely, machine learning approaches and cross-lingual approaches. Additionally, we specifically note the adaptation of traditional HRL text readability measures to LRL.

Our findings are aligned with the prognostication of Collins-Thompson (2014) that data-driven methods are one of the futures of computational text readability approaches. All the publications explored in this section, are dependent on data. This makes data the most important aspect of developing a readability measure for any language. Given the novelty of text readability in most languages discussed here, multiple strategies for collecting corpora for text readability assessment are employed. The use of textbook texts is prevalent in many studies explored. If we are to assume that there are some developments in the language(s) for which text readability measures are envisaged, then we hope that there are some textbooks in the language. School texts are favoured in this regard as some form(s) of gatekeeping ensure that the texts are accessible to learners in certain grades. Even though translated texts are also very useful in this regard, they are to be used with caution. When the helper language, that is the HRL, is structurally different from the LRL then the properties that make the texts easy to read might not be consistent.

When focusing on raw textual features such as word, syllable, and sentence length information and modelling the text readability measure on traditional readability measures, it is clear that the number of features is not so important as long as determinative features are identified. The range of features found in the literature (ranging from three to 406) allows for flexibility. If one is to correctly estimate text readability in a language, one would need to correctly assess the textual features that best indicate the differences in text readability. For instance, surface (or raw) features could be combined with semantic or syntactic features if needed. This, however, should be based on merit. Correlation tests should be used to determine these.

Cross-lingual approaches to text readability assessment have not been employed much. However, this should not deter future studies from embarking on such research as current studies show success.

Studies on text readability in LRLs included in this review are based only on the Google Scholar database and limited to those found by the “Publish or Perish” platform (Harzing, 2007). Future studies could look into alternative databases as well. Furthermore, they could consider also looking at search results outside the twenty-year period in this section (note that some of the traditional readability measures are over 50 years old). In this case, more studies may be identified. As mentioned earlier, many results from our different search words were identified as duplicates when the search results were merged. Perhaps

a more carefully chosen variation of search terms could be implemented.

2.4 Overview of English readability measures

2.4.1 Introduction

As previously indicated, traditional readability measures, such as grade-level formulas, generate numerical results in absolute scores or grade levels (Yan et al., 2006; Janan, 2011). In this way, these measures measure the readability of a text by assessing the level of education required for reading the text with ease (Nassiri et al., 2022a). However, such measures provide a general measure for user groups rather than tailoring the analysis for individuals (Kriukova, 2022). Thus, readers' attributes beyond the expected years of education are not considered.

To evaluate the readability of the different texts, we selected nine well-known readability measures, namely, FKGL (Kincaid et al., 1975), FRE (Flesch, 1948), SMOG (Mc Laughlin, 1969), GFI (Gunning, 1952, 1969), LIX and RIX (Anderson, 1983), ARI (Smith and Senter, 1967; Kincaid and Delionbach, 1973), CLI (Coleman and Liao, 1975), and the DCI (Dale and Chall, 1948). These measures are used in the *Python 3* readability package⁹ (Van Cranenburgh, 2013; Python, 2020). The existing traditional readability measures that we explore in this thesis are briefly described below. Note that this is a summarised discussion, in-depth discussions of these measures are provided elsewhere, for instance, see Heydari's (2012) and Zamanian and Heydari's (2012) studies.

The brief discussion of the measures below classifies the measures into three categories based on their distinguishing features, specifically: (i) syllable information, (ii) lengths of words measured by letters per word, and (iii) the consideration of frequently used words.

2.4.2 Syllable-based measures

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

2.4.2.1 Flech-Kincaid Grade Level (FKGL)

The Flech-Kincaid Grade Level (FKGL) uses United States (US) grades for labelling readability levels (Kincaid et al., 1975). For example, a score of ten corresponds to the

⁹See <https://github.com/andreasvc/readability/> for the readability application and its script.

tenth-grade (Toyama et al., 2017). The FKGL measure uses the following formula:

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

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

Although the FKGL measure was developed for the United States Navy, it has been proven to be suitable for use in multiple contexts including educational contexts (Zhang et al., 2019). As such, it remains relevant to this thesis and for use in evaluating the readability of different genres of text in Sesotho. The English FKGL formula (just like the FRE formula mentioned below) has 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).

2.4.2.2 Flesch Reading Ease (FRE)

The FRE is one of the most used traditional readability formulas (Waltl and Matthes, 2014). Furthermore, of all traditional readability formulas, FRE has undergone the most extensive adaptation to other languages (Bendová and Cinková, 2021). Notable examples of successful adaptations include the French (Kandel and Moles, 1958; Henry, 1975), Italian (Franchina and Vacca, 1986), Russian (Oborneva, 2006), Dutch (Douma, 1960; Brouwer, 1963), Bangla (Sinha et al., 2012; Sinha and Basu, 2016), and Czech (Bendová, 2021) adaptations. Moreover, the FRE measure has been adapted to at least one Southern African official language, namely Afrikaans, for which three variations have been proposed thus far (Jansen et al., 2017). Furthermore, when combined, the FKGL and FRE can be used for both first and second-language texts (Greenfield, 2004). Additionally, both FRE and FKGL are integrated into Microsoft Office (Bendová, 2021). As a result, they can be easily used by anyone who uses Microsoft Office products (see Figure 2.7 for an example of readability statistics that are freely available on Microsoft Word). In fact, according to Fischer et al. (2021, p. 183), the South African National Ethics Committee outlines that all informed consent forms should be assessed using the Microsoft FRE and FKGL readability tools to ensure that the forms are at the 8th-grade level.

FRE is calculated using the following formula:

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

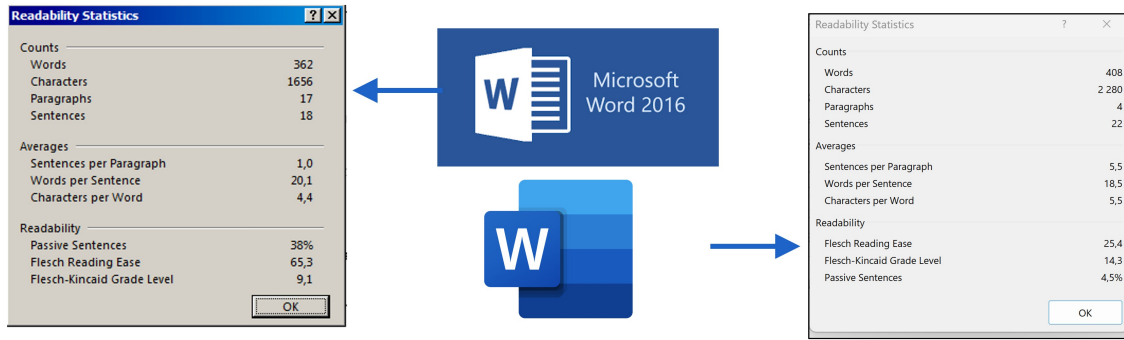


Figure 2.7: The readability statistics are readily available on Microsoft Word.

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

There have been criticisms of the FRE formula. For instance, Hartley (2016) criticises the FRE formula for being used inappropriately, ignoring word meanings, and producing inconsistent results across different computer programs. Even so, we posit that this thesis aims to determine the minimum educational level required for readers to read texts with ease rather than determining the comprehension that texts provide. Furthermore, discrepancies in the outputs across different computer programs are caused by the differences in the underlying computer programs rather than issues with the formula itself. Therefore, the FRE is a useful measure for assessing the readability of texts despite its limitations.

The FRE formula uses all counts of syllables from the text under analysis. Therefore there are no special measure-specific syllable information considerations for the FRE.

2.4.2.3 Gunning Fog index (GFI)

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

$$\text{GFI} = 0.4 \left(\left(\frac{\#words}{\#sentences} \right) + 100 \left(\frac{\#complex_words}{\#words} \right) \right)$$

The readability score generated by the English equation above typically falls within the range of six to twenty. A score of six indicates that the text is suitable for a sixth-grade

reading level, while a score of twenty or higher suggests that the text is appropriate for advanced readers such as those in university postgraduate programs.

2.4.2.4 Simple Measure of Gobbledygook (SMOG)

The SMOG measure also outputs United States school grade levels. The SMOG formula is also one of the most used traditional readability formulas (Janan, 2011). It estimates readability based on polysyllabic words and sentence complexity. Polysyllabic words as indicated in the formula refer to words with more than two syllables (Mc Laughlin, 1969; Kasabwala et al., 2012; Eltorai et al., 2015; Zhou et al., 2017; Harden, 2018). When calculating the SMOG for long texts, three samples are used, one from the beginning of the text, one from the middle, and one from the end of the text (Mc Laughlin, 1969; Zhou et al., 2017). Each sample comprises ten sentences. The samples are used to calculate SMOG using the following formula:

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

According to Rules (1969), words combined by the use of hyphens should be counted as one word. Additionally, abbreviations are counted as full words Hedman (2008). Although counting syllables in abbreviated words would be simpler for a human annotator, it might be challenging for a machine. It would be easier to train the machine on a list of abbreviations used in current corpora, however, the list would need to be a living document as new abbreviations can be anticipated. In this way, the counting of syllables from abbreviated words may be inconsistent.

2.4.3 Word-length-based measures

Four of the selected traditional readability measures are based on word lengths. The four measures are described below.

2.4.3.1 Läsbarhetsindex (LIX) and Readability Index (RIX)

LIX was originally developed for Swedish (Björnsson, 1968; Björnsson, 1983). For access, we use the English formula as a point of reference (Anderson, 1983). The RIX measure is an adaptation of the LIX measure (Anderson, 1983; Courtis, 1987). The LIX and the RIX formulas pay special attention to “long words”, that is, words that have more than six letters. For Swedish and English texts, words with more than six letters are considered long. Both the LIX and the RIX formulas output numbers that are then converted to grade levels and the fractions are ignored (Anderson, 1983). Ignoring the

decimals is particularly important in instances where adjusting the scores changes the predicted grade. For instance, when adjusting a LIX score of 47.99 to 48.0, the predicted grade level changes from the eleventh to the twelfth grade. The LIX and the RIX are computed using the formulas presented below:

$$\text{LIX} = \left(\frac{\#words}{\#sentences} \right) + \left(\frac{\#long_words}{\#words} * 100 \right)$$

$$\text{RIX} = \left(\frac{\#long_words}{\#sentences} \right)$$

The RIX formula considers the ratio of long words to the number of sampled sentences. While shorter texts may be considered as a whole, longer texts may use sentence sampling methods (Anderson, 1981, 1983). It is suggested that ten samples comprising ten sentences each be analysed when estimating both LIX and RIX (Anderson, 1983). When sampling is applied, the RIX formula considers the ratio of long words to the number of sampled sentences.

2.4.3.2 Coleman-Liau index (CLI)

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

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

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

2.4.3.3 Automated Readability Index (ARI)

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

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

Letters as used in CLI and ARI, refer to all letters and numbers that build words (Zhang et al., 2019). Thomas et al. (1975) describe it as strokes representing each word. In a similar study, the CLI formula was adapted to Croatian and only the weights were adjusted in that instead of deducting 21.43 from the totals, 23.23 is deducted (Kusec et al., 2002; Alić, 2020). When adapted to Czech, the weights and constants are drastically adapted so that it results in the following formula (Bendová and Cinková, 2021):

$$\text{Czech ARI} = 0.63\left(\frac{\#letters}{\#words}\right) + 3.66\left(\frac{\#words}{\#sentences}\right) - 19.491$$

According to Bendová and Cinková (2021), when the weights in the English formula are used, some scores indicate difficulty levels comparable to infancy levels. It is therefore important to adapt the weights for the text properties as it is clear that there may be some form of complexity trade-offs. For instance, sentence length seemingly affects readability more than the lengths of words in Czech while the opposite is true for Croatian and English texts.

2.4.4 Frequency-list-based measure

One frequency-list-based measure was identified from *Python 3*'s readability package. The measure is described below.

2.4.4.1 Dale-Chall Index (DCI)

The DCI measure uses a frequency list (Dale and Chall, 1948). The frequency list is based on a list of 3,000 words that a grade four learner is expected to be familiar with (Stocker, 1971). Difficult words are considered as those that do not appear in the list. Variations include words in plural forms, verbs that end in -s, -ed, -ing, and -ied, adverbs that end in -ly, abbreviations, and compound words (Barry and Stevenson, 1975). Note that the names of people and organisations are assumed as common. As such, they are counted with common words although they are not listed in the list of common words. they are counted only two times per 100-word sample. DCI is computed using the following formula:

$$\text{DCI} = 0.0496\left(\frac{\#words}{\#sentences}\right) + 0.1579\left(\frac{\#difficult_words}{\#words}\right) + 3.6365$$

It is advisable to use the whole text when texts are too short for sampling. For longer texts, one may sample four sets of 100 words per 2,000 words (Dale and Chall, 1948; Barry

and Stevenson, 1975; Barry, 1980). The list of words was updated in 1995 (Chall and Dale, 1995).

2.4.5 Example use of readability formulas

In Table 2.4, a machine translation from Sesotho to English is used, and the readability levels of seven readability measures are presented, along with the average sentence length (ASL) and average word length (AWL). The outcome shows that the readability levels are not consistently uniform. As a result, all nine traditional readability measures discussed in this thesis are relevant for cross-comparison.

Table 2.4: Example application of readability measures and their outputs.

GOOD WAYS TO MAKE CUSTOMERS HAPPY									
Experts say that treating customers well is not difficult, it just requires the store owner to know different ways to make them happy. It is important to make sure that you know the names of customers who buy from your store regularly. When they arrive at the store, greet them with their names or surnames. This makes them proud and realize that you care about them. Remember that customers are not all the same. Some will be angry and burning with flames. When you talk to such a customer, don't get angry with him. Be humble and talk to him in a cool spirit. Learn to listen to your customers. If you practice listening to them, you will be able to meet their needs. Another important thing is to accept their complaints regardless of what they are complaining about. If you do this, tomorrow they will keep coming back to your shop. Ask customers what goods or products they would like to have in your store. Maybe if you can get it for them, that can save them time and money. Be available when customers need you. Don't hide and disappear from the office. Even if the customer has left a message saying call him, call him. Never make promises to customers that you cannot fulfill. If you told the customer to come on Friday to get his goods, do so.									
Metrics	ASL	AWL	FRE	FKGL	DCI	CLI	GFI	SMOG	ARI
Readability	12.32	4.32	79.00	5.30	6.95	7.24	9.20	9.68	5.08

2.4.6 Summary

Two important things can be noted from the nine traditional readability measures briefly overviewed in this section. First, the measures use specific processes for estimating appropriate readability and grade levels. It is important to consider these processes. For instance, this is useful when considering the minimal corpus size necessary for adapting the measures to Sesotho. Although sampling addresses the issue of long texts, it does

not account for shorter texts except in the case of SMOG which caters for shorter texts through the SMOG for shorter texts. As a guideline, the texts used in this thesis contain a minimum of 200 each.

Second, the measures use specific weights that may need to be adapted to Sesotho. As was demonstrated by the discussion of the differences between the English, Croatian, and Czech ARI, more weight is attributed to the feature of the text that affects the readability of the text the most. As such, special consideration of aspects such as syllable information and the effect thereof on text readability needs to be addressed per measure.

At least the following properties should be extracted from texts to enable the development of the adaptation of classic readability measures to Sesotho:

- word properties (including lengths, counts, and averages),
- sentence properties (including lengths, counts, and averages),
- syllable information (including counts, multiplicity, and averages), and
- a list of frequently used words in Sesotho.

Moreover, it is important to consider the expected outputs from the formulas. This is most important for formulas that do not output grade levels, for instance, the FRE, LIX and RIX. For these measures, we may need to redefine the conversions to suit the context of Sesotho and to reflect South Africa's grade levels.

In spite of being used in multiple contexts, the traditional readability measures have been criticised for failing to measure comprehension (Tanprasert and Kauchak, 2021). Furthermore, the use of frequency lists, such as in the DCI list of common words, has been criticised for failing to account for specialised meanings (Yan et al., 2006). Even so, the traditional readability measures remain relevant to our project since our focus is not on meaning or comprehension but on the ease with which the text can be read.

2.5 An overview of resources and tools available to Sesotho

This section is adapted from: Sibeko, J. and Setaka, M., 2022. Sesotho BLARKs content. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11. <https://doi.org/10.55492/dhasa.v4i01.4440>.

2.5.1 Introduction

There is a growing interest in Human Language Technologies (HLTs) for LRLs (Strassel and Tracey, 2016). Accordingly, a number of HLT audits have been conducted on South

African official languages (Grover et al., 2010, 2011; Moors et al., 2018a,c). The language audits are aimed at two objectives that are (i) determining resources that need to be developed, and (ii) opportunities for multidisciplinary research. South Africa currently recognizes twelve official languages, namely, Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, South African Sign Language¹⁰, English, isiNdebele, isiXhosa and isiZulu (Republic of South Africa, 2023). In this section, we pay special attention to digital language resources available for Sesotho.

Of the eleven long-standing official languages of South Africa, two languages, namely; Afrikaans and South African English, have the most digital language resources followed by Setswana, Sepedi, isiZulu, isiXhosa and Sesotho in no particular order (Moors et al., 2018a,c). This is a reality which is propelled by the lack of investment in the remaining nine indigenous languages. This image of the inequality of languages in terms of the development of digital language resources was shown at the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (Pretorius et al., 2014). Nonetheless, although there are digital language resources in Sesotho, the language remains marginalised and under-resourced as it is yet to have enough high-quality supervised data (Koai and Fredericks, 2019; Magueresse et al., 2020). As a result, multiple studies identify Sesotho as an LRL (Mahloane and Trausan-Matu, 2015; Chiguvare and Cleghorn, 2021; Hanslo, 2021; Sibeko and Van Zaanen, 2022a). Moors et al. (2018c) go as far as identifying it as a severely under-resourced language. That is, it persists in being less studied for computerisation and thus remains less privileged and of low density in digital language resources (Cieri et al., 2016; Magueresse et al., 2020). Nonetheless, Sesotho has developed as both a spoken and written language (Moeketsi, 2014; Koai and Fredericks, 2019), and is used in various domains.

Although there have been language audits of all eleven official languages of South Africa, as far as could be ascertained for this thesis, no study has paid in-depth attention to Sesotho digital language resources. Ensuingly, this section seeks to describe basic digital language resources, modules, and applications used in written and speech applications for Sesotho as a way of inventorising Sesotho digital language resources.

2.5.2 Literature review

Natural Language Processing (NLP) has been a very active area of research (Drake, 2003). It is aimed at both automatically analysing and representing human language using computational technologies (Cambria and White, 2014; Kang et al., 2020). These computational techniques are based on both theory and available technologies that enable them to learn,

¹⁰It is worth noting that South African Sign Language, often abbreviated as SASL, has been omitted from a significant number of Human Language Technology (HLT) audits. This omission can be attributed to its relatively recent approval as an official language in July 2023. Additionally, these audits have predominantly centred their attention on the eleven written languages.

from existing human language content, understand how the human language is produced, and produce human language content independent of humans (Liddy, 2001; Hirschberg and Manning, 2015). NLP research gathers knowledge on how humans use language and in turn, develop applications to enable computers to simulate and handle natural languages (Chowdhury, 2003). As such, NLP is a discipline within Artificial Intelligence (AI) and linguistics since it is characterized by human-like language processing capabilities (Drake, 2003; Nadkarni et al., 2011). Frequent applications of NLP include (i) information retrieval, (ii) information extraction, (iii) question answering, (iv) summarisation, (v) machine translation, and (vi) dialogue systems (Drake, 2003). Additionally, some NLP models cater for more than one language (Castellucci et al., 2021).

Carrying out NLP tasks relies on the availability of HLT resources developed using digital language resources. Unfortunately, none of the South African official languages has official Basic Language Resource Kits (BLARKs) as conceptualised in Krauwer (2003). Generally, a BLARK defines the minimal required resources for performing pre-competitive research in both spoken and written language for a specific language (Arppe et al., 2010). BLARKs are categorised into three, namely, definition, specification, and content (Maegaard et al., 2006).

First, the definition category indicates what should be regarded as basic in the given language. For instance, Arppe et al. (2016) present a matrix that indicates the resources needed for Plains Cree (an indigenous language of central Canada) together with their levels of importance.

Second, the specification category prescribes the quantities of the resources defined such as the Arabic basic resource specification by Maegaard et al. (2006) that lists the quantities of resources needed for the Arabic language.

The content category of a BLARK describes what already exists. One example of this is a survey of language applications available for Swedish (Elenius et al., 2008). In this section, we limit our discussion of basic digital language resources to BLARK content. That is, we inventorize written and spoken digital language resources available in and to Sesotho. Although the definition of BLARK content is language-independent, the resources identified and the level of importance allotted to the resources are specific to Sesotho (Krauwer, 2003; Arppe et al., 2010).

Amongst others, industry and education institutions benefit from the availability of BLARKs (Maegaard et al., 2005). In other words, the utilisation of BLARK enables the identification of existing content, as exemplified in this section, as well as the identification of areas for further development, aligning with our primary goal in this section. This ability empowers both industrial and academic developers to allocate necessary resources more effectively. Additionally, BLARK content facilitates the discovery of a wide range of resources, including paid and freely accessible ones, thereby simplifying the process for researchers and educators to access the materials available to them. When carrying out

BLARK content analysis, three BLARK aspects are evaluated, namely, (i) data, (ii) modules, and (iii) applications (Krauwer, 2003; Arppe et al., 2010).

2.5.3 Methodology

It is recommended that the South African Centre for Digital Language Resources (SADiLaR) purposefully manages all publications of language resources in South African languages (Moors et al., 2018c). SADiLaR is supported and funded by the South African National Department of Science and Innovation (Sefara et al., 2021). SADiLaR supports development and innovation in the official languages of South Africa. According to Wilken et al. (2018), SADiLaR also aims to facilitate access to digital data and software applications. To this end, it has a publicly accessible repository¹¹.

The BLARK content presented in this section is based on digital language resources indexed in the aforementioned repository. In the first stage of our research, we used the search functionality on SADiLaR’s website to extract the indexed language resources. A very broad search query: “Sesotho” was used to yield a total of 119 resources. A summary of these results is presented in Table 2.5. In the second stage, we searched Google Scholar for literature related to the digital language resources identified from SADiLaR’s repository.

Table 2.5: The results of our search queries.

Results	Sum
Total results	119
Modules and applications	41
Spoken corpora	15
Written corpora	21
Irrelevant results	42
Relevant results	77

Our analysis includes both general language resources that were developed for other languages and can be applied to Sesotho and resources that were specifically developed for Sesotho. We do this since overviews of this kind ought not to be limited to monolingual resources (Arppe et al., 2010). Search results were independently evaluated by the researchers and the results were comparatively analysed.

¹¹See <https://repo.sadilar.org/handle/20.500.12185/9> for the online repository.

2.5.4 Findings

As indicated above, we limit our discussion to BLARK content as defined by Maegaard et al. (2006). We discuss written resources, speech-based resources, modules, and applications available to and for Sesotho.

2.5.4.1 Written resources

This section discusses all digital language resources developed for Sesotho and those that are applicable to Sesotho. A total of 21 resources are described.

2.5.4.1.1 Monolingual lexicon Few written monolingual corpora were identified. One, the National Centre for Human Language Technology (NCHLT) produced four data sets focused on plain corpora, annotated corpora (Eiselen and Puttkamer, 2014), phrase chunks (Eiselen, 2016b), and named entity (Eiselen, 2016a). Two, a customised government domain-specific dictionary was identified (Bosch and Griesel, 2017). Furthermore, a genre classification corpus for Afrikaans, Sepedi, Setswana, isiXhosa, isiZulu, and Sesotho was also identified. The corpus is composed of poetry, advertisements, informational pamphlets, instructions, news, official texts like policies, and speech texts (Snyman et al., 2011).

2.5.4.1.2 Translated corpora Many language pairs lack enough parallel texts (Koehn and Knight, 2002). This is seemingly the same case in South African indigenous languages including Sesotho. Even so, there are few translations that use English as a pivot language. We identified a total of nine such word lists that are translated from English to the other official languages of South Africa. The word lists are presented in Table 2.6.

The word lists were commissioned by the National Language Services under South Africa’s Department of Arts and Culture. The word lists present singular word translations on different subject matters such as politics, education, and sports. Language and translation experts in different official languages of South Africa gather in a workshop setting and words are translated in groups. Group quality assurance workshops are then held to ensure quality translations. The authors of the article which this section is based on have been involved in these translation projects.

We also identified the Mburisano Covid-19 multilingual corpus that contains screening and triage vocabulary of Covid-19-related multilingual corpora in all official languages. The corpus was developed in response to the Covid-19 pandemic as an attempt to ensure access to information for people of different linguistic repertoires. English is used as the source text for all the other official languages. Unfortunately, the corpus is not clearly marked for specific languages.

Table 2.6: Translated word lists available in Sesotho.

Word list	Size
Arts and Culture for the intermediate phase	550
Election	559
HIV/AIDS	586
Information and Communication Technology	132
Life Orientation for the intermediate phase	1,628
Mathematics	984
Natural Sciences and Technology list	2,756
Parliamentary jargon	502
Soccer	297

2.5.4.1.3 Machine Translation We identified three Machine Translation (MT) corpora. Two of the corpora were produced as part of the Autshumato Machine Translation project. Both the multilingual word and phrase translations and the MT evaluation set contain aligned translations from English to the other ten official written languages. The third MT corpus, the Centre for Text Technology (CTeXT) multilingual text corpora, provides document-level aligned texts for MT purposes.

2.5.4.1.4 Thesauri, ontologies, wordnets Two written dictionary data sets were identified. One, the Bukantswe Sesotho-English bilingual dictionary word list contains a total of 10,085 bilingual segments in Sesotho and English. Each line presents the Sesotho term followed by the English translation and the relevant part of speech where available. Two, the Sesotho custom dictionary for the government domain is a word list that contains two types of words, namely, those that are exclusive to the government domain and those that do not follow official Sesotho orthographic conventions.

2.5.4.2 Spoken resources

We initially identified a total of eighteen speech-related resources. However, upon close scrutiny, we found that three of these resources were false results. One such instance is the Lwazi II Cross-lingual Proper Name corpus that is meant for Northern Sotho and not Sesotho (Kgampe and Davel, 2010, 2011). In the end, we discuss a total of fifteen spoken language resources.

2.5.4.2.1 Text to speech and automatic speech recognition Seven resources were identified in this category. Three sets of corpora were produced by the Lwazi project. Two of the corpora are purposed for text-to-speech (TTS) while one is aimed at auto-

matic speech recognition (ASR). The Lwazi and Lwazi II Sesotho TTS corpora contain transcriptions annotated with phonemic and orthographic information. The training sentences contain approximately equal speech sounds (Badenhorst et al., 2011). The Lwazi Sesotho ASR corpus contains audio and transcriptions used for the Lwazi speech recognition systems. The transcriptions are annotated with orthographic information for each word.

The NCHLT project also produced three TTS corpora. The inlang Pronunciation Dictionary for Sesotho contains an associated rule for generating pronunciations for unseen words (Davel et al., 2013; Barnard et al., 2014b). The speech corpus contains orthographically transcribed broadband speech corpora including a text suite of eight speakers (De Vries et al., 2014). We also identified the auxiliary speech corpus (Barnard et al., 2014b; De Vries et al., 2014; Badenhorst et al., 2019).

Lastly, we identified the Sesotho multi-speaker TTS corpus that contains audio and annotated transcriptions created for investigating the implementation of a high-quality TTS system that uses a low-cost process. The data sets were quality-checked. However, the read.me file indicates that there might still be some errors. Unfortunately, accuracy results were not reported.

2.5.4.2.2 Sound-based corpora In this category, we discuss pronunciation-based and tone-based corpora. Four tone-based corpora were identified. First, the Sesotho vowel speech data set contains a collection of words that represent five Sesotho orthographic vowels (a, e, i, o, u). The speech data was recorded with seven female and three male participants. Second, the metadata indicates that the intonation model for Bantu tone languages contains a model for isiZulu, Sepedi, Setswana, and Sesotho. The model is intended for theoretical linguistic intonation rules in prose. Unfortunately, the model has not been uploaded to the repository. As such, we are unable to report on its contents. Third, the Sesotho tone data set also contains male and female audio recordings. The participants were sourced from a specific region of South Africa called Qwaqwa. Fourth, the Sesotho function word speech data corpora contains audios and annotated transcriptions aimed at studying the role of tone in *ke* and *o* as function words.

One pronunciation-based corpus was identified. The South African Multilingual Proper Names Corpus (Multipron) was developed in response to accent-based variation in the pronunciation of personal names (Giwa et al., 2011). This corpus uses different speakers from isiZulu, Afrikaans, English, and Sesotho. Four participants read Sesotho words. Sesotho words comprise 15% of the total data collected in this corpus. Each directory consists of the orthographic transcription in a text file, phonemic transcription containing phoneme strings and an audio file consisting of an acoustic representation of each word.

2.5.4.2.3 Dictation and telephony data We identified a total of three digital language resources in this category. First, the South African Directory Enquiries (SADE) Name corpus (Thirion et al., 2020), uses a Sesotho home language speaker for accent, but the words used for training the voice dictation platform do not contain Sesotho names. Second, we identified two telephony speech data sets. The high-quality TTS data for Afrikaans, Setswana, isiXhosa, and Sesotho contains multi-speaker TTS audio data and transcription files. The African Speech Technology Sesotho speech corpus contains speech spoken by Sesotho mother tongue speakers (Roux et al., 2004). Third, the SADE municipality hotline IVR prompts corpus contains audio and corresponding transcriptions in English, isiZulu, and Sesotho (Van Heerden et al., 2014). The produced recognition system recognises pronunciations with Afrikaans, English, isiZulu, and Sesotho accents. The interface can also be customised to isiZulu or Sesotho.

2.5.4.3 Modules and applications

This section discusses a total of 40 modules and applications available to Sesotho.

2.5.4.3.1 General applications Six non-language-specific applications were noted. First, the Autshumato PDF Text Extractor is used for extracting texts for translation using the Autshumato automatic translation machine. It functions as a plugin for the OmegaT computer-assisted translation application. The translation system is named after Autshumato, possibly South Africa’s first official translator and interpreter (Groenewald and Fourie, 2009; Skosana and Mlambo, 2021). Autshumato is one of the South African government’s initiatives for improving multilingualism through an increase in both quantity and quality of translation services (Groenewald and Fourie, 2009). Unfortunately, the quality and accuracy of the machine translation web service is low (Skosana and Mlambo, 2021). Nonetheless, if a translator translates similar documents, they can save the translation memories (TMs) and rely on them for further translations.

Second, we identified the Autshumato translation memory exchange (TMX) integrator which works as a utility that enables merging multiple TMs over networks through subversion (Schlemmer and Fourie, 2013). Unfortunately, translation memory sharing is not yet common practice. To this end, we cannot estimate the possible contribution that individual translation projects can make towards building a bigger and more reliable translation memory for Sesotho translations. Furthermore, the TMX integrator only supports translation from English to Sesotho and not from Sesotho to English (Reina et al., 2013).

Third, the Dictionary Maker (Davel and Barnard, 2003), has been evaluated in German, but we hope that it can also be applied to Sesotho texts. The Dictionary Maker allows the user to develop a pronunciation dictionary. When used, the human effort needed to develop such a dictionary is marginally decreased. Moors et al. (2018c) indicate that there are three pronunciation dictionaries for Sesotho, one identified in 2009 and two identified in the

2014 audit, namely, the NCHLT-inlang, Lwazi, and Lwazi II pronunciation dictionaries.

Fourth, we identified corpus-related applications such as (i) the CorpusCatcher, designed to crawl the web for data using seed documents for constructing queries for document retrieval, (ii) Spelt, used in the creation of classified word lists that are used in spell checking, and (iii) TurboAnnotate1.0, used for manual creation of gold-standard and annotated lists. According to Van Huyssteen and Puttkammer (2007), this application lowers human effort and improves the accuracy of annotation. It uses the Tilburg Memory-Based Learner machine learning system (see Daelemans et al. (2004)). It allows mother tongue speakers with limited and no experience with computational linguistics to annotate texts. Machine learning then learns from the user-generated annotations.

2.5.4.3.2 South African official languages In this section, we discuss applications that were developed for all eleven South African official languages. We identified at least 24 applications in this category with some applications having newer versions and others having basic and professional versions.

Five NCHLT products were identified, namely, the Optical Character Recognition (OCR), language identifier, text web service, tagger, and POS tagger. The OCR for South African Languages (Hocking and Puttkammer, 2016) enables the user to convert scanned documents into editable texts. It can reproduce almost any character or image. The language identifier uses both a graphical user interface and a command line interface for automatically identifying official South African languages. The text web services provide access to tokenisers, sentence separators, POS taggers, phrase chunkers, named entity recognisers, and OCR in South African languages. The tagger can be used either through the command line or a user interface. It annotates running text with either POS, named entity, noun phrase chunks, or nouns. Finally, the POS tagger was developed using a minimum of one million government-published tokens per language (Eiselen and Puttkamer, 2014).

Four CText products were identified. First, the Alignment Interface and the Alignment Interface Pro are utility applications used for aligning source texts. The Pro version allows for editing the segments. These products work with the CText Applications 1 that allows for automatic corpus query and manipulation for tokenisation and sentencisation, frequency and word list extraction, searching and extracting collocations. Additionally, CText Applications 2 adds POS tagging, named entity recognition, and phrase chunking.

We also identified six independent applications, namely, (i) the AStudio, a software that incorporates a graphic interface for developing flowcharts for speech applications, (ii) the Automatic Oral Proficiency Assessment application, developed as part of the Development of Resources for Intelligent Computer-Assisted Language Learning project, (iii) the Language Identifier (LID) classifier token level classification for all official languages, (iv) a Combination Tagger that uses memory-based tagger (MBT), support vector machines (SVM), Mobotix part of speech tagger (MXPOST), and Trigrams'n'Tags (TnT) for de-

ciding on tags, (v) the South African Fonts collection that contains fonts representing all alphabets, letters, and characters used in South African official languages, and (vi) the Format Normaliser 1.0. for normalising input files to UTF-8 txt, replacing smart quotes, and removing empty lines.

Four translation resources were also identified. One, the Rhonda machine translation system can handle speech-to-speech translation or speech-to-text translations. Two, the Translate application kit 1.4.0 is a collection of applications and parsers that handles various localisable and translatable formats. It composes modules for segmentation, authentication, and text enumeration. Three, the Autshumato Translation Management Systems web applications allow for capturing, editing, exporting, and importing terminologies. Four, the Autshumato Text Anonymiser classifies and replaces sensitive information. For instance, if one wishes to anonymise sensitive information such as study participants' names, this application replaces them with pseudo names.

Five TTS-related applications were also identified in this category. One, the Lwazi Telephony Platform combines Asterisk with MobillVR *Python* interface in one unified control interface. In this way, the process of developing experimental applications is accelerated. Two, the Qfrenzy TTS phone mappings application maps are used with Lwazi and NCHLT pronunciation dictionaries of the official languages. Three, the multilingual TTS Speect system provides a full service of decoding and encoding texts, that is, text analysis and speech synthesis with various APIs. Furthermore, it can be used for research and development of TTS system voices. Four, the Phonetic aligner contains scripts for automatic phonetic alignment of speech corpora using the Hidden Markov Model. Finally, the Text Selection scripts for ASR/TTS, use phonetic rules to phonetise texts, then the diphones are used for TTS and triphones are used for ASR on a per-language basis.

2.5.4.3.3 Language specific resources We identified four language-specific applications that also incorporate Sesotho. One, the EtsaTrans translation system user interface is available in English, Afrikaans, isiXhosa, and Sesotho. Two, the Multilingual Illustrated Dictionary with interactive games application is available for seven official languages, namely, Afrikaans, South African English, isiXhosa, isiZulu, Sepedi, Setswana, and Sesotho. Three, the NWU TransTips 1.0 is a PHP programming script that browses the web page for terms in the database. The translations appear when the user hovers over a certain word. Unfortunately, the user still has to decide on the correct translation between those presented. Lastly, the Automated multilingual telephone access to financial services is a prototype that allows switching between isiZulu, isiXhosa, English, and Sesotho.

2.5.4.4 Sesotho-only

We identified five applications that are specifically for Sesotho. First, we identify two NCHLT products. One, the lemmatiser was developed using a rule-based approach (Eiselen and Puttkamer, 2014). Two, the Morphological Decomposer that splits tokens into morphemes was developed after the POS tagger (Eiselen and Puttkamer, 2014). Second, the Lwazi Sesotho Pronunciation Dictionary includes audios for phonemes and letter-to-sound rule set based on generic words (Davel and Martirosian, 2009). Although accuracy is not guaranteed, practical usability is assured. Third, the Lwazi II Sotho Pronunciation Dictionaries (Du Plessis et al., 1974) are based on the Lwazi dictionaries. Fourth, Spelling Checker 1.0 is an application that checks word spelling and provides hyphenated words. It is compatible with some versions of Microsoft Office. CTeX continues to improve this application.

2.5.4.5 Games

We identified one game, the Open Spell (v1.0), a spelling game that contains spelling exercises aimed at teaching spelling skills to school children. The source code is also freely accessible.

2.5.5 Discussion and conclusions

Many LRLs are underrepresented in NLP tasks because of insufficient corpora needed to complete NLP tasks (Mahloane and Trausan-Matu, 2015). The limited availability of corpora as indicated in this section shows a great need for curating more corpora for Sesotho. This section considered basic digital language resources available to Sesotho from a BLARK content perspective.

The listed resources were listed without an in-depth analysis of each resource such as considering how each resource works or the levels of accuracy and practical issues such as user-friendliness. This is typical of BLARK content studies. Although this is basic, it functions as a starting point for investigations of BLARKs specification and definition, i.e., determining what should be regarded as basic in the language of Sesotho and the quantities that should be developed. Nonetheless, it is clear from this section that written resources are focused only on very basic functions. For instance, there are no resources for higher-level basic functionalities such as semantic analysis and term extractors. Even so, only the handwritten OCR and ontologies are missing from the written resources as identified in the BLARKs (Krauwier, 2003; Maegaard et al., 2005, 2006). There have been even fewer attempts at speech technologies, especially those that are specifically aimed at Sesotho or the Sesotho language group. Even so, it is interesting to see how much work has been achieved.

A narrowed investigation into the current resources and an evaluation of each technology should be considered for future studies. SADiLaR is currently conducting its routine language resource audit, more resources might be added to its repository after their investigation. Perhaps we may gain more insight into digital language resources available to Sesotho. Even so, there are two main issues with BLARKS. First, availability is not a binary distinction in that some existing resources may still be inaccessible due to financial or copyright issues (Krauwert, 2003). Second, the multiplicity of resources may not be equated to usability, quality output, or user-friendliness.

This discussion focused on three key aspects. Firstly, it should be noted that the resources surveyed in this review were not extensively evaluated in terms of their usability, accuracy, and precision. Secondly, while some of the resources were listed and indexed in the repository, their actual accessibility was limited. Thirdly, it is important to acknowledge that the section exclusively reviewed resources indexed by SADiLaR.

It is important to recognize that the current index reflects variations in the existing collection of resources. It is, therefore, recommended that future developments of digital language resources for Sesotho take this inventory into consideration when deciding which resources to prioritize. This way, attention can be given to the creation of new and currently unavailable resources.

2.6 Summary

This chapter delved into the methods, techniques, and approaches employed for measuring text readability in under-resourced languages, with a specific emphasis on Sesotho. Initially, an overview of the writing system in Sesotho was presented, considering the relative recency of writing in the language and the variations in orthographies between South African Sesotho (SAS) and Lesotho Sesotho (LS). Furthermore, this overview underscored the necessity for the development of foundational language resources, as demonstrated by the limited accessibility of existing digital language resources in Sesotho.

We also conducted a systematic literature review to identify the main approaches used for measuring text readability in LRLs like Sesotho. We identified three main approaches: machine learning, cross-lingual, and the modification of existing readability measures. However, due to the lack of resources and annotated data needed for Sesotho, we opted to adapt existing readability measures instead of using machine learning or cross-lingual approaches.

To inform our adaptation of existing measures, we provided an overview of nine traditional English readability measures and the text properties they used to assess readability. We then surveyed existing resources for Sesotho to identify the tools and resources that are available and those that are still needed. Our survey revealed that there are very few

corpora with annotations in Sesotho, and most available corpora are translations rather than original texts. Additionally, two resources were missing, namely, a list of frequently used words in Sesotho, and a system for identifying syllables from Sesotho words.

In summary, this chapter provided an overview of the methods and approaches used in measuring text readability in LRLs like Sesotho, highlighting the challenges and limitations of using existing measures and resources. Our reviews emphasise the need for the adaptation of existing measures and the development of new resources for Sesotho.

This chapter serves as the foundation for a significant conclusion. It becomes evident that even when utilising generic or non-language-specific resources to extract text readability properties from Sesotho texts, at least five of the measures examined in this thesis are rendered inaccessible. Specifically, the four measures reliant on syllable information and the DCI face challenges due to the absence of reliable systems for extracting syllable information and a comprehensive list of frequently used words in Sesotho. Thus, these limitations underscore the need for tailored digital language resources to accurately assess text readability in Sesotho.

Chapter 3

Developing digital language resources

This chapter discusses the development of digital language resources produced with the specific aim of developing readability measures for Sesotho. The resources include corpora, a lexicon, and applications. The measurement and comparison of the text readability levels of English texts that are machine-translated and post-edited serve as a means to determine the usability of the machine translation texts.

3.1 Introduction

As exemplified in Chapter 2, Sesotho as a LRL presents unique challenges for the adaptation of existing readability measures, including the availability of written resources such as corpora, the divergent nature of orthographies, and the lack of syllabification systems applicable to Sesotho text. In the previous chapter, we explored the general methods, techniques, and approaches used for measuring text readability in LRLs, with a particular focus on the potential application of these methods in Sesotho. The previous chapter also highlighted the limitations of existing resources and the need for further development of these resources to effectively measure text readability in Sesotho.

This chapter builds on the foundation established in Chapter 2 by addressing the second research question. That is, this chapter describes the development of basic digital language resources needed for developing traditional readability measures in the low-resource Sesotho. To structure the discussion herein, the discussion of the digital language resources developed is divided into five sections.

First, we describe a collection of reading comprehension and summary writing texts in Section 3.2. The texts comprise the final year high school exam texts for Home Language (HL) and First Additional Language (FAL) subjects written in South Africa between 2008 and 2020. The text collection contains texts from eleven official South African language subjects: Afrikaans, English, isiNdebele, isiXhosa, isiZulu, Sesotho, Setswana, Sepedi, Siswati, Tshivenda, and Xitsonga. PDF versions of the texts were downloaded from South Africa’s Department of Basic Education online public access repository. The data collection contains a total of 429 text files comprising a total of 1,314,551 tokens with 131,847 types (i.e., unique tokens). Of these, 223 are HL texts that have 689,730 tokens and 88,137 types, whereas the 206 FAL text documents contain 624,821 tokens with 73,562 types. In addition to the full exam texts, the reading comprehension and summary writing texts are extracted manually. This results in a total of 935 texts (485 for HL and 450 FAL texts). The extracted HL texts consist of 263,497 tokens and 58,853 types, whereas the extracted FAL texts consist of 195,739 tokens and 45,845 types.

Second, we describe the development of a list of frequently used words in written Sesotho in Section 3.3. The list is created with the aim of incorporating it into the adaptation of the DCI which uses a list of commonly used words. Accordingly, the section is contextualised against the background of the need for readable texts. That is, it discusses the issues of reading that are facing learners in South African basic education schools. Such an overall low literacy of learners exacerbates the challenges of choosing texts in a language that is already lacking in reading passage choices. Even so, it is assumed that some words are consistently used in the different domains and should thus be easier to read as language learners are used to seeing them in texts. Since we could not access school-graded texts beyond exam texts, we utilise translated texts. Three corpora comprising Bible text (818,311 tokens), machine translation texts (4,600,267 tokens), and high school exam

texts (47,233 tokens) are used in the creation of a corpus-based list of 3,000 frequently used words in Sesotho.

Third, we describe the development of a corpus of translated texts in Section 3.4. To overcome the challenge posed by the scarcity of gold-standard annotated corpora for readability assessment in LRLs, specifically focusing on Sesotho, we propose using Google Translate to construct a readability-labelled corpus and evaluate the effectiveness of the translations by comparing them to post-edited versions. The translations are evaluated using BLEU, NIST, and RIBES scores. Furthermore, the Ghent University readability demo is used to extract readability levels. The translation evaluation measures indicate substantial similarity between the machine translations and the corresponding post-edited versions. Furthermore, the results demonstrate a high level of consistency between machine translations and human-post-edited machine translations in terms of readability measures and text properties. The findings in this section suggest that Google Translate show promise in addressing challenges in developing readability-labelled data sets in LRLs like Sesotho, highlighting the potential of leveraging machine translation techniques to develop translated corpora for such languages. The evaluation of Google Translate in the context of educational texts in Sesotho and the demonstration of the feasibility and potential of using machine translations for enhancing readability in LRLs, specifically Sesotho, will aid in the quest for developing readability measures in Sesotho.

Finally, we describe the development of two syllabification systems in Section 3.5. First, we implemented a rule-based syllabification system based on three syllabification rules proposed by Guma (1982) and explained with examples in Sibeko (2022b). The rules govern the placement of vowels, consonants, simple and complex nasal consonants, and the lateral *l*. Second, we used $\text{\TeX}$'s pattern-based hyphenation system as a syllabification system. We also provide a brief discussion of the development of a word list that contains Sesotho words and their syllable information in Section 3.5.2.1. The syllable boundaries in the words from the dictionary were manually checked and, where needed, adjusted to maintain a consistent orthography. In the end, the pattern-based system was trained on the syllabified words from the Sesotho dictionary (Chitja, 2010). Both systems were evaluated on 13,551 tokens using ten-fold cross-validation (as the pattern-based system requires training data). For consistency, we used the same folds for the evaluation of the rule-based system as well. The rule-based system yields a 99.69% word level accuracy, which indicates that Sesotho syllabification is highly regular. The pattern-based system leads to 78.97% accuracy.

Through this chapter, we aim to highlight our contribution to the development of Sesotho's basic language resources. The resources developed in this chapter are then used in the adaptation of English traditional readability measures into Sesotho. We also use the corpora to analyse text readability in both English and Sesotho in Chapter 4.

3.2 Corpus 1: Final year reading comprehension and summary writing examination texts

This section is adapted from: Sibeko, J. and 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(1), p. 9. <https://doi.org/10.5334/johd.108>.

3.2.1 Introduction

Eleven of South Africa's official languages are offered as language subjects from pre-school (grade R) to final year (grade twelve) in South African schools (although not all schools offer all languages). Each official language can be taken either as a home language (HL), first additional language (FAL), or second additional language (SAL)¹ (Department of Basic Education, 2012, 2013). The HL subject is aimed at learners who start school with language competency skills such as listening, speaking, reading, and writing (Department of Basic Education, 2011b). The FAL subject is proposed for learners who start school with some exposure to the language (Department of Basic Education, 2011a, 2012, 2016), whereas the SAL subject is intended for learners who start school with no exposure or competency skills in the language (Department of Basic Education, 2011c; Horne and Ferreira-Meyers, 2017). The content, teaching schedule, and overall curricula for the language subjects are governed by the Curriculum and Assessment Policy Statements (CAPS). Curricula for other official languages are translated from the generic English CAPS (Van der Walt, 2010; De Vos et al., 2014; Tshesane, 2014; Probert and De Vos, 2016; Van der Walt, 2018).

At the end of the high school curriculum, grade twelve summative examinations are written for the different language subjects. First, these examinations are written around November, which is at the end of the academic year. Second, until 2018, supplementary examinations were written around February or March following the final year. Third, as a pilot, an additional examination opportunity was provided in May for learners who could not write all subjects in the final year November examinations (Department of Basic Education, 2019).

In the data collection, exam texts of the November examination opportunities are provided for the years 2008 to 2020. For February opportunities, texts are available for the years 2009 to 2018, excluding 2010, while for the May opportunities texts range from 2016 to 2019. Note, however, that for some languages, exam texts for certain years are not available, see Table 3.1 for an overview.

¹HL is often called first language, and FAL second language, while SAL is considered a foreign language (Horne and Ferreira-Meyers, 2017). However, any learner can enrol for any language level. As such, using terminology such as L1, L2, and foreign language, is not suitable.

Table 3.1: The distribution of texts per language and subject level.

	# of exam texts			# of extracted texts		
	HL	FAL	Total	HL	FAL	Total
Afrikaans	21	22	43	57	59	116
English	24	25	49	56	60	116
IsiNdebele	20	16	36	43	34	77
IsiXhosa	19	21	40	39	40	79
IsiZulu	18	18	36	36	39	75
Sepedi	22	20	42	48	42	90
Sesotho	22	19	41	49	39	88
Setswana	17	14	31	34	29	63
Siswati	21	18	39	43	38	81
Tshivenda	19	17	36	39	37	76
Xitsonga	20	16	36	41	33	74

Unfortunately, SAL exam texts are excluded from this data collection for the following reasons:

1. SAL exam texts are scarce and most could not be located for the purpose of this collection,
2. SAL examinations are set at the provincial level, while the HL and the FAL examinations are set at the national level. As a result, it is not clear whether the same rigorous processes followed in the preparation of the exams for HL and FAL are also followed (as such, unlike the grade twelve HL and FAL exam texts, the SAL exam texts cannot be generalised (Sibeko, 2021)), and
3. In most instances, second additional languages are non-official languages (Department of Basic Education, 2011c).

At each of the HL and FAL examination opportunities, three papers are written (Department of Basic Education, 2011a,b). The first paper assesses reading comprehension, summary writing skills, visual literacy and language conventions (Department of Basic Education, 2011a,b). The second paper evaluates literature analysis skills and covers genres such as poetry, drama, novels, and folklore (Department of Basic Education, 2011a,b). The third paper assesses creative writing skills and covers pieces such as essays and transactional texts (Department of Basic Education, 2011a,b,c).

The data collection described here contains both the full exam texts as well as the specific texts used for reading comprehension and summary writing (both are part of the first

paper).

3.2.2 Methods

The data collection process consisted of four steps. First, PDF files of the examination papers were downloaded from South Africa’s Department of Basic Education website². These files (like all other files in the data collection) are manually organised per language, per subject level, and per exam opportunity. The question papers on the online repository are in PDF format. Second, plain text from the PDF files was extracted using *pdftotext* (version 0.86.1) on a Linux platform. Third, the plain texts were tokenised (and sentence-cised) using *Ucto* (version 0.14). Fourth, the reading comprehension and summarisation texts were manually extracted from the tokenised plain text files. The names of all text files contain relevant metadata (language, subject level, year, month, and file type).

3.2.3 Data description

The data collection contains 429 PDF files and corresponding UTF-8 Unicode-encoded text files. Additionally, tokenised versions of the text files are provided (where tokens are separated by spaces and sentences using a <utt> marker). From these files, reading comprehension and summary writing texts are extracted and provided as separate text files (comprising a total of 935 UTF-8 Unicode files). All files are first divided into their file type (**pdf** for PDF files, **txt** for the extracted UTF-8 Unicode files, **tok** for the corresponding tokenised files, and **ext** for the manually extracted reading comprehension and summarisation texts). Next, the files are divided into their corresponding languages. Files for eleven official South African languages are provided: **Afrikaans**, **English**, **IsiNdebele**, **IsiXhosa**, **IsiZulu**, **Sesotho**, **Setswana**, **Sepedi**, **Siswati**, **Tshivenda**, and **Xitsonga**. Next, the files are divided into two subject levels, namely, Home Language (HL) and First Additional Language (FAL). Next, the files are divided into three examination months, namely, February or March, May or June, and November, respectively in **Feb**, **May**, and **Nov** directories. The files have a consistent naming scheme: *lang_subj_P1_month_year.ext* with *lang* the name of the language, *subj* the name of the subject level, *month* either **Feb-March**, **May-June**, or **Nov**, depending on the months the exams were written, *year* ranging from 2008–2020. *ext* represents the type of the filename, either **txt** for text files, or **pdf** for PDF files. For the extracted reading comprehension and summarisation files (found in the **ext** directory), before the extension (*.ext*), *_type* is present. This *type* can take the values **RC1**, **RC2** for the first and second reading comprehension texts respectively, or **SUM** for the summarisation texts.

Table 3.1 provides an overview of the distribution of the files in the data collection. Table 3.2 provides an overview of the token and type counts of the full exam texts, whereas

²See <https://www.education.gov.za>.

Table 3.2: Token and type count per language and subject level for the full exam texts.

	# tokens in exam texts			# of types in exam texts		
	HL	FAL	Total	HL	FAL	Total
Afrikaans	77,787	90,731	168,518	8,941	6,940	12,820
English	80,252	86,113	166,365	8,497	7,489	12,325
IsiNdebele	48,931	37,518	86,449	12,430	9,718	18,902
IsiXhosa	50,480	53,518	103,998	13,529	13,488	23,136
IsiZulu	43,456	44,082	87,538	11,738	11,071	19,721
Sepedi	66,846	56,594	123,440	5,709	5,253	8,374
Sesotho	80,592	66,934	147,526	6,897	5,811	9,735
Setswana	52,836	42,026	94,862	5,587	4,580	8,105
Siswati	54,597	43,845	98,442	14,608	10,787	21,863
Tshivenda	62,726	52,881	115,607	5,830	4,767	8,101
Xitsonga	71,227	50,579	121,806	5,831	4,446	7,933

Table 3.3: Token and type count per language and subject level for the extracted reading comprehension and summary writing texts.

	# tokens in exam texts			# of types in exam texts		
	HL	FAL	Total	HL	FAL	Total
Afrikaans	28,000	23,650	51,650	5,764	3,931	8,008
English	32,300	26,204	58,504	6,053	4,823	8,571
IsiNdebele	17,312	12,174	29,486	7,854	6,126	12,250
IsiXhosa	18,792	16,539	35,331	8,879	8,170	14,963
IsiZulu	17,709	14,790	32,499	8,165	6,748	13,188
Sepedi	26,573	18,565	45,138	3,952	3,413	5,820
Sesotho	28,856	20,249	49,105	4,428	3,375	6,219
Setswana	20,856	13,919	34,775	3,659	2,723	5,280
Siswati	19,957	12,705	32,662	8,854	6,171	13,276
Tshivenda	27,610	19,857	47,467	3,732	2,826	5,127
Xitsonga	25,532	17,087	42,619	3,880	2,870	5,401

Table 3.3 provides the same information for the extracted reading comprehension and summarisation texts.

Once downloaded, organised, and processed, the data collection of the PDF and raw (UTF-8 Unicode) text files is combined into one ZIP file, which is made available under a Creative Commons License Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)³. The collection can be accessed on the South African Centre for Digital Language Resources online repository⁴.

3.2.4 Summary

The data comprises final year (grade twelve) high school exam texts for eleven official languages of South Africa. Exam texts on reading comprehension and summary writing from both the home language and first additional language subjects for each language are provided. Currently, no such corpora in the South African education context exist. The data collection was developed to enable the evaluation of the readability of both English and Sesotho texts by making use of longer reading comprehension and summary writing texts in this thesis. Beyond this thesis project, the data can be used for a range of investigations, including comparative analyses between subject levels (home language versus first additional language), exam dates, as well as between languages. The investigations may focus on aspects such as text readability, text difficulty, or text complexity. The data consists of a specialised genre (exam texts) consistent for all eleven official South African languages and allows for a controlled comparison between the languages. The plain texts can be annotated further (e.g., on part of speech, partial parsing, named entities, etc). This allows for investigations on several textual properties. Finally, texts of languages of both disjunctive and conjunctive orthographies are represented. Researchers investigating the influence of orthography on, for instance, readability and comprehension can use these data.

For the purpose of this thesis, we pay special attention to reading comprehension and summary writing exam texts in both Sesotho and English.

3.3 Corpus 2: A list of common words in Sesotho

This section is adapted from: Sibeko, J. and De Clercq, O., 2023. A corpus-based list of frequently used words in Sesotho. In *Proceedings of the Fourth Resources for African Indigenous Languages workshop*, 2–6 May 2023, Dubrovnik, Croatia. pages 32–41. Association for Computational Linguistics.

³See <http://creativecommons.org/licenses/by-sa/4.0/>.

⁴See <https://repo.sadilar.org/handle/20.500.12185/568>.

3.3.1 Introduction

This section is premised on the issues surrounding learners' reading abilities and the importance of developing measures to aid assessors in selecting appropriate texts for learners. To overcome the challenges associated with participant accessibility, we have opted to utilise freely accessible corpora to compile a list of commonly used words in Sesotho.

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 (2019a,b), 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 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-than-desired 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 measures 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 measures, 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 measures for Sesotho. One

solution to the lack of readability measures may be the development of traditional readability formulas. Traditional readability measures 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). As indicated in Chapter 2, 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.

The development of the systems for extracting syllable information is presented in Section 3.5 of this current chapter. The presence of the systems entails that syllables can be extracted from Sesotho texts thus enabling the adaptation of syllable-based measures such as the FKGL, FRE, SMOG, and GFI into Sesotho. Even so, there is still no list of frequently used words to use in readability measures such as Dale and Chall’s (1948) DCI. Their strategy for compiling their list of frequently used words was to directly present these words in reading to fourth-grade learners, a word made it to the list when it was known by at least 80 percent of the participating 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. However, relying solely on a user-based technique may lead to limited outcomes due to the low reading abilities and limited vocabulary knowledge of the majority of learners in indigenous African language classes (Stoffelsma, 2019a,b). This user-based approach would primarily result in a list of basic entry-level words suitable for early readers. Additionally, we have insufficient information regarding the methodology employed in compiling the initial list of candidate words for inclusion in the Dale-Chall list.

3.3.2 Background

This section presents our efforts to create a list of the 3,000 most frequently used words in written Sesotho using a corpus analysis approach. We position our work in research on adapting readability measures. We then present our methodology for creating the list (Section 3.3.3) and subsequently discuss the results (Section 3.3.4). We conclude this section by highlighting the strengths of the list and by formulating a set of suggestions for further improvements (Section 3.3.5).

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., 2014a; Moors et al., 2018b,d). More specifically, we reviewed the basic language resource kit for Sesotho in Chapter 2 and concluded that many necessary digital language resources are still lacking.

An investigation by Van Oosten et al. (2010) revealed that the outcomes of traditional formulas developed for English, Dutch, and Swedish on Dutch data yield strong correla-

tions, which is explained by the formulas' strong reliance on certain language-independent properties, such as average word or sentence length. This implies that readability measures 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 traditional formulas also comprise variables based on frequency lists. The most well-known formula in this respect is the DCI (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 by Chall and Dale (1995).

The DCI is one of the most used measures 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 explore the creation of a frequency list 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.

Currently, there exist at least two lists of frequently used words in Sesotho, namely the MostCommonWords⁵ and the Waston Chen⁶ 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

⁵See <https://3000mostcommonwords.com>.

⁶See <https://wastonchen.com/6417.html>.

list is needed and opt for a corpus-based approach in this paper.

3.3.3 Methodology

We collected a total of three corpora which comprise original Sesotho text material, see Table 3.4 for some corpus statistics.

Table 3.4: The number of tokens and sentences present in the three Sesotho corpora used for this section.

	# 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

3.3.3.1 Corpus 1: Bible

We extracted text from the Sesotho (Southern Sotho) Bible version SSO89⁷. All texts were extracted from the SQL3lite format using *bash* scripts. The Bible texts are divided into three sections, namely, (i) Bible books containing 66 rows of data, (ii) verses containing 31,171 rows of data, and (iii) info containing ten 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 tokenised 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 one million tokens. We acknowledge the inherent bias associated with using religious bible texts. Even so, as exemplified in studies such as Agic and Vulic (2019) and Ifeoluwa Adelani et al. (2022), we leverage bible texts to augment the deficiency of corpora for Sesotho. We also acknowledge that the original orthographies of Sesotho were developed using bible texts (Makutoane, 2022).

3.3.3.2 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

⁷The Bible texts were downloaded in SQL3lite format from https://www.ph4.org/b4_index.php#google_vignette.

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 section, 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 four text types, namely magazines, policies, newsletters, translation works, and documents crawled from the government ([gov.za](https://www.gov.za)) domain. We are therefore confident that the corpus comprises different genres.

3.3.3.3 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 eleven 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](https://www.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 corpus⁸ (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 tokenised all text material with *Ucto*. Basic settings were used with the sentence segmentation option. In the end, the NCHLT corpus comprises around two million tokens.

3.3.3.4 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 normalised 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 3,000 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

⁸The corpus is available from <https://hdl.handle.net/20.500.12185/336>.

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 to end up with a clean list.

3.3.4 Results and discussion

As mentioned above the three frequency lists were merged and relative frequencies were calculated for duplicate entries. Afterwards, a qualitative analysis was carried out on all entries, which is presented next.

3.3.4.1 Proper names

The DCI 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.

3.3.4.2 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 contained ‘v’ letters such as *thelevishene* ‘television’.

Taljad (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 (Taljad, 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 3,000 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.

3.3.4.3 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.

3.3.4.4 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* appeared in the final list of 3,000 words and were thus 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 (refer back to Section 3.3.4.2 for our treatment of varied spellings).

Acronyms concatenate words into one. For instance, the World Health Organisation is counted as one word when abbreviated as WHO (Funk, 1968). For consistency, we removed all instances of abbreviations and acronyms. 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 3,000 most frequently used words before manual editing. Both AIDS and ‘*eitsi*’ are counted as one entry with varied spellings.

3.3.4.5 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.

3.3.4.6 Specific variations

Since our aim is to adapt the English DCI measure into Sesotho, we also consider the composition of the Dale-Chall list used in the English measure⁹. Even though not all words and their variations are listed in the list of frequently used words when computing the scores, specific variations are excluded from this list of difficult words. We focus on two such variations, namely: verbs and adverbs.

First, Barry and Stevenson (1975) indicate that 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 *getwa/getuwa* ‘finished’ is derived from *geta* ‘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.

⁹See <https://github.com/words/dale-chall>.

3.3.4.7 Final list of common words

After all these manual steps, we end up with a list of 3,000 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).

Please note that in order to allow for spelling variation of certain words (please refer back to Section 3.3.4.2 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 3,037. The list is made available to the research community at the repository of SADiLaR <https://repo.sadilar.gov>. The list is also attached as Appendix A.

3.3.5 Conclusions

This section discussed the development of a list of frequently used words in Sesotho using a corpus-based approach. Three different corpora were employed to this end. Although possibly unconventional, 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). 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 SAS and LS 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

section allows for taking further steps towards developing traditional readability measures which use frequency lists, such as the DCI. 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 *mopotso* ‘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.

Nevertheless, we are exploring other user tests that could make our list applicable to other target audiences, such as adult readers, authors, and publishers.

3.4 Corpus 3: Machine translation texts

This section is adapted from: Sibeko, J. 2023., Harnessing Google Translations to Develop a Readability Corpus for Sesotho: An Exploratory Study. *Under consideration by the Journal of the Digital Humanities Association of Southern Africa, submitted 15 August 2023.*

3.4.1 Introduction

A gold-standard corpus with clear levels of text difficulty is needed to develop an automated readability model (Van Oosten et al., 2010; François and Fairon, 2012). In fact, previous studies have emphasised the necessity of gold-standard corpora when adapting text readability measures from HRLs to LRLs (Van Oosten and Hoste, 2011; François and Fairon, 2012; Curto et al., 2014, 2015). Unfortunately, Sesotho, like other LRLs, does not have corpora readily annotated with levels of difficulty (Filho et al., 2016). As was highlighted in Chapter 2, the use of translated texts may provide a solution to this lack of levelled texts. For instance, texts translated to English from other languages can be assumably levelled according to grades in English. However, in our search for this section, we found a lack of English-Sesotho-aligned corpora labelled for text readability, thus hindering the availability of gold-standard corpora. In light of this, we propose an alternative approach by utilising translated texts to measure text readability in Sesotho. While human-translated and edited texts are considered ideal, they can be costly (Serhani et al., 2011). Therefore, we suggest employing machine-translated texts to create 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.

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 section, we have chosen Google Translate, 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). Therefore, while Google Translate continues to improve its grammatical accuracy, its translation accuracy, especially for African languages, remains a concern (Patil and Davies, 2014; Tsai, 2019). Consequently, human translations are preferred for their higher level of accuracy (Way and Hearne, 2011).

Nevertheless, several studies have demonstrated the utility of Google Translate for translating texts in various research domains, such as health (Patil and Davies, 2014), English Foreign Language learning (Tsai, 2019), academic language improvement (Groves and Mundt, 2015), and English teaching (Medvedev, 2016). In this section, we employ Google Translate to translate exam texts. The working hypothesis is that leveraging the existing readability measures in English allows us to translate texts into English, enabling us to assess their readability levels. These measurements can then serve as a foundation for developing new measures tailored to evaluate the readability of texts in Sesotho.

Since we hope that education stakeholders such as teachers, parents, learners, textbook authors, and examiners can use our envisaged readability measures for analysing Sesotho text readability, we are training our models on educational texts. We rely on texts collected as part of our corpus described in Section 3.2 above. We identified texts from grade twelve Sesotho HL and FAL examinations from the collection. As a rule of thumb, we followed Zamanian and Heydari’s (2012) guideline that a text should have at least 200 words for measures like FRE and FKGL to be applied successfully. As a result, we identified texts of no less than 200 words each. In the end, we could only use longer reading comprehension and summary texts.

3.4.2 Method

3.4.2.1 The texts

This section utilises the Sesotho reading comprehension and summary writing texts from the corpus of grade twelve examination texts (Sibeko and Van Zaanen, 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 and Van Zaanen, 2021), as well as linguistics complexity (Sibeko, 2021).

3.4.2.2 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 section is presented in Table 3.5. 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 Translate¹⁰, which facilitates whole-file translations. A comprehensive summary of the resulting machine translations is provided in Table 3.5. These translations form the first data set under discussion in this section.

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 3.5. These post-edited versions of the machine translations form the second data set under scrutiny in this section.

The comparison of the accuracy of translations between Sesotho HL and FAL Sesotho shows no clear pattern based on the extracts in Table 3.6.

Further investigation is necessary to determine if FAL texts produce more accurate translations than HL texts. Furthermore, the results in Table 3.6 suggest that translated English sentences tend to be shorter than their Sesotho counterparts. Further investigation is necessary to understand the underlying factors contributing to this pattern and its implications for training readability models tailored for Sesotho.

3.4.2.3 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,

¹⁰Google Translate is an online machine translation platform accessible at <https://translate.google.com>.

Table 3.5: An overview of Sesotho source texts, machine translations, and human-post-edited machine translations.

Year	Sesotho original texts						Machine translated texts						Post-edited translations					
	Tokens	Types	TTR	Sentences	ASL	AWL	Tokens	Types	TTR	ASL	AWL	Tokens	Types	TTR	ASL	AWL	Tokens	Types
2009	592	281	0.48	20.5	14.38	3.94	724	328	0.45	16.57	4.45	501	277	0.55	15.59	4.39		
2011	1,804	736	0.42	19.8	18.08	3.92	1,984	978	0.49	13.92	4.36	1,571	766	0.49	15.59	4.39		
2012	1,168	543	0.47	18.5	16.08	4.06	826	482	0.58	14.49	4.46	1022	543	0.53	14.24	4.37		
2013	2,022	845	0.44	25.6	16.81	4.05	1,100	616	0.56	12.85	4.28	1,805	875	0.48	14.57	4.24		
2014	1,898	803	0.44	22.0	18.00	4.19	1,121	613	0.55	15.99	4.61	1,693	829	0.49	16.2	4.39		
2015	949	396	0.42	17.0	19.34	4.08	845	457	0.54	15.64	4.54	806	385	0.48	16.47	4.39		
2016	1,811	834	0.46	17.5	17.27	4.33	1,819	872	0.48	15.27	4.57	1,535	828	0.54	14.54	4.66		
2017	1,232	552	0.45	17.6	18.33	4.19	1,234	595	0.48	18.73	4.30	1,056	537	0.51	16.05	4.43		
2018	1,167	515	0.45	15.0	19.53	4.06	1,255	581	0.46	20.35	4.31	1,002	524	0.52	17.03	4.32		
2019	1,150	535	0.47	15.5	19.45	4.15	1,231	620	0.50	15.57	4.55	1,014	566	0.56	17.24	4.46		
Total	13,793	6,040					12,139	6,142				12,005	6,130					
Average	1,379.3	604	0.45	18.9	17.727	4.10	1,213.9	614.2	0.51	15.94	4.44	1,200.5	613	0.52	15.75	4.40		

Table 3.6: Examples of Sesotho source text (ST) and English target text (TT) sentences.

Sesotho ST	English TT	Source
Liane o hatella taba ya hore ho bohlokwa ho ruta bana molemo wa tjhelete , kaha seo se tla etsa hore ba se e utswe.	Liane emphasises the fact that it is important to teach children the value of money, as that will make them not steal it.	(FAL Nov 2017)
Mehla eo ho phelwang ho yona kajeno e qobella batswadi ho se be dikgokgo baneng ba bona.	The times we live in today force parents not to be strict with their children.	(FAL Nov 2017)
Kgalekgale kgokahano e ne e le bothata.	Communication used to be a problem.	(HL Nov 2012)
Mekotlana ya bona ya tjhelete.	Their wallets.	(HL Nov 2013)

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 University¹¹. An example of the original end-user screen of the LT3's demo is presented in Figure 3.1.

In this form can you enter a text (max. 6000 characters) in the input field or as an UTF-8 text file. After the text is processed you will see, in different ways, how readable your text is.

Wait

Processing texts, especially large ones or many at the same time puts a heavy load on our server, so we put all the submissions in a queue and process them one by one. If you do not have the time to wait for the results you have the option to give us your e-mail address. As soon as the text is completely processed you will receive a link to the results. We do not use this e-mail address for anything else.

Scientific purposes

With the results we can improve our algorithms. If we can use your particular texts to do so, enable the checkbox. Else, your text will be deleted after one week.

Text or text-file

You have option to provide an UTF8-text file or fill in a text manually. When copying from a Word-processor, go via a text editor like notepad++ to first convert everything to UTF-8.

Browse... No file selected.

Text

Email

Email

When given, a link to the results is submitted when ready.

☐ **Allowed for scientific purposes**

When checked, this experiment is not deleted after one week.

Figure 3.1: The LT3 readability end-user interface.

The results obtained through the LT3 readability demo were then extracted and organised

¹¹This tool can be accessed online at the following URL: <https://lt3.ugent.be/readability-demo/>.

into a spreadsheet for further analysis using *RStudio*.

3.4.3 Results

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.

3.4.3.1 Machine Translation evaluation

The comparison of the data sets exhibited a commendable level of similarity. More specifically, as illustrated in Figure 3.2, 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.

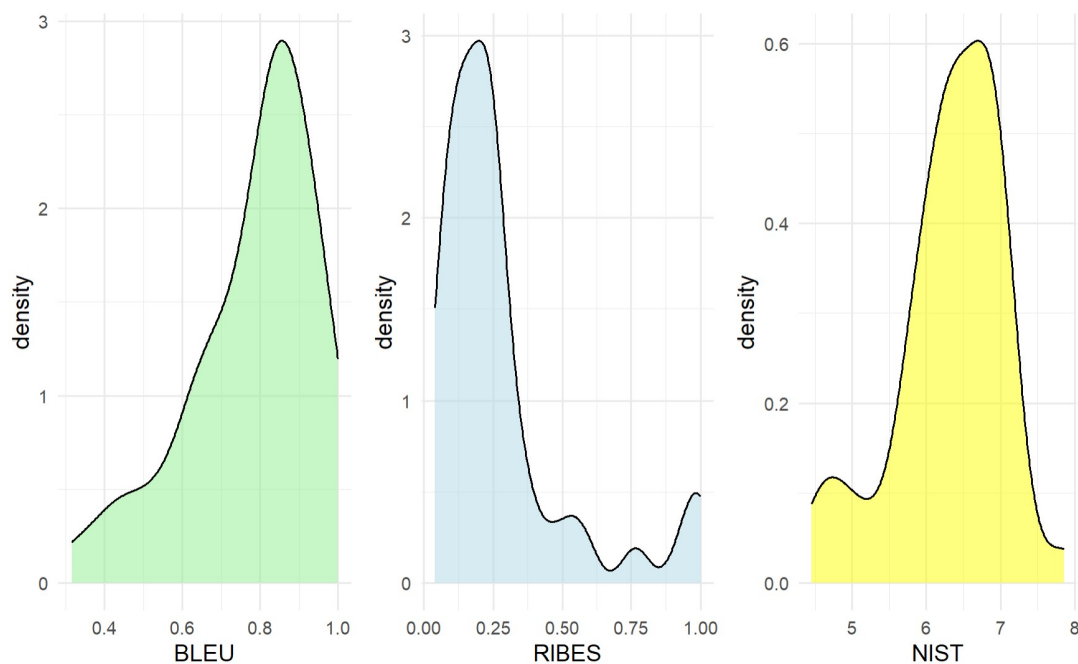


Figure 3.2: The distribution of machine translation evaluation scores.

Similar to the BLEU score, the RIBES scores range from zero to one. As can be observed from Figure 3.2, 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 3.2 visually demonstrates that the NIST scores predominantly cluster around six and seven, with a minimum of 4.45 and a 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 3.7. As can be observed in Table 3.7, 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 3.7: 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.

3.4.3.2 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 3.3. 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 3.8, 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.

The post-edited data sets exhibit slightly higher means which are associated with harder-

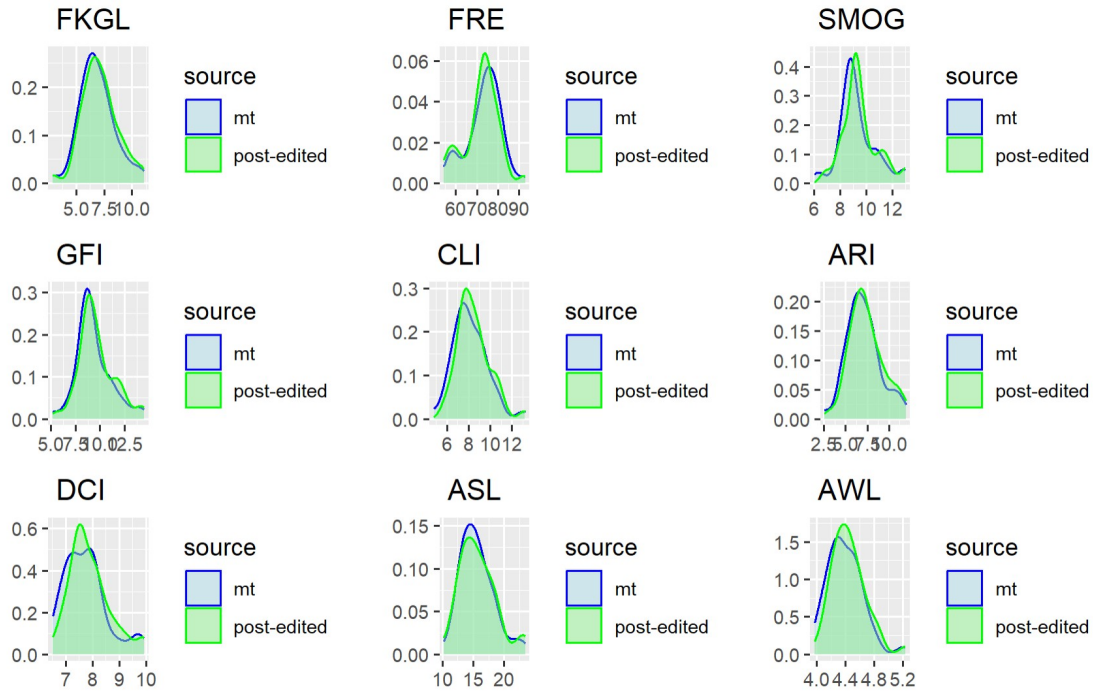


Figure 3.3: The distribution of readability scores for the machine translated (mt) and post-edited data sets.

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 ($M_{\text{machine translated}} = 15.59, M_{\text{post-edited}} = 15.76, p = .78$) and average word lengths ($M_{\text{machine translated}} = 4.39, M_{\text{post-edited}} = 4.45, p = .32$). In fact, the mean differences between the two data sets remain consistently within a negligible range.

3.4.4 Discussion and conclusions

This section 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

Table 3.8: 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

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 section aimed to investigate 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 section, 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 section 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 FAL texts and 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 section highlight the need for future research. For instance, while this section 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.

3.5 Developing Sesotho syllabification systems

This section is adapted from: Sibeko J. and Van Zaanen, M., 2023. Developing syllabification systems for Sesotho. *Under consideration for publication by Language Resources and Evaluation*. Submitted: 14 December 2022.

3.5.1 Introduction

Syllabification, i.e., extracting syllables from words (Goldsmith, 2011), is essential to many natural language processing tasks (Jucksriporn and Sornil, 2011). Unfortunately, not all languages have reliable tools that can automatically syllabify words.

As exemplified in Chapter 2, Sesotho, like many other indigenous languages in Southern Africa, remains an LRL that lacks many basic language resources, including automated syllabification systems (Krige and Reid, 2017). In short, a syllabification system segments words into syllables through a process called syllabification (Krisnawati and Mahastama, 2018; Suyanto, 2020).

The lack of resources for NLP tasks in Sesotho has led to researchers manually annotating data when investigating the Sesotho language. For instance, Krige and Reid (2017) use manual syllable annotations by Sesotho speakers in their investigation of text readability using traditional readability measures. These manual annotations are problematic. Krige and Reid (2017) identify time and cost factors as constraining similar investigations in indigenous languages. We assume that these time and cost factors were consequent to the manual retrieval of syllable-related information since the extraction processes of other

textual properties were automated.

To resolve this problem, we developed an automated system to identify Sesotho syllables, which can be easily integrated into NLP tasks for Sesotho. We believe that the research described in this section can lead to similar approaches for (at least) the other languages in the Sotho-Tswana language family. It was imperative that a syllabifier be developed in this thesis as four traditional readability measures depend on the automatic identification of syllables.

According to Goldsmith (2011), most syllabification theories started with rules. Evidently, even objective manual human syllabification would have to be governed by sets of rules that standardise their practice instead of relying on intuition. We use the structural method of syllabification in our rules-based syllabification system. The structural method investigates constituents of onsets, nuclei, and offsets for syllables and the rules that govern the identification of each (Bigi et al., 2010). Despite the absence of a universal definition for syllables (Krishna, 2020), a syllable nucleus is accepted across different languages (Bigi et al., 2010; Goldsmith, 2011; Kovač and Marković, 2018).

Sesotho syllables are identified based on syllable nuclei information. However, they do not contain codas (Motjope-Mokhali et al., 2020). That is, they can contain nuclei only, or nuclei and onsets. As a result, there are no ambi-syllabic consonants where offsets for one syllable also function as onsets for the next syllable (Kahn, 1976), this makes syllable boundaries easier to predict in Sesotho. As such, rules become easier to describe. Furthermore, the absence of codas in Sesotho syllables ensures that all non-syllabic consonants function as onsets. In addition, Sesotho syllables do not contain extrasyllabic consonants. In this way, they do not fall under theories advocating for extrasyllabic consonants (for instance, see Clements and Keyser (1983); Rubach (1993), and Duanmu (2009)) as there are no extra-syllabic consonants in Sesotho.

The longest syllable in Sesotho builds on the longest onset, which can be formed by a maximum of three consonants and one semi-vowel. Madigoe (2003) identifies such a syllable as a CCCWV type (where C represents a consonant, W is a semi-vowel and V is a vowel). Note that the syllable types that are used in the approaches described in this section are discussed in Section 3.5.2.2.

3.5.2 Methodology

Syllabification systems are typically implemented using either rule-based or machine-learning approaches (Bigi et al., 2010; Gregová, 2016; Kovač and Marković, 2018, 2019). In this section, we describe the development of both a rule-based as well as a machine-learning (pattern-based) approach to syllabification for Sesotho. For the training of the machine learning approach as well as for the evaluation of both systems, annotated data is essential. We will first describe the available annotated data, followed by descriptions

of the rule-based and pattern-based systems and their implementations.

3.5.2.1 Syllabification training data for Sesotho (Corpus 4)

Despite being considered an LRL, Sesotho benefits from the existence of a dictionary available in electronic and PDF formats, known as the Sesotho International Dictionary (Chitja, 2010). This comprehensive dictionary consists of over 13,000 entries, although the exact number is difficult to identify as there are some duplicates in the dictionary. We will provide exact information of the word list used for evaluation purposes, after the cleanup, below. The dictionary provides essential information such as headwords, pronunciation, English translations, definitions, and occasional related words. Moreover, the pronunciation information contained in the dictionary allows for the extraction of syllable details.

Consider the following example entries (the English translation in between brackets is not present in the dictionary):

- **Acheke** (a-chi-ki) /exclamation: oh no!/ Lekgotso le bontshang ho teneha ka ho toboketsa haholo. (*An exclamation indicating annoyance with great emphasis*).
- **Adima** (a-di-ma) /borrow/ Ho fa ka sepheo sa ho kgutlisetsoa ka mora nakoana; ho neha motho seo a se hlokang nakoana. (bap. kadima). Mael. Ho adima ditsebe: ho kopa ho mameloa. (*To give with an intention of receiving back after a while; to give someone something that they need for a while. (comp. lend). Idiom. To lend an ear: to ask to be listened to*).

In the example dictionary entries above, we provide the headword in bold. This is followed by pronunciation information (in brackets), which also indicates syllable boundaries using dashes. Next, an English translation is provided between slashes and a definition is provided.

We took the dictionary in PDF format, converted it to plain text using *pdftotext* (version 0.86.1) and then applied a combination of GNU *sed* (version 4.7) and GNU *grep* (version 3.4) commands to extract only the pronunciation information including the syllable information. This formed the basis for a manual correction step.

Although the dictionary is largely consistent in content presentation, it is unfortunately not quite consistent in the use of orthography. It uses both LS and SAS orthographies interchangeably. To resolve this issue, the orthography was manually adapted and standardised to the SAS orthography. In summary, the uses of vowels in place of semi-vowels were modified to the SAS orthography in this way:

- *oe* constructions were changed to *we*,

- *oa* constructions were changed to *wa*,
- *oo* constructions were changed to *wo*,
- *ea* constructions were changed to *ya*,
- *ee* constructions were changed to *ye*,
- *eo* constructions were changed to *yo*,
- *c* constructions were changed to *tj*.

Since the representation of semi-vowels in the LS orthography is different from that of the SAS orthography, the rules on consecutive vowels would result in inflated syllable counts and imprecise syllable boundaries when LS is used. For instance, the singular form of the word “you” has three syllable boundaries in LS *oena o-e-na (you)* while it only has two in SAS *wena we-na (you)*.

Additionally, the SAS orthography uses the digraph *tj* whereas the LS orthography uses the letter *c*. As such, we changed, where needed, the *c* to *tj* (manually checked on a case-by-case basis).

The phonological syllables indicated in the dictionary were also converted to an orthographic representation where needed. Given that the dictionary aimed at providing guidance for pronunciation, higher tones and lower tones were indicated morphologically. Such efforts affect the syllable boundaries as they allow some syllables to end in consonants. For instance, lower *o* in the word *Yuropa* is represented as *aw*, resulting in *Yu-raw-pa* instead of *Yu-ro-pa*. Therefore, although such changes do not affect syllable counts, they affect syllable information by blurring the boundaries and will render our system unusable if researchers investigate more than syllable counts. Moreover, we are reproducing the word list used for training and testing our systems. As such, consistency in orthography is important. The following changes were effected on the pronunciation information extracted from the dictionary entries:

- *u* indicating a higher tone *o* were changed to *o*,
- *i* indicating a higher tone *e* were changed to *e*,
- *aw* indicating a lower tone *o* were changed to *o*.

Additionally, we relied on the pronunciation information for the identification of syllable boundaries. Unfortunately, there were inconsistencies in the dictionary syllable boundaries. For instance, the word *abela* is presented as *abe-la* while the word *abelwa* is syllabified as *a-be-lwa*. To this end, some syllable boundaries from the dictionary pronunciation information needed to be updated according to the syllabic structure rules in Guma (1982).

Finally, orthography adaptations and multiple dictionary entries with different definitions resulted in some duplicates in the dictionary word list. These duplicates were removed.

The complete word list including the syllable information can be found on the repository of SADiLaR (see Sibeko and Van Zaanen (2022b)).

3.5.2.2 Rule-based approach

In general, a rule-based approach is preferred and best suited to languages that have insufficient resources and annotated corpora (which is the case for Sesotho), instead of the data-driven approach which would require segmented corpora (Kovač and Marković, 2018). In short, manually created rules provide clear specifications on regularities, such as where syllable breaks can be expected. Although rules-based approaches to syllabification are largely ineffective when used on languages with orthographies like that of English (Bigi et al., 2010), they may very well be effective on languages which do not contain silent, null or unpronounced letters. For these reasons, we believe that for Sesotho rules can be effectively designed and applied even in the absence of annotated corpora. We are not aware of any publicly available Sesotho syllable annotated word lists or corpora except the ones we developed in this thesis.

The rules that are used in a rule-based syllabification system may very well be based on linguistic theories. As such, the results of such a system not only lead to consistent results but at the same time may provide information on the reason why a particular syllable boundary was introduced. When describing the syllabification rules, linguistic constructs, such as vowels, fricatives, nasals, glides, occlusives, and liquids may be used (as an example, consider the system developed by Bigi et al. (2010), who developed a syllabification system for speech). Similar constructs are important in identifying syllable boundaries in Sesotho.

Our rule-based system is grounded on the three types of Sesotho syllables described by Guma (1982), namely, the *vowel-only* syllable (V-syllable), the *consonant-vowel* syllable (CV-syllable) and the *consonant-only* syllable (C-syllable). The V-syllables and the CV-syllables are affected by the limitations of vowels per syllable. According to Madigoe (2003), Northern Sotho syllables can only comprise one vowel. This rule extends to other Sotho languages as well.

Following Guma’s (1982) syllable structures¹², we developed three rules for V-syllables:

1. when a vowel is at the beginning of a word, insert a syllable boundary after it,
2. insert a syllable boundary between two consecutive vowels in a word,
3. no syllable boundary is inserted if a word consists only of a vowel.

¹²See Sibeko (2022b) for a more extensive description of the rules.

The constraint of having only one nucleus per syllable is what leads to the limitation of only one vowel per syllable. In Sesotho, each vowel serves as a nucleus in a syllable, and due to this constraint, it is not possible to have more than one vowel within a single syllable.

The CV-syllable rules indicate that when a vowel is preceded by a consonant (or more), a syllable boundary is inserted after the vowel. The number of consonants before the vowel does not affect the syllable boundary. For instance, words such as {*taba*, *thata*, *tjhata*, *tjhwatla*} (*situation*, *difficult*, *marry*, *break*) all have a syllable boundary after the first vowel *a*, i.e., {*ta-ba*, *tha-ta*, *tjha-ta*, *tjhwa-tla*}.

Note that Sesotho is a tonal language (Sekere, 2004; Mohasi et al., 2011), which carries tone by vowels and nasal consonants (Guma, 1982; Mohasi et al., 2011). Therefore, the nasals are also syllabic in that they can carry the peak of prominence in a word thereby allowing them to function as syllabic nuclei. Because of this, C-syllable rules are introduced.

For our C-syllable rules, we identified four nasal consonants consisting of two simple nasal consonants {*m*, *n*} and two complex nasal consonants {*ng*, *ny*}. Additionally, we identify the lateral consonant *l*. According to Demuth (2007), nasal syllables can occur word-initially *n-ka* (*take*), word-medially *ba-n-ka* (*bank*) or word-finally in the form *ng*, e.g., *ba-n-ke-ng* (*in the bank*). The C-syllable rules handle these cases as follows:

1. insert a syllable boundary after the nasal consonant if it is followed by any consonant,
2. insert a syllable boundary between two consecutive lateral consonants.

Unlike the nasal consonants, no syllable boundary is inserted between the lateral consonant and a consequent consonant. Furthermore, for both the nasal consonants and the lateral consonant, no syllable boundary is inserted before the consequent semi-vowel.

Table 3.9: Rules of syllabification examples.

Type	Rule	Input	Syllabified
V	1	<i>aba</i>	<i>a-ba</i>
	2	<i>boela</i>	<i>bo-e-la</i>
	3	<i>ba</i>	<i>ba</i>
CV	1	<i>banana</i>	<i>ba-na-na</i>
	1	<i>kgutshwane</i>	<i>kgu-tshwa-ne</i>
C	1	<i>nka</i>	<i>n-ka</i>
	1	<i>mme</i>	<i>m-me</i>
	1	<i>ngwana</i>	<i>ngwa-na</i>
	1	<i>nniyala</i>	<i>n-nya-la</i>
	2	<i>lla</i>	<i>l-la</i>
	2	<i>lwana</i>	<i>lwa-na</i>

Table 3.9 provides several examples of the syllabification rules. Note that the system never introduces syllable boundaries at the end of the word (even though the word may end with a vowel after which normally a syllable boundary is placed according to the V-syllable and CV-syllable rules).

We implemented the rules-based syllabification system using *Python 3*. The version of the system described in this section can be found on SADiLaR’s repository (see Sibeko and Van Zaanen (2022c)). The system expects plain text input with one word per line. The syllabified version of the input is provided on the output with spaces indicating syllable boundaries¹³. The system has three parameters, namely, the input, the output and a help and debug function.

3.5.2.3 Pattern-based approach

Bigi et al. (2010) state that data-driven approaches to syllabification are only effective if (enough) training data is available. Fortunately, the Sesotho International Dictionary provides syllabification information (after cleaning up). This allowed us to experiment with systems that automatically learn patterns from the example data.

The core of the system described here is *Patgen* (version 2.4) (see Liang (1983) for an extended description of the system). *Patgen* is the *de facto* standard hyphenator for $\text{\TeX}$ (and hence $\text{\LaTeX}$) (Németh, 2006). This system can be used to build hyphenation patterns (Haralambous, 2021). It has also been used to introduce boundaries in noun compounds in Afrikaans (Van Huyssteen and Van Zaanen, 2004).

Inserting potential hyphenation breaks is a similar process to inserting syllabification boundaries. Whereas the hyphenation system in $\text{\TeX}$ is trained on words with their potential hyphenation breaks, here we train the same system using words with their syllabification boundaries.

Our pattern-based syllabification system consists of a few GNU *bash* (version 5.0.17(1)) scripts that convert the training data (a word list with the syllabification boundaries) to the correct input format for training of *Patgen*. A *Python 3* script then takes the input (a plain word, one per line), applies the *Patgen* patterns (using *hyphenate.py* written by Ned Batchelder in 2007) and converts the output to the same format as the rule-based system (that is, space-separated syllables of the input words).

Five files were used in the development of the system. One, the *translate* file (.tr) was created. The *translate* file allows the developer to indicate similarities between the symbols and alphabets used in the system. For instance, we used the *translate* file to indicate similarities between upper and lowercase letters. In this way, the use of either upper or

¹³Note that the dash (-) is used to signify syllable boundaries for clarity in the examples provided herein even though the system provides spaces as syllable boundaries).

lowercase letters will not affect the system’s capabilities. We used the lowercase letters as primary representatives of the 26 letters of the alphabet used in the system. This was important to indicate since the text files only use primary representatives for letters (Haralambous, 2021). Two, a patterning file (.pat) was created. For this, *Patgen* is used as a pattern generator (Sojka and Sojka, 2003). We used the *do_all.sh* bash script to create our pattern. This script converts the plain text training data file to the correct format and then passes it on to the *make_full_patterns.sh* bash script where a pattern file is developed. Our plain text training data file contained one word per line. Each line contains one word followed by a space and the annotated word where syllable boundaries are indicated by the use of hyphens. For instance:

- A (A)
- Aha (A-ha)
- Amang (A-ma-ng)

Three, the dictionary file (.dic) was used to test the pattern at the end of the pattern creation process. The dictionary uses only the primary representatives for formal letters (lowercase letters). Syllable breaking points are indicated by the use of hyphens. Four, the system automatically creates an output (.txt) file with only the syllabified words. Syllables of each word are presented in a new line. For instance:

- A
- A ha
- A ma ng

Five, a test file (.pattmp) is created. The test file indicates wrong breaking points, right breaking points, and missed breaking points (Haralambous, 2021). We do not assume our word list to be free of human error. As such, any errors in the patterning file will result in patgen generating inconsistent and incorrect patterns (Sojka and Sojka, 2003). The system can be accessed from the repository of the South African Centre for Digital Language Resources (SADiLaR) (Sibeko and Van Zaanen, 2022c).

To use this system, three arguments must be satisfied. First, one must insert a plain text input file with one word per line. The white space at the end of the word will be deleted. An output file should also be indicated. The syllabification results will be printed on this document. A pattern file should also be indicated. The default pattern file is our pattern (.pat) file described above.

3.5.3 Results

For the experiments described in this section, we used the cleaned training data based on the Sesotho International Dictionary (Sibeko and Van Zaanen, 2022b). The annotated word list contains a total of 13,551 words with their corresponding syllable boundaries.

As the pattern-based system requires training data, we set up ten-fold cross-validation experiments. The same setup was used for both systems to allow for a direct comparison, even though the rule-based system does not use the training folds. The training data is simply the annotated data from the word list. For testing, we remove the annotations and let the systems add the syllabification boundaries, which can then be evaluated by comparing the output of the systems to the original, annotated version.

Table 3.10: Total number of words, number of erroneously syllabified words, and accuracy for both pattern-based and rule-based systems for each of the folds as well as their means and standard deviations.

Fold	Total	Pattern-based		Rule-based	
		Error	Accuracy	Error	Accuracy
0	1,356	254	81.27	2	99.85
1	1,355	270	80.07	4	99.70
2	1,355	293	78.38	5	99.63
3	1,355	289	78.67	3	99.78
4	1,355	290	78.60	6	99.56
5	1,355	271	80.00	1	99.93
6	1,355	296	78.15	7	99.48
7	1,355	299	77.93	6	99.56
8	1,355	286	78.89	6	99.56
9	1,355	302	77.71	2	99.85
Means	1,355.10	285.00	78.97	4.20	99.69
Standard Deviation	0.10	4.82	0.36	0.66	0.05

The results of both pattern-based and rule-based systems are presented in Table 3.10. This table contains the number of words per (test) fold, the number of erroneously annotated words and the accuracy per fold for both systems. Note that accuracy is computed on the word level, so the entire word needs to be correctly syllabified. Additionally, the means and standard deviations for the different measures are provided.

3.5.4 Discussion and conclusions

The rule-based system leads to much higher accuracy (99.69) than the pattern-based system (78.97). There may be several reasons for this. First, the syllabification of Sesotho words is very consistent with hardly any exceptions. Second, the amount of training data

may not be enough for the pattern-based system to perform optimally. That is, even though the lower performance of the pattern-based system can be expected given that the system does not always yield accurate results (Németh, 2006), the fact that the rules used in the rule-based system are so simple and require only a very limited amount of context, suggests that the limiting factor of the pattern-based system is the small amount of available training data.

Table 3.11: The different types of errors in the rule-based system output. Here WL indicates the wrong location, MB indicates a missing boundary, NE indicates a non-existent word, and LW indicates a loan word.

Dictionary	Rules	Error
pot-ri-si	po-tri-si	WL
mos-pa-ne-she	mo-spa-ne-she	WL
e-ba-nya-neng	e-ba-nya-ne-ng	MB
setshwa-n-tshi-si	se-tshwa-n-tshi-si	MB
bo-i-te-ka-nyetso	bo-i-te-ka-nye-tso	MB
el	e-l	NE
kgo-l-kgo-tha	kgo-lkgo-tha	LW

An in-depth, qualitative analysis of the erroneously annotated words from the rule-based system indicates mostly problems with the information coming from the dictionary. Table 3.11 provides several examples where the rule-based system leads to output different from that of the dictionary.

The word list from the dictionary still contains errors resulting from human oversight. First, some words have syllabification boundaries in incorrect places. In Table 3.11, these are indicated by WL, wrong locations of boundaries. The syllables provided in the dictionary are aimed at assisting the reader with the correct pronunciation. As such, some words are separated in a way that highlights the pitches and not syllable boundaries. Separating pitches in this way results in some dashes being placed in ways that translate to syllable offsets. However, syllable structures in Sesotho cannot have offsets (Guma, 1982; Sibeko, 2022b). As such, these incidences are erroneous in the gold-standard word list.

Second, the dictionary word list contains instances with missing syllable boundaries. In Table 3.11 these are indicated by MB, missing boundaries. For instance, *setshwa...* should have a syllable boundary as correctly identified by the rule-based system as *se-tshwa*.

Third, there are some non-existent words in the dictionary word list. For instance, the entry *l* (which is then provided with pronunciation information and thereby incorrectly syllabified in the dictionary as *el*), is not a proper word, but provides phonological information of the pronunciation of the letter *l*.

Finally, the word *kgolkgotha* (*Golgotha*) is actually a loan word (indicated by LW in Table 3.11). As such, it does not have a standard Sesotho orthography. It is written as

Gologotha in the Sesotho version of the Bible (Matthew 27:33), which would be syllabified as *Go-lo-go-tha* according to the rule-based syllabification system). Note the addition of the letter o after *Gol* compared to the letter k in *kgolkgotha*. This addition of a vowel allows for proper pronunciation.

Summarising, it seems that most (if not all) of the errors found by the rule-based system stem from incorrect syllable information in the word list (which stems from human oversight when cleaning and converting the word list), with the exception of non-words and loan words with non-standardised spelling. Note that these exceptions may have a larger impact on the performance of the pattern-based system than the rule-based system as that system might overgeneralise these errors.

In this section, we described the development of two syllabification systems for Sesotho: a rule-based and a pattern-based system. The rule-based syllabification system uses rules derived from the syllable types described by Guma (1982). In this order, three main structure-based rules were identified, vowel-only rules, consonant and vowel rules, and consonant-only rules. First, the rules for the vowels only syllables are emphasised by the absence of silent vowels in Sesotho. Furthermore, the lack of a coda in Sesotho syllables indicates that after each vowel, there is a mandatory syllable boundary. To this end, syllable boundaries are clear as far as vowels are concerned. Second, our rules governing consonant and vowel syllables do not need to consider the number of consonants in the onset position. Finally, the consonants-only syllables consider syllable breaks induced by the appearance of nasal consonants (m, n, ŋ, and ɲ) and the lateral l.

The pattern-based system uses annotated training data (also taken from the word list which stems from the Sesotho International Dictionary). We used *Patgen* (which was originally developed to learn hyphenation patterns in T_EX) to learn patterns describing the potential location of syllable boundaries. These patterns are then applied to words to identify syllable boundaries.

The ten-fold cross-validation results indicate that both the rules-based and pattern-based systems work. However, the rules-based system leads to higher accuracy. We expect that this is because the syllabification of Sesotho is highly consistent with few exceptions. More and more precise training data may yield improved results for the pattern-based system as the rules used in the rule-based system are relatively simple and only use local information.

The initial reason for developing a syllabification system for Sesotho was that several traditional readability measures (developed for English) rely on syllable information. The evaluation of the systems in this section has been *in vitro*. We would like to experiment with *in vivo* evaluations in the context of the readability measures for Sesotho.

3.6 Summary

In this chapter, we have provided a detailed discussion of the development of basic language resources needed for enabling the adaptation of existing text readability measures into Sesotho. We have thus answered the question: What language resources are needed for measuring text readability in Sesotho?

To address the challenges of insufficient resources for measuring text readability in Sesotho, we have discussed the corpora that we developed, the list of frequently used Sesotho words, our syllabification system training data, the general comprehension machine translations, and the development of our syllabification systems. Through this chapter, we have contributed to the development of basic language resources in Sesotho. These resources are then used in the adaptation of English traditional readability measures into Sesotho.

The development of the basic language resources discussed in this chapter is not only crucial for the successful adaptation of readability measures into Sesotho but may also have a broader application in Sesotho language processing tasks. For instance, we discuss a broader application of the rule-based syllabification system to the Sotho-Tswana language group in Sibeko and Setaka (2023).

Furthermore, the corpora developed in this section will be used in the analyses of text readability in both English and Sesotho in the next chapter. In this chapter, we only describe the materials. To be able to put them together, we need to develop a linear regression model.

Chapter 4

Adapting readability measures to Sesotho

This chapter explores the development of measures for text readability. Initially, we focus on examining the disparities in linguistic complexity between HL and FAL texts administered in grades eleven and twelve. We thoroughly analyse the text readability of grade twelve texts in both HL and FAL. To determine the readability of Sesotho texts, we adopt English measures as a reference point. Finally, we leverage our machine translation data set to develop specific readability measures for Sesotho using simple linear regression modelling.

4.1 Introduction

The preceding chapters provide the necessary context for this thesis by examining the challenges and approaches to measuring text readability in LRLs, specifically focusing on Sesotho. They also encompass an exploration of available resources and the development of new digital language resources required for the creation of Sesotho text readability measures. In continuation, this chapter builds upon this foundation and addresses the third research question: What insights can be derived from existing traditional readability measures? To accomplish this objective, three studies were conducted, each outlined below.

We rely on the use of the grade twelve examination texts for our experiments in this chapter. Unfortunately, the examination texts have not been annotated with readability levels. As a result, we conducted two studies to ascertain whether the texts used in the grade twelve examinations exhibited variety in their levels of linguistic complexity. Note that these studies are based on work in this thesis but are not included in this thesis. For both studies, we utilised the Coh-Metrix online platform for analysing the texts (McNamara and Graesser, 2012; McNamara et al., 2010a, 2012).

In the first study (Sibeko, 2021), we examined a set of 24 grade twelve final examination texts encompassing both English HL and FAL subjects, spanning the years 2008 to 2019. Five primary indices, encompassing word-level attributes, sentence structure, readability, lexical diversity, and referential cohesion, were analysed to determine the linguistic complexity.

The findings underscored significant disparities in the linguistic complexities of the texts employed for reading comprehension and summary writing in the two subjects. Specifically, it emerged that the HL texts exhibit higher linguistic complexity compared to their FAL counterparts. Additionally, the FKGL measure indicated a two-grade divergence between the FAL and HL texts. Nonetheless, the linguistic complexity metrics employed within this study validate the anticipated distinction, affirming that HL texts indeed manifest a higher degree of complexity relative to their FAL counterparts.

In the second study, an investigation into the linguistic complexity of grade eleven and grade twelve examination texts was conducted (Sibeko, 2022a). This analysis drew upon a data set comprising 80 texts extracted from grade twelve and grade twelve question papers, from the year 2011 to the year 2020. A comprehensive evaluation of eighteen indices was undertaken, encompassing various text features, such as lexical diversity, syntactic complexity, employment of connectives, surface-level complexity, and text readability.

The outcomes of this investigation demonstrated a consistent pattern of higher linguistic complexity characterising grade twelve examination texts in contrast to their grade eleven counterparts. Notably, the grade twelve examination texts exhibited higher levels

of lexical diversity and syntactical complexity relative to the grade eleven examination texts. Moreover, the indices measuring readability highlighted a discernible trend, with grade twelve examination texts being harder to read than their grade eleven counterparts. Note that the Coh-Metrix gauges text readability levels using three readability measures, namely, the FRE, FKGL, and the Coh-Metrix measure.

We complement the findings in Sibeko (2021) and Sibeko (2022a) with a more in-depth analysis of text readability of English examination texts in Section 4.2. We compute readability scores using nine existing readability measures for the final year exams of English HL and English FAL. The results show that indeed the readability of the texts is consistent over the years and significantly different between the two subjects. Generalising these results, we expect that we can use final-year examination texts to develop readability measures for Sesotho.

In Section 3.4, we investigate the differences between the readability of machine translations from Sesotho to English and compare the results to the readability of human-post-edited versions of the translations. These results indicate similarities between the machine-translated and human-post-edited translations. Even so, we have yet to investigate the differences between the readability levels of the Sesotho texts.

To address this knowledge gap, we investigate the readability levels of the texts in Sesotho in Section 4.3. Given the lack of readability measures for Sesotho, we employ the English readability measures. Although these measures are not suited to Sesotho texts, they can provide a sense of differences and similarities between the HL and the FAL examination texts.

To enable the assessment of text readability in Sesotho we employed a few modifications to the readability tool. Among other modifications, we replace the Dale-Chall list of frequently used words with our list of commonly used words in Sesotho (see Section 3.3). We also replace the English syllabification system with our rule-based syllabification system (see Section 3.5). For consistency, we use reading comprehension and summary writing texts like in the English investigations. Our assumption is that if these texts exhibit different levels of readability, they will substantiate our utilisation of them.

To ensure methodological consistency, we maintain the use of reading comprehension and summary writing texts, mirroring the approach adopted in the investigations of text readability of English texts conducted in this thesis. Our underlying premise is that any observed divergence in readability levels within these texts will serve to reinforce the validity of the employment of the examination texts in the training of our readability formulas.

We end the chapter by addressing the fourth research question: How can existing English traditional readability measures be modified and adapted to Sesotho? For this, we develop readability measures for Sesotho based on the formulas of the nine traditional readability measures discussed thus far in this thesis. The analyses of text readability using both

Sesotho and English examination texts serve as a starting point for identifying useful text properties to consider in the development of readability measures for Sesotho.

We employ the use of the translated readability corpus to extract textual features from the Sesotho texts and readability levels from the English translations. Through simple linear regression analyses, we examine the impact of the text features from Sesotho texts on the overall readability levels gauged from the English translations. The linear regression models allow for the identification of new coefficients and new intercepts for the different variables considered in the readability formulas. In the end, we propose ten readability formulas (one more than the initial nine due to the two options we provide for the GFI). We also introduce new intercepts for the GFI, the LIX, and the RIX. Our Sesotho formulas also introduce coefficients for these three formulas.

4.2 Using traditional readability measures to evaluate English grade twelve examination texts

This section is adapted from: Sibeko, J. and van Zaanen, M., 2021. An analysis of readability measures on English examination texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11¹.

4.2.1 Introduction

The previous section implies that the readability of grade twelve reading comprehension and summary writing examination texts remains consistent within the context of the English FAL subject. However, this conclusion is based on a limited investigation that primarily focuses on linguistic complexity, utilising restricted measures to measure text readability. Furthermore, the investigation is constrained by the selection of specific texts from a single examination sitting, namely the November session, and its exclusive focus on the FAL subject.

Given that educational texts are designed for language learners at specific educational levels, we anticipate that they possess a known or consistent level of readability. Consequently, it is worth considering the use of educational texts, specifically reading comprehension and summary writing texts employed in final year exams. Nonetheless, it remains uncertain whether these examination texts truly exhibit consistent readability, thereby making them suitable for the development of readability measures. To address this gap, the exploratory research outlined in this section aims to investigate the readability of English comprehension and summary examination texts utilised in South Africa.

¹This section was improved based on the feedback from audiences at the Digital Humanities of South Africa 2021 biannual conference. We were awarded the best paper at the conference.

We focus on English since tried and tested measures for measuring text readability in English exist. If the results for the English examination texts are as we expect, then we assume that we can use the same type of examination texts for the indigenous South African languages. The analysis of text readability using English readability measures may also provide information on what text properties to investigate further when developing text readability measures for the indigenous South African languages.

Grade twelve teachers who are responsible for preparing learners for final examinations are encouraged to seek out diverse texts and customise them to suit their learners' proficiency levels (Department of Basic Education, 2017). However, the existing guidelines for text selection do not provide specific details regarding the characteristics of texts that teachers can modify for their students. Furthermore, the examination guidelines do not indicate whether readability measures are utilised in the preparation of examination texts.

In order to understand the readability of texts used in English high school subjects, this section sets out (i) to check whether the readability of the English reading comprehension and summary writing examination texts is consistent (that is, whether there are no differences between the readability of the texts of the different examination opportunities and whether there are differences between the HL and FAL examination texts), and (ii) to investigate whether different readability measures are consistent with these results in order to get an idea of what text properties (used in the measures) might be useful for the development of similar measures for other languages.

4.2.2 Background

In Chapter 1, it was mentioned that measuring text readability can be approached from various perspectives. One perspective considers the characteristics of the readers themselves (Nouwens et al., 2016; Duff, 2019; Kärbla et al., 2019, 2020; Phillips Galloway et al., 2020). According to this perspective, the readability of a text is contingent upon how well a reader can comprehend the literal meaning of the text, make inferences from the text, or employ evaluative techniques to understand its content (Basaraba et al., 2013; Tennent, 2014; Kärbla et al., 2020). Therefore, since text readability is perceived in relation to the specific reader, it is often used interchangeably with terms like text difficulty and reading difficulty (Collins-Thompson, 2014).

Another perspective, which relates to the readability measures used in this thesis, does not view text readability in relation to the reader. Instead, text readability is viewed as a subcategory of text complexity (Amendum et al., 2018), which focuses on independent linguistic factors that can be manipulated (Mesmer et al., 2012, p. 235) as opposed to how the text interacts with the reader (McNamara et al., 2002; Meyer, 2003; Stahl, 2003; Stenner et al., 2006; Benjamin, 2012; Spencer et al., 2019). Readability measures are described as mathematical formulas obtained through regression analysis that are used to

measure readability (Mc Laughlin, 1969; Heydari, 2012; Begeny and Greene, 2014). They focus on the style of writing (Courtis, 1987) as manifested, among others, through word and sentence lengths (Stevens et al., 1992), syllable counts (Kate et al., 2010), and word lists (Balakrishna and Meurers, 2014).

In South Africa, studies on text readability of health documents (Joubert and Githinji, 2014; Leopeng, 2019; De Wet, 2021) and textbooks evaluations (Sibanda, 2019; Wissing et al., 2016) using traditional readability measures have been conducted. However, according to our knowledge, there are no empirical studies investigating the readability of reading comprehension and summary writing texts in the domain of South African basic education.

4.2.3 Methodology

4.2.3.1 Materials

We employ the use of our grade twelve examination corpus discussed in Section 3.2. From the English set of examination texts, we extract 48 examination texts composed of twelve HL and twelve FAL November texts from 2008 to 2019, eight HL and eight FAL February/March texts from 2011 to 2018, and four HL and FAL May/June texts from 2016 to 2019. The texts were manually cleaned to remove ‘noise’ such as page numbers, line and paragraph numbers, sentence markers, footnotes, endnotes, and source references.

Text readability analysis was conducted using the *Python 3*’s readability package (Van Cranenburgh, 2013). The outcomes of this readability assessment were collated and documented within a single spreadsheet file. Subsequently, this file was exported to *RStudio* to facilitate the analysis.

4.2.3.2 Analysis

To analyse the performance of the different readability measures on the HL and FAL examination texts, we will first provide the mean and standard deviation values of each of the measures. Additionally, we investigate correlations between the results of the different measures. As the different measures aim to describe the same property of the text, we expect there to be relatively high significant correlations.

Once the descriptive statistics are provided and discussed, we calculate regression ANOVAs for each of the readability measures. This indicates the relationship between the readability of the HL and FAL texts, including the years and months of the exams. We expect these analyses to identify significant differences between the HL and FAL texts and we expect no significant differences based on the years and months.

4.2.4 Results

4.2.4.1 Descriptive statistics

To get a better understanding of the behaviour of the measures on the texts, first, an overview of the textual properties used in the measures discussed in this section is presented in Table 4.1. The means and standard deviations of the English HL and FAL texts indicate the spread of the data in terms of the text features considered in the readability measures employed in this section.

Next, we explore the distribution of the readability levels of the English examination texts indicated by the readability measures. We present the distribution boxplots in Figure 4.1. As can be observed from these figures, the data is fairly evenly distributed.

The boxplots in Figure 4.1 also indicate that the FKGL, ARI, CLI, GFI, LIX, SMOG, RIX and DCI of the English examination texts are consistently higher for the HL and lower for the FAL. As expected, the FRE is higher for the FAL and lower for the HL.

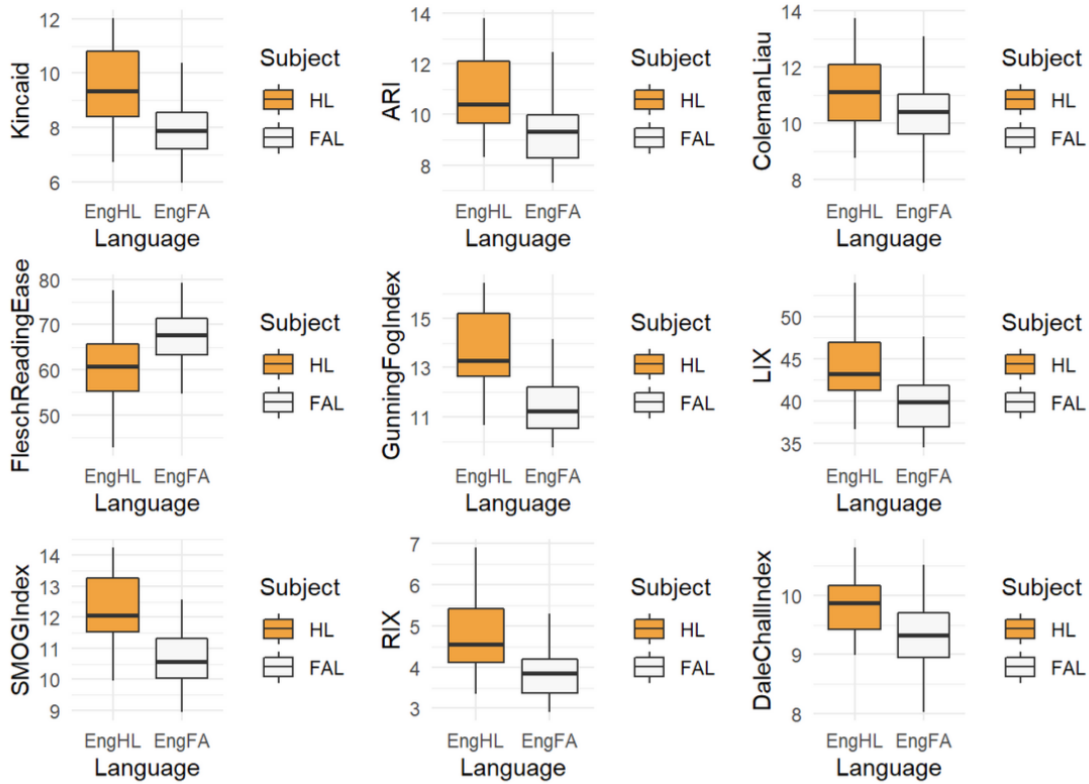


Figure 4.1: Data distributions box plots for all of the measures, separated by the HL and FAL texts.

Third, we provide mean and standard deviations for the different measures for both the HL and FAL texts in Table 4.2. In Table 4.2, we see consistent differences in the scores where FAL texts are considered to be easier than the HL texts. These results do not

Table 4.1: Mean and standard deviation (SD) scores for English examination text properties used in the measures.

Feature	February						May						November						Overall					
	FAL		HL		FAL		HL		FAL		HL		FAL		HL		FAL		FAL		HL		HL	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Words	915.25	37.72	1,173.75	26.83	891.00	17.42	1,157.00	40.35	962.42	72.80	1,134.67	83.04	934.79	63.27	1,151.42	65.35								
Syllables	1,347.25	65.01	1,726.50	84.87	1,315.00	45.23	1,774.75	67.93	1,373.42	75.46	1,727.00	206.78	1,353.87	71.09	1,734.79	157.70								
Syllable/word	1.47	0.08	1.47	0.07	1.48	0.03	1.54	0.11	1.43	0.07	1.52	0.10	1.45	0.07	1.51	0.09								
Sentences	55.63	1.51	64.38	3.08	54.00	0.59	63.75	6.42	59.17	4.46	59.08	8.10	57.13	3.62	61.63	7.02								
Words/sentence	16.71	2.22	18.27	0.81	16.53	0.87	18.30	1.55	16.45	1.69	19.48	2.54	16.54	1.79	18.88	2.05								
Letters	4,400.25	180.79	5,562.63	208.50	4,283.50	99.65	5,648.75	176.94	4,534.75	246.28	5,539.50	549.93	4,443.48	227.84	5,565.42	415.23								
Letters/word	4.81	0.23	4.74	0.16	4.81	0.05	4.89	0.28	4.72	0.19	4.87	0.20	4.77	0.19	4.83	0.22								
Long words	221.13	30.27	278.50	22.07	204.50	16.99	308.25	42.07	220.75	19.97	296.92	48.67	218.13	24.30	292.67	41.93								
Complex words	121.63	23.58	166.88	19.18	113.75	11.21	193.25	30.29	108.83	23.84	179.58	47.06	113.74	22.87	177.63	38.26								
Difficult words	281.75	35.74	375.00	26.19	269.50	25.27	395.25	27.63	294.00	26.69	388.33	47.71	285.09	31.17	385.04	39.37								

Table 4.2: Means and standard deviations (SD) of the different readability measures for the English HL and FAL examination texts.

Measure	HL		FAL	
	Mean	SD	Mean	SD
FKGL	9.53	1.49	7.99	1.10
FRE	60.18	8.60	67.18	6.20
SMOG	12.26	1.12	10.74	0.94
GFI	13.68	1.57	11.51	1.25
ARI	10.76	1.59	9.28	1.35
CLI	11.06	1.30	10.42	1.17
LIX	44.22	4.30	39.97	3.61
RIX	4.80	0.92	3.88	0.67
DCI	9.85	0.48	9.30	0.60

consider the influence of the different months or years. This will be investigated in more detail with the regression ANOVAs.

Table 4.3: Correlation between the different readability formulas. All correlations are significant ($p < .0001$).

Measure	FKGL	FRE	SMOG	GFI	ARI	CL	LIX	RIX	DC
FKGL	1	-0.95	0.94	0.96	0.85	0.96	0.95	0.96	.78
FRE	-0.95	1	-0.93	-0.92	-0.93	-0.87	-0.91	-0.88	-.78
SMOG	0.94	-0.93	1	0.99	0.81	0.87	0.89	0.89	.73
GFI	0.96	-0.92	0.99	1	0.79	0.9	0.91	0.92	.74
CLI	0.85	-0.93	0.81	0.79	1	0.85	0.89	0.84	.84
ARI	0.96	-0.87	0.87	0.9	0.85	1	0.96	0.97	.81
LIX	0.95	-0.91	0.89	0.91	0.89	0.9	1	0.99	.84
RIX	0.96	-0.88	0.89	0.92	0.84	0.97	0.99	1	.81
DCI	0.78	-0.78	0.73	0.74	0.84	0.81	0.84	0.81	1.00

The results of Pearson's correlations between the results of the different formulas are presented in Table 4.3. All of these correlations are significant ($p < .0001$). We see that most pairs of measures show strong positive correlations, except for the FRE measure, which (in contrast to the other measures) shows strong negative correlations as higher values mean easier-to-read texts. The lowest absolute correlations are found for DCI and SMOG ($r = .73$), and DCI and GFI ($r = .74$) measures (although these are still considered strong correlations). Overall, these results show that the measures provide very similar behaviour.

We also present density plots (see Figure 4.2) for the different measures for both HL and FAL texts. This shows that the readability scores are generally normally distributed.

4.2.4.2 ANOVA analyses

To investigate the influence of the subject level (English HL and English FAL), and the year and month of the examination on the readability, we performed ANOVA analyses for each of the readability measures. For this, we use Subject (HL vs FAL), Year, and Month as independent variables (we also consider the possibility of interaction between the Year and Month variables in the model) and the readability score as the dependent variable.

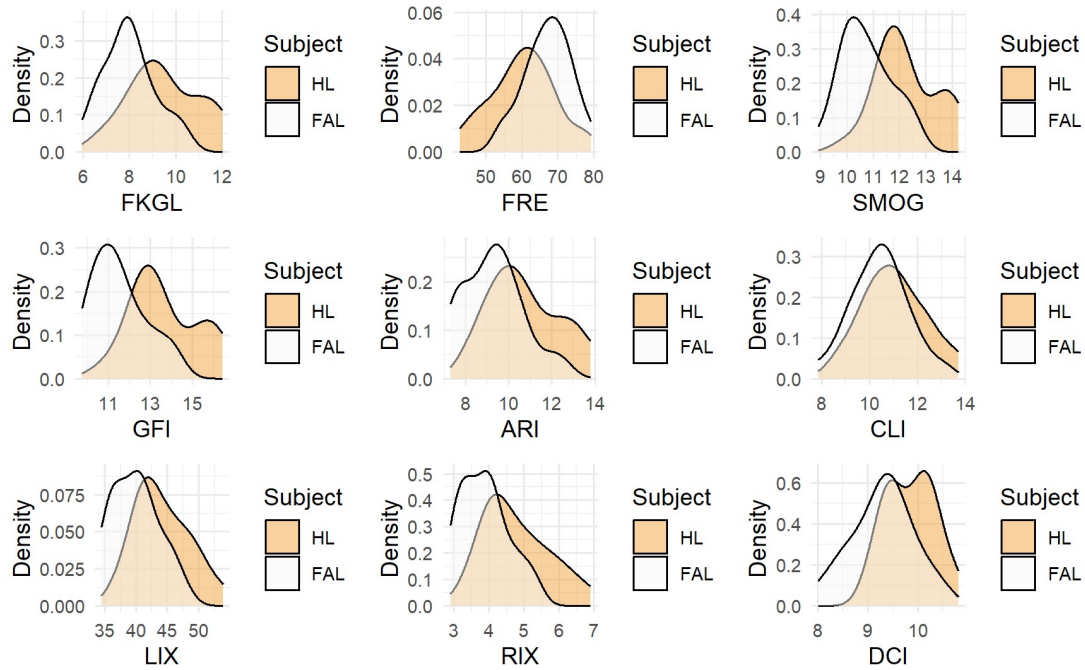


Figure 4.2: Data distribution density plots for all of the measures, separated by the English HL and FAL texts.

Both Year and Month are sum-coded, allowing investigation of the influence of these variables concerning the mean values. Note that although the readability scores are strongly correlated, they are not combined in any model.

For all measures, the residuals conform fairly well to the normality assumption. This can be observed in the Q-Q plots presented in Figure 4.3.

The outcomes of all the ANOVAs are presented in Table 4.4.

As can be observed from Table 4.4, all the ANOVA results indicate significant differences between the readability values of English HL and FAL texts for each of the nine measures. Note that the models (according to ANOVA analyses) did not show any significant differences resulting from the exam month ($p > .05$).

The results only showed a significant influence of the Year variable on the overall CLI scores ($F(11, 23) = 2.3303; p = .04$). Further analysis indicated, in at least two instances, significant differences influenced by the year of examination, specifically, YearSum 5 ($p =$

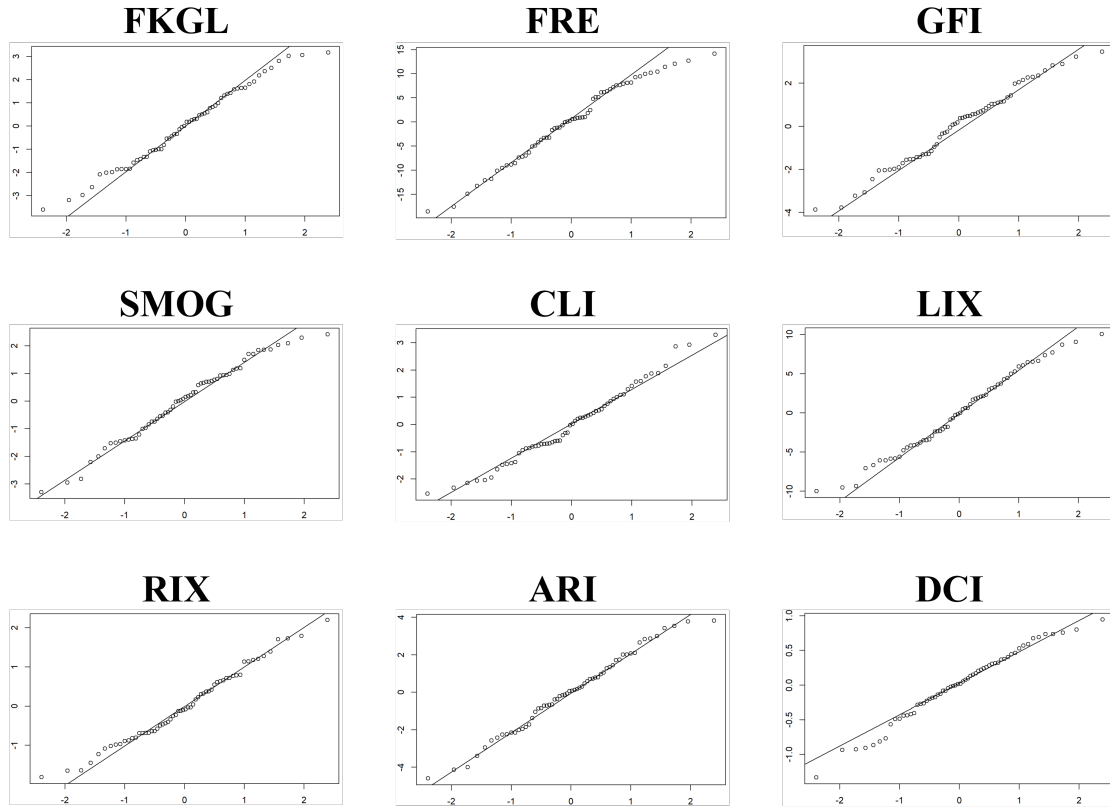


Figure 4.3: Data distribution Q-Q plots for all measures.

.047) and YearSum 11 ($p = .045$). The DCI also demonstrated an influence of the Year variable in YearSum 5 ($p = .043$). However, the overall ANOVA results did not indicate a significant influence of Year on the DCI results.

Furthermore, the ANOVAs for the DCI scores showed an overall significant interaction between the Year and Month variables ($F(10, 23) = 3.2075; p = .01$). Further analysis indicated, in at least two instances, significant interaction between the Year and Month variables, that is in YearSum 10 and MonthSum 1 ($p = .0021$) and YearSum 10 and MonthSum 2 ($p = .0066$). Note that DCI already showed the largest deviations in correlations with other measures, which may be due to the significant influence of the interaction between the Year and Month variables.

4.2.5 Discussion and conclusions

In this section, we explored the readability of the English HL and FAL high school examination texts used in South Africa. We used nine traditional readability measures to investigate the readability of each text. We showed that the nine measures are significantly and positively correlated with each other.

Table 4.4: ANOVA results. *Difference* indicates the estimated difference between the English HL and FAL values. *Error* indicates the standard error. $F(1, 23)$ provides the results of the F -test with the corresponding p -values between the texts.

	Difference	Error	$F(1, 23)$	p
FKGL	-1.54	0.38	16.8471	.0004
FRE	7.01	2.08	11.3126	.0027
SMOG	-1.53	0.32	23.2077	<.0001
GFI	-2.16	0.44	24.5655	<.0001
CLI	-0.65	0.31	4.3258	.0489
ARI	-1.48	0.42	12.4651	.0018
LIX	-4.25	1.13	14.0233	.0011
RIX	-0.92	0.23	15.3211	.0007
DCI	-0.55	0.12	21.5054	.0001

Our findings suggest that the texts used in the English HL and FAL grade twelve examinations show consistency over time. Specifically, our analysis reveals a lack of statistically significant differences among the examinations conducted in the different years and months with the exception of the occasional divergences observed for the DCI and the CLI scores.

Nonetheless, the significant differences between the HL and FAL texts correspond to the viewpoint that the HL subject is more complex and caters for learners with higher language competency than those in the FAL subject. From this result, we hope that this will also be the case for the examination texts of the indigenous languages, particularly Sesotho.

The nine measures investigated in this section all show similar behaviour. First, the correlation results indicate no significant differences between any of the measures. All measures correlate strongly when considering the readability of the examination texts. Second, the measures provide similar regression ANOVAs with only minor differences.

We observe two unexpected findings. First, the CLI shows a strong positive correlation with other measures in the study, but the ANOVAs indicate that there are significant differences in the CLI levels between some of the different years.

Second, the DCI ANOVA reveals statistically significant differences concerning Years and Months. Regrettably, this exploratory research does not delve into these specific nuances in detail. However, to adopt a measure based on the DCI, it is necessary to generate a list of commonly used words in the chosen language. Currently, two such lists exist for the official languages of South Africa. First, Jansen et al. (2017) discuss the list of 50 frequently used Afrikaans words extracted from the Media24 corpus. Second, we present a list of commonly used words in Sesotho in Section 3.3.

One of the aims of the exploratory research described in this section was to get a sense of

whether examination texts can be used in the development of text readability measures for South African indigenous languages. Given that the indigenous language curricula are translated from the English generic curricula we expect similar readability characteristics for the indigenous languages examination texts. That is, the HL and FAL texts for the indigenous South African languages can be expected to be significantly different while texts used at different examinations are expected to show no significant differences in readability.

4.3 Assessing Sesotho examination text readability with English traditional measures

4.3.1 Introduction

In this section, we explore the readability levels of Sesotho examination texts. We specifically focus on the analysis of the variations and similarities in the readability of the Sesotho HL and FAL texts.

As highlighted previously, the conventional assessment of the readability of Sesotho texts has traditionally relied on English readability metrics, as evidenced by the scholarly works of Krige (2009), Krige and Reid (2017), and Reid et al. (2019). Notably, the present research diverges from this precedent, opting not to attribute a definitive status to the readability levels determined through the use of English readability measures when evaluating the text readability of the Sesotho texts. Instead, we employ these measures to gain insight into possible dissimilarities in the text readability levels of Sesotho texts. We subject these texts to testing to ascertain that they do not possess identical levels of readability. Consequently, our focus lies not on the precise readability levels themselves, but rather on discernible distinctions in these levels.

The fundamental proposition presented here asserts that if the application of English readability measures reveals variations in the readability levels of Sesotho texts, then the rationale for employing these texts in the training of our readability formulas for Sesotho would be substantiated. In other words, the distinctions in readability levels identified through the use of English readability measures imply that the text attributes within our data set exhibit sufficient diversity to justify using these texts for the development of readability measures tailored to Sesotho.

4.3.2 Methodology

4.3.2.1 Materials

For our investigation, we extracted a total of 80 Sesotho examination texts from the corpus of grade twelve examination texts discussed in Section 3.2. We specifically focused on 40 Sesotho HL and 40 Sesotho FAL texts composing reading comprehension and summary writing texts for Sesotho ranging from the year 2011 to 2019. These texts, like the English ones, are also written in either November, February, or May. For our purpose, we cleaned the texts by removing ‘noise’ such as page numbers, line and paragraph numbers, sentence markers, footnotes, endnotes, and source references.

For computing the scores, we used the University of Ghent Language and Translation Team’s web-based readability tool (see Section 3.4.2.3). Since the web-based tool is based on English and Dutch traditional readability measures, a few moderations were warranted. For instance, the DCI uses a list of frequently used words (Gunning, 1969; Heydari, 2012), as such, we replaced the Dale-Chall list (Dale and Chall, 1948; Chall and Dale, 1995) with our list of commonly used Sesotho words. Furthermore, we replaced the English syllabifier with our rule-based Sesotho syllabification system for identifying syllables (see Section 3.5 for the description of the syllabification system). We did not implement any further changes to the measures. An example of the original end-user screen of the LT3’s demo is presented in Figure 3.1.

The results of the analysis of the readability levels of Sesotho texts were organised into distinct HTML files, each dedicated to one of the 80 files we analysed. We then used *bash*² to grab the readability statistics into a spreadsheet that we thereafter used for the analysis of the readability levels using *Rstudio*³.

4.3.2.2 Analysis

To assess the performance of various readability measures on the Sesotho HL and FAL examination texts, we commence by presenting the mean and standard deviation values for each respective measure. Furthermore, we explore the correlations of the readability scores from the different readability measures employed in this section. Based on our findings for the Sesotho-English translated texts in Section 3.4 and the English examination texts discussed in Section 4.2, we anticipate significant correlations between the readability

²Bash is a powerful scripting language used for automating tasks, managing files and directories, and interacting with the operating system. It is widely used in Linux and Unix for system administration and automation.

³RStudio provides a user-friendly interface for working with R which is a programming language and software environment specifically designed for statistical computing and graphics. RStudio includes a script editor, a console for direct R code execution, and various tools for data manipulation, visualisation, and statistical modelling.

measures.

Once the descriptive statistics are provided and discussed, we employ ANOVAs to investigate the relationship between the readability of the HL and FAL texts, including the years and months of the exams.

4.3.3 Results

4.3.3.1 Descriptive statistics

To get a better understanding of the behaviour of the measures on the texts, first, an overview of the textual properties used in the measures discussed in this section is presented in Table 4.5. The table provides an overview of various linguistic attributes, including the mean counts of words, unique words, and sentences. It encompasses the type-token ratio, average sentence length (ASL), and the average word length by both letter count (AWL, letters) and syllable count (AWL, syllables). Moreover, the percentages of uncommon words, those not present in our designated list of common Sesotho words, along with polysyllabic words consisting of three or more syllables, are also featured.

Table 4.5: Mean scores and standard deviations (SD) for the properties of Sesotho texts used in the measures.

	HL				FAL			
	SUM	RC1	Average	SD	SUM	RC1	Average	SD
Words	248.05	660.82	399.40	207.27	351.74	806.00	518.3	235.01
Unique words	120.63	249.64	167.93	65.34	150.53	303.18	206.5	76.89
TTR	0.49	0.38	0.45	0.06	0.43	0.38	0.41	0.04
Sentences	15.37	43.82	25.80	16.22	18.79	47.18	29.20	16.54
ASL	16.57	16.14	16.42	3.26	19.38	18.25	18.97	4.26
AWL (letters)	4.15	3.96	4.08	0.26	4.13	4.05	4.10	0.21
AWL (syllables)	2.02	1.97	2.00	0.09	2.02	1.98	2.00	0.10
% uncommon words	11.82	15.2	13.06	3.50	12.17	14.95	13.19	4.36
% polysyllabic words	30.83	28.82	30.09	3.53	29.12	28.65	28.95	3.11

As can be observed from Table 4.5, the averages of the Sesotho FAL text properties are largely higher than those of the Sesotho HL texts except for the TTR, the percentages of common words and polysyllabic words. The average word length by number of syllables is even for both the Sesotho HL and FAL texts.

Furthermore, the standard deviations for the Sesotho FAL texts are consistently lower than the Sesotho HL standard deviations. This means that the values of the text properties of the FAL texts are closer to the means than those of the Sesotho HL texts.

Next, we present the boxplots illustrating the distribution of the readability levels in

Figure 4.4. Unlike the text property data presented in Table 4.3, the distribution of the readability levels, as shown by the boxplots, reveals higher means for the Sesotho HL texts and comparatively lower means for the Sesotho FAL texts. In contrast, we observe higher FRE means for the Sesotho FAL texts compared to the Sesotho HL texts. Collectively, these findings suggest that the Sesotho FAL texts are easier to read than the Sesotho HL texts. Nevertheless, the data distributions around the means indicate a relatively even distribution of readability levels for both the HL and FAL texts respectively.

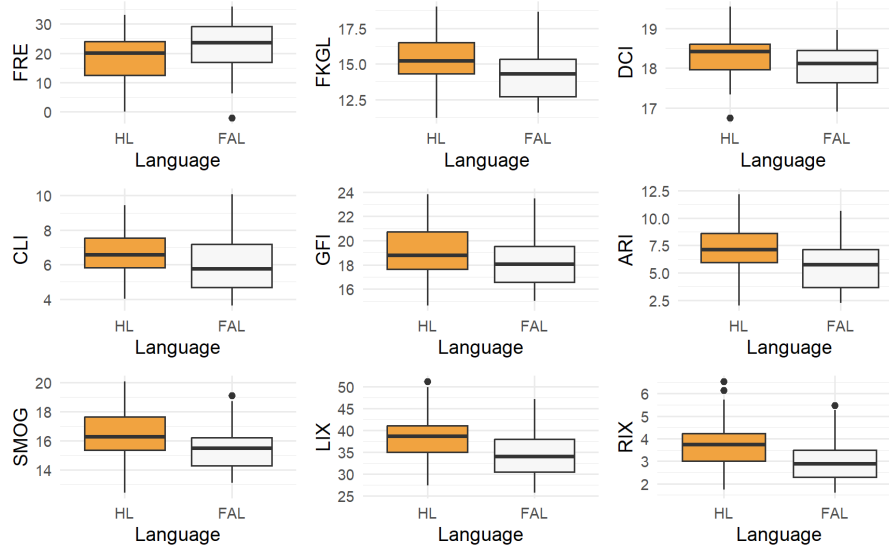


Figure 4.4: The distribution of readability levels depicted in box plots.

Table 4.6: Means and standard deviations (SD) of the different readability measures for the Sesotho HL and FAL examination texts.

Measure	Sesotho HL		Sesotho FAL	
	mean	SD	mean	SD
FKGL	15.49	1.9	14.48	1.83
FRE	17.81	8.79	20.54	9.01
SMOG	16.42	1.61	15.67	1.51
GFI	19.17	2.04	18.61	2.01
CLI	6.72	1.29	6.33	1.67
ARI	7.37	2.25	5.98	2.27
LIX	38.83	5.52	35.48	5.61
RIX	3.78	1.1	3.17	1.01
DCI	18.31	0.52	18.02	0.53

To gain a clearer understanding of the boxplots presented in Figure 4.4, we present the mean and standard deviations for the scores of the different readability measures for both the Sesotho HL and FAL texts in Table 4.6. We observe consistently higher means for the readability levels of the Sesotho HL texts, specifically for the FKGL, ARI, CLI, GFI,

LIX, SMOG, RIX, and DCI. Conversely, the FRE is higher for the Sesotho FAL texts and lower for the HL texts, which is in line with our expectations given that FRE indicates easier-to-read texts with higher values (refer to Chapter 2 for the interpretation of the scores).

Table 4.7: Correlation between the different readability formulas. All correlations are significant ($p < .0001$).

	FKGL	FRE	SMOG	GFI	CLI	ARI	LIX	RIX	DCI
FKGL	1	-0.9	0.97	0.96	0.79	0.98	0.97	0.96	.57
FRE	-0.9	1	-0.8	-0.88	-0.93	-0.81	-0.84	-0.83	-.55
SMOG	0.97	-0.8	1	0.97	0.69	0.97	0.93	0.94	.56
GFI	0.96	-0.88	0.97	1	0.79	0.93	0.92	0.92	.56
CLI	0.79	-0.93	0.69	0.79	1	0.75	0.77	0.73	.58
ARI	0.98	-0.81	0.97	0.93	0.75	1	0.96	0.96	.58
LIX	0.97	-0.84	0.93	0.92	0.77	0.96	1	0.99	.51
RIX	0.96	-0.83	0.94	0.92	0.73	0.96	0.99	1	.50
DCI	0.57	-0.55	0.56	0.56	0.58	0.58	0.51	0.5	1

Table 4.7 presents the results of Pearson's correlations between the outcomes of the different formulas. All of these correlations are significant ($p < .0001$). The results show that most pairs of measures exhibit strong positive correlations, except for the FRE measure which displays a strong negative correlation as higher values indicate easier-to-read texts. The FRE measure has a strong negative correlation with all other measures except with the DCI where it has a moderate negative correlation ($r = -.55$). The DCI also has moderate positive correlations with RIX ($r = .50$), LIX ($r = .51$), CLI ($r = .58$), ARI ($r = .58$), GFI ($r = .56$), SMOG ($r = .56$), and FKGL ($r = .57$). Near-perfect correlations can also be observed for the LIX and the RIX ($r = .99$) as well as ARI and FKGL ($r = .98$). Overall, these results demonstrate that the measures exhibit very similar behaviour. Note that although the FRE measure indicates negative correlations, it is reversed as higher FRE scores indicate easier-to-read texts while the other scores indicate easier-to-read texts through lower scores.

4.3.3.2 ANOVA analyses

To investigate the influence of the subject (HL and FAL), as well as the year and month of the examination on readability levels, ANOVAs were calculated for each of the readability measures. For this purpose, we used Subject (HL versus FAL), Year, and Month as independent variables (we also considered the possibility of interaction between the Year and the Month variables in the model), and the readability score as the dependent variable.

The results of the ANOVAs presented in Table 4.8 reveal significant differences between the Sesotho HL and FAL texts in at least eight readability measures ($p < .05$). Specifically,

Table 4.8: ANOVA results. *Difference* indicates the estimated difference between the Sesotho HL and FAL values. *Error* indicates the standard error. $F(1, 75)$ provides the results of the F -test with the corresponding p -values.

Measure	Difference	Error	$F(1, 75)$	HL vs FAL	Year	Month	p -value
FKGL	-1.01	0.45	1.54	0.0033	0.0015	0.9646	.6378
FRE	2.72	2.19	4.9596	0.0491	0.0007	0.8803	.4320
SMOG	0.66	0.37	3.1066	0.0094	0.0016	0.9359	.5021
GFI	-0.57	0.48	1.3677	0.0421	0.0008	0.8586	.3442
CLI	-0.39	0.37	1.0986	0.0638	0.0012	0.7115	.5080
ARI	-1.39	0.56	6.2555	0.0013	0.0029	0.9815	.7502
LIX	-3.39	1.36	6.0135	0.0018	0.0041	0.7038	.6125
RIX	-0.61	0.26	5.6562	0.0024	0.0025	0.6421	.5020
DCI	-0.29	0.14	4.5433	0.0233	0.0070	0.4934	.5980

significant differences were observed for the FRE ($p = .049$), FKGL ($p = .003$), SMOG ($p = .009$), GFI ($p = .042$), ARI ($p = .001$), LIX ($p = .002$), RIX ($p = .002$), and DCI ($p = .023$) measures. However, no significant differences were found for CLI measures ($p > .05$).

We also investigated the possible influence of the interactions between the years of examination and the months of examination. None of the results indicated an influence of the interaction between Year and Month variables ($p > .05$). Similarly, there were no significant influences of Month on the overall readability levels ($p > .05$).

Finally, we investigated the influence of the Year on the readability levels. As can be seen in Table 4.8, our findings indicate that all measures are influenced by the year of examination at ($p < .05$). In other words, examinations in some years are easier to read than examinations in other years.

4.3.4 Discussion and conclusions

In this section, we comparatively investigated the readability of Sesotho HL and FAL high school grade twelve examination texts used in South Africa. We employed nine traditional readability measures that were developed for English to gain insights into the complexity of these assessments. The primary objective of this exploratory research was to determine if there are variations in the readability levels of our collection of Sesotho grade twelve examination texts.

As previously mentioned, it is important to emphasise that the readability levels are not being posited as conclusive determinants of school grades in Sesotho. Therefore, while Table 4.6 showcases mean grade levels and readability levels for Sesotho HL and FAL subjects, it is crucial to understand that these values solely underscore the existence of discrepancies between the two subjects, without implying any other conclusions.

The results of this investigation revealed a significant and positive correlation between the nine measures. Furthermore, the ANOVAs demonstrated a consistent difference between the Sesotho HL and FAL texts when utilising eight measures. We do not account for the non-significant differences in the readability of the CLI measure in this investigation.

Additionally, no significant differences were identified in the HL or FAL texts administered during different examination months. Even so, the Year variable indicated a significant effect on the readability levels.

These results underscore the variations in readability levels present within the texts, particularly when employing English measures for assessment. As a result, the outcomes derived from the analysis presented in this section serve to validate our use of Sesotho examination texts.

4.4 Adapting traditional readability measures into Sesotho

This section discusses the development of readability measures for Sesotho.

4.4.1 Introduction

Earlier chapters have laid the foundation for measuring text readability in LRLs and have identified a range of existing and new resources for enabling the assessment of text readability in Sesotho. In the earlier sections of this chapter, we particularly highlighted valuable insights that can significantly contribute to the field of measuring text readability in Sesotho.

In this section, we present our efforts at addressing the research gap identified by Leopeng (2019) that traditional readability measures ought to be adapted to South African indigenous languages. By so doing, we are also addressing our fourth research question: How can traditional readability measures be effectively modified and adapted to suit the specific characteristics of Sesotho? To achieve this, we adapt existing traditional readability measures to Sesotho using English as a higher-resource helper language for the low-resource Sesotho.

As stated earlier, traditional readability measures focus on shallow text properties (Van Oosten et al., 2010; Zamanian and Heydari, 2012). According to Janan (2011, p. 108), these

textual features encompass both semantic features such as sentence and word lengths, together with syntactic features such as the identification and quantification of syllable information.

It is worth noting that a consensus regarding the optimal text features for predicting text readability remains elusive (De Clercq et al., 2014). Therefore, although the focus on using traditional readability measures for educational purposes has decreased in recent years (Wray and Janan, 2013), traditional readability measures have not been rendered obsolete although they rely on limited raw or shallow text features. Furthermore, traditional readability measures remain easier to compute. For instance, Microsoft Suite has made it easy to compute traditional readability measures such as the FRE and FKGL. As such, traditional readability formulas continue to be used in multiple domains (Janan and Wray, 2014) despite the multiple criticisms against their use.

4.4.2 Background

This section presents an exploration of the nine traditional readability measures discussed in Chapter 2. These measures encompass the FKGL, FRE, SMOG, and GFI, which are reliant upon syllable-related information, as well as the CLI, ARI, RIX, and LIX measures which are based on word-length information, while the DCI draws from a list of commonly used words.

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

Given that the number of syllables in typically long words is language-specific, the English requirement of 3+ syllables may not be indicative of long words as measured by the number of syllables per word in Sesotho. For this reason, Kusec et al. (2002), adjusted the minimum syllables counted from the English helper language to the Croatian LRL. They use the top 100 frequently used words in English and Croatian to determine the differences between syllable counts in order to determine the number of syllables that are typical in Croatian long words. In the end, they adjusted the requirement for polysyllabic words to 4+ syllables.

Orthographic word length, that is, the lengths of words as measured by the number of letters per word (Ziegler et al., 2001), has been a topic of interest in language studies, with research indicating variations across languages and over time. For instance, Bochkarev et al. (2015) investigate the evolution of word lengths in English and Russian as observed through e-libraries, Google Books, and the system Ngram Viewer. Their findings indicate an increase in the average length of words in both languages, with English increasing from

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

Similar to word length by the number of letters, sentence length has also been a topic of interest in English language studies for many years. Above all, it is generally accepted that longer texts are harder to understand and may not interest people to read (Vincent, 2014). The average sentence length of English across four genres (magazines, newspapers, non-fiction, and fiction) was 16.46 words in the year 2000, which is down from 19.23 words in 1900 (Rudnicka, 2018). Comparably, Nirmaldasan (2016) indicates that the current standard for English sentences is 17 words with a total of 25 syllables. These averages fall within the recommended 15–20 words per sentence (Hult, 2015).

4.4.3 Methodology

De Clercq et al. (2014) list three steps for developing readability measures. That is, (i) the development of a readability corpus, (ii) describing a methodology, and (iii) undertaking the prediction tasks (François and Fairon, 2012; Collins-Thompson, 2014). We structure the discussion of our methodology for adapting the traditional readability measures into Sesotho using these three steps below.

4.4.3.1 Step 1: A readability corpus

The use of training texts holds great significance in both the development and adaptation of traditional readability measures. However, as highlighted in Chapter 2, it is unclear what the minimum size of text corpora needed for the development and adaptation of text readability measures to new languages is. As a consequence, previous studies which have effectively investigated the development and the adaptation of text readability to LRLs, reveal variations in corpus sizes.

The differences in the linguistic features between HRLs and LRLs can be investigated through the use of parallel translated corpora. Texts are parallel at different levels, including (i) the word level, as in the case of translated word lists, (ii) the level of the sentence, as exemplified by Cardon and Grabar (2020), (iii) the level of paragraphs, as demonstrated by El-Haj and Rayson (2016), and (iv) the complete documents.

Two consistent categories of parallel texts have emerged: intra-lingual translations, encompassing original texts and their simplified counterparts, akin to the German and simplified German texts featured in Battisti et al. (2020), and inter-lingual translations composed

of texts from distinct languages, exemplified by the English-Arabic text corpus discussed in El-Haj and Rayson (2016). Furthermore, the process of translation which is instrumental in the adaptation of readability measures, assumes different methods, inclusive of manual translation, machine-generated translation, or a fusion of both methodologies. For instance, the adaptation of traditional readability measures to Czech, as discussed by Bendová and Cinková (2021), embraces both manually translated texts and machine-translated texts crawled from the web.

However, in the context of Sesotho as an LRL, the absence of readily available corpora annotated with readability levels underscores the imperative to resort to freshly devised or repurposed corpora for training readability measures. In this way, the corpus developed and discussed in Section 3.4 is used in this thesis.

4.4.3.2 Step 2: A methodology

First, we extract relevant text features from Sesotho texts. Second, we use the English machine translations that correspond to the Sesotho texts to determine readability levels for the texts. Third, we use linear regression models to determine the impact of the text features on the overall scores.

We examined the text characteristics employed in traditional readability measures. These text features, which form the basis of the traditional readability measures explored within this section, are outlined in Table 4.9. This list provides a clear and insightful description of the essence of each feature.

The following brief discussion outlines some of the text features we are focusing on in this thesis.

4.4.3.2.1 Word lengths by letters As an agglutinative language, words may be expected to be long in Sesotho (Blanchard, 2011). On the other hand, monosyllabic words which may comprise between one and four letters, and single-letter words may result in shorter averages for Sesotho texts (Messerschmidt et al., 2003). Furthermore, overall text word length by the letters may be affected by the use of subject concords in Sesotho. Nonetheless, given that there is no consensus on the normal length of words in Sesotho as measured by letters, we follow the English guideline for the LIX and RIX measures and thus consider words with more than six letters as long words.

4.4.3.2.2 Word length by syllables Polysyllabic words refer to words with more than one syllable. However, as indicated in Chapter 2, traditional measures used in this thesis, particularly the SMOG and the GFI consider only words with three or more syllables as polysyllabic, foggy, and complex. Although Sesotho words tend to have more

Table 4.9: Text features used in the traditional readability formulas adapted in this thesis.

Attribute	Definition
Letters	Number of letters within a word
Words	Distinct semantic units, demarcated by spaces
Sentences	Grammatical constructs encapsulating thoughts or ideas, concluding with punctuation
Difficult words	Words not among the 3,000 most common words
Long words	Words containing seven or more letters
Polysyllabic words	Words comprising three or more syllables. Also referred to as polysyllables
Multisyllabic words	Words encompassing four or more syllables
Word length in letters	Average number of letters per word
Word length in syllables	Average number of syllables per word
Sentence length	Average number of words per sentence
Average word length (letters)	$\# \text{ letters} / \text{by total } \# \text{ words}$
Average word length (syllables)	$\# \text{ syllables} / \text{by total } \# \text{ words}$
Average sentence length in words	$\# \text{ words} / \text{by total sentence count}$
Average sentences per word	$\# \text{ sentences} / \text{by total sentence count}$
Percentage complex words	$(\# \text{ words with 3 or more syllables} / \# \text{ words}) \times 100$
Percentage long words	$(\# \text{ words with 7+ letters} / \# \text{ words}) \times 100$
Percentage difficult words	$(\# \text{ words not among common list} / \# \text{ words}) \times 100$

syllables than English words, we also consider words with more than two syllables. Unfortunately, we are not aware of any system for identifying abbreviations in Sesotho. As a result, we forego counting the syllables of abbreviated words as full words.

4.4.3.2.3 Difficult words As indicated in Chapter 2, the DCI measure is based on the assumption that there are words that are commonly used and should therefore be easy to read. According to this method, words that do not appear on the list of frequently used words are considered difficult.

For our experiment, we make use of our list of common Sesotho words⁴ (see Chapter 3).

⁴We do, however, use the list with caution since it was not derived from educational texts.

4.4.3.2.4 Samples We forgo the sampling steps used in the formulas such as the DCI and the CLI. In this way, the number of sentences refers to all sentences in the text instead of sampled sentences. As can be observed in Table 4.12, the formula focuses only on word lengths as counted in letters, and sentences as either sampled or in the whole text as will be the case in our Sesotho formula.

We pre-process texts to consistently extract text features from the texts. One of the steps taken is tokenisation which allows us to identify words used in the texts. We opted for a non-language-specific tokeniser, specifically *Ucto*⁵.

4.4.3.3 Step 3: Prediction tasks

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

We then further explore the relationships between the scores of the syllable-based measures determined by the English readability metrics on the English translations and the syllable information extracted from Sesotho texts. These insights are important in our understanding of textual features that affect the readability of Sesotho texts.

Finally, we create linear regression models to determine the coefficients of the different textual features using our Sesotho and English training data.

The exploration of the interrelationships among the text features used in the traditional readability formulas was carried out through the application of the Pearson correlation measure. The outcomes of these correlations are presented in Table 4.10. Note that all correlations are significant at $p < .05$.

The correlation analysis reveals a perfect alignment between the count of words and syllables ($r = 1$) and a similar alignment between the number of syllables and letters ($r = 1$). A lower correspondence is observed between the number of sentences and the number of long words ($r = .72$), with this correlation, though slightly lower, remaining strong and positive.

Furthermore, a compelling association comes to light between syllables per word and the frequency of polysyllabic words per word ($r = .90$), underscoring that a higher number of syllables is associated with more words that contain three or more syllables. This resonance extends to the strong correlations observed between the number of polysyllabic words per word and both long words per word ($r = .80$) and the number of syllables per

⁵Note that the South African CTeXTools project developed tokenisation tools for all South African official languages. The tool is accessible at <https://repo.sadilar.org/handle/20.500.12185/423>. For consistency between the analyses for the translated English texts and the Sesotho original texts, we have opted to use the language-independent *Ucto*.

Table 4.10: The correlation of text features used in the readability measures computed from the Sesotho texts.

Feature	Label	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
sentences	V1	1.00	.89	.87	.91	.72	.86	.86	.82	.84	.54	-.34	-.29	.29	-.48	.58	-.34	-.27
tokens	V2	.89	1.00	1.00	.94	.91	.98	.99	.96	.99	.14	-.25	-.25	.25	-.10	.46	-.23	-.24
words	V3	.87	1.00	1.00	.93	.92	.98	1.00	.97	.99	.10	-.24	-.24	.24	-.07	.43	-.21	-.23
difficult_words	V4	.91	.94	.93	1.00	.84	.94	.93	.91	.92	.28	-.20	-.18	.17	-.24	.71	-.21	-.16
long_words	V5	.72	.91	.92	.84	1.00	.91	.95	.97	.96	-.05	.11	.10	-.10	.08	.38	.17	.07
types	V6	.86	.98	.98	.94	.91	1.00	.98	.96	.98	.11	-.20	-.20	.19	-.09	.50	-.18	-.19
syllables	V7	.86	.99	1.00	.93	.95	.98	1.00	.99	1.00	.08	-.15	-.15	.15	-.05	.44	-.13	-.15
polysyllables	V8	.82	.96	.97	.91	.97	.96	.99	1.00	.99	.05	-.04	-.03	.03	-.03	.45	-.02	.00
letters	V9	.84	.99	.99	.92	.96	.98	1.00	.99	1.00	.06	-.12	-.13	.13	-.03	.43	-.10	-.13
sentences/word	V10	.54	.14	.10	.28	-.05	.11	.08	.05	.06	1.00	-.24	-.16	.16	-.95	.48	-.32	-.11
letters/word	V11	-.34	-.25	-.24	.20	.11	-.20	-.15	-.04	-.12	-.24	1.00	.93	-.93	.21	-.01	.89	.85
syllables/word	V12	-.29	-.25	-.24	.18	.10	-.20	-.15	-.03	-.13	-.16	.93	1.00	-.100	.14	.05	.88	.90
words/syllable	V13	.29	.25	.24	.17	-.10	.19	.15	.03	.13	.16	-.93	-.100	1.00	-.13	-.05	-.89	-.90
words/sentence	V14	-.48	-.10	-.07	.24	.08	-.09	-.05	-.03	-.03	-.95	.21	.14	-.13	1.00	-.44	.31	.08
difficult_words/word	V15	.58	.46	.43	.71	.38	.50	.44	.45	.43	.48	-.01	.05	-.05	-.44	1.00	-.08	.06
long_words/word	V16	-.34	-.23	-.21	.21	.17	-.18	-.13	-.02	-.10	-.32	.89	.88	-.89	.31	-.08	1.00	.80
polysyllables/word	V17	-.27	-.24	-.23	.16	.07	-.19	-.15	.00	-.13	-.11	.85	.90	-.90	.08	.06	.80	1.00

word ($r = .89$).

Contrarily, negative correlations emerge between the number of letters per word and the number of sentences ($r = -.34$), words ($r = -.24$), and long words ($r = -.20$). This denotes that increased letter counts per word are likely to result in a lower number of sentences, words, and long words.

Further analysis unveils strong positive correlations between syllables per word and letters per word ($r = .93$), long words per word ($r = .88$), and polysyllables per word ($r = .90$). Moreover, there are no strong correlations between words per sentence and other variables. Even so, we observe a moderate positive correlation between words per sentence and the number of sentences ($r = .54$), as well as the number of difficult words per word ($r = .48$). Notably, the strong negative correlation between words per sentence and sentences per word ($r = -.95$) is expected as it aligns harmoniously with the inverse nature of these two variables.

Our investigation extended to investigating the correlations between syllable-based formulas and pertinent syllable-related variables, yielding the outcomes presented in Table 4.11. The findings reveal a weak negative correlation between the number of syllables and the scores of the FKGL ($r = -.04$), the GFI ($r = -.02$) and the SMOG ($r = -.03$). This observation indicates that a higher number of syllables within the text is likely to result in lower scores for these three measures which would translate to harder-to-read texts. Additionally, a weak negative correlation emerges between the percentage of polysyllabic words that contain three or more syllables and the FRE ($r = -.45$).

In further deliberation, the prospect of modifying the criteria for defining polysyllabic words was entertained, encompassing a potential increase to a minimum of four syllables (4+ syllables). However, such an adjustment yielded lower correlations with the overall syllable count ($r = .90$) in contrast to the nearly perfect positive correlation observed with the minimum of three syllables (3+ syllables) and the overall syllable count ($r = .99$). Furthermore, the relationships between the final readability scores and both the 4+ syllables and 3+ syllables criteria consistently demonstrated lower correlations for the 4+ syllables variable as opposed to the 3+ syllables variable. Consequently, we did not change the minimum number of syllables for the syllable per word count in the GFI and SMOG formulas in order to stay as true as possible to the original formulas.

We then create linear regression models for the different measures. The formulas are presented in Table 4.12. Our proposed readability formulas for Sesotho maintain a degree of structural preservation for the DCI, CLI, SMOG, FRE, and FKGL formulas. Note that a more simplified version of the CLI formula would use the actual counts and not percentages and result in $CLI_{Sesotho} = -3.683470 + 3.8782(\frac{\#letters}{\#words}) - 72.7569(\frac{\#sentences}{\#words})$.

Overall, our analysis resulted in a few observations. Initially, we observed a reduction in the coefficients of syllables per word within the Sesotho formulas. For example, this

Table 4.11: The correlation of syllable-based measures and syllable information computed on the Sesotho texts.

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

manifests as a heightened and negative weighting for syllables per word within the Sesotho FRE formula.

Noteworthy is the proposal of two structures for the GFI formula. Both versions introduce an intercept for the formula, involving a deduction of 0.177916, a departure from the original formulation. The first proposed formula, $GFI(1)_{Sesotho}$, follows the structure of the original English formula more closely except the added intercept. On the other hand, the second formula, $GFI(1)_{Sesotho}$, adds a coefficient to the percentage of complex words, thereby deviating from the original structure.

We also introduce weights to the LIX and RIX formulas. The English counterparts of LIX and RIX, as discussed in Chapter 2, do not include weights for these measures. To align the readability values that were acquired through the application of English readability measures on the translated examination texts, with the text features obtained from the Sesotho texts, it was necessary to introduce weighting factors. This adjustment ensured a more accurate correspondence between the readability values and the adapted formulas. Thus, we introduce weights to both formulas while adhering to the considered variables. For the LIX, the impact of words per sentence is accorded weight, while the percentage of long words remains unaltered. However, in the RIX formula, we ascribe weight to the fraction of long words per word. Note that we also introduce intercepts for both the $LIX_{Sesotho}$ and $RIX_{Sesotho}$.

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

Table 4.12: Readability measures and corresponding adapted Sesotho formulas.

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

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

4.4.4 Evaluation

The evaluation of the efficacy of our formulas was conducted through the application of three distinct metrics: Adjusted *R*-squared, *F*-statistic, and residual standard error. The outcome of the evaluations is presented in Table 4.13.

First, the lower residual standard errors observed across all formulas serve as an indicator of the models' precision and accuracy in approximating the anticipated outcomes.

Second, the Adjusted *R*-squared metric indicates how much the independent variables

Table 4.13: Evaluations of the adapted linear regression models for Sesotho. Note that the p -values are all significant at $p < .001$

	FKGL	FRE	SMOG	GFI1	GFI2	CLI	ARI	LIX	RIX	DCI
F -statistic	293.3	118.3	113.0	124.7	109.1	117.4	433.4	144	269.5	23.3
Adjusted R -squared	.881	.748	.586	.610	.732	.746	.916	.784	.773	.361
Residual std. error	0.647	4.806	0.873	1.166	0.966	0.848	0.626	2.758	0.441	0.631

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

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

4.4.5 Discussion and conclusions

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

Furthermore, an examination of the correlation between sentence length and word length

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

Among other things, we explored whether four syllables would correlate more with the GFI and SMOG readability scores which rely in part on the number of words with three or more syllables. Our attempts to modify the minimum syllables criterion for the polysyllabic variable were contradicted by our findings which indicated that three syllables or more syllables correlated more with the readability scores than four or more syllables.

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

Furthermore, an empirical examination involving human participants could be undertaken as our proposed readability measures enable the readability assessment of texts. The empirical evaluation involving human participants would involve selecting and grading texts using our proposed formulas, and subsequently administering these texts to learners within the grade levels indicated by the formulas. Such an approach would facilitate the evaluation of our formulas' societal impact. However, this necessitates the delineation of explicit criteria to discern success or failure in the tests that would be administered to participants for evaluation of the readability levels suggested by our formulas for Sesotho, particularly given the prevailing challenges South African learners encounter in reading for comprehension.

Moreover, the utilisation of existing grade levels, albeit untested within the South African education context, underscores an avenue for future research. Subsequent inquiries could investigate the applicability of the FRE, CLI, LIX, and other indicators to South African grade levels, refining the contextual relevance of these metrics.

Our reliance on the prevailing adaptation of readability scores, particularly pertaining to formulas such as FRE, LIX, and RIX, within the South African context, is substantiated by the works of Bargate (2012), and Leopeng (2019). Perhaps future works can consider recalibrating such scores to the South African grades through human-based evaluation methods.

4.5 Summary

The aims of this chapter were two-fold. First, we investigated the use of traditional English text readability measures in order to answer the third research question: What can we learn from existing traditional text readability measures? Second, we discussed the adaptation of nine readability metrics to Sesotho in order to answer the fourth research question of this thesis: How can existing English traditional readability measures be modified and adapted to Sesotho?

To achieve our aim, we conducted two investigations. First, we investigated the readability of English grade twelve examination texts using the English readability measures in order to gain insight into the trends of readability in the examination texts. Our findings indicate consistency in the readability of texts at the HL and FAL levels. Particularly, the HL texts are consistently harder to read than the FAL texts.

Second, we investigated the readability of Sesotho grade twelve examination texts using the English traditional readability measures. The aim of this investigation was to gain insight into the disparities in the text readability levels of the Sesotho examination texts. Our findings indicate that there are indeed significant differences between the Sesotho HL and FAL texts. We also uncover that the levels of readability are impacted by the year of examination.

Finally, we extracted text features from Sesotho examination texts and matched them to text readability levels gauged through the use of English measures on English translations of the same Sesotho examination texts. Simple linear regression models were then built based on the matched text features and readability levels. The results of this experiment resulted in the reassignment of coefficients for the variables considered in the different readability measures developed for Sesotho.

Overall, we make a few observations. For instance, 1. the readability measures for Sesotho indicated a strong favour for the sentence length by the number of words feature, 2. the new formulas introduced intercepts for formulas such as the LIX and the RIX, and 3. coefficients were introduced for formulas such as the LIX, the RIX, and the GFI.

Chapter 5

General discussion

In this discussion chapter, we analyse the results of this thesis on how text readability can be measured in Sesotho. We also compare our findings to previous research results and then propose future research directions. The chapter also reviews the assumptions on which this thesis was based.

5.1 Introduction

This chapter provides a discussion of the entire thesis, highlighting its main contributions and limitations. The chapter begins by recapping the research assumptions and aims, followed by an overview of the thesis in the context of previous research, which highlights the specific gap addressed by this thesis thereby explicating the new contributions. The strengths and weaknesses of the research design, methodology, and analyses are then critically evaluated. Finally, potential future directions for research are suggested. The chapter concludes with a summary of the key points discussed in the remainder of the chapter.

5.2 Research assumptions

This thesis aimed to investigate how text readability could be measured in LRLs, with a particular focus on Sesotho (Southern Sotho). We employed the use of the South African orthography and therefore excluded texts written in the Lesotho orthography.

We made three main assumptions upon which we based our investigation. The assumptions are presented in Section 1.4. We revisit these assumptions in this section. First, this thesis assumed that Sesotho texts have different levels of readability. Furthermore, it was assumed that the texts can be ranked according to different levels of readability.

The second assumption was that the adaptation of traditional readability measures can enable their application in the assessment of Sesotho texts' readability. To evaluate this assumption, we investigated the applicability of general methods, techniques, and approaches that are used in measuring text readability in under-resourced languages on Sesotho texts. Furthermore, we also considered the methods that could be transferred from HRLs such as English to LRLs such as Sesotho. A review of existing methods revealed many options, and while some were deemed stronger, the decision was made to adopt traditional readability measures which are well-researched and commonly used methods. Moreover, it was posited that insights from English readability measures could be informative, as English is the most resourced language (Tsvetkov, 2017), and many readability measures have been developed for English (Zamanian and Heydari, 2012).

The third assumption was that an automated process would standardise and simplify the procedure of ranking Sesotho texts according to their different levels of readability. To assess this assumption, we examined Sesotho as an LRL and the implications this had on the efforts to automate the process of assessing readability in Sesotho. We, therefore, surveyed available resources and those that could be developed to enable measuring text readability in Sesotho without ignoring the unique structure of written Sesotho texts. Our survey of available resources revealed several key gaps, including the lack of a gold-standard corpus with readability assessment information.

To address the lack of resources, we employed a few strategies for curating and developing missing resources. Such resources also allowed us to implement a machine-learning approach to the development of readability measures for Sesotho texts.

In summation, the results of our study provide strong evidence substantiating our assumptions concerning the classification of Sesotho texts based on different levels of readability. The successful implementation of these assumptions holds significant implications for the field of text readability in LRLs, especially those in Southern Africa and those related to Sesotho.

5.3 Methodology overview

The research project utilised different methods. First, the systematic literature review allows for reproducibility as all methods are stated. Furthermore, the inclusion and exclusion criteria streamlined the purpose of the research. We also made use of the basic language resource kits method for surveying existing resources in Sesotho. In this way, we managed to identify existing methods as well as those that needed to be developed.

We also contributed to the advancement of computational research on Sesotho by developing a few digital language resources using both manual and computational approaches. These resources are made available for free. As such, they can be freely used for future research.

Our evaluations primarily made use of the nine text readability measures selected for adaptation in this thesis. We analysed the correlation of the different measures given that the measures consider different text properties. We also relied on ANOVAs to identify differences between our examination texts.

Finally, we developed readability measures for Sesotho based on information extracted from Sesotho exam texts which are aligned with readability values that are computed on the translated English versions of these texts. The resulting measures describe a large amount of variability in the values. This indicates that these are strong measures.

5.4 Comparison to previous research

This section provides a comparative analysis of the current thesis with existing literature on text readability research in LRLs. The primary objectives of this section are twofold. Firstly, it situates this thesis within the broader literature on text readability by highlighting its significance and contribution to the field. Secondly, it validates the thesis by identifying similarities and differences between its findings and those of previous research. To achieve these objectives, we first summarise the relevant literature on text readability

in LRLs and then compare the findings of this thesis to existing research in this area. Finally, we discuss the unique contributions of this thesis to the field of text readability research in LRLs.

5.4.1 Literature review summary

Crossley et al. (2023) make three important observations about text readability measures, that is, they rely on the estimations of lexical complexity by measuring word lengths, and they estimate syntactic complexity by measuring sentence length by the number of words per sentence. Additionally, they note that although the DCI is seemingly based on a frequency word list, it also uses syntactic complexity by measuring the number of words per sentence.

Our findings indicate that, unlike Chinese texts for which traditional readability measures such as those used in this thesis are not applicable (Feng et al., 2022), the measures are transferable to Sesotho as it uses the same writing system as English. In this way, our research also relies heavily on lexical complexity as proposed by Crossley et al. (2023). Furthermore, we followed the steps set out by De Clercq et al. (2014) for developing readability measures. The steps resulted in the successful adaptation of text readability formulas for Sesotho.

5.4.2 Comparison to previous research

According to Gkikas et al. (2022), text readability studies focus on longer texts. To address this skewed focus, their research explores the use of the FKGL and the GFI to measure the text readability of social media posts. In this thesis, we followed the guidelines of using texts with at least 200 words for assessment (Zamanian and Heydari, 2012), although Sazzed (2022) indicates a minimum of 100 words.

Our application of English readability measures to Sesotho texts aligns with the approach taken by Bendová and Cinková's (2021) in adapting readability metrics to Czech. In their study, they use English readability measures for both English and Czech texts, subsequently establishing correlations between the outcomes. Similarly, we conduct a correlation analysis between the results of the readability measures on translated English texts with the text features in the original Sesotho texts. As observed in Bendová's (2021) study, the use of already existing readability formulas on text features of the new language can help determine which formula closely matches the structure of the new language. In this way, Bendová (2021) evaluates English, Russian, Italian, and French formulas for the adaptation of readability measures to Czech.

Oborneva (2006) introduces adaptations of both FRE and FKGL measures for the Russian language. The proposed formulas allocate lower coefficients for the ASL and AWL

compared to our proposed FRE and FKGL formulas for Sesotho. However, as observed by Solovyev et al. (2018), the use of the FKGL formula proposed by Osborneva (2006) yields inflated scores when assessing Russian academic texts. In response, they propose new FRE and FKGL adaptations for the Russian. Similar to our proposed Sesotho FRE formula, their revised FRE formulation assigns reduced weights to average word lengths by syllables and higher weights to sentence lengths within the Russian FRE, deviating from both the English FRE and Osborneva’s (2006) Russian FRE.

According to Jansen et al. (2017), four readability measures have been developed for Afrikaans, one of the official languages of South Africa. Notably, McKellar (2008) developed a readability measure for Afrikaans drawing from a pool of 200 texts to develop a readability metric for Afrikaans. While rooted in the FRE formula, this novel formula introduced innovative elements, including a compilation of 50 frequently used words and considerations for numerical figures and symbols. In contrast, our proposed formulas for Sesotho do not introduce novel variables to the adapted readability measures.

In another study, McDermid Heyns (2007) introduced a distinct readability measure for Afrikaans by incorporating average sentence lengths and the average word length based on syllable counts, emphasizing words with three or more syllables. According to Jansen et al. (2017), this approach bears resemblance to an adaptation of the FRE formula. Notably, McDermid Heyns’s (2007) adaptation assigns reduced weight to the ASL compared to our proposed formulas. Moreover, McDermid Heyns (2007) suggests a shift in the variables considered within the FRE measure, by incorporating the polysyllabic words variable. However, our adaptation of the FRE formula for Sesotho remains faithful to the original text features considered in the formula.

Van Rooyen (1986), on the other hand, used cloze tests and multiple-choice assessments to formulate two readability metrics for Afrikaans, based on a corpus of 30 texts. This dataset contains much fewer texts than the collection of examination texts we employed for training Sesotho readability measures in this thesis. Van Rooyen’s (1986) work contributed two additional Afrikaans readability measures, derived from cloze tests and multiple-choice texts respectively. In contrast, our approach solely uses text-based methods, without incorporating human evaluations.

Interestingly, among the Afrikaans readability measures, only the formula proposed by McKellar (2008) demonstrated a significant correlation with comprehensibility, as reported by Jansen et al. (2017). In our thesis, we refrain from assessing comprehensibility, focusing exclusively on text features that affect the ease with which a text is read. As such, we abstain from incorporating user evaluations in our methodology.

5.4.3 Contribution to the field

Although readability research dates back to the 1920s for languages such as English (Wilkins et al., 2022), and more than 200 measures have been developed since then (DuBay, 2004; Lee and Vajjala, 2022), the problem of text readability remains unsolved (Jian et al., 2022). Even worse, text readability research in South African indigenous languages has not seen any development of readability measures. As far as could be ascertained for this thesis, this is the first investigation that looks into the development of text readability measures for a South African indigenous language.

Furthermore, this thesis makes a notable contribution to the scholarship of text readability in LRLs by expounding upon the strategies employed to address the challenges posed by the lack of resources needed for the development of readability measures for LRLs like Sesotho. In fact, a number of resources are developed and made publicly available. These resources are not specifically designed for readability only and as such may boost further computational linguist research on Sesotho as well.

5.5 Strengths and weaknesses

This section provides a transparent assessment of this thesis by discussing its strengths and weaknesses. Identifying and addressing the limitations of the research is crucial for ensuring the validity and reliability of the findings. By examining the strengths and weaknesses of the research design, methods, and analysis, we can provide insights into the limitations of this thesis and guide future research in measuring text readability in LRLs like Sesotho.

5.5.1 Critical evaluation of research design

The use of traditional readability measures for developing language-specific measures of text readability is a well-established approach and was therefore a natural choice for the current research design. Furthermore, alternative approaches were hindered by the lack of large data.

The scope of the thesis was also limited by the non-existent readability corpus for Sesotho. Furthermore, the use of examination texts led to a limited variety of readability levels given that the texts are from the same school grade.

Additionally, the selection of the nine measures used in this thesis was based on their inclusion in the *Python 3* readability package and their compatibility with LT3's readability platform. Although other measures used in HRLs may also be adaptable to Sesotho, they were not included in this thesis because they are not part of the systems used in

the experiments forming this thesis. While this approach has its strengths, it would be interesting to explore more African indigenous language-oriented methodologies.

Overall, the research design demonstrated a clear understanding of established methods for measuring text readability but could benefit from broader and more diverse data sources and consideration of additional readability measures. Furthermore, a more text feature-based approach where all raw text features are considered simultaneously instead of limited to those considered in the specific formulas could yield a more accurate readability analysis of Sesotho texts.

5.5.2 Methodological strengths and weaknesses

The strengths of this thesis lie in its use of multiple traditional readability measures, which include the FKGL, FRE, SMOG, GFI, CLI, ARI, LIX, RIX, and DCI. The combination of these measures allows for a comprehensive evaluation of text readability from various angles, such as syllables, word lengths, sentence lengths, and the use of a frequency-based list. In this way, varied levels are expected as the measures are based on different premises (Krige, 2009, p. 106).

Moreover, the multiple measures approach helps to counter the adverse effects of using only one type of text, such as the high school examination texts, as the measures are best suited for varied text types.

Although the survey of methodologies for measuring text readability in LRLs indicates no specific pattern for the number of texts as measured by sentences, documents, genres, or levels of readability, this thesis could have benefited from more extensive data and diverse data types. Nevertheless, a notable advantage of our current text collection lies in its consistency within a specific context.

5.5.3 Data analysis strengths and weaknesses

This thesis utilised various methods of data analysis, including literature reviews using systematic state-of-the-art reviews, and a thematic review of resources for Sesotho using the BLARK approach. The combination of these literature analysis methods provides a comprehensive review of the literature.

In Chapter 4, evaluations of text readability were conducted using similar methods, employing the use of statistical approaches, including correlation, ANOVA, and linear regression. Correlation analyses were used to identify relationships, which was crucial for analysing text readability using different measures as it facilitated a better understanding of the factors that contribute to the levels of text readability. Furthermore, correlation results helped validate the properties used in each of the different results. They also in-

formed decisions regarding further analysis, such as selecting which texts to use in linear regression analyses.

ANOVAs were employed to understand the relationships between text properties, such as word lengths, and the final readability levels. The ANOVAs were also used to identify relationships, such as the effect of the exam year on the final readability output. Finally, linear regression was used to develop the Sesotho readability measures.

The weaknesses of these analysis methods include the limited number of data points, which were a result of using a small corpus for the investigation. In this order, further research may benefit from using larger sample sizes by the number of documents and incorporating multiple statistical techniques to enhance the validity and generalisability of the findings.

Furthermore, it is important to acknowledge that our approach to crafting readability measures for Sesotho relied on a limited number of statistical techniques. It remains a possibility that the inclusion of additional methodologies could potentially unveil more connections between text features and readability outcomes. Nevertheless, our confidence in the suitability of the chosen statistical techniques remains unwavering, thereby affirming the validity and substantiation of our research findings.

5.6 Future directions

While this thesis has provided valuable contributions to the understanding of measuring text readability in the low-resource Sesotho, it has also brought to light a few unanswered questions and identified avenues for future research. This section explores the limitations of the thesis and highlights areas in which further investigation could advance our understanding of text readability in LRLs.

5.6.1 Unanswered questions

While this thesis has shed light on the readability of Sesotho texts, there are still important questions that require further investigation.

One limitation of this thesis is that it relied solely on traditional readability measures to estimate text readability. These measures do not take into account reading comprehension or text understandability, which could significantly impact the reading experience for Sesotho speakers. Therefore, it is important for future research to develop text readability measures that account for these factors to provide a more nuanced idea of text readability in LRLs.

Another unanswered question from this thesis is related to the target audience. The thesis used models similar to those in English, where the outputs are ranked in order of difficulty,

but it is unclear to whom the texts are easier or more difficult to read. This gap could be addressed through user-evaluation studies in the future. Unfortunately, due to budget and time constraints, we were unable to include participants in the research described in this thesis. In this way, we are unable to ascertain the practical effectiveness of our proposed readability measures within specific contexts. In essence, we lack the means to determine the extent to which these metrics would prove valuable in practical applications, such as educational settings. Consequently, we remain unable to definitively establish whether our formulated measures would indeed facilitate teachers' ability to identify appropriate texts for their learners.

5.6.2 Future research directions

The readability measures developed for Sesotho in this thesis may be used to aid future research on text readability in Sesotho. For instance, the formulas have the potential to be utilised in the development of readability corpora. Even so, the measures may be biased towards educational texts such as reading comprehension and summary texts. Nonetheless, future work might develop better gold-standard corpora by relying on other approaches such as human readability assignment.

To determine the accuracy of the measures in other genres such as literary texts, future research could evaluate the metric formulas in such contexts. For instance, novels, dramas, poetry, folktales, and prose are also taught in South African schools. In fact, teachers are expected to expose their learners to all these genres (Zano, 2020). Such texts are unfortunately copyrighted and permission is needed to access them. Unfortunately, we were not able to obtain copyright permission to access such texts. Nonetheless, the evaluation of the measures in other genres would reveal whether the measures are reliable in these contexts or if modifications are necessary.

The syllabification system developed in this thesis has demonstrated applicability to the Sotho-Tswana language group (Sibeko and Setaka, 2023). The presence of resources such as these reveal avenues for possible expansion of the readability research to the larger Sotho-Tswana language group for enabling cross-lingual text readability analyses. Moreover, the list of common words such as the one we described in Sibeko and De Clercq (2023) could be developed for Sepedi and Setswana to enable the adaptation of the DCI text readability metric to the two languages. Note that all the corpora used for deriving the common words in Sesotho are also available in Sepedi and Setswana.

Additionally, the current measures could be evaluated on Sepedi and Setswana texts to determine their applicability to the entire Sotho-Tswana language group. This approach could also enable cross-lingual techniques for text readability that cover the larger Sotho-Tswana language group.

5.7 Summary

This chapter has provided a comprehensive discussion of this thesis, highlighting its main contributions and limitations. The chapter began with a recap of the research assumption and aims, followed by an overview of the thesis in the context of previous research, which clearly demonstrated the specific gap addressed by this thesis and the new contributions made to the field. The strengths and weaknesses of the thesis design, methodology, and analyses were critically evaluated, and potential future directions for research were suggested. This chapter has provided a reflective summary of the thesis, emphasising the key findings and their implications for future research.

Overall, the chapter has presented a discussion of our attempts to address the absence of text readability measures in the low-resource Sesotho. In this way, our assumptions that texts can be ranked according to their levels of readability in Sesotho have been validated.

Chapter 6

Summary and conclusions

This chapter presents the summary and conclusions of the research covered in this thesis. We first present a summary of our findings by revisiting the research questions presented in Chapter 1, which is then followed by a discussion on the implication of our findings and results.

6.1 Introduction

In this thesis, we set out to investigate how readability could be analysed in Sesotho. The literature survey of methods for LRLs revealed a few options. This thesis adopted the adaptation of existing traditional readability measures.

In this concluding chapter, we summarise the main findings by revisiting the problem statement and research questions set out in Section 1.4, discuss the implications of the findings of this thesis by outlining the practical implications, and discuss the limitations of our research.

6.2 Summary of findings

This thesis started with the problem statement: “How might one measure text readability in Sesotho?”. This problem statement was then subdivided into four research questions. The findings from each of the questions are summarised below.

What methods, techniques, and approaches are used in measuring text readability in LRLs?

Three main approaches to measuring text readability in LRLs were identified, including (i) cross-lingual approaches, which have not been employed much, (ii) crowd and expert sourcing, which rely on the availability of texts and evaluators, and (iii) the adaptation of traditional readability measures from HRLs such as English. One of the common methods for adapting traditional readability measures was the use of translated texts between HRLs as helper languages and LRLs. Traditional readability measures use shallow textual features such as (i) the number of words, (ii) the lengths of sentences, both of which can easily be counted, (iii) syllabic information, which requires a syllabification system, and (iv) a frequency word list of commonly used words, which is language specific.

What language resources are needed for measuring text readability in Sesotho?

As exemplified by our analysis of readability methods, the development of text readability measures necessitates a corpus (De Clercq et al., 2014). Such a corpus may be already graded for difficulty or be assumed to have different levels of readability levels. Unfortunately, a pre-existing text readability corpus for Sesotho could not be located.

In fact, beyond the corpora requirement, other basic digital language resources that are essential for facilitating the development of text readability measures were also unavailable.

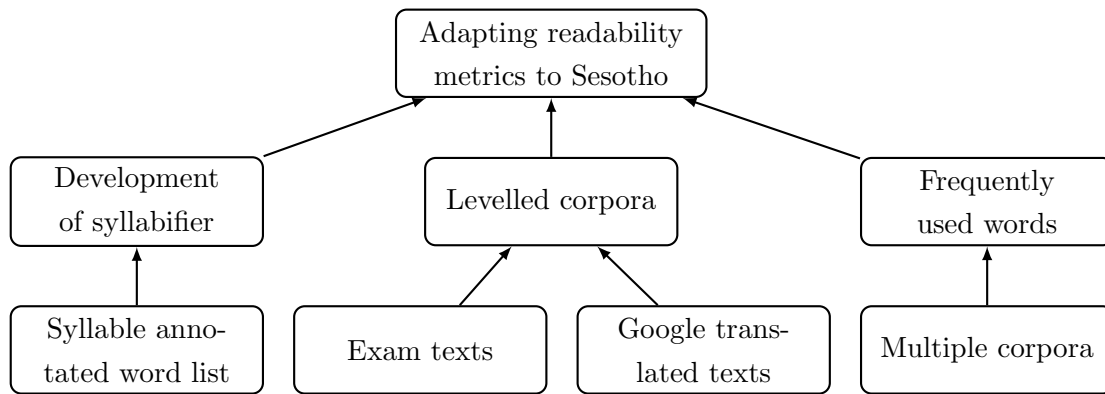


Figure 6.1: The integration of language resources and their relevance to the study.

In particular, a Sesotho syllabification system and a word list of frequently used Sesotho words are required. As a result, a substantial part of this thesis, specifically Chapter 3, was directed towards the establishment of imperative basic language resources, which are outlined below.

Each of the resources produced in this thesis was aimed at facilitating the adaptation of readability measures. These resources are shown in Figure 6.1. Note that although the list of common words was created using multiple corpora, these corpora were not generated as part of this thesis and, therefore, are not discussed in this section (see Section 3.3 for a discussion of the secondary data used in the development of the list of common Sesotho words). Nonetheless, the secondary corpora are also listed in Figure 6.1 to provide a context for the creation of the list of common Sesotho words.

In total, five basic digital language resources were created as part of this thesis. The resources include (i) a corpus consisting of examination texts, (ii) a list of common Sesotho words, (iii) a data set of machine and human-post-edited translation texts, (iv) a syllable-information annotated lexicon, and (v) the rule-based and pattern-based syllabification systems. These resources are briefly outlined below.

We extracted reading comprehension and summary writing texts from the grade twelve language examination question papers between the years 2008 and 2020. We extracted these texts from examination texts of eleven written official languages of South Africa. The texts in our corpus are split into two categories, that is, the home language (HL) and the first additional language (FAL). Previous research indicated that the English exam texts show consistently lower readability levels for the HL texts as opposed to the FAL texts (Sibeko, 2021). As such, the use of these Sesotho and English texts guaranteed some level of variability needed for a readability corpus.

We employed a corpus-based methodology for constructing a list of frequently used words

in Sesotho. Our approach drew upon three distinct corpora to extract a compilation of 3,000 common Sesotho words. This procedure further entailed a manual validation process to cross-reference words exhibiting diverse spellings, alongside filtering out elements such as numerical characters and invalid entries.

The examination texts for English and Sesotho that were collected for our corpus of examination texts were also used in creating a corpus of translated texts which were then used for our analysis of readability in Sesotho texts. Furthermore, the readability corpus encompassing original Sesotho texts and the machine-translated texts was then used in the linear regression modelling for developing formulas for Sesotho.

The development of the syllabification systems necessitated the construction of a gold-standard syllabification lexicon that is annotated with syllable information first. For the syllabification lexicon, we extracted dictionary entries and syllable information from *Bukantswe ya Machaba ya Sesotho* ‘The international dictionary of Sesotho’ (Chitja, 2010).

We used the syllable-annotated word list for the development of two syllabification systems. In the end, we developed a T_EX-based pattern-oriented system as well as a rule-based syllabification system. The process for extracting syllable information from the dictionary and creating a word list is described thoroughly in Section 3.5.2.1. The availability of the syllabification systems enabled the adaptation of traditional readability measures, specifically, the syllable-information-based formulas, namely, the FRE, FKGL, SMOG, and GFI.

What can we learn from existing traditional text readability measures?

Traditional measures of text readability focus on various factors that influence the complexity of a text. These factors encompass aspects such as the number of words per sentence, which is taken into account across six of the adapted measures in this thesis. Additionally, the syllables per word feature that is integral to the FKGL and FRE formulas plays a crucial role. Polysyllabic or complex words, defined by their composition of three or more syllables and used in the SMOG and FOG measures, as well as long words that exceed six letters, and difficult words which are not found in the list of common words used in the DCI measure, are also key attributes examined in the measures tailored to Sesotho (refer to Table 4.9 for the comprehensive list of features considered in the Sesotho-adapted measures).

The review of established traditional text readability measures has facilitated the identification of digital language resources essential for crafting readability measures in new languages such as Sesotho. For instance, our analysis revealed the necessity for an automated syllabification system to extract syllabic information, as four measures depend on this data. Similarly, the exploration of the DCI underscored the requirement for curating a lexicon of commonly used words in Sesotho.

Nevertheless, our investigation was restricted to exploring only the text features already used in existing traditional readability metrics. In this way, this thesis does not encompass the exploration of distinct, novel attributes beyond those already integrated into existing traditional readability measures. This stance is grounded in the recognition of prior analyses of the readability of Sesotho texts as well as other indigenous South African languages. By using English readability measures, previous studies have leaned on shallow text features focusing on word, sentence, and syllable information which are used in the English text readability measures. Such features serve as the foundation for assessing text readability, aligning with the conventions of English traditional readability formulas.

Overall, our findings underscore the higher influence of the words per sentence feature within the Sesotho formulas. The higher weights allocated to the sentence lengths indicate that sentence length has more impact on the overall readability of texts in Sesotho than in the English texts.

How can existing English traditional readability measures be modified and adapted to Sesotho?

Nine traditional readability formulas were adapted to Sesotho. This adaptation process involved the consideration of the aspects measured by each of the text attributes employed in the pre-existing formulas, as detailed in Section 4.4.4. Among others, we focused our investigation on shallow text properties with a specific focus on their applicability to Sesotho texts.

To recalibrate the Sesotho formulas, we employed a simple linear regression modelling approach. This procedure entailed the redefinition of coefficients and intercepts based on the Sesotho texts. The foundation for this adjustment was laid by extracting pertinent text features from Sesotho texts and correlating them with the readability levels derived from English texts, which had been translated from their Sesotho counterparts as detailed in Section 3.4. In essence, the readability levels in the texts translated to English from Sesotho served as reference points for the Sesotho texts, utilising the extracted properties. The decision to employ this methodology stems from the absence of pre-existing formulas (and readability annotated corpora) for Sesotho, thereby necessitating the utilisation of English as a higher-resource helper language.

The linear regression models were developed using the same features as those employed in the English formulas, ensuring a comparative basis with established research. It is important to note that we refrain from asserting a definitive interpretation of readability in Sesotho due to a lack of established criteria. Consequently, our adaptation is confined to leveraging pre-existing comprehension of text readability. Intriguingly, our findings underscore that the attributes already embedded within the English formulas retain predictive efficacy even when transposed to the Sesotho texts. This phenomenon is substantiated by

the evaluations of our formulas, as detailed in Section 4.4.4.

Beyond the adjustment of coefficients, we introduce select refinements to the formulas. For example, we incorporate intercepts into the formulas to ensure accurate values for certain measures, and we advocate for the incorporation of weights in formulas like the LIX and the RIX. The inclusion of such weights assumes significance as these formulas may generate results that deviate from established parameters when the underlying features lack appropriate weighting.

6.3 Implications of research

As far as could be ascertained, this investigation represents the initial formulation of readability measures specifically tailored for a South African indigenous language. Before this project, only the Afrikaans language (which originated from European antecedents as a variant of Dutch) had established text readability measures within the South African context.

The rest of this section expounds upon the primary contributions stemming from this thesis, as informed by the research questions in the introductory chapter. Principally, the practical applications engendered by the devised measures are elucidated, followed by a discussion of the limitations inherent to this thesis.

Practical applications

The widespread application of traditional readability measures as a means of assessing text readability has found utility across diverse domains. One particularly receptive sector is the realm of education, spanning from elementary to advanced stages, encompassing foundational, basic, intermediate, and higher education levels. As expounded in the introductory chapter, our principal aim revolves around streamlining the process of identifying correctly levelled texts for readers with varying levels of reading proficiency. This pursuit stands to offer substantial benefits to education stakeholders, including those entrusted with text selection and consumption. For instance, teachers can leverage our formulas to effectively choose texts aligned with their students' respective reading capacities. In this way, teachers can augment textbook texts with alternative texts that best suit their students' individual reading abilities.

Furthermore, the implications of this thesis extend to academic research, particularly within the field of readability in LRLs such as Sesotho and other African indigenous languages. Historically, studies in this domain have leaned upon traditional English text readability measures. The present investigation, however, offers a novel foundation upon which subsequent research can build, ushering in a new era of readability exploration for

cognate languages, including those within the Sotho-Tswana language group.

Moreover, the potential applications of our adapted readability measures extend to translation endeavours. Translators can harness these measures to adjust their translations, ensuring that the target audience encounters a reading experience similar to that of the original English texts. This prospect assumes significance given that a substantial portion of available Sesotho texts comprises translations, thus presenting a dearth of authentically authored newspapers, magazines, and online texts. Furthermore, publicly accessible text corpora predominantly consist of translations sourced from governmental platforms. By juxtaposing the readability levels of English source texts with their translated counterparts, translators can identify areas necessitating further refinement in their translations.

Anticipating further implications, the establishment of text readability measures within Sesotho could catalyse the development of text simplification systems tailored for Sesotho and the broader Sotho-Tswana language group. For instance, the use of common words can serve as a strategic means to limit the inclusion of less commonly used vocabulary in new simplified texts, while simultaneously introducing trending vocabulary progressively. This approach which integrates commonly used words possesses the potential to streamline texts and enhance their accessibility to readers.

Limitations

The limitations specific to the individual studies encompassed within this thesis are discussed within their respective sections. Beyond these particularities, the overarching constraints inherent to this thesis are expounded upon below, supplemented by suggestions for potential enhancements to inform future investigations where applicable.

A prominent obstacle, delineated in the introductory chapter and expounded in greater detail within the third chapter, pertains to the absence of an established gold-standard corpus that is annotated with readability levels. While we addressed this issue by using grade-level texts, it is acknowledged that alternative methodologies, such as crowdsourcing and expert sourcing, have exhibited superior outcomes according to previous scholarship (De Clercq et al., 2014). It is noteworthy, however, that the financial and time constraints associated with the adoption of these methodologies constrained our investigation. Future research may benefit from the application of these more precise techniques.

Another important limitation pertains to the potential genre bias intrinsic to our formulated measures. Given that the construction of these formulas depended on training data sourced from reading comprehension and summary writing texts, their efficacy may be compromised when extended to other genres such as literary compositions. This limitation suggests a promising avenue for future research which could involve an exploration of diverse textual genres to establish a more comprehensive framework.

Lastly, it is prudent to underscore that this thesis, per the delineation of its scope (refer to Section 1.5), does not address the conventional limitations associated with established traditional readability measures. Specifically, this research does not entail the enhancement of any of the traditional readability measures that are used in this thesis. Rather, our focus centres on the adaptation of these measures to Sesotho, thereby rendering them pertinent and more accurate for Sesotho texts. For example, the adapted Sesotho readability measures do not address text comprehension or the challenges related to reading comprehension. In this way, the criticisms poised against the use of readability formulas are not addressed by this thesis.

6.4 Summary

In essence, this thesis constitutes a significant contribution by developing nine readability formulas for Sesotho, a low-resource indigenous language of South Africa. The outcomes of this research not only advance our understanding of gauging text readability within LRLs but also address the central research inquiry: how to effectively measure text readability in Sesotho.

The thesis establishes that prevailing text readability measures can be harnessed as a foundation for constructing readability metrics in languages with inadequate digital language resources. Moreover, it underscores the viability of employing data-driven methodologies, particularly in cases where HRLs serve as helper languages. Additionally, the research highlights the use of practical approaches, such as machine translations and pre-existing lexicons, in curating essential resources that facilitate the development of readability assessment measures for languages with limited resources.

Significantly, this thesis contributes to bridging the research gap identified by Leopeng (2019, p. 35), wherein the adaptation of existing traditional readability measures to South African languages was identified as an imperative. By effectively addressing this gap, the thesis advances the understanding and practical implementation of readability assessment within the context of indigenous South African languages.

Bibliography

- Agic, Ž. and Vulic, I. (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.
- Ai, H. and Lu, X. (2010). A web-based system for automatic measurement of lexical complexity. In *27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10)*. Amherst, MA. June, pages 8–12.
- Al-Khalil, M., Habash, N., and Jiang, Z. (2020). A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3053–3062. European Language Resources Association (ELRA).
- Al Shaye, S. and Al Dhefeiri, M. D. (2021). Readability level of Arabic language textbook of the sixth grade in the state of Kuwait. *Journal of Educational and Social Research*, 11(4):197–212.
- Al-Thanyyan, S. S. and Azmi, A. M. (2021). Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.
- Alahyari, S., Hafner, M., and Knittle, S. (2022). Ensuring privacy in natural language processing: A PRISMA oriented. In *Workshops at 18th International Conference on Intelligent Environments*, pages 180–190.
- Alić, M. (2020). Privacy policy understandability analysis of croatian electronic publications. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1488–1492, New York. IEEE.
- Alkaldi, W. (2022). *Enhancing text readability using deep learning techniques*. PhD thesis, Université d’Ottawa/University of Ottawa.
- Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 fifth workshop on innovative use of NLP for building educational applications*, pages 1–9.
- Amendum, S. J., Conradi, K., and Hiebert, E. (2018). Does text complexity matter in the elementary grades? a research synthesis of text difficulty and elementary students’ reading fluency and comprehension. *Educational Psychology Review*, 30:121–151.

- Anderson, J. (1981). Analysing the readability of English and non-English texts in the classroom with LIX. In *The Annual Meeting of the 'Australian Reading-Association' (Darwin, Australia)*. ERIC.
- Anderson, J. (1983). LIX and RIX: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.
- Andova, A. (2017). *Assessment of text readability using statistical and machine learning approaches*. PhD thesis, University of Ljubljana.
- Anghelescu, A. (2016). Distinctive transitional probabilities across words and morphemes in Sesotho. *University of British Columbia Working Papers in Linguistics*, 3(1):1–17.
- Ansaldi, U. and Nordhoff, S. (2009). Complexity and the age of languages. In *Complex processes in new languages*, pages 345–363. John Benjamins, Amsterdam.
- Antunes, H. and Lopes, C. T. (2019a). Analyzing the adequacy of readability indicators to a non-English language. In *Proceedings of the international conference of the Cross-Language Evaluation Forum for European Languages*, pages 149–155, New York. Springer.
- Antunes, H. and Lopes, C. T. (2019b). Readability of web content. In *Proceedings of the 2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–4, New York. IEEE.
- Arppe, A., Beck, K., Branco, A., Camilleri, V., Caselli, T., Cristea, D., Hinrichs, E., Liin, K., Nissinen, M., Parra, C., Rosner, M., Schuurman, I., Skadina, I., Quochi, V., Van Uytvanck, D., and Vogel, I. (2010). Description of the BLARK, the situation of individual languages. Report, CLARIN, SFS University of Tübinge, Germany.
- Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8, Slovenia.
- Azpiazu, I. M. and Pera, M. S. (2020). Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.
- Badenhorst, J. and De Wet, F. (2022). NCHLT auxiliary speech data for ASR technology development in South Africa. *Data in Brief*, 41:107860.
- Badenhorst, J., Martinus, L., and De Wet, F. (2019). BLSTM harvesting of auxiliary NCHLT speech data. In *Proceedings of South African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/ROBMECH/PRASA 2019)*, pages 123–128, New York. IEEE.
- Badenhorst, J., Van Heerden, C., Davel, M., and Barnard, E. (2011). Collecting and evaluating speech recognition corpora for 11 South African languages. *Language resources and evaluation*, 45:289–309.

- Bailin, A. and Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301.
- Balakrishna, S. V. (2015). *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Phd thesis, Eberhard Karls Universität Tübingen.
- Balakrishna, S. V. and Meurers, D. (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 194(2):194–222.
- Banerjee, S., Kumar, N., and Madhavan, C. E. V. (2013). Text simplification for enhanced readability. In Fred, A., Liu, K., and Filipe, J., editors, *International Conference on Knowledge Discovery and Information Retrieval*, volume 2, pages 202–207, Pennsylvania. ScitePress.
- Bansal, R., Choudhary, H., Punia, R., Schenk, N., Dahl, J. L., and Pagè-Perron, E. (2021). How low is too low? a computational perspective on extremely low-resource languages. *arXiv:2105.14515*, preprint. <https://aclanthology.org/2021.acl-srw.5>.
- Bardill, J. E. and Cobbe, J. H. (2019). *Lesotho: dilemmas of dependence in Southern Africa*. Routledge, Cape Town, SA.
- Bargate, K. (2012). The readability of managerial accounting and financial management textbooks. *Meditari Accountancy Research*, 20(1):4–20.
- Barnard, E., Davel, M. H., Heerden, C. V., Wet, F. D., and Badenhorst, J. (2014a). The NCHLT corpus of the South African languages. In *Proceedings of the 4th International Workshop Spoken Language Technologies for Under-resourced Languages, St Petersburg, Russia*, pages 194–200, Russia. St Petersburg, Russia.
- Barnard, E., Davel, M. H., Van Heerden, C., F. De Wet, F., and Badenhorst, J. (2014b). 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.
- Barnard, E. and Wissing, D. (2008). Vowel variation in Southern Sotho: An acoustic investigation. *Southern African Linguistics and Applied Language Studies*, 26(2):255–265.
- Barry, J. G. (1980). Computerized readability levels. *IEEE Transactions on Professional Communication*, 23(2):88–90.
- Barry, J. G. and Stevenson, T. E. (1975). Using a computer to calculate the Dale-Chall formula. *Journal of Reading*, 19:218–222.
- Basaraba, D., Yovanoff, P., Alonzo, J., and Tindal, G. (2013). Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26:349–379.

- Batinić, D., Birzer, S., and Zinsmeister, H. (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.
- Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (2020). A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3302–3311, Paris. European Language Resources Association.
- Begeny, J. C. and Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, 51:198–215.
- Bendová, K. (2021). Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Journal of linguistics*, 72(2):477–487.
- Bendová, K. and Cinková, S. (2021). Adaptation of classic readability metrics to Czech. In Ekštejn, K., Pártl, F., and Konopík, M., editors, *Proceedings of the International Conference on Text, Speech, and Dialogue: 24th International Conference*, pages 159–171, New York. Springer.
- Bengoetxea, K. and Gonzalez-Dios, I. (2021). MultiAzterTest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv:2109.04870*, preprint. <https://arxiv.org/abs/2109.04870>.
- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.
- Benzahra, M. and Yvon, F. (2019). Measuring text readability with machine comprehension: A pilot study. In Yannakoudakis, H., Kochmar, E., Leacock, C., Madnani, N., Pilán, I., and Zesch, T., editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 412–422, New York. Association for Computational Linguistics.
- Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., and Trautwein, U. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4):518–543.
- Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and linguistic computing*, 5(4):257–269.
- Bigi, B., Meunier, C., Nesterenko, I., and Bertrand, R. (2010). Automatic detection of syllable boundaries in spontaneous speech. In *7th International conference on Language Resources and Evaluation (LREC 2010)*, pages 3285–3292.
- Björnsson, C.-H. (1968). *Läsbarhet*. Bokförlaget Liber, Stockholm, Sweden.

- Björnsson, C.-H. (1983). Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.
- Blanchard, D. (2011). Unsupervised word segmentation: An investigation of sub-word features. Online on github. <https://dan-blanchard.github.io/papers/proposal.pdf>.
- Bochkarev, V. V., Shevlyakova, A. V., and Solovyev, V. D. (2015). The average word length dynamics as an indicator of cultural changes in society. *Social Evolution and History*, 14(2):153–175.
- Boles, C. D., Liu, Y., and November-Rider, D. (2016). Readability levels of dental patient education brochures. *American Dental Hygienists' Association*, 90(1):28–34.
- Bolli, G. G., Rini, D., and Spina, S. (2017). Predicting readability of texts for Italian L2 students: A preliminary study. In Eugenio, E. G., editor, *Proceedings of the Learning and Assessment: Making the Connections - ALTE 6th International Conference*, pages 272–279, United Kingdom. Association of Language Testers in Europe.
- Bosch, S. and Griesel, M. (2017). Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator*, 38(1):1–12.
- Bott, S. and Saggion, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, 48(1):93–120.
- Brouwer, R. (1963). Onderzoek naar de leesmoeilijkheden van Nederlands proza (Research into reading difficulty in Dutch prose). *Pedagogische studiën*, 40:454–464.
- Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. In Housen, A., Kuiken, F., and Vedder, I., editors, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, pages 23–46. John Benjamins Publishing, Philadelphia.
- Bunch, G. C., Walqui, A., and Pearson, P. D. (2014). Complex text and new common standards in the united states: Pedagogical implications for English learners. *Tesol Quarterly*, 48(3):533–559.
- Buse, R. P. L. and Weimer, W. R. (2008). A metric for software readability. In Ryder, B. G., editor, *Proceedings of the 2008 international symposium on Software testing and analysis*, pages 121–130, New York. Association for Computing Machinery.

- Buthlezi, M. (2017). A corpus analysis of readability and the degree of complexity of isiZulu texts. *Uploaded on Research Gate*, pages 1–22. Available at: <https://doi.org/10.13140/RG.2.2.22908.74888>.
- Cambria, E. and White, B. (2014). Jumping NLP curves: A review of Natural Language Processing research. *IEEE Computational intelligence magazine*, 9:48–57.
- Cardon, R. and Grabar, N. (2020). French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online)*, New York, USA. International Committee on Computational Linguistics.
- Carel, D. (2019). 4 ways to stay safe online-developing a text difficulty indicator for isiXhosa early grades. In *Policy Commons*. Research on Socio-Economic Policy. Available at: <https://policycommons.net/artifacts/2105074/4-ways-to-stay-safe-online/2860372/>.
- Castellucci, G., Filice, S., Croce, D., and Basili, R. (2021). Learning to solve NLP tasks in an incremental number of languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)*, pages 837–847, New Brunswick, New Jersey, USA.
- Castilhos, S. (2016). *Pylinguistics: An open source library for readability assessment of texts written in Portuguese*. Bachelor thesis, Universidade Federal Do Rio Grande Do Sul.
- Cavalli-Sforza, V., Saddiki, H., and Nassiri, N. (2018). Arabic readability research: Current state and future directions. In *Proceedings of the 4th International Conference on Arabic Computational Linguistics*, volume 142, pages 38–49.
- Chakraborty, S., Nayeem, M. T., and Ahmad, W. U. (2021). Simple or complex? Learning to predict readability of Bengali texts. In Yang, Q., editor, *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-2)*, pages 12621–12631, Washington. Association for the Advancement of Artificial Intelligence.
- Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, Brookline, USA.
- Chen, Y.-H., Tsai, Y.-H., and Chen, Y.-T. (2011). Chinese readability assessment using TF-IDF and SVM. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 705–710.
- Chiguvare, P. and Cleghorn, C. W. (2021). Improving transformer model translation for low resource South African languages using BERT. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, New York. IEEE.

- Chitja, M. (2010). *Phatlamantsoe ya Sesotho ya Machaba*. Mazenod Publishers, Maseru, Lesotho.
- Chokoe, S. (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.
- Chowdhury, G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology*, 37:51–89.
- Cieri, C., Maxwell, M., Strassel, S., and Tracey, J. (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.
- Clements, G. N. and Keyser, S. J. (1983). *CV Phonology: A Generative Theory of the Syllable*. MIT Press, Cambridge, Massachusetts.
- Cockcroft, K. (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.
- Coleman, M. and 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.
- Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American society for information, science and technology*, 56(13):1448–1462.
- Comrie, B. (1992). Before complexity. In *The evolution of human languages*, page 193–211. New York: Addison-Wesley.
- Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in society*, 1:104–126.
- Courtis, J. K. (1987). Fry, SMOG, LIX and RIX: Insinuations about corporate business communications. *The Journal of Business Communication (1973)*, 24(2):19–27.
- Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., and Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.
- Crossley, S. A., Dufty, D. F., McCarthy, P. M., and McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29, Seattle, Washington. Cognitive Science Society.

- Cucu, H., Buzo, A., Besacier, L., and Burileanu, C. (2014). SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian. *Speech Communication*, 56:195–212.
- Curto, P., Mamede, N., and Baptista, J. (2014). Automatic readability classifier for European Portuguese. *System*, 5(1):309–324.
- Curto, P., Mamede, N., and Baptista, J. (2015). Automatic text difficulty classifier: Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching. In Helfert, M., Restivo, M. T., Zvacek, S., and Uhomoibhi, J., editors, *Proceedings of the 7th International Conference on COmputer Supported Education (CSEDU-2015)*, volume 1, pages 36–44, Portugal. Science and Technology Publications.
- Daelemans, W., De Clercq, O., and Hoste, V. (2017). Stylen: An environment for stylometry and readability research for Dutch. In *CLARIN in the Low Countries*, pages 195–210. Ubiquity Press, London.
- Daelemans, W., Zavrel, J., Van Der Sloot, K., and Van den Bosch, A. (2004). Timbl: Tilburg memory-based learner. Technical report, Tilburg University.
- Dahl, O. (2004). *The growth and maintenance of linguistic complexity (Studies in Language Companion Series)*, volume 71. John Benjamins Publishing, Amsterdam.
- Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.
- Das, S. and Roychoudhury, R. (2006). Readability modelling and comparison of one and two parametric fit: A case study in Bangla. *Journal of Quantitative Linguistics*, 13(1):17–34.
- Davel, M. and Barnard, E. (2003). Bootstrapping in language resource generation. In *Fourteenth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 97–100.
- Davel, M., Basson, W., Charl, V. H., and Barnard, E. (2013). NCHLT dictionaries: Project report. May. Available at: <https://sites.google.com/site/nchltspeechcorpus/home>.
- Davel, M. and Martirosian, O. (2009). Pronunciation dictionary development in resource-scarce environments, brighton, uk. In *Proceedings of the Interspeech*, pages 2851–2854, Ireland.
- De Almeida, D. and Aluísio, S. (2009). Text readability analysis with Natural Language Processing tools: The adaptation of Coh-Metrix metrics for Portuguese. In Ceballos, S., editor, *2009 Seventh Brazilian Symposium in Information and Human Language Technology*, pages 53–62, New York. IEEE.

- De Clercq, O. and Hoste, V. (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.
- De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., and Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.
- De Vos, M., Van der Merwe, K., and Van der Mescht, C. (2014). A linguistic research programme for reading in African languages to underpin CAPS. *Journal for Language Teaching*, 48:149–177.
- De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E., and De Waal, A. (2014). A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, pages 119–131.
- De Wet, A. (2021). *The development of a contextually appropriate measure of individual recovery for mental health service users in a South African context*. Thesis, Stellenbosch University.
- Dell’Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Dell’Orletta, F., Montemagni, S., and Venturi, G. (2014a). Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2):163–193.
- Dell’Orletta, F., Wieling, M., Cimino, A., Venturi, G., and Montemagni, S. (2014b). Assessing the readability of sentences: which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Demuth, K. (2007). Sesotho speech acquisition. In McLeod, S., editor, *The international guide to speech acquisition*. Thomson Delmar Learning, pages 526–538. Thomson Delamar Learning, New York, USA.
- 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.

- Department of Basic Education (2011c). *English Second 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 (2012). *National policy pertaining to the programme and promotion requirements of the National Curriculum Statement Grades R – 12*. Government printing works, Pretoria. Policy Document.
- Department of Basic Education (2013). *The incremental introduction of African languages in South African schools*. Government Printing Works, Pretoria. Draft policy, September 2013.
- Department of Basic Education (2016). *Curriculum and Assessment Policy Statement Foundation Phase Grades 1–3: English Second Additional Language*. Government printing works, Pretoria. Policy Document.
- Department of Basic Education (2017). *Official Languages: First Additional Language Examination Guidelines*. Government printing works, Pretoria. Policy Document.
- Department of Basic Education (2019). *The discontinuation of the multiple examination opportunities (MEO) dispensation with effect from 2020*. Government Printing Works, Pretoria. Circular E 29 of 2019.
- Dickens, P. (1978). A preliminary report on Kgalagadi vowels. *African Studies*, 37(1):99–106.
- Ditsele, T. (2014). Why not use Sepitori to enrich the vocabularies of Setswana and Sepedi? *Southern African Linguistics and Applied Language Studies*, 32(2):215–228.
- Douma, H. (1960). Readability of Dutch farm papers: a discussion and application of readability-formulas. *Wageningen: Afdeling Sociologie en Sociografie van de Landbouwhogeschool*, 17:433–470.
- Drake, M. (2003). *Encyclopedia of Library and Information Science, Second Edition*. Taylor & Francis, New York, USA.
- Drndarevic, B. and Saggion, H. (2012). Towards automatic lexical simplification in Spanish: An empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations, Montréal, Canada*, pages 8–16, New York, USA. Association for Computational Linguistics.
- Du Plessis, J. A., Gildenhuys, J. G., and Moilola, J. J. (1974). *Moilola. Bukantswe ya maleme-pedi Sesotho-Seafricanse / Tweetalige woordeboek Afrikaans-Suid-Sotho*. Via Afrika Beperk, Cape Town.
- Duanmu, S. (2009). *Syllable Structure. The Limits of Variation*. Oxford University Press, Oxford.

- DuBay, W. H. (2004). The principles of readability. Technical report, Impact Information, Costa Mesa.
- Duff, D. (2019). The effect of vocabulary intervention on text comprehension: Who benefits? *Language, speech, and hearing services in schools*, 50:562–578.
- Duvenhage, B. (2019). Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.
- Eiselen, E. R. and Puttkamer, M. J. (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).
- Eiselen, R. (2016a). Government domain named entity recognition for South African languages. In *Proceedings of the 10th Language Resource and Evaluation Conference, Portorož, Slovenia*, page 3344–3348, Paris. European Language Resources Association (ELRA).
- Eiselen, R. (2016b). South African language resources: phrase chunkers. In *Proceedings of the 10th Language Resource and Evaluation Conference, Portorož, Slovenia*, pages 689–693, Paris. European Language Resources Association (ELRA).
- El-Haj, M. and Rayson, P. (2016). Osman: A novel Arabic readability metric. *Procedia Computer Science*, 142:38–49.
- El-Kassas, W. S., Salama, R. C., Ahmed, R., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with applications*, 165(4):1–46.
- Eldredge, E. A. (2002). *A South African kingdom: The pursuit of security in nineteenth-century Lesotho*, volume 78. Cambridge University Press, Cambridge.
- Elenius, K., Forsborm, E., and Megyesi, B. (2008). Language resources and tools for swedish: A survey. In *Proceedings of the LREC 2008, Marrakesh Morocco*, Paris. European Language Resources Association (ELRA).
- Eleyan, D., Othman, A., and Eleyan, A. (2020). Enhancing software comments readability using Flesch Reading Ease score. *Information*, 11(9):430.
- Eltorai, A. E., Naqvi, S. S., Ghanian, S., Eberson, C. P., Weiss, A. P. C., Born, C. T., and Daniels, A. H. (2015). Readability of invasive procedure consent forms. *Clinical and translational science*, 8:830–833.
- Falkenjack, J. and Jönsson, A. (2014). Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 114–122, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

- Falkenjack, J. and Mühlenbock, K. H. (2012). Using the probability of readability to order Swedish texts. In *Proceedings of the Fourth Swedish Language Technology Conference*, pages 27–28.
- Falkenjack, J., Mühlenbock, K. H., and Jönsson, A. (2013). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–47. Association of Computer Linguistics (ACL).
- Feng, H., Hou, S., Wei, L.-Y., and Zhou, D.-X. (2022). CNN models for readability of Chinese texts. *Mathematical Foundations of Computing*, 5(4):351–362.
- Feng, L. (2010). *Automatic readability assessment*. New York: City University of New York.
- Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In Huang, C.-R. and Jurafsky, D., editors, *Coling 2010: Posters*, pages 276–284, New York, USA. Association of Computational Linguistics.
- Filho, J. A. W., Wilkens, R. S., Zilio, L., Idiart, M., and Villavicencio, A. (2016). Crawling by readability level. In Silva, J., Ribeiro, R., Quaresma, P., Adami, A., and Branco, A., editors, *Proceedings of the 12th International Conference on Computational Processing of the Portuguese Language*, volume 1, pages 306–318.
- Fischer, A., Venter, W., Collins, S., Carman, M., and Lalla-Edward, S. (2021). The readability of informed consent forms for research studies conducted in South Africa. *South African Medical Journal*, 111(2):180–183.
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.
- Flesch, R. (1974). *The art of readable writing*. Harper, New York, 2nd edition.
- Forsyth, J. N. (2014). *Automatic readability detection for modern Standard Arabic*. Master of arts thesis, Brigham Young University.
- François, T. (2015). When readability meets computational linguistics: A new paradigm in readability. *Revue Française de linguistique appliquée*, 20(2):79–97.
- François, T. and 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*, pages 466–477, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Franchina, V. and Vacca, R. (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.

- Frantz, R. S., Starr, L. E., and Bailey, A. L. (2015). Syntactic complexity as an aspect of text complexity. *Educational Researcher*, 44(7):387–393.
- Funk, L. C. (1968). *The Reading Level of Selected Magazines As Determined by the Dale-Chall Readability Formula*. PhD thesis, University of Minnesota: Minnesota.
- Gauthier, M. and Johnson, N. (2019). PNS373 identification and recommendations of readability tests for the evaluation of clinical outcome assessments. *Value in Health*, 22:S828.
- Giwa, O., Davel, M. H., and Barnard, E. (2011). A southern African corpus for multi-lingual name pronunciation. In *Pattern Recognition Association of South Africa and Mechatronics International Conference*, pages 49–53, New York. IEEE.
- Gkikas, D. C., Tzafilkou, K., Theodoridis, P. K., Garmpis, A., and Gkikas, M. C. (2022). How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in facebook. *International Journal of Information Management Data Insights*, 2(1):100067.
- Glazkova, A., Egorov, Y., and Glazkov, M. (2021). A comparative study of feature types for age-based 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, New York, USA. Springer.
- Goldsmith, J. (2011). The syllable. In Goldsmith, J., Riggle, J., and Yu, A. C. L., editors, *The handbook of phonological theory*, volume 2, pages 162–196. Blackwell Publishing Ltd, New Jersey, USA.
- Gonzalez-Dios, I. (2016). *Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures*. Phd thesis, University of the Basque country.
- Gonzalez-Dios, I., Aranzabe, M. J., De Ilarraza, A. D., and Salaberri, H. (2014). Simple or complex? Assessing the readability of Basque texts. In Tsujii, J. and Hajic, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 334–344, New Jersey. Association for Computational Linguistics.
- Gonzalez-Garduno, A. V. and Søgaard, A. (2017). Using gaze to predict text readability. In Tetreault, J., Burstein, J., Leacock, C., and Yannakoudakis, H., editors, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Gopal, R., Maniam, M., Madzlan, N. A., binti Shukor, S. S., and Neelamegam, K. (2021). Readability formulas: An analysis into reading index of prose forms. *Studies in English Language and Education*, 8(3):972–985.

- Grabar, N., Van Zyl, I., De la Harpe, R., and Hamon, T. (2014). The comprehension of medical words. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 334–342.
- Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., and Pennebaker, J. (2014). Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):210–229.
- Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, and computers*, 36(2):193–202.
- Greenfield, J. (2004). Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1):5–24.
- Gregová, R. (2016). *The Generative and the Structuralist Approach to the Syllable: A Comparative Analysis of English and Slovak*. Cambridge Scholars Publishing, Pietermaritzburg, South Africa.
- Groenewald, H. J. and Fourie, W. (2009). Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 190–196.
- Grover, A. S., Huyssteen, G. B. V., and Pretorius, M. W. (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.
- Grover, A. S., Huyssteen, G. B. V., and Pretorius, M. W. (2011). The South African Human Language Technology audit. *Language resources and evaluation*, 45:271–288.
- Groves, M. and Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, 37:112–121.
- Guma, S. M. (1982). *An outline structure of Southern Sotho*. Shooter and Shuter Publishers, Pietermaritzburg, South Africa, 2nd edition.
- Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill: New York.
- Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.
- Gunning, T. (2003). The role of readability in today’s classrooms. *Topics in Language Disorders*, 23(3):175–189.

- Hansen, H., Widera, A., Ponge, J., and Hellingrath, B. (2021). Machine learning for readability assessment and text simplification in crisis communication: A systematic review. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 2265.
- Hanslo, R. (2021). Evaluation of neural network transformer models for named-entity recognition on low-resourced languages. In *16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 115–119, New York. IEEE.
- Haralambous, Y. (2021). A revised small tutorial on patgen, 28 years after. ONLINE. Available at: <https://ctan.org/pkg/patgen2-tutorial>.
- Harden, S. (2018). *Comparison of readability indices with grades 1–5 narrative and expository texts*. Doctoral thesis, Wayne State University.
- Hartley, J. (2016). Is time up for the flesch measure of reading ease? *Scientometrics*, 107(3):1523–1526.
- Harzing, A.-W. (2007). Publish or perish. Available from: <https://harzing.com/resources/publish-or-perish> [Accessed: 21 Sep. 2022].
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2020). A survey on recent approaches for Natural Language Processing in low-resource scenarios. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistic.
- Hedman, A. S. (2008). Using the SMOG formula to revise a health-related document. *American Journal of Health Education*, 39(1):61–64.
- Hefer, E. (2013). Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English. *Southern African Linguistics and Applied Language Studies*, 31(3):359–373.
- Heilman, M., Collins-Thompson, K., and Eskenaz, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Henry, G. (1975). *Comment Mesurer La Lisibilité (How to Measure Readability)*. Labor, Brussels, Belgium.
- Heydari, P. (2012). The validity of some popular readability formulas. *Mediterranean Journal of Social Sciences*, 3:432–423.

- Hiebert, E. H. and Mesmer, H. A. E. (2013). Upping the ante of text complexity in the common core state standards: Examining its potential impact on young readers. *Educational Researcher*, 42(1):44–51.
- Hirschberg, J. and Manning, C. D. (2015). Advances in Natural Language Processing. *Science*, 349:261–266.
- Hocking, J. and Puttkammer, M. (2016). Optical character recognition for South African languages. In *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–5, New York. IEEE.
- Horne, F. and Ferreira-Meyers, K. A. F. (2017). Multilingualism and the language curriculum in South Africa: contextualising french within the local language ecology. *Stellenbosch Papers in Linguistics Plus*, 51(1):23–40.
- Hult, C. A. (2015). *The handy English grammar answer book*. Visible Ink Press, Canton.
- Humphreys, A. H. and Humphreys, J. T. (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.
- Ifeoluwa Adelani, D., Oluwadara Alabi, J., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., et al. (2022). A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv e-prints*, pages arXiv–2205.
- Islam, Z., Mehler, A., and Rahman, R. (2012). Text readability classification of textbooks of a low-resource language. In Adriani, M. and Arka, I. W., editors, *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*, pages 545–553. Association of Computational Linguistics.
- Islam, Z. and Rahman, R. (2014). Readability of Bangla news articles for children. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, pages 309–317. Association of Computational Linguistics.
- Ivanov, V., Solnyshkina, M. I., and Solovyev, V. (2018). Efficiency of text readability features in Russian academic texts. In *Proceedings of the Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”*, pages 403–413.
- Jahan, M. S. and Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546(August):126232.
- Jakobsen, T. and Skardal, T. (2007). Readability index. Report, Agder University.
- Janan, D. (2011). *Towards a new model of readability*. PhD thesis, University of Warwick.

- Janan, D. and Wray, D. (2012). *Guidance on the principles of language accessibility in National Curriculum assessments: research background*. Office of Qualifications and Examinations Regulation, London.
- Janan, D. and Wray, D. (2014). Reassessing the accuracy and use of readability formulae. *Malaysian journal of learning and instruction*, 11:127–145.
- Jansen, C., Richards, R., and Van Zyl, L. (2017). Evaluating four readability formulas for Afrikaans. *Stellenbosch Papers in Linguistics Plus*, 53:149–166.
- Jian, L., Xiang, H., and Le, G. (2022). English text readability measurement based on convolutional neural network: a hybrid network model. *Computational Intelligence and Neuroscience*, 2022:1–9.
- Jiang, Z., Sun, G., Gu, Q., and Chen, D. (2014). An ordinal multi-class classification method for readability assessment of Chinese documents. In Buchmann, R., Kifor, C. V., and Yu, J., editors, *Proceedings of the International Conference on Knowledge Science, Engineering and Management*, volume 8793, pages 61–72, New York. Springer.
- Joubert, K. and Githinji, E. (2014). Quality and readability of information pamphlets on hearing and paediatric hearing loss in the gauteng province, South Africa. *International journal of pediatric otorhinolaryngology*, 78(2):354–358.
- Jucksriporn, C. and Sornil, O. (2011). A minimum cluster-based trigram statistical model for thai syllabification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 493–505, New York. Springer.
- Kadenge, M. and Mugari, V. (2015). The current politics of African languages in Zimbabwe. *Per Linguam: A Journal of Language Learning*, 31:21–34.
- Kahn, D. (1976). *Syllable-based generalizations in English phonology*. Doctoral thesis, MIT, Cambridge, Massachusetts.
- Kamwangamalu, N. (2013). Language-in-education policy and planning in Africa’s monolingual kingdoms of lesotho and swaziland. In *Language Policies in Education: critical issues*, pages 156–169. Routledge, New York.
- Kandel, L. and Moles, A. (1958). Application de l’indice de Flesch à la langue Française. *Cahiers Etudes De Radio-Télévision*, 19(1958):253–274.
- Kang, Y., Cai, Z., Tan, C. W., Huang, Q., and Liu, H. (2020). Natural Language Processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7:139–172.
- Karpov, N., Baranova, J., and Vitugin, F. (2014). Single-sentence readability prediction in Russian. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts*, New York. Springer.

- Kasabwala, K., Agarwal, N., Hansberry, D. R., Baredes, S., and Eloy, J. A. (2012). Readability assessment of patient education materials from the American academy of otolaryngology—head and neck surgery foundation. *Otolaryngology–Head and Neck Surgery*, 147(3):466–471.
- Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *23rd International Conference on Computational Linguistics’ (Coling 2010), Beijing*, pages 546–554, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Kaur, S., Kaur, K., and Kaur, P. (2018). The influence of text statistics and readability indices on measuring university websites. *International Journal of Advanced Research in Computer Science*, 9(1):403–414.
- Kgampe, M. and Davel, M. H. (2010). Consistency of cross-lingual pronunciation of South-African personal names. In *Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)*, pages 123–127.
- Kgampe, M. and Davel, M. H. (2011). The predictability of name pronunciation errors in four South African languages. In *Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)*, pages 85–90.
- Khallaf, N. and Sharoff, S. (2021). Automatic difficulty classification of Arabic sentences. *arXiv:2103.04386*, preprint. Available at: <https://arxiv.org/abs/2103.04386>.
- Kincaid, J. P. and Delionbach, L. J. (1973). Validation of the automated readability index: A follow-up. *Human Factors*, 15(1):17–20.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease formula) for navy enlisted personnel. Report, Defense Technical Information Center.
- Koai, M. and Fredericks, B. G. (2019). Sesotho is still a marginalised language. *Southern African Linguistics and Applied Language Studies*, 37:303–314.
- Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia*, pages 9–16, New Brunswick, New Jersey, USA.
- Kondru, J. (2006). *Using part of speech structure of text in the prediction of its readability*. PhD thesis, The University of Texas at Arlington.
- Kosch, I. (2013). An analysis of the Oxford bilingual school dictionary: Northern Sotho and English (De Schryver 2007). *Lexikos*, 23:611–627.
- Kovač, A. and Marković, M. (2018). A rule-based syllabifier for Serbian. In *Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia*, pages 140–146, Ljubljana, Slovenia. University of Ljubljana.

- Kovač, A. and Marković, M. (2019). A mixed-principle rule-based approach to the automatic syllabification of Serbian. *Contributions to Contemporary History*, 59(1):120–1395.
- Krauwer, S. (2003). The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of SPECOM*, volume 2003, pages 15–22.
- Krige, D. (2009). Readability of the South African patient package insert (ppi). *Communitas*, 14:105–124.
- Krige, D. and Reid, M. (2017). A pilot investigation into the readability of Sesotho health information pamphlets. *Communitas*, 22:113–123.
- Krishna, D. (2020). Syllable structure and two word stress in Dogri. *IUP Journal of English Studies*, 15(3):115–132.
- Krisnawati, L. D. and Mahastama, A. W. (2018). A Javanese syllabifier based on its orthographic system. In *International Conference on Asian Language Processing (IALP)*, pages 244–249.
- Kriukova, A. (2022). *Measuring readability of technical texts*. Master thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.
- Kryściński, W., Paulus, R., Xiong, C., and Socher, R. (2018). Improving abstraction in text summarization. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium*, pages 1808–1817, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Kusec, S., Mastilica, M., Pavlekovic, G., and Kovacic, L. (2002). Readability of patient information on diabetes on the Croatian Web sites. In *Health Data in the Information Society, Netherlands*, pages 128–132. IOS Press, Amsterdam.
- Kusters, W. (2003). *Linguistic complexity, the Influence of Social Change on Verbal Inflection*. Thesis, University of Leiden.
- Kärbla, T., Uibu, K., and Männamaa, M. (2019). National Estonian-language tests: What is measured in text comprehension tasks? *New Trends and Issues Proceedings on Humanities and Social Sciences*, 6:8–16.
- Kärbla, T., Uibu, K., and Männamaa, M. (2020). Teaching strategies to improve students’ vocabulary and text comprehension. *European Journal of Psychology of Education*, 35:1–20.
- Land, S. (2015). Reading and the orthography of isiZulu. *South African Journal of African Languages*, 35(2):163–175.

- Laposhina, A., Veselovskaya, T., Lebedeva, M., and Kupreshchenko, O. (2018). Automated text readability assessment for Russian second language learners. In *Proceedings of the Computational Linguistics and Intellectual Technologies*, pages 396–406.
- Lasecki, W. S., Rello, L., and Bigham, J. P. (2015). Measuring text simplification with the crowd. In Carriço, L. and Mirri, S., editors, *Proceedings of the 12th International Web for All Conference, Florence Italy*, pages 1–9, New York. Association for Computing Machinery.
- Latief, M. R. A., Saleh, N. J., and 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.
- Lee, J. and Vajjala, S. (2022). A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*. Available at: <https://arxiv.org/abs/2203.07450>.
- Lehohla, P. (2011). Census in brief. *Statistics South Africa*. government printing works: Pretoria. Available at: https://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf [Accessed: 14 Sept. 2022].
- Lemeko, P. A. (2018). *Diachronic investigation into current issues in language variation. A case of Sesotho language*. PhD thesis, Bloemfontein: Central University of Technology, Free State.
- Leopeng, M. T. (2019). *Translations of informed consent documents for clinical trials in South Africa: Are they readable?* Thesis, University of Cape Town.
- Lesotho (1993). *The Constitution of Lesotho*. Government Printer, Maseru.
- Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-put-er*. Doctoral thesis, Stanford University, Stanford, USA.
- Liddy, E. D. (2001). *Natural Language Processing*. Marcel Decker, Inc, New York, 2nd ed edition.
- Lignos, C., Holley, N., Palen-Michel, C., and Sälevä, J. (2022). Toward more meaningful resources for lower-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland*, pages 523—532, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Lin, A., Kolluri, S., and Sheehan, D. (2021). Calculating Latin readability scores using linear regression. A Pre-Print. Available at: https://mehtaplustutoring.com/wp-content/uploads/2021/08/Latin_Readability_Project_Paper.pdf.
- Lissón, P. (2017). Investigating the use of readability metrics to detect differences in written productions of learners: A corpus-based study. *Bellaterra Journal of Teaching and Learning Language and Literature*, 10(4):68–86.

- Lloret, E. and Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review*, 37(1):1–41.
- Loukatou, G. (2019). From phonemes to morphemes: Relating linguistic complexity to unsupervised word over-segmentation. In Dubossarsky, H., McCarthy, A. D., Ponti, E. M., Vulić, I., Vylomova, E., Berzak, Y., Cotterell, R., Faruqui, M., Korhonen, A., and Reichart, R., editors, *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, Florence, Italy*, New Jersey, USA. Association of Computational Linguistics.
- Lu, X. and Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29:16–27.
- Lyashevskaya, O. and Sharov, S. (2009). *The frequency dictionary of modern Russian language*. Azbukovnik, Moscow.
- Madigoe, M. W. (2003). *Syllable structure processes in Northern Sotho: A linear and non-linear phonological analysis*. Stellenbosch University Press, Stellenbosch, South Africa.
- Maegaard, B., Choukri, K., Mokbel, C., and Yaseen, M. (2005). *Language technology for Arabic*. Center for Sprogteknologi, University of Copenhagen.
- Maegaard, B., Krauwer, S., Choukri, K., and Jørgensen, L. D. (2006). The BLARK concept and BLARK for Arabic. In *LREC*, pages 773–778.
- Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *ArXiv*, abs/2006.07264. <https://api.semanticscholar.org/CorpusID:219636137>.
- Mahloane, M. J. and Trausan-Matu, S. (2015). Metaphor annotation in Sesotho text corpus: Towards the representation of resource-scarce languages in NLP. In *The 20th International Conference on Control Systems and Computer Science*, pages 405–410, New York. IEEE.
- Makoni, S. (2003). From misinvention to disinvention of language: Multilingualism and the South African constitution. *Black linguistics, language, society, and politics in Africa and the Americas*, pages 132–153.
- Makutoane, T. J. (2022). ‘The people divided by a common language’: The orthography of Sesotho in Lesotho, South Africa, and the implications for Bible translation. *HTS Theologiese Studies/Theological Studies*, 78(1):9.
- Mandic, C. G., Rudd, R., Hehir, T., and Acevedo-Garcia, D. (2012). Readability of special education procedural safeguards. *The Journal of Special Education*, 45(4):195–203.
- Mankayi, S., Matenda, F. R., and Sibanda, M. (2023). An analysis of the readability of the chairman’s statement in South Africa. *Risks*, 11(3):54.

- Marupi, O. and Charamba, E. (2022). Revisiting the effects of -isms in the promotion, development, and revitalisation of indigenous languages in Zimbabwe: The position of Sesotho in Gwanda South, Zimbabwe. In *Handbook of Research on Teaching in Multicultural and Multilingual Contexts*, pages 32–46. IGI Global, Hershey.
- Matlosa, L. (2017). Sesotho orthography called into question: the case of some Sesotho personal names. *Nomina Africana: Journal of African Onomastics*, 31(1):51–58.
- Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.
- McDermid Heyns, J. (2007). Readability statistics for Afrikaans. In *LSSA/SAALT/SAALA Joint Annual Conference, North-West University, Potchefstroom, South Africa*.
- McKellar, C. (2008). A readability formula for Afrikaans. Poster presented at PRASA, 27–28 November, Cape Town, South Africa.
- McKellar, C. (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.
- McKellar, C. (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].
- McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010a). Linguistic features of writing quality. *Written Communication*, 27:57–86.
- McNamara, D. S. and Graesser, A. C. (2012). Coh-metrix: An automated tool for theoretical and applied Natural Language Processing. In McCarthy, P. M. and Boonthum, C., editors, *Applied Natural Language Processing: Identification, investigation and resolution*, pages 188–205. IGI Global, Hershey.
- McNamara, D. S., Graesser, A. C., Cai, Z., and Kulikowich, J. M. (2011). Coh-metrix easability components: Aligning text difficulty with theories of text comprehension. In *Annual meeting of the American Educational Research Association*, pages 1–12. SAGE.
- McNamara, D. S., Graesser, A. C., and Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In Sabatini, J., Albro, E., and O'Reilly, T., editors, *Measuring up: Advances in how we assess reading ability*, volume 83, pages 89–116. R&L Education.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, New York.

- McNamara, D. S., Louwerse, M. M., and Graesser, A. C. (2002). Coh-matrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Technical report, Institute for Intelligent Systems, University of Memphis.
- McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010b). Coh-matrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.
- McWhorter, J. (2012). Complexity hotspot. In *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, pages 243–246. Walter De Gruyter, Berlin.
- Medvedev, G. (2016). Google translate in teaching English. *Journal of teaching English for specific and academic purposes*, 4(1):181–193.
- Mesmer, H. A., Cunningham, J. W., and Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading research quarterly*, 47:235–258.
- Messerschmidt, H., Messerschmidt, J., and Thulo, D. (2003). A human-assisted computer generated LA-grammar for simple sentences in Southern Sotho. *Southern African Linguistics and Applied Language Studies*, 21(1–2):41–47.
- Meyer, B. J. (2003). ‘Text coherence and readability’. *Topics in Language Disorders*, 23:204–224.
- Miestamo, M. (2006). On the feasibility of complexity metrics. In Kerge, K. and Sepper, M., editors, *Finest linguistics, proceedings of the annual Finnish and Estonian conference of linguistics*, pages 11–26, Tallin. Tallinn University Press.
- Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In *Language complexity: Typology, contact, change*, pages 23–41. John Benjamins Publishing Company, Amsterdam.
- Miestamo, M. (2017). Linguistic diversity and complexity. *Lingue e linguaggio*, 16(2):227–254.
- Mikulecky, D. C. (2001). The emergence of complexity: science coming of age or science growing old? *Computers & chemistry*, 25(4):341–348.
- Moeketsi, V. S. M. (2014). The demise of Sesotho language in the democratic South Africa and its impact on the socio-cultural development of the speakers. *Journal of Sociology and Social Anthropology*, 5:217–224.
- Mohammadi, H. and Khasteh, S. H. (2020). A machine learning approach to Persian text readability assessment using a crowdsourced dataset. In *Proceedings of the 28th Iranian Conference on Electrical Engineering*, pages 1–7. ICEE.
- Mohammed, L. A., Aljaberi, M. A., Anmary, A. S., and Abdulkhaleq, M. (2023). Analysing English for science and technology reading texts using Flesch Reading Ease online formula: the preparation for academic reading. In *International Conference on Emerging Technologies and Intelligent Systems*, pages 546–561, New York. Springer.

- Mohasi, L. and Mashao, D. (2005). Phonetization for text-to-speech synthesis in Sesotho. In *The sixteenth annual symposium of the Pattern Recognition Association of South Africa*, pages 121–122, Langebaan, South Africa. Citeseer.
- Mohasi, L., Mixdorff, E., and Niesler, T. (2011). An acoustic analysis of tone in Sesotho. In *17th International Congress of Phonetic Sciences*, pages 17–21, Hong Kong. Cambridge.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Reprint—preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Physical therapy*, 89(9):873–880.
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., and Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*, 4(1):1–9.
- Moher, D., Stewart, L., and Shekelle, P. (2016). Implementing PRISMA-P: recommendations for prospective authors. *Systematic reviews*, 5:1–2.
- Mojela, V. (2016). Etymology & figurative: The role of etymology in the lemmatization of Sotho terminology. In *The 10th International Conference of the Asian Association for Lexicography (AsiaLex2016)*, pages 93–100, New York. IEEE.
- Moleleki, M. (1999). The state of lexicography in Sesotho. *Lexikos*, 9(9):241–247.
- Moors, C., Wilken, I., Calteaux, K., and Gumede, T. (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.
- Moors, C., Wilken, I., Calteaux, K., and Gumede, T. (2018b). 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.
- Moors, C., Wilken, I., Gumede, T., and Calteaux, K. (2018c). Human Language Technology audit 2017/18. Available at: <https://sadilar.org/index.php/en/2-general/284-health-resources> [Accessed: 14 May. 2023].
- Moors, C., Wilken, I., Gumede, T., and Calteaux, K. (2018d). Human Language Technology audit 2017/18. Technical report, CSIR Meraka Institute.
- Motjope-Mokhali, T. L., Kosch, I. M., and Mafela, M. J. (2020). Sethantso sa Sesotho and Sesuto-English dictionary: A comparative analysis of their designs and entries. *Lexikos*, 30:1–17.
- Naderi, B., Mohtaj, S., Karan, K., and Möller, S. (2019). Automated text readability assessment for German language: A quality of experience approach. In Möller, S. and Egger-Lampl, S., editors, *Proceedings of the eleventh international conference on Quality of Multimedia Experience (QoMEX)*, New York. IEEE.

- Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural Language Processing: An introduction. *Journal of the American Medical Informatics Association*, 18:554–551.
- Nahar, K. M. O., Al Eroud, A. F., Barahoush, M., and Al-Akhras, A. M. (2019). SAP: Standard Arabic profiling toolset for textual analysis. *International Journal of Machine Learning and Computing*, 9:222–229.
- Narasinh, V. (2019). Readability analysis of Kannada language. In *Proceedings of the 1st International Conference on Advances in Information Technology*, pages 45–49. Curran Associates, Inc.
- Nassiri, N., Cavalli-Sforza, V., and Lakhouaja, A. (2022a). Approaches, methods and resources for assessing the readability of Arabic texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–30.
- Nassiri, N., Lakhouaja, A., and Cavalli-Sforza, V. (2022b). Arabic L2 readability assessment: Dimensionality reduction study. *Journal of King Saud University - Computer and Information Sciences*, 34:3789—3799.
- Ndhlovu, F. (2012). Book review the politics of language and nation building in Zimbabwe. *Southern African Linguistics and Applied Language Studies*, 30(1):125—129.
- Ndlovu, E. (2011). Mother tongue education in the official minority languages in Zimbabwe. *South African Journal of African Languages*, 31:229–242.
- Ndlovu, E. (2013). *Mother tongue education in official minority languages of Zimbabwe: A language management critique*. Thesis, University of the Free State.
- Németh, L. (2006). Automatic non-standard hyphenation in openoffice.org. *TUGboat*, 27(1):32–37.
- Newbold, N. (2013). *New Approaches for Text Readability*. Thesis, University of Surrey.
- Newbold, N. and Gillam, L. (2010). Populating a framework for readability analysis: Word frequency = word difficulty.
- Nirmaldasan (2016). Longer the sentence, greater the strain. ONLINE Blog. Available at: <https://strainindex.wordpress.com/> [Accessed: 1 Mar. 2023].
- Nkolola-Wakumelol, M., Rantsoz, L., and Matlhaku, K. (2012). Syllabification of consonants in Sesotho and Setswana. In Nginga-Koumba-Binza, H. S. and Bosch, S., editors, *Language Science and Language technology in Africa: Festschrift for Justus C. Roux*, pages 10–13. Sun Express, Stellenbosch, South Africa.
- Nouwens, S., Groen, M. A., and Verhoeven, L. (2016). How storage and executive functions contribute to children’s reading comprehension. *Learning and Individual Differences*, 47:96–102.

- Nyika, N. (2012). The politics of language and nation building in Zimbabwe: by Finex Ndhlovu. *Southern African Linguistics and Applied Language Studies*, 30(1):125–129.
- Oborneva, I. V. (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 sodержaniya i metodov obucheniya [RAS Institute of Content and Teaching Methods].
- O'Brien, S. (2010). Controlled language and readability. In Shreve, G. M. and Angelone, E., editors, *Translation and cognition*, volume 15, pages 143–168. John Benjamins Publishing Company, Philadelphia.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 88:1–11.
- Palotti, J., Goeuriot, L., Zuccon, G., and 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*, pages 965–968.
- Paterson, J. H. and Hartmann, K. (2016). Readability metrics for program code: How is reading ease reflected in gaze? In Bednarik, R., Busjahn, T., Schulte, C., and Tamm, S., editors, *Proceedings of the Third International Workshop on Eye Movements in Programming: Models to data*, volume 23, pages 19–22. University of Eastern Finland.
- Patil, S. and Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. *British Medical Journal*, 349:1–3.
- Phani, S., Lahiri, S., and Biswas, A. (2019). Readability analysis of Bengali literary texts. *Journal of Quantitative Linguistics*, 26(4):287–305.
- Phillips Galloway, E., Uccelli, P., Aguilar, G., and Barr, C. D. (2020). Exploring the cross-linguistic contribution of Spanish and English academic language skills to English text comprehension for middle-grade dual language learners. *AERA Open*, 6:1–20.
- Pires, C., Cavaco, A., and Vigário, M. (2017). Towards the definition of linguistic metrics for evaluating text readability. *Journal of Quantitative Linguistics*, 24(4):319–349.
- Pitler, E., Louis, A., and Nenkova, A. (2010). Automatic evaluation of linguistic quality in multi-document summarization. In Hajič, J., Carberry, S., Clark, S., and Nivre, J., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Brussels, Belgium*, volume 1, pages 544–554, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Piu, R. K., Hossain, K. R., Sabab, N. H., and Rafi, R. B. (2020). *Hatred Message Identification using DistilBERT*. PhD thesis, United International University.

- Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., and Thompson, W. H. (2017). The readability of scientific texts is decreasing over time. In Rodgers, P. A., editor, *Meta-Research: A Collection of Articles*, volume 6, pages 1–14. eLife.
- Prah, K. K. (2007). *Challenges to the promotion of indigenous languages in South Africa*. The Center for Advanced Studies of African Society, Cape Town.
- Prasanna, P. and Vaishnavh, N. (2014). Readability estimation: A recursive formulation. *Natural Language Processing*, pages 1–14.
- Pretorius, E., Mohohlwane, N., and Spaull, N. (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.
- Pretorius, E. J. and Stoffelsma, L. (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.
- Pretorius, L., Soria, C., and Baroni, P., editors (2014). *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*. LREC, Iceland.
- Probert, T. and De Vos, M. (2016). Word recognition strategies among isiXhosa/English bilingual learners: The interaction of orthography and language learning and teaching. *Journal of the Reading Association of South Africa*, 7:1–10.
- Python (2020). Python online. SOFTWARE. Available at: <https://www.python.org> [Accessed: 22 October 2020 2020].
- Qiu, X.-Y., Deng, K., Qiu, L., and Wang, X. (2018). Exploring the impact of linguistic features for Chinese readability assessment. In Huang, X., Jiang, J., Zhao, D., Feng, Y., and Hong, Y., editors, *Proceedings of the NLPCC 2017: Lecture Notes in Computer Science*, volume 10619, pages 771–783, New York. Springer.
- Razon, A. and Barnden, J. (2015). A new approach to automated text readability classification based on concept indexing with integrated part-of-speech n-gram features. In Mitkov, R., Angelova, G., and Bontcheva, K., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria*, volume 1, Shoumen, Bulgaria. INCOMA Ltd.
- Reid, M., Nel, M., and 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.
- Reina, A., Robles, G., and González-Barahona, J. M. (2013). A preliminary analysis of localization in free software: How translations are performed. In *IFIP International Conference on Open Source Systems*, pages 153–167.

- Republic of South Africa (1996). *Constitution of the Republic of South Africa*. Department of Justice, Pretoria.
- Republic of South Africa (2023). *Constitution Eighteenth Amendment Bill*. Department of Justice and Correctional Services, Pretoria.
- Rescher, N. (1998). *Complexity: A philosophical overview*. Transaction, New Brunswick.
- Riep, D. M. (2013). Seeing Sesotho: Art, history, and the visual language of South Sotho identity. *Southern African Humanities*, 25:217–244.
- Rosen, A. (2016). *InterCorp—a look behind the façade of a parallel corpus*, pages 21–40. Instytut Lingwistyki Stosowanej, Warszawa.
- Roux, J. C. and Bosch, S. E. (2019). Preserving and developing indigenous languages in the South African context. In *Proceedings of the Language Technologies for All (LT4All), Paris, France*, pages 97–100, Paris. European Language Resources Association.
- Roux, J. C., Louw, P. H., and Niesler, T. R. (2004). The African speech technology project: An assessment. In *Proceedings of LREC*, pages 93–96.
- Rowley, J. and Slack, F. (2004). Conducting a literature review. *Management research news*, 27:31–39.
- Rubach, J. (1993). *The Lexical Phonology of Slovak*. Oxford University Press, New York.
- Rudnicka, K. (2018). Variation of sentence length across time and genre. In Whitt, R. J., editor, *Diachronic corpora, genre, and language change*, pages 220–240. John Benjamins, Amsterdam.
- Rules, W. C. (1969). The SMOG readability formula. *Journal of Reading*, 12(8):639–646.
- Saddiki, H., Bouzoubaa, K., and Cavalli-Sforza, V. (2015). Text readability for Arabic as a foreign language. In *Proceedings of the IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, Morocco*, pages 1–8, New York. IEEE.
- Sai, S., Mishra, A., Anirban, L., Jain, P., and Sankaranarayanan, K. (2019). Unsupervised neural text simplification. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, volume 1, pages 2058–2068, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Sarimanah, E., Soeharto, A. R., and Suhendra, R. E. (2021). Investigating Indonesian textbooks readability using Fry Graph formula. *Jurnal Keguruan dan Ilmu Tarbiyah*, 6(1):207–218.
- Sazzed, S. (2022). Influence of language proficiency on the readability of review text and transformer-based models for determining language proficiency. In *Companion Proceedings of the Web Conference 2022*, pages 881–886.

- Scarton, C. and Aluisio, S. M. (2010). Coh-metrix-port: A readability assessment tool for texts in Brazilian Portuguese. In Pardo, T. A. S., Branco, A., Klautau, A., Vieira, R., and De Lima, V. L. S., editors, *Proceedings of the Computational Processing of the Portuguese Language, 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil*, volume 6001, New York. Springer.
- Scarton, C., De Oliveira, M., Candido, A., Gasperin, C., and Aluísio, S. M. (2010). SIMPLIFICA: A tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In Rosé, C. P., editor, *Proceedings of the NAACL HLT2010: Demonstration Session, Los Angeles, California*, pages 41–44, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Schlemmer, M. and Fourie, W. (2013). Autshumato TMX Integrator. Southern African Centre for Digital Language Resources, <https://repo.sadilar.org/handle/20.500.12185/416> [Accessed: 1 Jun. 2022].
- Sefara, T. J., Mokgonyane, T. B., and Marivate, V. (2021). Practical approach on implementation of wordnets for South African languages. In *Proceedings of the 11th Global Wordnet Conference, Pretoria, South Africa*, pages 20–25, Netherlands. Global Word-Net Association.
- Sekere, N. B. (2004). *Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa*. Thesis, University of South Africa.
- Selepe, T. (1997). Some implications of media policy and ethics in literary production: A preliminary survey of Sesotho literature. *South African Journal of African Languages*, 17(2):80–84.
- Serhani, M. A., Jaffar, A., Campbell, P., and Atif, Y. (2011). Enterprise web services-enabled translation framework. *Information Systems and e-Business Management*, 9:497–517.
- Shamseer, L., Moher, D., Clarke, M., Gherzi, D., Liberati, A., Petticrew, M., Shekelle, P., and Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *British Medical Association*, 349:1–25.
- Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.
- Si, L. and Callan, J. (2001). A statistical model for scientific readability. In Pu, C., Liu, L., and Paques, H., editors, *Proceedings of the tenth international conference on Information and knowledge management, Atlanta Georgia, USA*, pages 574–576, New York. Association for Computing Machinery.
- Sibanda, J. and Baxen, J. (2016). Determining ESL learners’ vocabulary needs from a textbook corpus: Challenges and prospects. *Southern African Linguistics and Applied Language Studies*, 34(1):57–70.

- 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. (2022a). Exploring linguistic complexity of English First Additional Language grade 12 and grade 11 exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 4(2):01–12.
- Sibeko, J. (2022b). Tshebediso ya melao kabong ya dinoko tsa Sesotho. *Southern African Linguistics and Applied Language Studies*, 40(4):494–506.
- Sibeko, J. and 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)*, Dubrovnik, Croatia, pages 32–41, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Sibeko, J. and Setaka, M. (2022). Sesotho BLARK content. *Journal of Digital Humanities Association of South Africa*, 4(2):1–11.
- Sibeko, J. and Setaka, M. (2023). Evaluating the Sesotho rule-based syllabification system on Sepedi and Setswana words. In *Proceedings of the Fourth workshop on Resources for African Indigenous Language (RAIL 2023)*, Dubrovnik, Croatia, pages 76–85, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Sibeko, J. and Van Zaanen, M. (2021). An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11.
- Sibeko, J. and Van Zaanen, M. (2022a). 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.
- Sibeko, J. and Van Zaanen, M. (2022b). *Raw and syllabified word list for Sesotho*. Southern African Centre for Digital Language Resources. <https://repo.sadilar.org/handle/20.500.12185/556>, [Accessed: 1 Jun. 2022].
- Sibeko, J. and Van Zaanen, M. (2022c). Sesotho syllabification systems. Southern African Centre for Digital Language Resources. Available at: <https://repo.sadilar.org/handle/20.500.12185/555> [Accessed: 3 Jan 2023].

- Sibeko, J. and Van Zaanen, M. (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.
- Sinha, M. and Basu, A. (2016). A study of readability of texts in Bangla through machine learning approaches. *Education and Information Technologies*, 21:1071–1094.
- Sinha, M., Dasgupta, T., and Basu, A. (2014). Text readability in Hindi: A comparative study of feature performances using support vectors. In Sharma, D. M., Sangal, R., and Pawar, J. D., editors, *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, pages 223–231, India. NLP Association of India.
- Sinha, M., Sharma, S., Dasgupta, T., and Basu, A. (2012). New readability measures for Bangla and Hindi texts. In Kay, M. and Boitet, C., editors, *Proceedings of COLING 2012, Mumbai, India*, pages 1141–1151, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Sinnemäki, K. (2011). *Language universals and linguistic complexity: Three case studies in core argument marking*. Thesis, University of Helsinki.
- Sjöholm, J. (2012). *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. Thesis, Linköpings Universitet.
- Skosana, N. J. and Mlambo, R. (2021). A brief study of the autshumato machine translation web service for South African languages. *Literator*, pages 1–7.
- Smith, E. A. and Senter, R. (1967). *Automated readability index*. Clearing house for Federal Scientific and Technical information, Cincinnati, Ohio.
- Smith, T. (2016). The problems with current readability methods and formulas: Missing that usability design. In Hart, H., editor, *2016 IEEE International Professional Communication Conference (IPCC), Austin, Texas*, pages 1–4, New York. IEEE.
- Snyman, D., Van Huyssteen, G. B., and Daelemans, W. (2011). Automatic genre classification for resource scarce languages. In *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa*, pages 132–137.
- Soh, K. C. (2019). Readability formula for Chinese as a second language: An exploratory study. *Frontiers of Education in China*, 14(4):551–574.
- Sojka, P. and Sojka, O. (2003). Towards new Czechoslovak hyphenation patterns. *Zpravo-daj Československého sdružení uživatelů TeXu*, 30(3):118–126.
- Solnyshkina, M. I., Guryanov, I., Gafiyatova, E., and Varlamova, E. (2018a). Readability metrics: The case of Russian educational texts. In Uslu, F., editor, *The Proceedings of the ADVED-4th International Conference on Advances in Education and Social Sciences, Istanbul, Turkey*, pages 676–681. Ocerints Publications.

- Solnyshkina, M. I., Harkova, E., and Kazachkova, M. (2020). The structure of cross-linguistic differences: Meaning and context of ‘readability’ and its Russian equivalent ‘chitabelnost’. *Journal of Language and Education Volume*, 6(1):103–119.
- Solnyshkina, M. I., Solovyev, V., Ivanov, V., and Danilov, A. (2018b). Studying text complexity in Russian academic corpus with multi-level annotation. In Loukachevitch, N. and Elizarov, A., editors, *Proceedings of the international workshop on Computational Models in Language and Speech, Kazan, Russia*, pages 126–136, Aachen, Germany. CEUR Workshop Proceedings.
- Solnyshkina, M. I., Zamaletdinov, R. R., Gorodetskaya, L. A., and Gabitov, A. I. (2017). Evaluating text complexity and Flesch-Kincaid Grade Level. *Journal of social studies education research*, 8(3):238–248.
- Solovyev, V., Ivanov, V., and Solnyshkina, M. I. (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of intelligent and fuzzy systems*, 34(5):3049–3058.
- Solovyev, V., Solnyshkina, M. I., Ivanov, V., and Batyrshind, I. (2019). Prediction of reading difficulty in Russian academic texts. *Journal of intelligent and fuzzy systems*, 36(5):4553–4563.
- Sowmya, V. and Detmar, M. (2014). Readability assessment for text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):194–222.
- Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., and Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Reading and writing*, 32:603–637.
- Stahl, S. A. (2003). Vocabulary and readability: How knowing word meanings affects comprehension. *Topics in language disorders*, 23(3):241–247.
- Stenner, A. J., Burdick, H., Sanford, E. E., and Burdick, D. S. (2006). How accurate are lexile text measures? *Journal of Applied Measurement*, 7:307–322.
- Stenner, A. J. and Stone, M. H. (2004). Does the reader comprehend the text because the reader is able or because the text is easy? In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner*. Springer Nature Singapore.
- Stevens, K. T., Stevens, K. C., and Stevens, W. P. (1992). Measuring the readability of business writing: The cloze procedure versus readability formulas. *The Journal of Business Communication (1973)*, 29:367–382.
- Stocker, L. P. (1971). Increasing the precision of the Dale-Chall readability formula. *Reading Improvement*, 8(3):87.

- Stoffelsma, L. (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.
- Stoffelsma, L. (2019b). From ‘sheep’ to ‘amphibian’: English vocabulary teaching strategies in South African township schools. *South African Journal of Childhood Education*, 9(1):1–10.
- Strassel, S. and Tracey, J. (2016). Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3273–3280.
- Sun, G., Jiang, Z., Gu, Q., and Chen, D. (2014). Linear model incorporating feature ranking for Chinese documents readability. In Dong, M., editor, *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Singapore, pages 29–34, Middlesex, New Jersey. Piscataway.
- Suyanto, S. (2020). Phonological similarity-based backoff smoothing to boost a bigram syllable boundary detection. *International Journal of Speech Technology*, 23:191–204.
- Taljard, E. (2008). Terminology practice in a non-standardized environment: A case study. In *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)*, pages 1073–1080.
- Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Computational linguistics*, 36(2):203–227.
- Tanprasert, T. and Kauchak, D. (2021). Flesch-Kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14.
- Tennent, W. (2014). *Understanding reading comprehension: Processes and practices*. Sage, London.
- Thirion, J. W., Van Heerden, C., Giwa, O., and Davel, M. H. (2020). The South African directory enquiries (SADE) name corpus. *Language Resources and Evaluation*, pages 155–184.
- Thomas, G., Hartley, R. D., and Kincaid, J. P. (1975). Test-retest and inter-analyst reliability of the automated readability index, Flesch Reading Ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.
- Tongtep, N., Coenen, F., and Theeramunkong, T. (2014). Content-based readability assessment: A study using a syllabic alphabetic language (Thai). In Walsh, T. and Governatori, G., editors, *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI-2014)*, Australia, Australia. Pacific Rim International Conferences on Artificial Intelligence (PRICAI).

- Toyama, Y., Hiebert, E. H., and Pearson, P. D. (2017). An analysis of the text complexity of levelled passages in four popular classroom reading assessments. *Educational Assessment*, 22(3):139–170.
- Tsai, S.-C. (2019). Using Google Translate in EFL drafts: a preliminary investigation. *Computer Assisted Language Learning*, 32(5-6):510–526.
- Tshesane, T. M. I. (2014). *Evaluating the functionality of the translated Sepedi Home Language CAPS for Grade 10–12*. MA report, University of the Witwatersrand.
- Tsvetkov, Y. (2017). Opportunities and challenges in working with low-resource languages. CMU, Powerpoint Slides, Part One.
- Vadlapudi, R. and Katragadda, R. (2010a). On automated evaluation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In Hachey, B. and Osborne, M., editors, *Proceedings of the NAACL HLT2010 Workshop on Computational Linguistics in a World of Social Media, Los Angeles, California, USA*, New Jersey, USA. Association for Computational Linguistics.
- Vadlapudi, R. and Katragadda, R. (2010b). Quantitative evaluation of grammaticality of summaries. In Gelbukh, A., editor, *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Iasi, Romania*, volume 1, pages 736–747, New York. Springer.
- Valera, H. A. and Luštrek, M. (2022). Ensuring privacy in natural language processing: A PRISMA oriented. In *Workshops at 18th International Conference on Intelligent Environments (IE2022), Biarritz, France*, volume 31, page 180, Amsterdam. IOS Press.
- Van Cranenburgh, A. (2013). *Readability*. Github. Available at: <https://github.com/andreascv/readability/> [Accessed: 20 Feb 2023].
- Van der Walt, C. (2010). Of shoes-and ships-and sealing-wax: A dynamic systems approach to language curriculum orientation. *Southern African Linguistics and Applied Language Studies*, 28:323–337.
- 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 der Walt, M., Maree, K., and Ellis, S. (2008). A mathematics vocabulary questionnaire for use in the intermediate phase. *South African Journal of Education*, 28(4):489–504.
- Van Heerden, C., Kleynhans, N., Barnard, E., and Davel, M. (2010). Pooling ASR data for closely related languages. In Besacier, L. and Castelli, E., editors, *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU 2010), Penang, Malaysia*, pages 17–23, Malaysia. Universiti Sains Malaysia.

- Van Heerden, E., Davel, M., and Barnard, E. (2014). Performance analysis of a multilingual directory enquiries application. In *Proceedings of the Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, pages 258–263.
- Van Huyssteen, G. and Van Zaanen, M. (2004). Learning compound boundaries for Afrikaans spelling checking. In *Pre-Proceedings of the Workshop on International Proofing Tools and Language Technologies*, pages 101–108.
- Van Huyssteen, G. B. and Puttkammer, M. (2007). Accelerating the annotation of lexical data for less-resourced languages. *Interspeech*, pages 1505–1508.
- Van Oosten, P. and 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*, pages 120–129.
- Van Oosten, P., Tanghe, D., and 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)*, Valletta, Malta, pages 775–782, Paris. European Language Resources Association (ELRA).
- Van Rooyen, R. (1986). Eerste Afrikaanse leesbaarheids formules. *Commucatio: South African Journal for Communication Theory and Research*, 12(1):59–69.
- Čermák, F. and Rosen, A. (2012). The case of Intercorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.
- Venugopal, G., Pramod, D., Saini, J. R., Pramod, D., and Saini, J. (2021). Analyzing complex words in Hindi using parameters of classical readability formulae (part 1). *Computer Science Journal of Moldova*, 29(3):366–387.
- Verma, P. and Om, H. (2019). A novel approach for text summarization using optimal combination of sentence scoring methods. *Sādhanā*, 44:1–15.
- Verma, P. and Verma, A. (2020). A review on text summarization techniques. *Journal of Scientific Research*, 64(1):251–257.
- Vincent, S. (2014). Sentence length: why 25 words is our limit. ONLINE Blog. Available at: <https://insidegovuk.blog.gov.uk/2014/08/04/sentence-length-why-25-words-is-our-limit/> [Accessed: 1 Mar. 2023].
- Škvorc, T., Krek, S., Pollak, S., Holdt, S. A., and Robnik-Šikonja, M. (2018). Evaluation of statistical readability measures on Slovene texts. In Lenardič, J., editor, *Proceedings of the Conference on Language Technologies and Digital Humanities, University of Ljubljana*, pages 240–247, Slovenia. Prispevki.
- Škvorc, T., Krek, S., Pollak, S., Holdt, S. A., and Robnik-Šikonja, M. (2019). Predicting Slovene text complexity using readability measures. *Prispevki za novejšo zgodovino*, 58(3):197–218.

- Štajner, S., Calixto, I., and Saggion, H. (2015). Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international international conference recent advances in Natural Language Processing, Hissar, Bulgaria*, pages 618–626, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Waltl, B. and Matthes, F. (2014). Towards measures of complexity: Applying structural and linguistic metrics to German laws. In *Legal knowledge and information systems, Kraków, Poland*, pages 153–162. IOS Press, Amsterdam.
- Wang, Z., Zhao, X., Song, W., and Wang, A. (2019). Readability assessment of textbooks in low resource languages. *Computers, Materials and Continua*, 61:213–225.
- Way, A. and Hearne, M. (2011). On the role of translations in state-of-the-art statistical machine translation. *Language and Linguistics Compass*, 5(5):227–248.
- Webb, V., Lafon, M., and Pare, P. (2010). Bantu languages in education in South Africa: An overview. Ongekho akekho!—The absentee owner. *Language learning journal*, 38(3):273–292.
- Wilken, I., Gumede, T., Moors, C., and Calteaux, K. (2018). Human Language Technology audit 2018: Design considerations and methodology. In *International Conference on Intelligent and Innovative Computing Applications (ICONIC), Plaine Magnien, Mauritius*, pages 1–7, New York. IEEE.
- Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K. P., and François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.
- Wills, S., Uys, P., Van Heerden, C. J., and Barnard, E. (2020). Language modeling for speech analytics in under-resourced languages. In *Interspeech 2020*, pages 4941–4945.
- Wilsenach, C. and Schaefer, M. (2022). Development and initial validation of productive vocabulary tests for isiZulu, Siswati and English in South Africa. *Language Testing*, 39(4):567–592.
- Wissing, D. and Roux, J. C. (2017). The status of tone in Sesotho: A production and perception study. *Nordic Journal of African Studies*, 26:19–19.
- Wissing, G. J., Blignaut, A. S., and Van Den Berg, K. (2016). Using readability, comprehensibility and lexical coverage to evaluate the suitability of an introductory accountancy textbook to its readership. *Stellenbosch Papers in Linguistics*, 46:155–179.
- Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In Barzilay, R. and Johnson, M., editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association of Computational Linguistics.

- Wray, D. and Dahlia, J. (2013). Exploring the readability of assessment tasks: The influence of text and reader factors. *REMIE: Multidisciplinary Journal of Educational Research*, 3(1):69–95.
- Wray, D. and Janan, D. (2013). Readability revisited? The implications of text complexity. *The Curriculum Journal*, 24(4):553–562.
- Wubben, S., Van Den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In Li, H., LIne, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1015–1024, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Xia, M., Kochmar, E., and Briscoe, T. (2019). Text readability assessment for second language learners. In Tetreault, J., Burstein, J., Leacock, C., and Yannakoudakis, H., editors, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, volume 1, pages 12–22, New Brunswick, New Jersey, USA. Association for Computational Linguistics.
- Xie, Z., Man, W., Liu, C., and Fu, X. (2023). A PRISMA-based systematic review of measurements for school bullying. *Adolescent Research Review*, 8(2):219–259.
- Xu, W., Yao, Z., and Chen, D. (2019). Chinese annual report readability: Measurement and test. *China Journal of Accounting Studies*, 7(3):407–437.
- Yan, X., Song, D., and Li, X. (2006). Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 540–549.
- Zamanian, M. and Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.
- Zano, K. (2020). Reading comprehension strategies of the EFAL learners in the FET phase: Teacher perspectives. *e-BANGI*, 17(7):1–12.
- Zhang, Y., Lin, N., and Jiang, S. (2019). A study on syntactic complexity and text readability of ASEAN English news. In *2019 International Conference on Asian Language Processing (IALP)*, pages 313–318, New York. IEEE.
- Zhou, S., Jeong, H., and Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 6:97–111.
- Ziafar, M. and Namaziandost, E. (2019). A formulaic approach to propositional density and readability. *International Journal of Innovation and Research in Educational Sciences*, 6(6):816–821.

- Ziegler, J. C., Perry, C., Jacobs, A. M., and Braun, M. (2001). Identical words are read differently in different languages. *Psychological science*, 12(5):379–384.
- Zouhair, T. (2021). *Automatic Speech Recognition for low-resource languages using Wav2Vec2: Modern Standard Arabic (MSA) as an example of a low-resource language*. Thesis, Dalarna University.
- Zulu, N. (2003). The literary history of Sesotho literature in social transition: A survey of prominent themes and trends after 1990. *South African Journal of African Languages*, 23(2):77–94.
- Zupon, A., Crew, E., and Ritchie, S. (2021). Text normalization for low-resource languages of Africa. *preprint*, arXiv:2103.15845.

Appendix A

Common Sesotho words

A				
a	abela	abele	abelwa	abo
afrika	aha	ahe	ahile	ahileng
ahlolela	ahlolele	ahlolwa	ai	aids
akasiya	akgela	ako	aletare	aletareng
amehang	amoha	amohela	amohele	amohelehang
ana	ananela	ananelwa	ananetse	ao
araba	arabe	arabela	areka	arola
atametse	aterese	athe	atisa	atleha
afe	afidavite	afidaviti	aforika	atlehileng
ahisaneng	ahlola	ahlolang	ahlole	atolosa
ajwa	akanya	akaretsa	akaretsang	aubuti
ama	amana	amanang	amang	atamela
amohlehleng	amohelwa	amohetse	amohetsweng	asiria
apara	aparela	apere	apesa	asa
arolelana				
B				
baa	baabi	baahi	baahlodi	baamehi
babatsehang	babedi	babelona	babo	babusi
badudi	badumedi	baegepeta	baemedi	baetapele
baflesita	bafu	bafung	bahalaledi	bahale
bahlanka	bahlankana	bahlokomedi	bahlolohadi	baholo
bajete	bajuda	baka	bakalatiya	bakanana
bakwang	bala	balake	balaodi	bale
balekgotla	balemi	baleng	balevi	balla
bamoabe	bana	banana	baneng	bang
bao	baokamedi	bapala	bapalang	baphedisuwa
baprista	bara	baradi	bararo	baratuwa
basadi	basebedisi	basebeletsi	basebetsi	bashane
bathopuwa	batjha	batla	batlang	batle
batsi	batswadi	be	bea	bee
beilwe	beilweng	beke	bekeng	bene
betere	bethele	betlelehema	betlilweng	beuweng
bili	bilione	bina	bitsa	bitsang
boahi	bobe	bobebe	bobedi	bobodu
boemedi	boemo	boemong	boetapele	boetse
bofuma	bofutsana	boha	bohahlaudi	bohale
bohano	bohlaswa	bohle	bohloko	bohlokwa
bohuwang	boikarabelo	boikemelo	boikemisetso	boiketlo
boiphihlelo	boipiletso	boiqapelo	boiteko	boithabiso
boitsebelo	boitsebiso	boitshepo	boitshwaro	bojalefa
bokella	bokelle	bokgabane	bokgabo	bokgoni

bolaile	bolailweng	bolaodi	bolawa	bolawe
bolele	bolelele	boleletse	bolella	bolellang
boleng	boletse	boletsweng	bolo	boloka
bolotsana	bolwetse	bomadimabe	bomasepala	bommasepala
bone	boneng	bong	bongaka	bongata
bonono	bontsha	bontshang	bontshe	bontshitse
bootswa	bopa	bopaki	bophahamo	bophara
bopilwe	bopilweng	boporofeta	boramesebetsi	boramosebetsi
bosiu	bosiung	bosupa	botala	botebo
bothopuweng	bothoto	botjha	botjhaba	botjhabela
botsa	botsamaisi	botshelela	botshepehi	botsitso
bue	bueng	buile	buileng	buisana
bula	busa	busang	buse	butle
baamohedi	baapostola	babadi	baballa	boya
bademona	badilweng	badisa	baditjhaba	buka
baeti	baetsadibe	bafani	bafarisi	buuwa
bahete	bahlabani	bahlahisi	bahlalefi	boyena
baholoholo	bahwebi	baiseraele	baithuti	bukana
bakang	bakeng	bakudi	bakwa	buuwang
balebedi	baleha	balehela	balekane	bua
balletsweng	balwa	balwang	bamidiane	bukantswe
bangata	banka	bankakarolo	banna	uang
baphetwa	bapisa	bapiswa	baporofeta	bukeng
bareki	barutuwa	barutwana	barwetsana	borapedi
basotho	bathahaselli	batho	bathong	botekgeniki
batona	batsamaisi	batshehetsi	batsho	botlalo
beha	behilweng	behwa	beile	boraro
beng	beno	beso	bete	boteng
bewa	bi	bile	bileng	botle
bitso	bitswa	bitswang	bo	borobedi
bodulo	bodumedi	boela	boele	bothata
bofe	bofelo	bofifi	bofokodi	botlokotsebe
bohareng	bohata	bohlahlobi	bohlale	borwa
bohlola	bohobe	boholo	bohohoholo	botho
boikgathollo	boikgethelo	boikwetliso	boima	boto
boithaopo	boithuto	boitlamo	boitokisetso	bophelo
boka	bokae	bokajeno	bokamoso	bophelong
bokgopo	bokgosi	bokolla	bokwe	bophirima
bolaya	bolaye	bolela	bolelang	bophirimela
bolelle	bolelwa	bolelwang	bolemi	boo
boloke	bolokehileng	bolokolohi	bolokwa	bonyane
bona	bonahala	bonahalang	bonang	bonwang

bonnete	bonngwe	bonokwane	bonolo	bonwa
C				
covid				
D				
damaseka	datha	demokerasi	demokrasi	di
dialetare	diaparo	diatla	dibaka	dibakeng
dibenefiti	dibini	dibolayalehola	dibopeho	dibopuwa
diesel	dieta	difahleho	difate	dife
dihlahiswa	dihlomo	dihlong	dihlooho	dihlopha
dijo	dijothollo	dijotholong	dikahare	dikahlolo
dikarabo	dikarete	dikariki	dikarolo	dikarolong
dikepe	dikeramo	dikerubime	dikete	diketsahalo
dikgetho	dikgoka	dikgokahano	dikgokahanyo	dikgomo
dikgwebo	dikgwedi	dikgweding	dikgwerekgwere	dikhampani
dikobong	dikoduwa	dikokwanyana	dikolo	dikoloi
dikopano	dikopanong	dikopo	dikotsi	dilemo
dimethara	dimilione	dimpho	dinaha	dinaheng
ding	dingaka	dingolwa	dinkgwa	dinku
dinonyana	dinotshi	dintho	dinthong	dintlha
dipaki	dipalangwang	dipale	dipalesa	dipallo
dipatlisiso	dipehelo	dipelo	dipelong	dipeo
diphateng	dipheleu	dipheo	diphepetso	diphephetso
dipholo	diphoofolo	diphooko	diphoso	diphuleng
dipolelo	dipolelong	dipolotiki	dipono	diporeisi
diprofense	diprojeke	diprojekte	diprovense	dipuisano
diqueto	diquha	dira	diranta	diraporoto
disebediswa	disenthimethara	ditaba	ditabatabelo	ditabeng
ditefello	ditefiso	ditefo	ditekanyetso	diteko
diterekere	diterompeta	dithaba	dithabeng	dithaere
ditherisano	dithethefatsi	ditho	ditholwana	dithoto
ditjeho	ditjeo	ditjhaba	ditjhelete	ditlaleho
ditlhahiso	ditlhahlobo	ditlhaku	ditlhaloso	ditlhapi
ditloholo	ditlolo	ditlwaelo	ditokelo	ditokiso
ditoropong	ditsamaiso	ditsebe	ditsebi	ditsebisio
ditshebetso	ditshehetso	ditshenyehelo	ditshila	ditshisinyo
ditsiane	ditsie	ditsietsi	ditumellano	diyuniti
dumedisa	dumela	dumelang	dumele	dumella
diabolosi	diahelo	diahelong	diakhaeve	dutseng
dibapadi	dibatana	dibe	dibeke	ditokomane
dibuka	didiba	dieha	diepollo	ditsele
difofu	diforomo	dihekthara	dihela	ditshiu
dihora	dihwai	dijalo	dijana	dubilweng

dikamano	dikamele	dikamore	dikano	dumellana
dikarolwana	dikedare	dikeledi	dikeletso	ditone
diketso	dikgato	dikgau	dikgeo	ditselelang
dikgoro	dikgosi	dikgothaletso	dikgutshwane	ditshiya
dikhamphani	dikhemikhale	dikhopi	dikilo	dula
dikolong	dikoloto	dikomiti	dikonyana	dumellane
dilemong	dilopotsiya	dimakatso	dimela	ditonki
dinako	dinaledi	dinatla	dinawa	ditsha
dinohe	dinoka	dinoko	dinomoro	ditshwantsho
dintlheng	dinyehelo	dinyewe	dipakeng	dulang
dipalo	dipalopalo	dipapadi	dipaterone	dumellwa
dipere	diperesente	dipersente	diphate	ditopo
diphetho	diphetoho	diphofu	dipholisi	ditsehebeletso
diphuphutso	diphuputso	dipina	dipohwana	ditso
diporofense	diporofensi	diporojeke	dipotso	dule
dipuisanong	dipula	dipuo	dipuong	dutse
diratswana	direkoto	disabole	disampole	ditlhapiso
ditaelo	ditakatso	ditataiso	ditau	ditlhokeho
ditema	ditemeng	ditemoso	diterateng	ditlhoko
ditharollo	dithebe	ditheko	ditheo	ditlhophiso
dithupa	dithuso	dithuto	dithutong	
ditlamorao	ditlatse	ditleleime	ditletlebo	
E				
e	eba	ebang	ebe	ebile
eitse	eitsi	eja	eka	ekaba
eketsehileng	eketswa	ela	ele	eleasare
elwa	ema	emang	eme	emela
eneji	eng	enngwe	ente	entse
eo	epela	epelwa	ere	erekaha
eshwa	esita	esitana	eso	esong
etlare	etsa	etsahala	etsahalang	etsahetse
etseditse	etseng	etsetsa	etsetsang	etsetse
edomo	ee	e	e	etsetswang
ekare	ekasitana	eketsa	eketseha	etswa
elellwa	eleng	eletsa	eli	etswang
emetse	emisa	empa	ena	etswe
entseng	entswe	entsweng	enwa	eyang
esale	esele	eseng	esere	eya
etela	eteletse	etella	etla	
etsahetseng	etsang	etse	etseditse	
F				
fa	faele	fafatsa	fahleho	fana

fanweng	fapafapaneng	fapana	fapane	fapaneng
fedisa	fedise	fediswa	feela	feiye
fepa	feshene	feta	fetang	fete
fetisetswa	fetisisa	fetiswa	fetoha	fetohe
fetolela	fetolwa	fifi	fihla	fihle
filweng	fodisa	fofang	foka	fokola
foreisetata	foromo	foromong	fosahetseng	freistata
fumaneha	fumanehang	fumanehe	fumaneng	fumantshwa
fuperweng	fuputsa	furaletse	furalla	furalle
fuwa	fuwang	fuwe	fuweng	fumanwa
fanang	fane	fanwa	fanwang	futsanehileng
fate	fatshe	fe	fedile	fumanwang
fela	feleng	feletseng	felletseng	futuha
fetile	fetileng	fetisa	fetisetsa	fumanwe
fetohela	fetohile	fetola	fetole	futuhela
fihlela	fihlella	fihlellwa	fihlile	fumanweng
fokolang	fokotsa	fola	foramo	futuhele
fula	fumana	fumanang	fumane	
G				
gauteng				
H				
ha	habedi	habo	habohloko	hae
haellwa	haesale	haese	haha	hahe
haholwanyane	hajwale	hakaakang	hakaalo	halalela
halelele	halofo	hamane	hamate	hammoho
hang	hangata	hanghang	hanngwe	hanong
hao	hapa	hape	hapile	hara
hasupa	hatakela	hatella	haufi	hauhela
heletsa	heno	heshebone	heso	hilekiya
hlaba	hlaha	hlahang	hlahella	hlahellang
hlahisitswe	hlahisitsweng	hlahiswa	hlahiswang	hlahlama
hlajwa	hlajwe	hlaka	hlakile	hlakileng
hlalose	hlalositswe	hlalositsweng	hlaloswa	hlaloswang
hlatswa	hle	hlile	hlohonolofatsa	hlohonolofatse
hlokahalang	hlokang	hloke	hlokeha	hlokehang
hlokomelang	hlokomele	hlokomelwa	hlokometse	hlokwang
hloma	hlompha	hlomphe	hlompho	hlooho
hlophiswa	hloya	hlwahlwa	hlwaya	hlweka
ho	hoba	hobane	hobaneng	hodima
hohela	hohle	hohlehohle	hoja	hokae
holang	hole	hona	hong	hoo
hore	hosasane	hoseng	howa	howeletsa

haeba	haebe	haella	haellang	hodimo
hahela	hahola	haholo	haholoholo	hokahana
halalelang	halefa	halefela	halefile	hopola
hamonate	hamorao	hampe	hana	hula
hantle	hantlentle	hanyane	hanyenyane	hodimodimo
hare	harepa	harese	hase	hokahaneng
hauhele	he	heborone	hekthara	hopolang
hira	hirame	hiv	hla	hwa
hlahisa	hlahisang	hlahise	hlahisitse	hodisa
hlahlamang	hlahloba	hlahlobo	hlahlojwa	hokahanya
hlakisa	hlakola	hlalosa	hlalosang	hopole
hlano	hlaola	hlapanya	hlapanyetsa	hohang
hloile	hloileng	hloka	hlokahala	hola
hlokehe	hloko	hlokolosi	hlokomela	hora
hlola	hlole	hloleha	hlollwa	
hloohong	hlopha	hlophisa	hlophisitsweng	
hlwekile	hlwekileng	hlwekisa	hlwekisitsweng	
I				
ikamahanya	ikarabellang	ikela	ikemetseng	ikemiseditse
ikonomi	ikopanya	ikopanye	ikutlwa	ikutlwahatsa
ilo	indasteri	indastering	inkela	inotshi
iphumana	ipiletsa	iponahatsa	ipotsa	isa
iteanya	ithorisa	ithuta	ithutileng	itihela
itshetlehile	itshetlehileng	itshetlehlwe	itshilafatsa	itima
ikemiseditseng	iketsetsa	ikgethang	ikgethileng	itse
ikwetlisa	ila	ile	ileng	itseng
inshoreense	inthanete	ipapisa	ipata	itshetleha
isang	ise	iseraele	iswa	
J				
ja	jakileng	jala	jalwa	jang
jesu	jewa	joko	judea	jwala
jwalokaha	jwang	jeriko	jwalojwalo	jwalo
jara	je	jwaloka	jerusalema	jwale
K				
ka	kabelo	kabinete	kabo	kadimo
kahisano	kahlolo	kaho	kahobane	kahodima
kakaretso	kalafo	kalaka	kalebe	kamano
kamorao	kamoreng	kanana	kananelo	kang
kantoro	kantorong	kaofela	kapa	kapela
karete	kariki	karolo	karolong	karolwana
katlasa	katlase	katleho	katoloso	ke
keletso	kelohlolo	kena	kenang	kene

kengwa	kenngwa	kenngwang	kentse	kentsweng
kenyeleditswe	kenyeletsa	kenyeletsang	kenyeletswa	kenyelletsa
kereiteng	kereiti	kereiting	kereke	keteka
kgabane	kgaello	kgafetsa	kgahlamelo	kgahlano
kgaitsedi	kgale	kgalefo	kgalemela	kgama
kgaolo	kgaolong	kgaotsa	kgaotse	kgapo
kgato	kgauta	kgauteng	kgeloha	kgelosa
kgethilweng	kgetho	kgethollo	kgethwa	kgiro
kgohlo	kgohlong	kgokahano	kgolo	kgomo
kgonang	kgone	kgonehang	kgopo	kgora
kgothaletsa	kgothaletso	kgothaletswa	kgothalletsa	kgothatsa
kgumamelang	kgutla	kgutle	kgutlela	kgutlele
kgutshwane	kgutshwanyane	kgutsufatso	kgwebisano	kgwebo
khabinete	khampane	khampani	khamphani	khansele
kila	kile	kileng	kobo	kodi
komello	komiti	konteraka	kontinente	konyana
kopaneng	kopanetsweng	kopang	kopano	kopanong
kora	koranteng	koro	kotara	kotareng
kotula	kotulo	kreste	kula	kuno
kwano	kwazulu	kwetliso	koporasi	kotsing
kae	kaha	kahara	kahare	kwana
kahodimo	kahohle	kahoo	kajeno	kopantsweng
kamehla	kamohelo	kamoo	kamora	kotjwa
kankere	kannete	kano	kantle	kutlwisiso
kapele	kappa	karabelo	karabo	kopanya
karolwaneng	katamelo	kateng	kathoko	kotlo
keke	keketseho	kelello	kelellong	kwahela
kenela	kenella	keneng	keng	kopo
kenya	kenyang	kenye	kenyeleditse	kotsi
kenyelletsang	kenyelletswa	kepa	kereite	kwala
ketsahalo	ketsamelao	ketsamolao	ketso	koloi
kgahlanong	kgahlanyetsa	kgahlisa	kgahlwa	kopanelo
kgang	kganna	kganya	kganyapa	kolobe
kgatello	kgatelopele	kgathala	kgatiso	kopane
kgema	kgetha	kgethehileng	kgethile	kokwanahloko
kgobokanya	kgodisang	kgoeletso	kgohlano	kopana
kgona	kgonahala	kgonahalang	kgonahalo	koduwa
kgoro	kgorong	kgosana	kgosi	kopa
kgotso	kgotsofatsa	kgotsofatsang	kgumamela	khoutu
kgutlisa	kgutlisetsa	kgutlisetse	kgutsa	khopi
kgwebong	kgwedi	kgweding	kgwele	
kharikhulamo	khomishene			

L				
la	laela	laelang	laele	laesense
lakatsa	lang	laola	laolang	laolwa
lata	latela	latelang	latelwa	le
lebaka	lebala	lebaleng	lebatowa	lebatoweng
lebella	lebelletswe	lebelletsweng	lebelo	lebese
lebitso	lebitsong	leboha	lebokose	lebokoseng
leeme	leeto	leetong	lefa	lefapha
lefeela	lefella	lefifi	lefifing	lefika
lehae	lehlabula	lehlaka	lehlakore	lehlakoreng
leholo	lehong	lehwatata	lehwatateng	leihlo
lekana	lekanang	lekane	lekaneng	lekanya
lekgatheng	lekgetho	lekgoba	lekgolo	lekgotla
lekotswe	lelapa	leleka	leleme	leloko
lena	lenaka	lenane	lenaneng	lenaneo
lengolo	lentswe	lenyalo	leo	leoto
leraba	lerako	lerata	lerato	lere
leseae	lesedi	lesela	leshano	leshome
letho	lethose	lethule	letjha	letlalo
letsa	letsatsi	letsatsing	letsetsa	letsha
letshwelehad	letsibolo	letsoho	letsohong	letswai
lewatleng	lla	llela	loetse	lohle
lokelang	lokelwa	lokelwang	loketse	loketseng
lokiswa	lokolla	lomoswang	lona	lopolla
laetse	laetseng	lahla	lahlehelwa	lwantsha
laolwang	lapa	lapeng	lapile	lohothwang
lea	leaho	leano	leba	lokile
lebela	lebeletse	lebeletswe	lebeletsweng	lopolle
lebisa	lebisang	lebitla	lebitleng	loka
lebone	lebopong	lebotho	leboya	lokileng
lefapheng	lefatshe	lefatsheng	lefe	lota
lefu	lefung	lefuwa	leha	lokan
lehlohonolo	lehodimo	lehodimong	lehola	lokisa
lejwe	leka	lekala	lekaleng	lotho
leke	lekeno	lekgabunyane	lekgathe	lokela
lekgotleng	lekola	lekolwa	lekoteng	loketsetsa
lema	leme	lemo	lemoha	lwana
lenaneong	lene	leng	lengelo	letswalo
lepera	leqephe	leqepheng	leqosa	levi
lerole	leru	lerumo	leruo	lewa
leshwa	lesira	lesupi	lethathama	lewatle

letlole	letloleng	letona	letoto	
letshehadi	letsholo	letshwao	letshwele	
M				
ma	maaforika	maafrika	maafrika	maano
mabatoweng	mabedi	mabenkeleng	mabitla	mabitso
madimabe	madinyane	maemo	maemong	madungwadungwana
mafeela	mafelong	mafisa	mafu	mafura
mahareng	mahe	mahlaahlela	mahlabelo	mahlakore
mahlatsipa	mahlo	mahloko	mahlomola	mahlong
mahosana	mahwashe	maikemisetso	maikutlo	majoro
makgetha	makgethe	makgetlo	makgoba	makgolo
malakabe	malapa	malapeng	malebana	maleme
mamele	mamella	mamello	manaka	mananeo
manganga	mangata	mangeloi	mangolo	mano
manyolo	maoto	maotong	maphelo	mapheo
maqheku	maqosa	marako	maralla	marangrang
maruo	masalla	masapo	masea	masela
masimong	masupi	maswabi	matamo	matha
matjha	matjhaba	matjhabeng	matla	matlafala
matlo	matlole	matlotlo	matlung	matlwana
matsete	matshwafo	matshwao	matshwele	matshwenyeho
mawa	meaho	mebala	mebaraka	mebila
medumo	meedi	mefeng	mefuta	mehala
mehlano	mehlape	mehlang	mehlodi	mehloding
mekete	mekga	mekgahlelo	mekgatlo	mekgatlong
melaetsa	melao	melata	melawana	melemo
meng	mengata	meno	menontsha	menyako
merafong	meralo	merari	meraro	merero
mesebetsing	metha	metheo	metjha	metjhine
metse	metseng	metsi	metsing	metso
meutlwa	meya	milione	miliyone	million
mmalwa	mmamela	mmao	mmaapa	mmaraka
mmele	mmeleng	mmesa	mmetse	mmino
mmona	mmopi	mmotsa	mmusi	mmusisi
moahlodi	mobu	mobung	modimo	modisa
moedi	moelelo	moemedi	moena	moetapele
mofumahadi	mofuta	mofutsana	mohala	mohato
mohlabani	mohlahisi	mohlahlobi	mohlahlobuwa	mohlala
mohlodi	mohllokomedi	mohlolohadi	mohlomong	mohlwaare
mohwang	moifo	moikopedi	moithuti	mojaki
mokga	mokgabiso	mokgahlelo	mokgahlelong	mokgatlo
mokgweng	mokopi	mokubetso	mola	mokgwatshebetso

molaong	molaotseo	molato	molawana	molebedi
molemo	molemong	moleng	mollo	mollong
momahano	mona	monahano	monate	mong
mongodi	mongodisi	mongolo	monko	monna
monyane	monyenyane	monyetla	moo	moofisiri
mopholosi	moporesidente	moporofeta	mopresidente	moprista
mora	morabe	morabeng	moradi	moralo
moratuwa	moreki	morena	morero	morethetho
moruong	moruti	morutuwa	morwetsana	mosa
mosebeletsi	mosebetsi	mosebetsing	moshanyana	moshemane
mothamo	motheho	motheo	motho	motjha
motlotlo	motlotsuwa	motsamaisi	motsamao	motse
motso	motswadi	motswako	motswalle	moya
mphe	mpho	mpolaya	mphalane	motshwayi
mabaka	mabaleng	mabapi	mabatowa	moriana
mabokelle	mabone	mabotho	madi	mosadi
maetsi	mafa	mafapha	mafapheng	moshwelella
mafutsana	mahae	mahaeng	mahala	motjheso
mahlakoreng	mahlaku	mahlale	mahlano	motseng
mahodimo	mahodimong	mahola	maholo	moyeng
majwe	makala	makgabane	makgabunyane	moriri
makgopo	makgotla	makgulo	makoloi	mosadimoholo
malwetse	mamati	mamela	mamelang	moshweshwe
mananeong	mane	mancha	mang	motlakase
mantlha	mantsiboya	mantswe	manyala	motsheanong
mapolasi	mapolasing	mapolesa	maqephe	mpa
mararo	mareo	mariha	maru	morui
masepala	mashano	mashome	masimo	mose
mathata	mathateng	matheba	matitjhere	mosing
matlafatsa	matlafatso	matlapa	matle	motlatsi
matona	matsapa	matsatsi	matsatsing	motsheare
matsibolo	matsoho	matsohong	mpe	matshwenyehong
mebileng	mebuso	medimo	medingwana	moruo
mehato	mehla	mehlala	mehlaleng	mosebedisi
mehlolo	meholo	mehoo	mehopolo	mosireletsi
mekgwa	mekutu	mela	motle	mekgwatshebetso
meloko	melokong	melomo	mene	mookamedi
menyetla	meputso	merabe	merafo	moputso
meriana	meroho	meru	mesebetsi	moralong
metjhini	metle	metsamao	metsana	mookotaba
metsotso	metsu	metswako	metswalle	moqhaka
mma	mmadi	mmae	mmala	morao

mmasepala	mme	mmea	mmedi	mopeleto
mmoho	mmolai	mmolaya	mmolaye	moqoqo
mmuso	mmusong	mo	moahi	moraorao
modise	moditjhaba	modulasetulo	modumo	mophetwa
mofapahlooho	mofepedi	mofilesita	mofu	moqosuwa
mohatong	mohatsa	mohau	mohla	morara
mohlang	mohlanka	mohlankana	mohlape	monongwaha
moholo	moholwane	mohope	mohopolo	monwana
mojalefa	mojuda	mokete	moketjana	monyako
mokgeng	mokgetho	mokgosi	mokgwa	monyakong
molaetsa	molamu	molao	molaodi	mongobo
molefi	molekane	molekgotla	molemi	monghadi
moloko	molomo	molomong	molora	
monga	mongata			
N				
na	naha	nahana	nahanang	naheng
nama	namane	namolela	nanaola	nang
nee	neela	neelana	neelane	neha
neileng	neilweng	neng	nepahalo	nepahetse
ngaka	ngata	ngatanyana	ngodisa	ngodisitswe
ngole	ngoleng	ngolla	ngolwa	ngolwang
ngwanabo	ngwanaka	ngwanana	ngwaneno	ninife
nkgauhele	nkgono	nkile	nku	nkuwa
nnee	nnete	nnetefatsa	nngwe	nnile
nomoro	nonyana	nosetsa	nowe	nqa
ntatao	ntate	ntho	nthomileng	ntjha
ntle	ntlha	ntlheng	ntlo	nto
ntsha	ntshe	ntshetsa	ntshetsapele	ntshetsopele
ntwa	ntweng	nwa	nwe	nyakalla
nyedisa	nyefola	nyehela	nyehella	nyehelo
nako	nakong	nakwana	nalane	nyakalle
natal	ne	nea	neang	nyenyane
nehelana	nehelano	nehwa	neile	nyala
nepahetseng	nepo	netefatsa	newa	nyewe
ngodisitsweng	ngodiso	ngodiswa	ngola	nyalanya
ngotswe	ngotsweng	ngwahola	ngwana	nyoloha
nka	nkang	nke	nkelwa	nyane
nkwa	nna	nnang	nne	nyolohela
no	noha	noka	nokeng	ntsu
nqalong	nqane	ntata	ntatae	ntshitseng
ntjhafatso	ntlafatsa	ntlafatso	ntlafetseng	ntshitse
ntsa	ntsane	ntse	ntseng	ntshetsopeleng

O				
o	ofe	ofisi	ofising	ohle
ona	oo	osele	otla	otle
oho	ole	oli	ommeng	otlolohileng
otlolla	otloloha			
P				
paballo	padi	paka	pakeng	paki
paleng	palesa	palo	palohare	pampiri
patsi	pedi	pehelo	pejana	pela
peo	pere	perese	pesalema	phaello
phahello	phakisa	phaladisa	phallo	phano
phatlalatsa	phatlalatse	phatlalatso	phatlalatswa	phato
phelang	phele	pheletso	pheleu	phemiti
pheta	phetapheto	phetela	phetha	phethahala
phethe	phethehileng	phethela	phethisa	phethola
phihlelo	phirimana	phofa	phoka	phokotso
pholose	pholoswa	phomola	phomolo	phonyoha
phuleng	phumantsho	phupu	phuputso	phutha
polao	polasi	polasing	polelo	polelong
pontsheng	pontsho	poone	popeho	poraefete
porovense	poso	potapota	potapotileng	potlakileng
potsong	praevelte	profeshenale	projeke	projekte
puo	puong	puso	pula	potlakisa
palame	palamente	palamenteng	pale	provense
papadi	papatso	paseka	pata	potlako
pele	pelo	pelong	penshene	provenseng
phahama	phahame	phahameng	phahamisa	potoloha
phapang	phapantsho	phaposi	phaposing	puisano
phedisa	phedisano	pheha	phela	potso
phepelo	phephetso	phepo	pherekong	porojeke
phethahale	phethahatsa	phethahatso	phethahetseng	porofeta
phetoho	phetolelo	phetse	phihlello	porofense
pholisi	pholo	pholoho	pholosa	poraevete
phoofolo	phooko	phoso	photo	pono
phutheho	plantere	pohwana	pokello	poloko
polokeho	polokelo			
Q				
qadile	qaka	qala	qalang	qale
qetelle	qetellong	qetelo	qetile	qeto
qosa	qotso	qhobosheane	qoba	qolla
qalong	qeta	qetela	qetella	qoqa

R				
ra	rala	ralewa	ralwa	rama
rapella	rapellwang	raporoto	rarolla	rata
rea	realo	rejisetara	rejistara	reka
rephaboliki	rera	reretswe	reretsweng	rerile
robala	robedi	robetse	robong	rohakehe
rona	rorisa	rorise	rubene	ruri
ramosebetsi	rapela	rapelang	rapele	roma
ratana	ratang	rate	re	ruta
rekisa	rekota	rekoto	reng	romela
rialo	riphaboliki	ripitla	roba	rute
romelwa	rometse	ruthe	rutwa	
S				
sa	sabatha	sabole	saena	saense
se	seabo	seaparo	seatla	sebaka
sebedisa	sebedisana	sebedisang	sebedise	sebedisetswa
sebediswang	sebediswe	sebele	sebeletsa	sebeletsang
sebetsanang	sebetsang	sebetse	sebetseng	sebetswa
sebopehong	sebopi	sebui	sedi	sediba
sefafatsi	sefahleho	sefate	sefateng	sefe
sehalalelo	sehalalelong	sehalo	sehla	sehlabelo
sehlohong	sehlomathiso	sehlooho	sehloohong	sehlopha
sejwalejwale	seka	sekaseka	seke	sekema
sekgeo	sekgobo	sekoboto	sekolo	sekolong
selekane	selelekela	selemo	selemong	sello
sena	seng	sengolwa	seno	senwelo
seotswa	sepetlele	sepha	sephéo	sephetho
seqha	sera	seratswana	seriti	sesa
setaele	setatemente	setereke	sethatho	setheo
setjha	setjhaba	setjhabeng	setjheso	setle
setseng	setsha	setshabelo	setshwantsho	setshwantshong
sewelo	sheba	shebahala	shebana	shebane
shejwa	shwa	shwe	shwela	shwele
simolla	sinai	sione	sireletsa	sisingwang
sitwa	siya	siyo	so	sodoma
soya	supa	supang	supileng	shebaneng
sala	samaria	saruri	satane	shweleng
sebakeng	sebapadi	sebatana	sebe	sitelwa
sebedisitse	sebedisitswe	sebedisitsweng	sebediswa	sohle
sebeletse	sebeta	sebeta	sebetsana	shebella
seboka	sebokeng	sebolelo	sebopeho	sidone

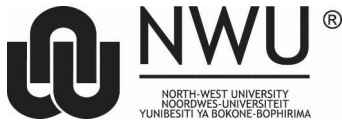
sedibeng	sedika	sedikadikwe	sedulwana	sitetswe
sefelo	sefela	sefuba	seha	soma
sehlahiswa	sehleng	sehlohlolong	sehloho	shebile
sehlopheng	seholo	sehwai	sejothollo	silafatsa
sekepe	sekepeng	sekete	sekgahla	sitilwe
sekoloto	sekotlolo	sekwahelo	selehae	sona
semela	semelo	semmuso	semolao	shebisisa
senya	senyesemane	seo	seotlwana	silefera
sephiri	sepikara	sepolesa	sepolotiki	sitisa
sesebediswa	sesila	sesosa	sesotho	soneblomo
setho	sethoto	setifikeiti	setipe	
setloholo	setopo	setse	setsebi	
setsi	setso	setswalle	sewa	
T				
taba	tabeng	tadima	tadime	tadimile
tamare	taolo	taolong	tataisa	tataiso
tebello	tebileng	teboho	tefello	tefelo
tekanyo	tekatekano	teko	tekolo	telele
temo	temong	tempele	tempeleng	teng
teroneng	thaba	thabang	thabe	thabela
thaere	thahasello	thapedi	thapelo	tharo
thebe	theepa	theha	thehilweng	thehwa
theko	thekolohelo	thelevishene	theoha	theohela
thibella	thibelo	thobalano	thohako	thoholetsa
thonya	thopa	thopuwa	thoriso	thothokiso
thupello	thupelo	thusa	thusang	thuse
thutong	thwe	tiileng	tiisa	tiisang
timelo	timetsa	timetse	timetswa	timetswe
tjhadimo	tjhaelletse	tjhankana	tjhankaneng	tjhe
tjhesang	tjhese	tjheseho	tjhesetsa	tjheswa
tlama	tlamang	tlameha	tlamehang	tlamehile
tlatsetso	tlatsitsweng	tlatswa	tle	tlelaemete
tlhahiso	tlhahisoleseding	tlhahisong	tlhahlobo	tlhaho
tlhodisano	tlhokeho	tlhoko	tlhokomelo	tlholo
tlilo	tlisa	tlise	tlisetsa	tlisetse
tlohela	tlokoma	tlola	tlolo	tloo
tlotlehang	tlotlisa	tlotlontswe	lotsa	lotse
to	tobana	toka	tokelo	tokiso
tonki	topollo	tora	toro	totobatsa
tsamaisana	tsamaisang	tsamaiso	tsamaisong	tsamaiswa
tse	tseba	tsebang	tsebe	tsebeng
tsebo	tsee	tsejwa	tsejwang	tseko

tsepanisa	tsepaniso	tsetela	tshaba	tshabang
tshebedisano	tshebediso	tshebedisong	tshebeletso	tshebeletsong
tshehetsa	tshehetsang	tshehetso	tshehetswa	tshekamelo
tsheletseng	tshenyo	tshepa	tshepahala	tshepahalang
tshimo	tshimolohong	tshimong	tshireletsa	tshireletso
tsholla	tshosa	tshusumetso	tshwaetso	tshwana
tshwanelehang	tshwanelehileng	tshwanelo	tshwanetse	tshwanetseng
tshwarele	tshwarellang	tshwarelo	tshwarwa	tshwaya
tsitsitseng	tsofetse	tsoha	tsohe	tsohle
tsuba	tswa	tswala	tswalla	tswalo
tswekere	tswela	tswelang	tswelapele	tswele
tswetsweng	tswile	tswileng	tukang	tumellano
taelo	tafole	tahlehelo	taka	tshwantshiso
tatelano	tatellano	tau	tb	tshwere
tefiso	tefo	tekano	tekanyetso	tsoho
tema	temana	temaneng	temeng	tswalwa
tente	tenteng	terekere	terone	tswella
thabeng	thabile	thabisa	thabo	tumello
tharollo	thata	thato	the	tshwara
theilweng	thekenoloji	thekiso	theknoloji	tshwereng
theola	thepa	thetsa	thibela	tsona
thoko	thola	tholwana	thontsweng	tswalwe
thothomela	thoto	thumisehileng	thupa	tswellang
thuso	thuswa	thutapuo	thuto	tumelo
tiisetso	tikoloho	tikolohong	timela	tshwarehang
titjhere	tjee	tjena	tjhaba	tshweu
tjhefo	tjhelete	tjhentjha	tjhesa	tsosa
tla	tlala	tlaleha	tlaleho	tswang
tlang	tlasa	tlase	tlatsa	tswelopele
tletlebo	tletse	tletseng	tlhabollo	tumelong
tlhaku	tlhaloso	tlhaselo	tlhekefetso	tshwarela
tlhompho	tlhophiso	tlhophollo	tlile	tsietsi
tliswa	tlo	tloha	tlohe	tsoseletso
tloong	tlosa	tlosang	tlotla	tswe
tlung	tlwaelehang	tlwaelehileng	tlwaelo	tswetseng
tokoloho	tokomane	tona	tonakgolo	tv
tsa	tsamaelana	tsamaelanang	tsamaisa	tshitwe
tsamaya	tsamayang	tsamaye	tsatsi	tshwanang
tsebile	tsebisa	tsebisio	tsebiswa	tshoha
tsela	tseleng	tsena	tseo	tshwane
tshabe	tshabehang	tshabela	tshabo	tshohanyetso
tshebetso	tshebetsong	tshedimosetso	tshedisa	tshwanela

tshekatsheko	tshela	tshele	tshelela	tshohla
tshepe	tshepile	tshepiso	tshepo	tshwanelang
U–Z				
utlwa	utlwahala	utlwahalang	utlwahale	utlwang
utlwisisa	utlwisise	utlwe	utlwela	utlwile
utlwileng	veine	vaal	wona	wohle
wa	wane	wela	wele	wena
ya	yaba	yang	yare	ye
yuniti	yena	yo	yohle	yona

Appendix B

Ethics approval letters of study



Private Bag X1290, Potchefstroom
South Africa 2520

Tel: 018 299-1111/2222
Fax: 018 299-4910
Web: <http://www.nwu.ac.za>

Senate Committee for Research Ethics
Tel: 018 299-4849
Email: nkosinathi.machine@nwu.ac.za

12 October 2021

ETHICS APPROVAL LETTER OF STUDY

Based on the verification of the NWU's Ethics Committee for Language Matters (ECLM), your study is **approved** as indicated below. This implies that the North-West University Senate Committee for Research Ethics (NWU-SCRE) grants its permission that, provided the special conditions specified below are met and pending any other authorisation that may be necessary, the study may be initiated, using the ethics number below.

Study title: Measuring text readability in Sesotho																																
Student (applicant): Sibeko J (22260390)																																
Supervisors: Professor M Van Zaanen																																
Ethics number:	<table border="1"> <tr> <td>N</td><td>W</td><td>U</td><td>-</td><td>0</td><td>0</td><td>7</td><td>2</td><td>9</td><td>-</td><td>2</td><td>1</td><td>-</td><td>A</td><td>7</td> </tr> <tr> <td colspan="3">Institution</td> <td></td> <td colspan="5">Study Number</td> <td></td> <td colspan="2">Year</td> <td colspan="2">Status</td> </tr> </table>			N	W	U	-	0	0	7	2	9	-	2	1	-	A	7	Institution				Study Number						Year		Status	
N	W	U	-	0	0	7	2	9	-	2	1	-	A	7																		
Institution				Study Number						Year		Status																				
<u>Status:</u> S = Submission; R = Re-Submission; P = Provisional Authorisation; A = Authorisation																																
Application Type: Single Study																																
Commencement date: 12/10/2021		Risk Category: No risk																														
Expiry date: 11/10/2022																																
Approval of this no risk text-based study of reading texts in the public domain (website of the Department of Basic Education) and copyrighted material (such as textbooks, provided with permission and rules for use in the study by the publishers), with no additional human participation, is provided for a year, after which continuation of the study is dependent on receipt and review of the annual (or as otherwise stipulated) monitoring report and the concomitant issuing of a letter of continuation.																																

Special in process conditions of the research for approval (if applicable):

No additional conditions.

General conditions:

While this ethics approval is subject to all declarations, undertakings and agreements incorporated and signed in the application form, the following general terms and conditions will apply:

- *The study leader/supervisor (principle investigator)/researcher must report in the prescribed format to the ECLM-REC:*
 - *annually (or as otherwise requested) on the monitoring of the study, whereby a letter of continuation will be provided, and upon completion of the study; and*
 - *without any delay in case of any adverse event or incident (or any matter that interrupts sound ethical principles) during the course of the study.*
- *The approval applies strictly to the proposal as stipulated in the application form. Should any amendments to the proposal be deemed necessary during the course of the study, the study leader/researcher must apply for approval of these amendments at the ECLM-REC, prior to implementation. Should there be any deviations from the study proposal without the necessary approval of such amendments, the ethics approval is immediately and automatically forfeited.*
- *Annually a number of studies may be randomly selected for an external audit.*
- *The date of approval indicates the first date that the study may be started.*
- *In the interest of ethical responsibility, the NWU-SCRE and ECLM-REC reserves the right to:*
 - *request access to any information or data at any time during the course or after completion of the study;*
 - *to ask further questions, seek additional information, require further modification or monitor the conduct of your research or the informed consent process;*
 - *withdraw or postpone approval if:*
 - *any unethical principles or practices of the study are revealed or suspected;*
 - *it becomes apparent that any relevant information was withheld from the ECLM-REC or that information has been false or misrepresented;*
 - *submission of the annual (or otherwise stipulated) monitoring report, the required amendments, or reporting of adverse events or incidents was not done in a timely manner and accurately; and / or*
 - *new institutional rules, national legislation or international conventions deem it necessary.*
- *ECLM-REC can be contacted for further information or any report templates via Theresa.Louw@nwu.ac.za or 018 910 3442.*

The ECLM-REC would like to remain at your service as scientist and researcher, and wishes you well with your study. Please do not hesitate to contact the ECLM-REC or the NWU-SCRE for any further enquiries or requests for assistance.

Yours sincerely



Prof A.S. Coetzee-Van Rooy
Chairperson NWU ECLM-REC

Original details: (22351930) C:\Users\22351930\Desktop\ETHICS APPROVAL LETTER OF STUDY.docm
8 November 2018

File reference: 9.1.5.4.2



Private Bag X1290, Potchefstroom
South Africa 2520

Tel: 018 299-1111/2222
Fax: 018 299-4910
Web: <http://www.nwu.ac.za>

Senate Committee for Research Ethics
Tel: 018 299-4849
Email: Feziwe.mseleni@nwu.ac.za

26 October 2022

ETHICS APPROVAL LETTER OF STUDY: LETTER OF CONTINUATION

As stipulated by the Standard Operating Procedures (SOPs) (items 8.6 and 8.8), requests for extensions of approved projects are made on the Annual Monitoring form received and reviewed by the chairperson of the ECLM (NWU Ethics Committee for Language Matters). Based on the verification of the chairperson of the ECLM of the Annual Monitoring report received from the applicant on 24 October 2022, the ECLM chair approves the extension of the research project for the period documented below. This implies that the North-West University Senate Committee for Research Ethics (NWU-SCRE) grants its permission that, provided the special conditions specified below are met and pending any other authorisation that may be necessary, the ethics approval for the study is extended and the study may continue, using the ethics number below.

Study title: Measuring text readability in Sesotho

Student/Researcher (Principal Investigator): Sibeko J (22260390)

Study Leader/Supervisor: Professor Menno van Zaanen

Ethics number:

N	W	U	-	0	0	7	2	9	-	2	1	-	A	7
Institution			Study Number						Year		Status			

Status: S = Submission; R = Re-Submission; P = Provisional Authorisation; A = Authorisation

Application Type: Single

Commencement date: 26 October 2022

Risk Category:

No risk, Text-based

Expiry date: 31 December 2023

Approval of the extension of this no risk text-based study of reading texts in the public domain (website of the Department of Basic Education) and copyrighted material (such as textbooks, provided with permission and rules for use in the study by the publishers), with no additional human participation, is provided for the period stipulated above, after which continuation of the study is dependent on receipt and review of the annual (or as otherwise stipulated) monitoring report and the concomitant issuing of a letter of continuation.

Special in process conditions of the research for approval (if applicable):

- The ECLM chairperson verified that the study remains a no risk text-based study with no human participation.

- The study progressed well and several publications have already been made.
- The student aims to complete the study by the end of 2023.

General conditions:

While this ethics approval is subject to all declarations, undertakings and agreements incorporated and signed in the application form, the following general terms and conditions will apply:

- *The study leader/supervisor (principle investigator)/researcher must report in the prescribed format to the ECLM-REC:*
 - *annually (or as otherwise requested) on the monitoring of the study, whereby a letter of continuation will be provided, and upon completion of the study; and*
 - *without any delay in case of any adverse event or incident (or any matter that interrupts sound ethical principles) during the course of the study.*
- *The approval applies strictly to the proposal as stipulated in the application form. Should any amendments to the proposal be deemed necessary during the course of the study, the study leader/researcher must apply for approval of these amendments at the ECLM-REC, prior to implementation. Should there be any deviations from the study proposal without the necessary approval of such amendments, the ethics approval is immediately and automatically forfeited.*
- *Annually a number of studies may be randomly selected for an external audit.*
- *The date of approval indicates the first date that the study may be started.*
- *In the interest of ethical responsibility, the NWU-SCRE and ECLM-REC reserves the right to:*
 - *request access to any information or data at any time during the course or after completion of the study;*
 - *to ask further questions, seek additional information, require further modification or monitor the conduct of your research or the informed consent process;*
 - *withdraw or postpone approval if:*
 - *any unethical principles or practices of the study are revealed or suspected;*
 - *it becomes apparent that any relevant information was withheld from the ECLM-REC or that information has been false or misrepresented;*
 - *submission of the annual (or otherwise stipulated) monitoring report, the required amendments, or reporting of adverse events or incidents was not done in a timely manner and accurately; and / or*
 - *new institutional rules, national legislation or international conventions deem it necessary.*
- *ECLM-REC can be contacted for further information or any report templates via susan.coetzevanrooy@nwu.ac.za or 016 910 3442.*

The ECLM-REC would like to remain at your service as scientist and researcher, and wishes you well with your study. Please do not hesitate to contact the ECLM-REC or the NWU-SCRE for any further enquiries or requests for assistance.

Yours sincerely



Prof A.S. Coetzee-Van Rooy

Chairperson NWU Research Ethics Committee for Language Matters

Appendix C

Ghent University invitation

Johannes Sibeko
Heron Street 6
6070 Port Elizabeth
South Africa

DEPARTMENT OF EDUCATIONAL POLICY
REGISTRAR'S OFFICE

E international@ugent.be

Het Pand
Onderbergen 1
9000 Gent

www.ugent.be/en

LETTER OF INVITATION

GHENT UNIVERSITY - UNIVERSITEIT GENT

Academic Year 2022/23

Name of the student:	Sibeko, Johannes	Nationality:	South African
Date & place of birth:	20/08/1990, Frankfort (South Africa)		
Home University:	North-West University (Potchefstroom, South Africa)		
Period of stay in Ghent:	from 13/10/2022 until 30/11/2022		
Exchange program:	Other resources		
Faculty at Ghent University:	Faculty of Arts and Philosophy (<i>Faculteit Letteren en Wijsbegeerte</i>)		
Faculty contact person:	Focquaert, Carine (international.lw@ugent.be)		
Prospect number:	0328002		

We declare that the above mentioned student is registered as an exchange student at Ghent University. Ghent University is an institution recognized, organised and subsidized by the Belgian Government (*Ministerie van de Vlaamse Gemeenschap, Departement Onderwijs*). A work permit is not needed for a stay as exchange student at Ghent University.

Sincerely,



Frederik De Decker
Head International Relations Office



The institutional number for Ghent University (UGent) is 110114.
This number has been attributed by the Flemish Ministry of Education and Training.

Check the authenticity on <http://attesting.Ugent.be/>
with code c24c9-e19e6-35bf5-d8363

Printed on: 29/08/2022

Page 1/1

Johannes Sibeko
Heron Street 6
6070 Port Elizabeth
South Africa

DEPARTMENT OF EDUCATIONAL POLICY
REGISTRAR'S OFFICE

E international@ugent.be

Het Pand
Onderbergen 1
9000 Gent

www.ugent.be/en

LETTER OF INVITATION

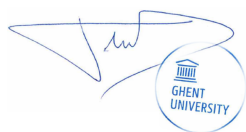
GHENT UNIVERSITY - UNIVERSITEIT GENT

Academic Year 2022/23

Name of the student:	Sibeko, Johannes	Nationality:	South African
Date & place of birth:	20/08/1990, Frankfort (South Africa)		
Home University:	North-West University (Potchefstroom, South Africa)		
Period of stay in Ghent:	from 30/01/2023 until 03/03/2023		
Exchange program:	Other resources		
Faculty at Ghent University:	Faculty of Arts and Philosophy (<i>Faculteit Letteren en Wijsbegeerte</i>)		
Faculty contact person:	Focquaert, Carine (international.lw@ugent.be)		
Prospect number:	0328002		

We declare that the above mentioned student is registered as an exchange student at Ghent University. Ghent University is an institution recognized, organised and subsidized by the Belgian Government (*Ministerie van de Vlaamse Gemeenschap, Departement Onderwijs*). A work permit is not needed for a stay as exchange student at Ghent University.

Sincerely,



Frederik De Decker
Head International Relations Office



The institutional number for Ghent University (UGent) is 110114.
This number has been attributed by the Flemish Ministry of Education and Training.

Check the authenticity on <http://attesting.Ugent.be/>
with code 19b30-b90d6-61lac-d7353

Printed on: 13/10/2022

Page 1/1

Appendix D

Permissions to use co-authored publications

Mmasibidi Setaka
Private bag X600, Potchefstroom,
North-West, 2520
mmasibidi.setaka@nwu.ac.za
2023 May 19

Johannes Sibeko
University Way, Summerstrand
Johannes.sibeko@mandela.ac.za
Eastern Cape, 6070

Permission to Include Co-Authored Article in Thesis

Dear Johannes Sibeko,

I hope this email finds you well. Thank you for reaching out and seeking permission to include our co-authored article, titled *An Overview of Sesotho BLARK Content*, in your thesis. I have carefully considered your request and am pleased to grant you permission to include our article in your thesis.

I believe that incorporating our co-authored article will enhance the depth and quality of your research, as it holds valuable insights and contributes significantly to the topic. I appreciate your assurance regarding the proper citations and references, and I trust that you will adhere to the academic conventions and guidelines while acknowledging our joint contribution.

Please proceed with including the article in your thesis, and I am glad to be a part of your academic endeavour. I wish you the best of luck with your thesis and look forward to seeing the final outcome. Kindly keep me informed of any updates or developments regarding the publication of your thesis.

Thank you for your dedication and collaboration. I am confident that your thesis will be a valuable contribution to the field.

Best regards,
Ms Mmasibidi Setaka



Signature

END OF THESIS