Investigating the use of articles, modal verbs and selected discourse markers in Zimbabwean English: A corpus-based analysis using the dynamic model

By

Faith Chiedza Chapwanya

Submitted in fulfilment of the requirements for the degree

Doctor of Philosophy in Linguistics

in the Faculty of Humanities,

UNIVERSITY OF PRETORIA

Supervisor: Dr J.H. Nel

March 2022
Declaration

I hereby declare that the thesis submitted for the degree Doctor of Philosophy in Linguistics, at the University of Pretoria, is my own work and has not previously been submitted to any other institution of higher education. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of bibliography.

Faith Chiedza Chapwanya

March 2022
Ethics statement

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that she has observed the ethics standards required in terms of the University of Pretoria’s Code of ethics for researchers and the Policy guidelines for responsible research.
Acknowledgements

Praise be to the Almighty God for giving me the strength and courage to pursue this project until the end, against all odds.

I am forever grateful for the assistance that I got from various individuals and institutions. Thank you to everyone mentioned below for contributing to the success of this project.

Firstly, thank you to my supervisor Dr. Joanine Nel for guiding me during this journey from the start of the project until its completion. Thank you for your patience and for the valuable feedback and explanations which enabled me to understand concepts better. Thank you for the moral support, especially in 2019 when it all seemed too much for me. Baie dankie!

This project would not have succeeded without input from anonymous participants who provided data for the study. Thank you all for setting your precious time aside to advance the frontiers of knowledge. I also want to thank the editor of Zimpapers for permitting me to access the Herald and Sunday Mail newspapers online.

I thank North West University, especially Prof. Bertus Van Rooy, for granting me permission to use part of Dr. Marungudzi’s corpus. Many thanks also go to the University of Zimbabwe for allowing me to do research at their institution.

To my amazing sisters Eunice and Theresa, I appreciate the trips you took to come and help me take care of the kids. If it wasn’t for your assistance, I would not have managed at all. Thank you for reassuring me when it all seemed too much for me. You were always there to listen to my concerns. You are stars!

To all my brothers and sisters, thank you for always encouraging me and for all the love and support that you gave me during this journey.

Kuna mai nababa Elizabeth naEdmon Mufanebadza ndinotenda nerudo rwenyu nekundidzidzisa kukosha kwefundo kwamakaita. Baba, thank you for leading my way whenever I was taking a step further in my educational journey. You taught me endurance. I did not understand it then, but looking back I always smile when I relive all the moments you were by my side supporting me.
To my children Nyasha, Noku and Nashe, thank you for making me laugh and for always asking me about my “homework”. You are the reason I did this!

Lastly, to my husband Mike, the man who dares me to always dream big, I say thank you for believing in me and for supporting me throughout this interesting journey. You have always believed in my strengths. Thank you for showing me the way, for encouraging me to continue pushing and for being my anchor.

I am thankful to the University of Pretoria for the financial support provided.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.
Abstract

This study investigated the use of articles, modal verbs and selected discourse markers (so, well and but) in Zimbabwean English (ZE) to determine whether there are variations and innovations in their use. Different morphosyntactic features that are reported by Kortmann, Lunkenheimer and Ehret (2020) to show innovation were used as the basis for analysis. A comparative analysis was done on the ZE corpus totalling 356 007 words and the International Corpus of English-Great Britain (ICE-GB) totalling 608 235 words in order to determine whether there are statistically significant variations in the frequencies of articles, modal verbs and discourse markers (so, well and but). A log likelihood test was used to check whether the observed variations between the corpora were statistically significant. The spoken and written registers and the different genres were compared for variations.

The statistically significant variations observed in the frequencies of articles, modal verbs and DMs (so, well and but) in the spoken and written registers, and in different genres, pointed to the fact that there are differences in the frequencies of these constructions in the ZE corpus and the ICE-GB. No statistically significant variations in the spoken and written registers and in different genres in these constructions means that the frequencies of these constructions are not different in the two corpora. The statistically significant variation of these constructions in functions of these constructions point to the fact that ZE speakers and BrE speakers do not use the functions of articles, modal verbs and DMs (so, well and but) in a similar way. Where no statistically significant variations were recorded in the frequency of the functions of articles, modal verbs and DMs (so, well and but), this points to the fact that ZE speakers and BrE speakers use these functions the same way. These reported variations and no variations will aid in the determination of the stage of ZE in the DyM.

In the ZE corpus, there was attested absence of the use of the definite article where StE has indefinite article (feature 60), use of indefinite article where StE has definite article (feature 61), use of definite article where StE favours zero (feature 64), indefinite article one/wan (feature 66), and non-standard use of modals for politeness reasons (feature 127). In relation to the use of zero article where StE has definite article (feature 62), use of zero article where StE has indefinite article (feature 63), use of indefinite article where
StE favours zero (feature 65), double modals (feature 121), and present tense forms of modals used where StE has past tense forms (feature 123), L2 users of English who speak Shona as their L1 seemed to deviate slightly from L1 English conventions. This is because the features exist but are extremely rare. The use of demonstratives for definite article (feature 67) was attested to be neither pervasive nor extremely rare. Results from this study support the idea that ZE is at both stage 2 and stage 3 of Schneider's (2003, 2007) dynamic model.

**Keywords:** Zimbabwean English, dynamic model, log likelihood, statistically significant bilingualism, corpus-based analysis, World Englishes, English as a second language, innovation, linguistic proficiency, nativisation, norm, variety
# Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declaration</td>
<td>ii</td>
</tr>
<tr>
<td>Ethics statement</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>Abstract</td>
<td>vi</td>
</tr>
<tr>
<td>Table of contents</td>
<td>viii</td>
</tr>
<tr>
<td>List of tables</td>
<td>xv</td>
</tr>
<tr>
<td>List of figures</td>
<td>xviii</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>xix</td>
</tr>
</tbody>
</table>

## Chapter 1

**Introduction**

1.1 Introduction and background

1.2 Theoretical framework of this study

1.3 Zimbabwe’s language situation

1.4 The status of English in Zimbabwe

1.5 Problem statement

1.6 Motivation/ Rationale

1.6.1 Possible variations on articles

1.6.2 Possible variations on modals

1.6.3 Possible variations on discourse markers

1.7 Research questions

1.8 Aims and objectives of the study

1.9 Research design

1.10 Key terms

1.11 Structure of the thesis

## Chapter 2

**Theoretical framework**

2.1 Introduction
2.1.1 How World Englishes is conceptualised 21
2.1.2 Purposes and functions of English in the WE context 22

2.2 English language: A brief history 22

2.3 A historical overview of world Englishes 25
   2.3.1 Differing perspectives on variety formation 28
   2.3.2 Codification 29
   2.3.3 Misconceptions about users and uses of English 30
   2.3.4 Attitudes towards world Englishes 31
   2.3.5 Second language acquisition and world Englishes 33

2.4 Some models proposed for World Englishes 35
   2.4.1 McArthur- Circle of world English (1987) 35
   2.4.2 Görlach - Circle model of English 36
   2.4.3 Kachru- Three concentric circles of English (1985, 1988) 37
   2.4.4 Onysko (2016) Language contact typology of world Englishes 40

2.5 The Electronic World Atlas of Varieties of English 41

2.6 The dynamic model of postcolonial Englishes 43
   2.6.1 Stage 1: Foundation 44
   2.6.2 Stage 2: Exonormative stabilisation 44
   2.6.3 Stage 3: Nativisation 45
   2.6.4 Stage 4: Endonormative stabilisation 46
   2.6.5 Stage 5: Differentiation 47

2.7 Criticism of the dynamic model 48

2.8 Definition of concepts 49
   2.8.1 World Englishes/ New Englishes/ Global Englishes 49
   2.8.2 African English/ African Englishes 51
   2.8.3 Black South African English 54
   2.8.4 Some studies on Zimbabwean English 57
   2.8.5 Bilingualism 60
      2.8.5.1 Types of bilinguals 63
      2.8.5.2 How bilinguals are classified in this study 64
   2.8.6 Linguistic proficiency 64
      2.8.6.1 Why proficiency is difficult to measure 67
      2.8.6.2 How linguistic proficiency was ascertained/determined 67
   2.8.7 Innovation or error 68
   2.8.8 Nativisation 75

© University of Pretoria
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8.9 Norm</td>
<td>76</td>
</tr>
<tr>
<td>2.8.10 L1 or L2 speakers as teachers of English</td>
<td>78</td>
</tr>
<tr>
<td>2.8.11 Register</td>
<td>79</td>
</tr>
<tr>
<td>2.9 Conclusion</td>
<td>79</td>
</tr>
</tbody>
</table>

**Chapter 3**

**Corpora: An overview**

3.1 Introduction | 81 |
3.2 Corpora | 81 |
3.3 A historical overview of corpus linguistics | 82 |
3.4 Types of corpora | 83 |
3.4.1 General corpora | 83 |
3.4.2 Specialised corpora | 84 |
3.4.3 Written corpora | 84 |
3.4.4 Spoken Corpora | 85 |
3.4.5 Tagged and untagged corpora | 85 |
3.5 What to consider in corpus design | 85 |
3.5.1 Representativeness of the corpus | 86 |
3.5.2 Corpus size | 87 |
3.5.3 Balance in text size | 89 |
3.6 Corpus analysis tools | 89 |
3.7 Some of the studies done on African English using corpora | 90 |
3.8 The International Corpus of English | 92 |
3.9 Conclusion | 93 |

**Chapter 4**

**Methodology**

4.1 Introduction | 94 |
4.2. Differences between methodology and method in this study | 95 |
4.3 Research design: Corpus linguistics method | 95 |
4.3.1 Concepts of corpus-based and corpus driven approaches | 96 |
4.3.2 Introspective data versus observational data collection and analysis | 98 |
4.3.3 Quantitative versus qualitative data collection and analysis | 101 |

© University of Pretoria
4.4 Research methodology

4.4.1 Stage 1: Theoretical research

4.4.2 Stage 2: Empirical research: Data collection

4.4.2.1 Step 1: Pilot study and instrument development

4.4.2.2 Step 2: Main study

4.4.2.3 Step 3: Corpus linguistic data transcription and annotation

4.4.2.4 Step 4: Corpus linguistic data: The Zimbabwean English corpus

4.4.2.5 Step 5: Corpus linguistic data: The ICE-GB

4.4.2.6 Step 6: Data analysis

4.5 Ethical considerations

4.6 Assumptions

4.7 Conclusion

Chapter 5

Articles

5.1 Introduction

5.2 Types of English articles and their functions

5.2.1 Definite article

5.2.2 Indefinite article

5.2.3 The zero article

5.2.4 The contextual use of the three article types

5.3 Functions of English articles

5.3.1 Functions of the definite article

5.3.1.1a i Anaphoric reference: direct

5.3.1.1a ii Anaphoric reference: indirect

5.3.1.1b Cataphoric reference

5.3.1.1c Sporadic reference

5.3.1.2 Situational use

5.3.1.2a Immediate situation use

5.3.1.2b Larger situation use

5.3.1.3 The “logical” use of the

5.3.1.4 The use of the with reference to body parts

5.3.1.5 Generic use

5.3.2 Functions of the indefinite article

5.3.3 Functions of the zero article
5.4 Variations on articles reported in literature  
5.5 Attestation and pervasiveness of features  
5.6 Results  
5.6.1 Quantitative analysis of articles  
5.6.1.1 Occurrence frequencies of all three article types (a/an, the and the zero article) for both corpora with Log likelihood comparison  
5.6.1.2 Occurrence frequencies of all three article types across registers for both corpora with Log likelihood comparison  
5.6.1.3 Occurrence frequencies of all three article types across genres for both corpora with Log likelihood comparison  
5.6.1.4 Function frequency count of all three articles for both corpora with Log likelihood comparison  
5.6.2 Qualitative analysis of articles  
5.6.2.1 Functions of the definite article in both corpora  
5.6.2.2 Functions of the indefinite article in both corpora  
5.6.2.3 Functions of the zero article in both corpora  
5.6.3 Summary of quantitative and qualitative results for articles  
5.6.4 Attestation of innovation features according to Kortmann et al. (2020) for articles used in Zimbabwean English  
5.6.4.1 Use of definite article where Standard English has indefinite article: Feature 60  
5.6.4.2 Use of indefinite article where Standard English has definite article: Feature 61  
5.6.4.3 Use of zero article where Standard English has definite article: Feature 62  
5.6.4.4 Use of zero article where StE has indefinite article: Feature 63  
5.6.4.5 Use of definite article where Standard English favours zero – Feature 64  
5.6.4.6 Use of indefinite article where StE favours zero: Feature 65  
5.6.4.7 Indefinite article one/wan: Feature 66  
5.6.4.8 Demonstratives for definite article: Feature 67  
5.7 Conclusion  

Chapter 6  

Modal verbs  

6.1 Introduction  
6.2 Modal verbs in English  
6.3 Functions of modal verbs  
6.4 Studies of modals in L2 and New Englishes
6.5 Features that show variation in varieties of English 230

6.6 Results 232

6.6.1 Quantitative data analysis of modal verbs 233

6.6.1.1 Occurrence frequencies of the nine modal verbs 234

6.6.1.2 Occurrence frequencies of all nine modal verbs across registers for both corpora with Log likelihood comparison 235

6.6.1.3 Occurrence frequencies of all the nine modal verbs across genres for both corpora with Log likelihood comparison 238

6.6.1.4 Function frequency of modal verbs 248

6.6.2 Qualitative analysis of the modal verbs 253

6.6.3 Summary of quantitative and qualitative results for modal verbs 257

6.6.4 Attestation of innovation features according to Kortmann et al. (2020) for modal verbs used in Zimbabwean English 262

6.6.4.1 Double modals: Feature 121 263

6.6.4.2 Present tense forms of modals used where StE has past tense forms: Feature 123 267

6.6.4.3 Non-standard use of modals for politeness reasons: Feature 127 270

6.7 Conclusion 273

Chapter 7 276

Discourse markers 276

7.1 Introduction 276

7.2 Discourse markers 277

7.3 Characteristics of discourse markers 281

7.4 Functions of discourse markers 282

7.4.1 Functions of so 283

7.4.2 Functions of well 284

7.4.3 Functions of but 287

7.5 Discourse markers in L2 English 288

7.6 Results 291

7.6.1 Quantitative data analysis of discourse markers 291

7.6.1.1 Occurrence frequencies of the discourse markers 291

7.6.1.2 Occurrence frequencies of three discourse markers (so, well and but) for both corpora with Log likelihood comparison 293

7.6.1.3 Occurrence frequencies of discourse markers (so, well and but) across genres for both corpora with Log likelihood comparison 296

© University of Pretoria
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.6.1.4 Function frequency of discourse markers</td>
<td>301</td>
</tr>
<tr>
<td>7.6.2 Qualitative analysis of discourse markers</td>
<td>305</td>
</tr>
<tr>
<td>7.6.2.1 Functions of <em>so</em></td>
<td>306</td>
</tr>
<tr>
<td>7.6.2.2 Functions of <em>well</em></td>
<td>313</td>
</tr>
<tr>
<td>7.6.2.3 Function of <em>but</em></td>
<td>323</td>
</tr>
<tr>
<td>7.6.3 Summary of quantitative and qualitative results for discourse markers</td>
<td>327</td>
</tr>
<tr>
<td>7.7 Conclusion</td>
<td>329</td>
</tr>
</tbody>
</table>

Chapter 8  

Conclusion  

8.1 Introduction  

8.2 Revisiting the research questions  

8.2.1 Frequencies: Establishing whether there are variation in ZE  

8.2.2 Establishing whether there are variations in function frequencies  

8.2.3 Determining whether there is innovation in ZE  

8.2.4 The stage of ZE in the dynamic model  

8.3 Limitations and possibilities for future research  

8.4 Conclusion  

Bibliography  

Appendix A: Consent form for individuals  

Appendix B: Consent form for private business letters  

Appendix C: Biographical and language background questionnaire  

Appendix D: Semi-structured interview schedule  

Appendix E: Permission letter from North-West University  

Appendix F: University of Zimbabwe permission letter  

Appendix G: Permission letter from The Herald and The Sunday Mail newspapers  

Appendix H: Results from biographical and language background questionnaires  

Appendix I: Transcription conventions
List of tables

Table 1. Variations reported in literature (source: Kortmann, Lunkenheimer & Ehret’s (2020)  
Table 2. Eight features of ZE (source: Marungudzi, 2016a: 89)  
Table 3. A qualitative comparison of a text versus a corpus (Source: Tognini-Bonelli, 2010)  
Table 4. Summary of Linguistic backgrounds for participants in pilot study  
Table 5. Genres in pilot study  
Table 6. Summary of Linguistic backgrounds for private semi-scripted dialogue participants  
Table 7. Summary of Linguistic backgrounds for private dialogue participants  
Table 8. Summary of Linguistic backgrounds for private business letter participants  
Table 9. ZE corpus summary  
Table 10. ICE-GB summary (Source: www.ucl.ac.uk/english-usage/projects/ice-gb/)  
Table 11. Word count for text categories in ZE corpus and ICE-GB  
Table 12. Use of articles with different noun types in English (Source: Nel, 2015: 56)  
Table 13. Uses of the articles with count and noncount nouns (Source: Quirk et al. 1985: 253)  
Table 14. Functions of indefinite article (Source: Botha, 2012: 264-271)  
Table 15. Functions of the zero article (Quirk et al. 1985: 274-281)  
Table 16. Variations with articles reported in Kortmann, Lunkenheimer & Ehret (2020)  
Table 17. Summary of how features are attested in different L2 varieties  
Table 18. Absolute and normalised frequencies of articles  
Table 19. Frequency of the across registers  
Table 20. Frequency of a/an across registers  
Table 21. Frequency of the across genres  
Table 22. Frequency of a/an across genres  
Table 23. Frequency of functions of the definite  
Table 24. Frequency of functions of the indefinite article  
Table 25. Frequency of functions of the zero article  
Table 26. Functions of the zero article  
Table 27. Feature numbers and their descriptions  
Table 28. Distribution of feature 64 across genres  
Table 29. Distribution of feature 65 across genres  
Table 30. Distribution of feature 66 across genres  
Table 31. Distribution of feature 67 across genres  
Table 32. Summary of features and their categorisations in ZE  
Table 33. Modal verbs and their functions (Sources: Quirk et al. 1985; Swan, 1995; Collins, 2009)  
Table 34. Variations with modal verbs reported in Kortmann, Lunkenheimer & Ehret (2020)  
Table 35. Summary of how features are attested in different L2 varieties
Table 36. Summary and frequencies of modal verbs per 10 000 words

Table 37. Frequency of modal verbs in spoken and written registers

Table 38. Frequency of can across genres

Table 39. Frequency of could across genres

Table 40. Frequency of may across genres

Table 41. Frequency of might across genres

Table 42. Frequency of must across genres

Table 43. Frequency of shall across genres

Table 44. Frequency of should across genres

Table 45. Frequency of will/’ll across genres

Table 46. Frequency of would/’d across genres

Table 47. Frequencies of functions of should and must

Table 48. Frequencies of functions of modal verbs of possibility, permission and ability (may, can, might and could)

Table 49. Frequencies of functions of will, would and shall

Table 50. Functions of modal verbs in the ZE corpus

Table 51. Functions of modal verbs in the ICE-GB

Table 52. Feature numbers and their descriptions

Table 53. Summary of features and their categorisations in ZE

Table 54. Functions of so (Sources: Müller 2005; Vickov & Jakupčević, 2017)

Table 55. Functions of well (Source: Müller, 2005)

Table 56. Summary and frequencies of DMs per 10 000 words

Table 57. Frequency of so across registers and instances per 10 000 words

Table 58. Frequency of well across registers

Table 59. Frequency of but across registers

Table 60. Frequency of so across genres

Table 61. Frequency of well across genres

Table 62. Frequency of but across genres

Table 63. Frequencies of functions of so

Table 64. Frequencies of functions of well

Table 65. Frequencies of functions of but

Table 66. A summary of the Log likelihood values and associated p-values for each construction, subtype, register and associated genres to answer research question 1.2 and 1.3.

Table 67. Variation and variation direction for each construction, subtype, register and associated genres to answer research question 1.2 and 1.3.

Table 68. A summary of the Log likelihood values and associated p-values for each construction, subtype and function to answer research question 2.2 and 2.3.

Table 69. Variation and direction for each construction, subtype and function.
Table 70. Summary of features and their categorisations in ZE
List of figures

Figure 1. When English was learned (Source: Language questionnaire for this study) 7
Figure 2. The foundational concepts that are involved in the WE 21
Figure 3. McArthur’s circle of World English (source: McArthur 1987: 11) 36
Figure 4. Görlach’s circle model of English (source: Mesthrie & Bhatt, 2008: 29) 37
Figure 5. Three concentric circles of English (source: Kachru, 1988: 5) 38
Figure 6. The Language Contact Typology of world Englishes (Onysko, 2016) 40
Figure 7. Growth in corpus sizes over 50 years (Source: Anthony, 2013: 145) 88
Figure 8. Population pyramid of Zimbabwe 110
Figure 9. Percentages for spoken and written texts in the ZE corpus 124
Figure 10. Screenshot showing KWIC concordances of “the” in Sketch Engine 128
Figure 11. Screenshot showing KWIC concordances of “an” in Sketch Engine 129
Figure 12. The different subclasses of nouns in English 148
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE</td>
<td>African English</td>
</tr>
<tr>
<td>Ar</td>
<td>Arabic</td>
</tr>
<tr>
<td>BICS</td>
<td>Basic interpersonal communicative skills</td>
</tr>
<tr>
<td>BNC</td>
<td>British National Corpus</td>
</tr>
<tr>
<td>BrE</td>
<td>British English</td>
</tr>
<tr>
<td>BSAE</td>
<td>Black South African English</td>
</tr>
<tr>
<td>CALP</td>
<td>Cognitive academic language proficiency</td>
</tr>
<tr>
<td>DM</td>
<td>Discourse Marker</td>
</tr>
<tr>
<td>DMs</td>
<td>Discourse Markers</td>
</tr>
<tr>
<td>DyM</td>
<td>Dynamic Model</td>
</tr>
<tr>
<td>EFL</td>
<td>English as a foreign language</td>
</tr>
<tr>
<td>Eng</td>
<td>English</td>
</tr>
<tr>
<td>ENL</td>
<td>English as a native language</td>
</tr>
<tr>
<td>ENS</td>
<td>English as a native speaker</td>
</tr>
<tr>
<td>ESL</td>
<td>English as a Second Language</td>
</tr>
<tr>
<td>eWAVE</td>
<td>electronic World Atlas of Varieties of English</td>
</tr>
<tr>
<td>GloWbE</td>
<td>Global Web-Based English</td>
</tr>
<tr>
<td>ICE</td>
<td>International Corpus of English</td>
</tr>
<tr>
<td>ICE-GB</td>
<td>International corpus of English- Great Britain</td>
</tr>
<tr>
<td>IDG</td>
<td>Indigenous</td>
</tr>
<tr>
<td>Kal</td>
<td>Kalanga</td>
</tr>
<tr>
<td>KWIC</td>
<td>Key Word In Context</td>
</tr>
<tr>
<td>L1</td>
<td>First Language</td>
</tr>
<tr>
<td>L2</td>
<td>Second Language</td>
</tr>
<tr>
<td>LCT</td>
<td>Language contact typology</td>
</tr>
<tr>
<td>LL</td>
<td>Log likelihood</td>
</tr>
<tr>
<td>LOCNESS</td>
<td>Louvain Corpus of Native English Student Essays</td>
</tr>
<tr>
<td>Nd</td>
<td>Ndebele</td>
</tr>
<tr>
<td>NE</td>
<td>New Englishes</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>NICE</td>
<td>Negation, inversion, code, and emphasis</td>
</tr>
<tr>
<td>NNS</td>
<td>Non-native speakers</td>
</tr>
<tr>
<td>NNVE</td>
<td>Non-native varieties of English</td>
</tr>
<tr>
<td>NP</td>
<td>Noun Phrase</td>
</tr>
<tr>
<td>NPs</td>
<td>Noun Phrases</td>
</tr>
<tr>
<td>Ny</td>
<td>Nyanja</td>
</tr>
<tr>
<td>Por</td>
<td>Portuguese</td>
</tr>
<tr>
<td>S1</td>
<td>Sentence 1</td>
</tr>
<tr>
<td>S2</td>
<td>Sentence 2</td>
</tr>
<tr>
<td>SLA</td>
<td>Second Language Acquisition</td>
</tr>
<tr>
<td>Sh</td>
<td>Shona</td>
</tr>
<tr>
<td>Shan</td>
<td>Shangani</td>
</tr>
<tr>
<td>Skema</td>
<td>Sketch Engine manual annotation</td>
</tr>
<tr>
<td>StE</td>
<td>Standard English</td>
</tr>
<tr>
<td>STL</td>
<td>Settler</td>
</tr>
<tr>
<td>Swa</td>
<td>Swahili</td>
</tr>
<tr>
<td>TLEC</td>
<td>Tswana Learner English Corpus</td>
</tr>
<tr>
<td>Tso</td>
<td>Tsonga</td>
</tr>
<tr>
<td>UNESCO</td>
<td>United Nations Educational, Scientific and Cultural Organisation</td>
</tr>
<tr>
<td>WE</td>
<td>World Englishes</td>
</tr>
<tr>
<td>WSAE</td>
<td>White South African English</td>
</tr>
<tr>
<td>ZE</td>
<td>Zimbabwean English</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Introduction and background

This study investigates the use of articles, modal verbs and selected discourse markers (so, well and but) in Zimbabwean English (ZE) to determine whether there is variation and innovation in their use. Different morphosyntactic features that are reported by Kortmann, Lunkenheimer and Ehret (2020) to show variation are used as the basis for analysis. The study is situated within the field of World Englishes and focuses on innovations in L2 varieties and the nativisation of English to help shed light on the spread of English and the form and function of different characteristics of English.

The number of people who speak English as a second language\(^1\) (ESL) or additional language continues to increase worldwide (Graddol, 2000; Bruthiaux, 2002; Schneider, 2007). This increase may be attributed to the opportunities brought about by English both in individual countries and globally; for example, in the spheres of education, employment, trade, industry, and innovation (Kachru, 1992; Makoni, 1992; Thondhlana, 2002). There are approximately 6 000 languages spoken throughout the world, with around 275 languages spoken in Europe and the Middle East, 900 in the Americas, 1900 in Africa, and 3000 in the Pacific and Asia (Thomason, 2001). “A significant proportion of the world's population uses English as its other tongue - a second or foreign language” (Kachru, 1992: 2-3).

Research on L2 varieties of English has been growing since the 1960s with preliminary work by researchers such as Kachru’s (1965, 1976, 1980) work on Indian English and Sey’s (1973) work on the grammatical description of Ghanaian English. Other early studies on L2 varieties of English include Platt’s (1989) study on New Englishes, Bokamba’s (1992) report

\(^1\) The term “English as a second language” is used in this study to refer to the English spoken as a second, third or additional language. Ellis (2015) explains that the majority of researchers use the term ESL as a cover term to refer to the second, third, or any language learned after the first one.
on the Africanisation of English, and Bamgbose’s (1998) research on innovations in world Englishes. Schmied (1991) provided a comprehensive overview of English in Africa and Banjo (2000) reported on English in West Africa. There is a growing body of research on L2 English varieties as evidenced by scholarly work on the subject (e.g. De Klerk, 2005a, 2005b; Van Rooy & Terblanche, 2006; Schneider, 2007; Van Rooy, 2007, 2013; Kachru, 1992, 2006; Schmied, 2006; Makalela, 2007, 2013; Wasserman & Van Rooy, 2014; Deshors, Götz & Laporte, 2016; Kruger & Van Rooy, 2017). Due to the differences in culture, religion and politics, speakers of ESL may speak English differently (Kachru, 1992). These variations are of interest to researchers (Lange & Leuckert, 2020) and specifically in this study because the variations will show whether there is innovation in the way English is spoken in Zimbabwe.

Different terms have been proposed for L2 varieties of English. Some of the terms are “world Englishes” ("WE"), “global English”, “new Englishes” ("NE") (Schneider, 2003, 2007). An explanation of these terms is given in Section 2.8.1. Researchers have studied African English (AE) varieties (one of the varieties of WE) like Nigerian English (e.g. Werner & Fuchs, 2017), Black South African English (e.g. Mesthrie, 2006), Swazi English (e.g. Arua, 1998), and Kenyan English (e.g. Schmied, 2006). In this study, following a pluricentric approach to WE, the term “Zimbabwean English” (a variety of AE) is used to refer to the English that is spoken by Shona mother tongue speakers. While English has grown in use as a second or additional language throughout the world, it has done so particularly in Africa. The linguistic diversity of Africa has not helped to ward off the dominance of English (Ndhlovu, 2011) as indigenous languages often share the status of official language with English (Thomason, 2001; Ndhlovu, 2011). In countries where many languages are spoken, there is bound to be language contact (Thomason, 2001; Clyne, 2003; Ndhlovu, 2011). Africa, particularly Southern Africa (part of the global south) is home to many indigenous languages and English is also spoken as a second language (L2). For example, English is an official language in Zimbabwe, Zambia, South Africa, and Namibia. English is used “in various domains of society such as education, business, media, and government” (Chang, 2008: 1).

---

2 The term “world Englishes” is used in literature to show that there are different varieties of English that emerge from countries where English is used as a second or additional language.

3 The pluricentric approach recognises that there are different varieties of English around the world.

4 Zimbabwean English is oriented towards BrE which was introduced in Zimbabwe through colonialism. BrE provides the norm in teaching and learning and in print media (e.g. newspapers and publishing houses follow the BrE conventions).

© University of Pretoria
According to Arua (1998) the scarcity of research into new English varieties in Southern Africa (save for South Africa) can be attributed to the assumption that the English spoken in other Southern African countries is covered in research on South African English. Marungudzi (2016a) notes that there are differences in each variety of English in Southern Africa, which warrant further exploration on the topic with a specific focus on English in Zimbabwe. Kadenge (2009) and Marungudzi (2016a) noted that theoretical and applied linguists have not studied ZE to a great extent. Kadenge suggested that the term “Zimbabwean English” be adopted to refer to a variety of English spoken in Zimbabwe.

Given the economic and educational value placed on English, many people learn it as a second or additional language (Makoni 1992). Age, acculturation, attitude, motivation, learning style, first language proficiency, community, and socio-economic aspects are contributing elements to second language acquisition (SLA) and variety formation (Brown, 2014; Orosco & Hoover, 2009). The aforementioned elements are highlighted to show the different contributing elements in SLA and new language variety formation. They are not discussed further in this study. In the process of learning English, learners may adapt and adopt features from their first language (L1) in their speech, leading to innovation and nativisation (Fishman, 1992; Kachru, 1992; Schneider, 2003). Some of the features that may adapt in L2 varieties of English are discourse markers (DMs) (Müller, 2005), the article system, plural markings, mass and count nouns, the verbal system, determiners, quantifiers, and tag questions (De Klerk, 2003a). It is in part due to the identification of these features that this study will also include the feature of articles, modal verbs and discourse markers (for an exposition on the features included in this study (cf. Section 1.6).

This study seeks to use the dynamic model (DyM) of postcolonial Englishes to test whether there is innovation in the use of articles, modal verbs, and DMs (so, well and but). Kortmann, Lunkenheimer and Ehret’s (2020) description of innovation features (60, 61, 62, 63, 64, 65, 66, 67, 121, 123 and 127) was used as a basis for description. The features are explained in Section 1.6.

5According to Dewaele (2018), there is a discrepancy in the use of terms ESL and non-native varieties of English (NNVE). In this study, the term “ESL varieties” is used to refer to varieties of English that are spoken as second or additional languages. The term “NNVE” is used as it appears in literature.
Marungudzi (2016a) studied 8 morphosyntactic features of ZE using a corpus of 390 000 words. The corpus consisted of 53.4% spoken texts and 46.6% written texts. Part of Marungudzi's (2016a) corpus was used in this study in addition to the corpus that I compiled in order to examine whether ZE has features which show innovation in terms of Schneider’s (2003, 2007) framework on stage 2, 3, and 4 (cf. Section 2.6 for an exposition on Schneider's framework and the different stages).

1.2 Theoretical framework of this study

This study is based on the DyM (Schneider, 2003; 2007) because it accounts for the development of postcolonial varieties of English. Schneider suggested that speech communities where English is not a first language undergo five stages of development. These stages are foundation, exonormative stabilisation, nativisation, endonormative stabilisation, and differentiation. In order to test the applicability of the DyM, Schneider studied several countries and explained the current stages for the countries. A detailed overview of this theoretical framework is provided in Section 2.6.

A functionalist usage-based approach is also utilised in this study. The approach is suitable because this study investigates language as it is used by speakers. One of the useful tools used in the usage-based approach is a corpus (Partridge, 2019). Corpus data, which is an example of observational data, allows for the quantification of linguistic features, thereby enabling “a rather objective identification of what may be considered important and what may be considered rather marginal” (Gries, 2006a: 191). The quantification of features allows researchers to check whether an observed feature has been conventionalised or not. De Klerk (2003a) is of the opinion that the usage-based framework benefits from the quantifiable data provided by corpus data. According to Botha (2012: 14), "in functionalist approaches to language, meaning (semantics) and context (discourse/pragmatics) are important aspects of language function, entailing that morphology and syntax are not described without reference to meaning and

---

6 The initial plan was to obtain permission for the whole of Marungudzi’s (2016a) corpus. Unfortunately, he passed on in January 2018. Although I got permission to use Marungudzi's corpus from North-West University, I was able to access 150 000 words out of the 390 000 words. This meant that I had to adjust the data collection process to collect more samples than previously planned.
discourse”. By including the qualitative study of functions of articles, modal verbs and discourse markers, the study includes the semantic and pragmatic interpretations which aids in the overall identification of variations contextualised within the DyM.

The following section provides an overview of the languages used in Zimbabwe while contextualising the use of English in the country.

1.3 Zimbabwe’s language situation

The promotion of one language, which is typically a former colonial language, at the expense of another seems to be a hallmark of most countries in Africa (Gough, 1996a; Thondhlana, 2002; Kadenge, 2010). Multilingual societies like Zimbabwe have a difficulty when it comes to striking a balance between language promotion and linguistic rights. Consequently, some languages turn out to be marginalised in education, industry, and government business, to mention a few. According to Kadenge (2010) English is used in high domains such as education, sports, and parliamentary business whilst indigenous languages are rarely used in these domains. In situations where English is the official language, either on its own or with other local languages, there is language contact as mentioned in Section 1.1. As languages come into contact, the effects may be manifested in the forms and functions of specific features in the languages as well as the languages (overall) themselves.

According to Ndhlovu (2009), in Zimbabwe, English is an official language together with Shona (spoken by 75% of the population) and Ndebele (spoken by 16% of the population). Thondhlana (2002) puts the number of minority indigenous languages of Zimbabwe spoken by about 10% of the population at fourteen. These are Kalanga, Chewa, Tonga, Nambya, Hwesa, Shangani, Barwe, Sotho, Venda, Xhosa, Sena, Tshwawo or Khoisan, and Tswana. The Constitution of Zimbabwe Amendment (Number 20) Act of 2013 now officially recognises sixteen languages namely; Chewa, Barwe, Xhosa, Venda, Tswana, Tonga, Sotho, sign language, Shona, Shangani, Ndebele, Ndau, Nambya, Khoisan, Kalanga, and English. Although these languages have been elevated to official status in the Constitution, they are still not being used in official communication by government departments. In order to raise the status of indigenous languages in Zimbabwe, the Education Act of 1987 provided that children learn in their mother tongue from grade one to grade three in line with the United Nations Educational, Scientific and Cultural Organisation (UNESCO)’s 1951 education
policies. The Education Act has been revised to incorporate mother tongue learning until grade seven, an effort by the government to uplift the status of indigenous languages. From the discussion above, it can be noted that Zimbabwe’s linguistic landscape is complex with many languages being used throughout the country. The following section focuses specifically on the use of English in Zimbabwe and the status attributed to the language.

1.4 The status of English in Zimbabwe

Within the African context, English was introduced by means of colonisation in countries like Kenya, Malawi, Zambia, Nigeria, South Africa, Swaziland, and Zimbabwe. Zimbabwe was a British colony from 1890 to 1980 (Ndhlovu, 2009). During colonialism, English was introduced and promoted to an official language. Up until the present day, English enjoys a higher status than any other language in Zimbabwe because it is the language of education, trade, and government (Makoni, 1992; Kadenge, 2009). From the time learners enter the education system, they are exposed to English in the classroom. In order to attend tertiary education or to seek formal employment, one has to pass Form 4 (equivalent to Grade 11 in South Africa) with at least a “C” symbol. This requirement shows the high status placed on English in Zimbabwe’s education system.

Graddol (2000) and Jenkins (2003) make a distinction between three different types of English speakers. They use the term “English as a native speaker”\(^7\) (ENS) to refer to speakers whose first and native language is English. Another group is the English as a foreign language (EFL) speakers who learn English for tourism and commerce. ESL speakers are those L2 speakers who use English as a second or additional language. Mesthrie, (2004) notes that ESL is found in countries where a stable L2 is used in the schooling system and in government. In the case of Zimbabwe, it can be noted that the majority of speakers can be classified as ESL speakers.

\(^7\)Schmitz (2009) argues that the terms “native” and “non-native” sometimes have political and linguistic connotations with debate arising from their use. Schmitz uses the terms to distinguish between L1 and L2 users of a language.
In Zimbabwe, the majority of English teachers from Grade 1\(^8\) to Form 6 (equivalent to Grade 12 in South Africa) are L2 users themselves. Some teachers may have achieved what is referred to in literature as native-like accent, but others may not have done so. Thus, they will pass on their idiosyncratic linguistic features to learners (Makoni, 1992). Most learners have minimal to no English language exposure outside the classroom, leading to learners achieving varying levels of linguistic proficiency\(^9\). The fact that most learners start using English at school is supported by results from the biographical and language background questionnaires carried out in this study, as shown in figure 1 below.

![Figure 1. When English was learned (Source: Language questionnaire for this study)](image)

From figure 1 above, it can be noted that the majority of participants (52) learned English when they entered the school system.

In rural areas, where some of the poorest citizens live, Shona is used in everyday communication because most rural communities in Zimbabwe are made up of people who speak the same language and the economic activities are mainly agriculture and informal trade (Makoni, 1993). The situation is a bit different in towns and cities where English is used more often than in rural areas, due to the linguistic diversity of urban areas and due to the fact that most formal jobs and economic activities are available in urban areas (Thondhlana, 2002). The following section outlines the problem statement in conjunction

---

\(^8\) In Zimbabwe, primary education starts at grade 1 and ends at grade 7. Secondary education starts at form 1 to form 4. High school education starts at form 5 and ends at form 6. After passing form 6, learners can go to university.

\(^9\) Bialystok (1981) associates linguistic proficiency with the degree of mastery of a language. An overview of linguistic proficiency is given in Section 2.8.6.
1.5 Problem statement

Research on innovations in L2 varieties and nativisation of English has helped shed light on the spread of English and the form and function of different features of English. Linguistic innovation is one of the characteristic features of the DyM. This study looked at morphosyntactic features 60, 61, 62, 63, 64, 65, 66, 67, 121, 123 and 127 as described by Kortmann, Lunkenheimer and Ehret (2020). The features are highlighted in Section 1.6. Articles, modal verbs, and DMs (well, so, but), were studied focusing on adults aged 18 and above, who had at least 10 years of formal education and had been exposed to English. This is in line with Kachru (2006) and Xu’s (2017) assertion that there is a need to study the grammatical and lexicographic features of English as spoken by L2 speakers. The following is a list of problems that were addressed:

1. Although Kadenge (2009) carried out exploratory research to study how English pronunciation of vowels by Shona L1 speakers is affected by Shona phonology, he did not examine consonants and other grammatical features of the language. He noted that “more empirical research is required on the transfer of consonantal and prosodic characteristics of Shona into the spoken English of Shona-English bilinguals” (Kadenge, 2009: 169). Marungudzi (2016a) investigated some morphosyntactic characteristics of ZE but did not study articles, DMs, and modal verbs. Possible variations are pointed out by Kortmann, Lunkenheimer, and Ehret (2020) in their description of features 60, 61, 62, 63, 64, 65, 66, 67, 121, 123 and 127 (the features are described in Section 1.6 below). Therefore, this study firstly seeks to address the gap in literature by going beyond what Kadenge and Marungudzi have focussed on.

2. Schneider (2003, 2007) proposed the DyM to account for the stages in the development of new Englishes (NE) and suggested that the model needs to be tested in different settings. The model also casts light on the relationship between L1 and L2 varieties of English that emerged in contexts where the English language was introduced through colonialism. Although the DyM has been significant in the study

---

10 There are several terms used to refer to new varieties of English. Besides the term WE, new Englishes is also used by scholars.
of WE, the model has been applied minimally to ZE (Marungudzi, 2016a). The study aims to apply this model to ZE.

3. There are several perspectives available to researchers of ESL. ZE has not been adequately studied using a corpus-based analysis to better understand its features and characteristics. Kortmann, Lunkenheimer and Ehret (2020) propose a variety of features that show morphosyntactic variation in different varieties of English. Some of the features as described in Section 1.6 were used as basis for description in this study.

4. Louw and Jordan (1993) compiled a corpus of ZE using English language textbooks that were prescribed for secondary schools. However, these compiled texts do not reflect the actual use of English language in Zimbabwe because they do not reflect real language use contexts and are rather more prescriptive than descriptive. Marungudzi (2016a) compiled a corpus of 390 000 of ZE. As stated in Section 1.2, due to the unavailability of some parts of Marungudzi’s corpus, I, the researcher got access to about 150 000 words. In order to add to the corpus and to have a better representation of texts, a corpus was compiled from 10 private spoken dialogue samples, 45 private semi-scripted dialogue samples, 10 samples of editorials, 14 samples of newspaper reportage, and 5 samples of private business letters were collected to use in addition to Marungudzi’s corpus.

5. According to Müller (2005), although DMs have been studied extensively in literature, the majority of studies focus on native varieties of English (e.g. Schiffrin, 1987; Fraser, 1993, 2009; Blakemore, 2002). Therefore, research still needs to focus on L2 varieties of English, which are sometimes referred to as non-native varieties of English (NNVE)11 in literature. In addition, few studies have investigated the effects of text type on the use of DMs (Lam, 2009b). This study explores the frequency and functions of the DMs so well, and but in different texts in the ZE corpus.

---

11 Dewaele (2018) notes that there is controversy surrounding the use of terms “native speakers” and “non-native speakers” with some scholars preferring to use L1 users and ESL users respectively. In this thesis, the terms “native and non-native speakers” are used as they appear in literature but the terms “L1 and L2 speakers” are also used to refer to first language speakers and second language speakers respectively. The terms are used interchangeably.
While taking into account that the grammar, lexicology, phonetics and many other aspects should be paid attention to in further research (Kadenge, 2009), this research aimed to study these phenomena in terms of the five problems mentioned above. The specific motivations are outlined further in the section below.

### 1.6 Motivation/ Rationale

A great deal of knowledge exists about English as a language of learning and teaching in Zimbabwean schools. Research has shown that there is inadequate linguistic proficiency and competence in learners and teachers alike, leading to code-switching and code-mixing (e.g. Mashiri, 2002). Yet little is known about the nature and characteristics of English that is spoken by Shona mother tongue speakers. When Kachru (1992: 2-3) introduced the term “WE”, he noted that “a significant proportion of the world’s population uses English as its other tongue (a second or foreign language) . . . very little, if anything is said about the formal and functional characteristics of institutionalized World Englishes”. That is why this study seeks to build on and contribute to the study of WE, particularly English as it is spoken by Shona L1 speakers. The study firstly builds on the foundation laid down by researchers who opened the discussion about ZE (e.g. Makoni, 1992, 1993; Kadenge, 2009, 2010, 2012; Mlambo, 2009; Marungudzi, 2016a). In addition, the study also seeks to contribute to the literature on African Englishes since ZE is a variety of AE (cf. point one in Section 1.5). When Schneider (2003, 2007) posited the DyM to explain the evolution of NE, he indicated that innovations occur at stages 2, 3 and 4. Section 2.6 gives a detailed explanation of the characteristic features of each stage.

The novelty of this study is that firstly, the study sets out to investigate the characteristic features of ZE and to explore the frequency and functions of selected DMs. Secondly, the description of features was based on Kortmann, Lunkenheimer and Ehret’s, (2020) electronic World Atlas of Varieties of English (eWAVE). The eWAVE, an interactive database is most suitable to base the descriptions because it provides a list of possible morphosyntactic variations from which to base the analysis and to do comparisons with other varieties of English. For the purpose of this study, articles, modal verbs and DMs (so, well and but) were examined for any variation.
While many of the morphosyntactic features have been studied in AE varieties (e.g. Gough, 1996a; De Klerk, 2003a, 2003b; Mesthrie, 2006; Makalela, 2007, 2013 Botha, 2012, 2013, 2015; Kruger & Van Rooy, 2017), there are still gaps in terms of the morphosyntactic features of ZE. Aside from Marungudzi’s (2016a) study of some features of ZE and Kadenge’s study on the lexical and semantic features of ZE, no study has yet focused on articles, modal verbs and DMs in ZE within the WE framework. This study attempts to close the gap in research by focusing on selected morphosyntactic features, which have not been studied before.

This study is multidimensional in nature (cf. Biber, 1995) because different linguistic features are analysed in order to determine if there are any variations between ZE and British English (BrE). Biber (1995) conducted a multi-dimensional analysis of 67 different linguistic features in English. On Biber’s list, DMs were analysed under the lexical class, modal verbs under the modals category and articles were studied under the nominal forms category. The reason for selecting these features is that features 60, 61, 62, 63, 64 and 65 show variation with articles and belong to the noun phrase category in Kortmann, Lunkenheimer and Ehret’s (2020) classification of features whilst features 121, 123 and 127 belong to the modal verbs category. These features are reported to be attested in other AE varieties such as BSAE, Nigerian English, Ghanaian English and Cameroon English but have not been studied in ZE. Since ZE is a type of AE, examining these features will allow for comparisons with other AE varieties.

The DMs so, well and but were selected because they are among the most frequently studied DMs and research has shown that there is variation in the frequencies and use of the DMs in L2 varieties of English. For example, De Klerk (2005a) reported on variations in the use of well in Xhosa English, Müller (2005) reported that there are variations in the use of well and so in L2 English in Germany and Unuabonah (2019) reported variations in the use of but in Nigerian English. Studying these DMs will enable comparisons with other L2 varieties of English

12 Due to the scope of this study, it was not possible to study all the DMs. Future research could examine other DMs not included in this study.

© University of Pretoria
An analysis of the use of articles, modal verbs and selected DMs will help answer the research questions regarding whether there is variation between ZE and British English and to situate ZE in the DyM. The descriptions of features given by Kortmann, Lunkenheimer and Ehret on www.ewave-atlas.org are shown for articles, modals and DMs, respectively.

1.6.1 Possible variations on articles

Possible variation on articles and the features used in this study, in terms of descriptions of features given by Kortmann, Lunkenheimer and Ehret on www.ewave-atlas.org, are:

- Use of definite article where Standard English\(^{13}\) (StE) has indefinite article - feature 60.
- Use of indefinite article where StE has definite article - feature 61.
- Use of zero article where StE has definite article - feature 62.
- Use of zero article where StE has indefinite article - 63.
- Use of definite article where StE favours zero [article]\(^{14}\) - feature 64.
- Use of indefinite article where StE favours zero [article] - feature 65.
- Use of indefinite article one/wan - feature 66.
- Use of demonstratives for definite article - feature 67.

1.6.2 Possible variations on modals

Possible variation on modals and the features used in this study, in terms of descriptions of features given by Kortmann, Lunkenheimer and Ehret on www.ewave-atlas.org, are:

- Double modals - feature 121.
- Present tense forms of modals used where StE has past tense forms - feature 123.
- Non-standard use of modals for politeness reasons - feature 127.

\(^{13}\) There is no consensus about what constitutes Standard English. Trudgill (1999:118) states that Standard English is “the variety of English normally used in writing, especially printing; it is the variety associated with the education system in all the English-speaking countries of the world”. In this thesis, comparisons were made between ZE and BrE (which is a type of StE) in the study of DMs. This is because English was introduced in Zimbabwe through British colonialism. Therefore, mostly BrE provides the norm in the education system.

\(^{14}\) eWAVE does not include the word “article” after zero. But in this thesis, [article] is used where reference is made to the zero article to distinguish the use of the word “zero” that is used in other concepts.
For the purpose of this study, features that are categorised as being pervasive or obligatory (category A) or features which are neither pervasive nor obligatory (category B) were regarded as being characteristic features of ZE.

**1.6.3 Possible variations on discourse markers**

An examination was done on the frequency and functions of the DMs *(so, well and but)* in the corpus. Studies have shown variations in the use of DMs by L2 speakers of English (Müller, 2005). Some of the studies are in Spanish English (Trillo, 2002), Xhosa English (De Klerk, 2005a, 2005b), Hong Kong English (Lam, 2009a, 2009b), Pakistani English (Jabeen et al., 2011), Irish English (Murphy, 2015) and Nigerian English (Unuabonah, 2019).

According to Müller (2005) and Buysse (2010, 2012), the bulk of research conducted on DMs uses data from native or L1 speakers of English. De Klerk (2005a) adds that there is a dearth of research on DMs in L2 varieties of English. Although DMs have been studied from different perspectives, little attention has been given to DMs in L2 English used by Shona L1 speakers. This thesis attempts to fill the gap in research by investigating the use of DMs in an ESL context by examining the frequency and functions of DMs *so, well and but* in ZE in the WE context (Huddlestone & Fairhurst, 2013). Text type was used as a basis for analysing DMs. Studying DMs will help shed light on the variety status of a new English since variability is one of the characteristics of a new variety of English (Schneider, 2003, 2007).

Kachru (1992) notes that the development of particular discourse styles is one of the features of nativisation in new varieties of English. In addition, nativisation occurs at stage 3 of the DyM of postcolonial Englishes. By exploring the selected DMs, this study tests whether ZE is developing as a new variety or if it conforms to native speaker\(^{15}\) norms. This study adds to the research on DMs in L2 varieties of English (e.g. Gough, 1996a; De Klerk, 2005a, 2005b; Müller, 2005; Algouzi, 2014; Vickov & Jakupcevic, 2017; Oladipupo & Unuabonah, 2020; Unuabonah, 2019; Aşık & Cephe, 2013). A description of the frequency and functions of selected DMs "may provide a useful benchmark for comparison with mother tongue norms" (De Klerk 2005b: 1).

\(^{15}\) “Native speakers are those for whom English is the primary language, non-native speakers” are those for whom English is a secondary language” (Strevens, 1992:36).
Still discussing motivations for this study, De Klerk (2005a) studied the discourse marker (DM) *well* in Xhosa English and concluded that *well* was less frequent in Xhosa English than in L2 varieties of English and that there was diversified use of *well* by Xhosa English speakers. She further remarks that research needs to be conducted on other ESL varieties in order to ascertain whether the differences are universal across the new varieties. Therefore, by examining the DMs *so*, *well* and *but*, this study attempted to fill the gap in research on DMs in ESL noted by De Klerk (2005a).

Another motivation for studying selected DMs is that Lam (2009b) examined the situational context in which the DM *so* was produced in Hong Kong English and noted that further research needs to be done to determine whether there are similarities with other varieties of English. According to Lam (2009b) there is variability in the use of DMs in different text types. Another point made is that one of the least researched topics on DMs is contextual variation. That is why this study explored the frequency and functions of *so*, *well* and *but* in a corpus of ZE. Comparisons were also made between the ZE corpus and the International Corpus of English-Great Britain (ICE-GB). This was done in order to establish whether text type influences the frequency and functions of the DMs. Therefore, by examining the variation in the use of DMs in ZE using a corpus-based analysis, this study fills the research gap noted by Lam (2009b). The conditions under which variations occur in the DyM are outlined in Section 2.6. The criteria for examining innovations was by looking at some of the features contained in the eWAVE as discussed above.

A specialised corpus of ZE “is available on computer tapes at the University of Zimbabwe Computer Centre. Also, on tape is the dictionary file which lists in alphabetical order all the words contained in the books, together with a frequency file, which lists the words in order of use” (Louw & Jordan, 1993: 131). Although Louw and Jordan compiled a corpus of ZE consisting of 592 994 words, the corpus does not show a real picture of English as a L2 use. This is because English language prescribed textbooks for Form 1 to Form 4 (*Dawson, Structures and Skills in English* and, *English for Zimbabwe*) were used as a source of words for the corpus. Louw and Jordan compiled the corpus so that it would be used as a resource in different fields such as academia, teaching, curriculum planning and psychology. Marungudzi (2016a) compiled a corpus of ZE comprising of 390 000 words but I only got access to 150 000 words. As mentioned in Section 1.5, part of Marungudzi’s corpus was used in this study in addition to the corpus that the researcher compiled. Since ZE is a less
resourced variety of WE, a corpus was compiled to use as a linguistic resource that can be made available for future research (cf. point four in Section 1.5).

1.7 Research questions

As a guideline in this study, the following questions were posed:

1.1 What are the occurrence frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB? (Cf. point five in Section 1.5.)

1.2 How does the occurrence of articles, modal verbs and discourse markers in ZE compare to that of BrE?

1.3 Are the variations between the corpora statistically significant? (Cf. point one and three in Section 1.5.)

2.1 What are the function frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB? (Cf. point five in Section 1.5.)

2.2 How do the functions of articles, modal verbs and discourse markers in ZE compare to those of BrE?

2.3 Are the variations between the corpora statistically significant? (Cf. point one and three in Section 1.5.)

3.1 Which features of innovation are present with regards to articles and modal verbs in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE? (Cf. point one and three in Section 1.5.)

4 What is the stage of adult use of these features in ZE when considering the dynamic model of postcolonial English? (Cf. point two in Section 1.5.)

In this study the term “occurrence frequency” is used as a description with the meaning of “the number of times the construction or feature appears in the corpus” while the term “function frequency” is used as a description with the meaning of “the number of times the specific purpose of the construction or feature appears in the corpus”.

© University of Pretoria
(There is no corresponding research question related to point four in Section 1.5, the creation of a corpus for ZE as this point is addressed in chapter 4 in terms of the research method which will be employed in the study.)

1.8 Aims and objectives of the study

Flowing from the research questions above, the aim of this study is to do the following:

1. To establish the frequency and functions of articles, modal verbs and discourse markers; *so*, *well* and *but* in a corpus of ZE and British English.
2. To determine how variation with articles, modal verbs and DMs as outlined in Section 1.6 in ZE compares to that of British English.
3. To establish whether the variation with articles, modal verbs and DMs and the associated features as outlined in Section 1.6 in ZE shows innovation.
4. To determine the stage of adult use of these features when considering Schneider’s (2003, 2007) DyM of postcolonial English.

In order to do the above, a corpus of Zimbabwean English was compiled using texts from a variety of genres. A summary of the texts is provided in Section 4.4.2.4. The ZE corpus was compared to the ICE-GB (cf. Section 4.4.2.5) in order to determine whether there are variations in the frequency and functions of DMs.

1.9 Research design

This study employed a corpus linguistics method, which integrates qualitative and quantitative research designs in order to get a better understanding of the research problem (Biber, Conrad & Reppen, 1998). Using the quantitative method, data was gathered, queried for frequency of items and presented numerically. Data was also compared to the ICE-GB. A qualitative functional analysis was done on articles, modal verbs and DMs (*so*, *well* and *but*). Biographical and language background questionnaires (extralinguistic data) were used to confirm participants’ L1 and L2 and to get information that was used during data analysis. Data was collected from Shona L1 speakers who used English as a L2. Semi-structured interviews were used to guide the conversation in order
to prompt participants to speak. Additionally, part of Marungudzi’s (2016a) corpus was used.

Using a corpus of 356,007 words, the study investigated the features and characteristics of ZE and focused on articles, modals and DMs, so, well and but. Articles and modals were chosen for analysis because the eWAVE lists possible ways in which new varieties of English differ from StE varieties (cf. Section 1.6). In the case of DMs, some researchers report on variations in their use in WE (e.g. De Klerk, 2005a, 2005b; Müller, 2005; Algouzi, 2014; Vickov and Jakupcevic, 2017; Unuabonah, 2019; Aşık & Cephe, 2013). By examining DMs this study attempted to ascertain their frequency and functions in the ZE corpus.

Sketch Engine, a commercially available concordancer (https://www.sketchengine.eu/) was used to analyse the corpus. The corpus was queried for the features according to Kortmann, Lunkenheimer and Ehret’s (2020) list of morphosyntactic variation on articles and modal verbs. In addition, the corpus was analysed for frequency and functions of articles, modal verbs and selected DMs so, well and but. Results from the questionnaires were incorporated in the analysis. In order to determine whether the observed variations were statistically significant, a log likelihood test was used (cf. subsection c in Section 4.4.2.6).

1.10 Key terms

In this section, a list of key terms used in this thesis is provided in alphabetical order. The terms are discussed in detail in chapter 2.

Innovation: In WE, if a linguistic feature has gained acceptance and is used within a community, then it is called an innovation. This is seen from the definitions supplied in literature. According to Bamgbose (1998: 2), “an innovation is seen as an acceptable variant”. Bamgbose notes that there is no clear-cut distinction between an innovation and an error. Another definition given is that, “an innovation in the outer circle, refers to the linguistic formations which are contextually and/or formally distinct from language use in the inner circle” (Kachru, 1985: 18). “To assist in deciding if a linguistic feature is an innovation, an appeal may be made to an external standard or the decision may be based entirely on internal factors” (Bamgbose, 1998: 2). Authoritative, codification, geographic,
acceptability and demographic factors are factors that can be used to determine innovation status.

**Linguistic proficiency**: Brière (1971) defines proficiency as the degree of competence or how capable an individual is in speaking a given language. Some of the contentious issues include the degree of competence and the accuracy of different measuring instruments that are used to measure proficiency. According to Llurda (2000: 88), “proficiency is a term that suggests variability and it has traditionally been related to measurement and testing in L2 teaching and learning”.

**Nativisation**: One of the features of WE is nativisation which occurs when English is used by non-English mother language speakers (Fishman, 1992; Schneider, 2003). “Nativisation is the process through which a language is accommodated and adapted to its speakers and their circumstances” (Gough, 1996b: xx)\(^{17}\). Nativisation of English was once considered to be a sign of incompetence in the language but this attitude is slowly changing with studies focusing on creativity and ownership of English by diverse linguistic groups (Ndhlouv, 2011).

**Norm**: Norm orientation is a basic facet in L2 teaching and learning. According to Kruger and Van Rooy (2017: 207) in addition to editors, “norm orientation is also enforced by the education system and specifically by feedback that learners receive on their writing, which may further reinforce particular constructions and prevent the entrenchment of others”. Trudgill and Hannah (2017) state that StE is the norm in teaching and learning in ESL and EFL classes. They believe that there are two main Standard varieties of English namely BrE and American English.

**World Englishes**: “WE” is a cover term used to refer to all the varieties that exist in countries where English is used as a L2 or additional language (Kachru, 1992). For example, Trudgill and Hannah (2017) refer to Indian English (used in Pakistan, India, Sri Lanka, and Bangladesh), Singapore English and Philippines English. Hickey (2017) mentions Irish English as a type of WE and notes that Irish English has features that make it identify more with L2 varieties of English than with L1 varieties.

---

\(^{17}\) This sentence appears in the “Introduction” part of the *Dictionary of South African English on Historical Principles* (Gough, 1996b).
**Zimbabwean English:** Within the WE framework, ZE is a type of African English, which is a type of WE. There is no consensus on the use of the term “ZE” in literature. For example, Kadenge (2009) and Mlambo (2009) use the term to describe the variety of English that is spoken in Zimbabwe as a L2. Mlambo argues that although ZE has not been institutionalised, there are a variety of features that can be identified as ZE features. Makoni (1993) disagrees with adopting the term and classifies the English spoken in Zimbabwe as an interlanguage, a learner language that is developed by L2 speakers. According to Makoni (1993: 97), “the environment within which the L2 is acquired and the influence of the mother tongue”, affects the nature of an interlanguage. Literature on SLA refers to an interlanguage as “a product of a number of major processes, such as language transfer, transfer of training, second language learning strategy, second language communication strategy, overgeneralisation of target language structures and other minor ones, such as hypercorrection, spelling pronunciation, etc.” (Sridhar and Sridhar, 1992: 97).

Due to the many languages spoken in Zimbabwe, pinpointing a universal form of ZE is difficult (Kadenge, 2009). In this study, the term “ZE” is used to refer to the English that is spoken as a L2 in Zimbabwe by Shona mother tongue speakers. “However, the existence of a single unified ZE variety is debatable and disputable within the country in which there is a wide spectrum of usage of local languages” (Kadenge, 2009: 158).

**1.11 Structure of the thesis**

This thesis comprises of eight chapters. A comprehensive literature review is carried out in chapter 2 in order to contextualise the study. Chapter 3 discusses corpus linguistics and situates corpora in the context of WE. The research design and the methodology used in this study are unpacked in chapter 4. Chapters 5 and 6 deal with articles and modal verbs respectively by explaining the proposed variations and then analysing data for the possible variations. The ZE corpus and the ICE-GB are compared for variations in the frequency and functions of DMs so, well and but in chapter 7. Finally, chapter 8 provides a summary of the findings, a discussion about the limitations as well as suggestions for further research.
Chapter 2

Theoretical framework

2.1 Introduction

This chapter focuses on the literature review of WE in order to better understand how the subject of WE has evolved over the years and to situate ZE within the WE paradigm. Firstly, a summary of how WE are conceptualised in this study is provided. Then a brief history of the English language is given in order for us to comprehend the stages that English has gone through up to the present day, including how English has spread to different cultures and contexts. In addition, a historical overview of WE is provided in order to better understand the differing perspectives on variety formation, the misconceptions about users and uses of English and the attitudes of people towards WE. Since WE are associated with SLA and, considering that WE users are usually L2 users of English, it is important to investigate the link between SLA and WE. Thereafter, different models proposed for WE are highlighted. In order to inform on the theoretical framework of this study, a detailed overview of the DyM is provided, together with a summary of the studies that have utilised the DyM. The discussion will also highlight some of the criticism levelled against the DyM. Terms that help to better understand the subject of WE in general and ZE in particular are defined.
2.1.1 How World Englishes is conceptualised

How WE are conceptualised in this study is summarised in the following diagram where there are different perspectives in variety formation.

![Diagram showing the relationships between Innovation/error, Proficiency, and Variety/register.]

*Figure 2. The foundational concepts that are involved in the WE*

As shown in figure 2 above, nativisation, a process which occurs when English is used in different contexts and settings than the StE contexts, is a central process in the study of WE. This is because the process of nativisation\(^{18}\) is linked to proficiency and normative processes. Innovation leads to the formation of new varieties of English (Kachru, 1992; Schneider, 2003, 2007). Innovation (cf. Section 2.8.7) is also a result of nativisation. Whereas WE researchers see the different ways in which English is used in different parts of the world as innovation, some scholars use the term “error” to refer to any deviations from the norm (which is StE spoken by L1 speakers).

WE are learned as a L2 in different environments (usually at school) with learners having very limited or no interaction with native speakers (Kachru, 1992). Although teachers are expected to teach learners the norm, some of them may be L2 users of English themselves meaning that the input that learners receive is that of a L2 speaker of English.

---

\(^{18}\) Nativisation is discussed in detail in Section 2.8.8.
This leads to learners having different proficiency levels in different registers. Therein lies the debate over whether English spoken differently in different parts of the world can be considered as new varieties (WE). Proponents of the WE (McArthur, 1987; Kachru, 1992; Schneider, 2003, 2007; Onysko, 2016) believe that different contexts in which English is spoken throughout the world should be taken into consideration in order to better understand global diffusion of English. The following section deals with various purposes and functions of English.

2.1.2 Purposes and functions of English in the WE context

English has different purposes and functions in different countries around the world. In the context of WE where it is used as a L2, English is used in the higher domains such as education, government departments and business (Kadenge, 2010). As mentioned earlier in Section 1.3, one needs to pass English in order to proceed to tertiary education. In contrast, indigenous languages are used in low domains such as private communication. One is not required to have passed an indigenous language in order to be considered for tertiary education or for vocational training.

The fact that ESL learners will use English together with other languages (like the L1) means that English’s functions are different from situations where only English is used as a native language. “A native language is one spoken as the primary language of the community in question, that is, learned first in childhood, used exclusively or at least preferentially in all dealings public and private and supported again, probably exclusively - in statute and education” (Nelson, 1992: 329). Nelson explains that NNVE writers incorporate elements from their L1 in their style of writing and the way they arrange discourse.

Having explored the functions of English, the following section reviews how English came to be in a position that it is today.

2.2 English language: A brief history

The development of English as a language can be traced back to the 5th and 6th century when English grew out of Germanic dialects spoken in Britain (Trudgill & Hannah 1994; Mesthrie & Bhatt, 2008). Freeborn, French and Langford (1993) refer to English and German as cognate languages because they originated from a similar source. They are
both West Germanic languages. Mesthrie and Bhatt (2008) are of the opinion that between the years 450 and 1100, multiple standards of English existed wherever it was spoken and English also borrowed from other languages that it came into contact with. Few people in England spoke English with most people speaking Welsh and Cornish. According to Trudgill and Hannah (2017), the geographical and demographic growth of English beyond the British Isles started in the 17th century. The 1800s saw English expanding to the Southern Hemisphere. According to Trudgill and Hannah (1994), different varieties began to emerge when English was spreading to different parts of the world. Strevens (1992) discusses the changes that occurred between year 1750 and year 1900 in the colonised countries that helped shape the English language. Firstly, the number of people who spoke English as a L1 increased. Secondly, the end of colonialism in countries such as the United States and Australia "greatly reinforced the degree of linguistic difference" (Strevens, 1992:30). In addition, indigenous people began to learn English for use at the workplace. Therefore, the functions that English served in a society changed to higher functions in education, industry and trade, the media, government and the workplace (Makoni 1992).

Strevens (1992) observes that English has different varieties and is spoken in different ways by both native and non-native speakers and explains that, until the 18th century, there existed only the native speaker variety of English, that is, the variety used in Britain. American English came into existence during America’s fight for independence. According to Strevens, BrE and American English are the two major branches from which different varieties of English emerged worldwide. In the “derivational and linguistic sense”, Strevens (1992: 32) mentions that the English spoken in Canada, Puerto Rico and the Philippines is related to American English. Other varieties of English spoken in countries such as Australia, India and New Zealand are related to BrE.

The building of schools where education was offered in English aided the spread of English to non-native speakers between 1900 and 1950 according to Strevens (1992). He further states that when many former British colonies gained independence, the status of English changed from being a vehicle of superiority to being a tool of upward mobility and a status symbol.

© University of Pretoria
Over the course of the next millennium, English became the primary home language of the British Isles, with languages like Welsh and Irish in receding use today and others, including Cornish, have been completely replaced by English (Milroy, 2007). Strevens (1992) and Trudgill and Hannah (1994) note that colonialism played a major role in the spread of English when “explorers, merchant adventurers, buccaneers, traders, settlers, soldiers and administrators went out from Britain to begin settlements and colonies overseas” (Strevens, 1992: 29). Some of the initial areas where English spread to include North America, Bahamas, the Caribbean and Ireland.

Gorlach (1995) notes that the expansion of English started after the language had been standardised in its written form and to a greater extent the phoneme structure of its spoken form. He further states that American English became a Standard variety in the 19th century, Australian English and New Zealand English in the 20th century.

McArthur (1998) observes the uniqueness of English due to its spread and use for different purposes and in various domains throughout the world at the end of the second millennium. Kachru (1992) views English as a universal language, which has its roots in colonialism in many countries. In most former colonial countries, English is now spoken as a L2, “the other tongue” (Kachru, 1992). English has spread from its earliest stages during the 5th and 6th century (Trudgill & Hannah, 1994), when it was spoken by just a miniscule number of people, to the present day where English has a huge number of speakers from different linguistic backgrounds.

In the global context, English is used for different purposes by people in different parts of the world. Some people speak English as their first language, others speak it as their L2, third, or fourth language Graddol (2000).

In the process of its expansion into new territories, English has evolved to mirror the different cultures and linguistic backgrounds of its speakers. Graddol (2000) predicted that this process would continue in the 21st century. This is a prediction that seems to have come true, given the continued spread of English into different territories where there are few or no native speakers of English. The impact of the global spread of English can be discussed by looking at the different varieties that have emerged. Since the spread of English has been linked to the emergence of new varieties (cf. Kachru, 1992; Schneider,
2003, 2007), the following section describes the history of WE.

2.3 A historical overview of world Englishes

McArthur (1998) cites the early 1980s as the time when the term “WE” began to be used. He further states that the term is used to incorporate the diverse environments and cultural contexts in which English is spoken around the world and seems to suggest that English spread into different parts of the world due to colonialism. The plurality of the term “WE” mirrors the different types of English spoken throughout the world. Hence, the term encompasses the different types of English. Serious interest in new English varieties by researchers started in the early 1960s and by the mid-1970s, the notion of World Englishes was strongly established (e.g. Kachru, 1965, 1975, 1976; Smith, 1976; Smith & Rafiqzad, 1979). The early years of the study of WE were characterized by lack of theories and methodological frameworks (Schneider, 2003). Thus, studies concentrated on describing the features of the new varieties of English. Due to the novelty of the field, one finds different terms used to refer to the new varieties of English. Some of the terms used by researchers are: “New Englishes”, “World Englishes” and “Global Englishes”. The similarity in these terms lies in their use of the plural “Englishes, meaning ‘varieties of . . .’” (Schneider, 2003: 234). Schneider uses the term “NE” in the DyM.

English has spread globally and has broken international boundaries because of its use, even by people who consider it as the root cause of the decline in the use of indigenous languages (Ndhlovu, 2011). With the growing demand to learn English worldwide, it is inevitable that L2 English teacher numbers will increase (Moussu & Llurda, 2008). According to Orosco and Hoover (2009) sometimes there is pressure on teachers to produce results leading to English language teachers wrongly identifying learners (mostly L2 speakers of English) as needing special education. Orosco and Hoover advocate for a better understanding of these learners’ cultural and linguistic background in order to assist them with integration.

The diffusion of English into different cultures and contexts has led to the emergence of new English varieties. According to Kachru (1992) and Thomason (2001), the varieties emanating in different countries are a sign of linguistic innovation (cf. Section 2.8.7), a feature of WE and a result of language contact. This view is also shared by Paradowski and Jonak (2012: 131) who assert that “linguistic creativity is a manifestation of communities’
and cultures' innovativeness.” Chisanga and Kamwangamalu (1997) are of the opinion that lexical transfer, syntactic transfer and semantic extension are the procedures that lead to innovations by L2 speakers of English, which aided in the creation of the above mentioned Englishes (or varieties).

At the turn of the new millennium, Graddol (2000) remarks that English is undergoing some form of transformation as evidenced by the increase in the number of non-native speakers of English. The emergence of new varieties of English, which have been termed “WE' shows the global expansion of English. Kachru (1992) estimates the number of non-native users of English to be two billion.

Braj Kachru is one of the leading scholars in WE. According to Kachru (1992), L2 varieties of English have developed in communities where English is in contact with other languages leading to nativisation. He notes that English adapts to local contexts and adopts new features leading to the emergence of new varieties. He advocates for “a deeper understanding of the forms and functions of English within different sociolinguistic and cross-cultural contexts” (Kachru, 1992: 6). Kachru (1985, 1988) divides different speech communities where English is spoken into Inner Circle, Outer Circle and Expanding Circle. The three concentric circles are discussed in Section 2.4.3.

Regarding WE, Schneider (2011) observes that even though the English that is spoken in various parts of the world seems to be different, people are able to understand each other in these countries. Another claim made is that, in spite of the diversity of communities, cultures and languages, there are structural and sociolinguistic similarities in the NE. Schneider further states that migration, borrowing and variability are the main factors causing the emergence of different varieties of English worldwide. He also compares the variability that characterises NE to the way StE has Germanic roots and also borrowed from languages such as Latin and French. Just as language contact has influenced StE in the past, so is contact with indigenous languages influencing the WE. Another observation made by Schneider is that the sovereignty gained by former colonies should have resulted in the abandonment of English as an official language. Instead, the opposite happened with English retaining its official status in most countries. He mentions that Tanzania is an example of a country, which promoted an indigenous language (Swahili).
The belief that English is a means to gain power, prestige and economic prosperity has boosted the number of people who learn English as a L2 or additional language. “English has grown local roots in many countries and cultures. It has become nativised, i.e. it has developed indigenous forms and these local ways of speaking English have been adopted by many speakers as symbols of regional identities” (Schneider, 2011: 53). That is why there is a need to study whether there are innovation features in English as a L2 in Zimbabwe.

Schneider (2011) views the diversity that is evident in the way people speak English worldwide as proof that language contact between English and indigenous languages has played a role in the development of different varieties. Variability is given as a principal characteristic of languages and cultures by Schneider. This variability “generates novelty, socially successful communication and joint development - things we all need in a globalised world” (Schneider, 2011: 230).

WE operate in different contexts from StE (Kachru, 1985; McArthur, 1998). One of the differences is that WE are spoken by people who use it as a L2 or additional language. These speakers will have achieved varying degrees of linguistic proficiency. Linguistic proficiency is discussed in Section 2.8.6.

The teaching and learning of English in countries where it is not a L1 means that there are bound to be differences in the way English is acquired (Kachru, 1985; Richards, 1972). English learners may not be exposed to L1 speakers since the bulk (if not all) of the English teachers may be L2 speakers themselves.

A recent, elite form of L1 English usage in predominantly L2 contexts is emerging. Schneider (2011) describes some environments where English is a L1 of children from wealthy or educated families in Africa and Asia. These children grow up using English both at home and at school. These varieties are likely to be closer to StE than the indigenised varieties used by L2 speakers in the same contexts.

Still unpacking the historical overview of WE, the following section looks at different perspectives on variety formation that have emerged over the years.
2.3.1 Differing perspectives on variety formation

As discussed in the previous section, the use of English in different contexts from when it is used in traditional contexts (StE context) has resulted in the emergence of new varieties of English. There is no consensus on a number of issues arising from the global spread of English that have implications for WE. According to Kachru (1985), scholars do not agree on issues like (i) codification, a process whereby a language attains official acceptance because it has reference books such as grammar books, dictionaries and user guides (Schneider, 2003); (ii) innovation, referring to a linguistic feature which has gained acceptance and is used within a community; (iii) de-Englishisation (also referred to as nativisation in literature), which involves the extent to which English adjusts to the new environment in which it is spoken; and (iv) creativity, which deals with how appropriate different aspects of language are used by non-native speakers (Kachru, 1985).

Kachru (1985, 1992) and Nelson (1992) advocate for studies in the diffusion of English into different territories and cultures to consider the context in which English operates in the new territories. Sydney Greenbaum commented\(^{19}\) in Kachru’s (1985: 32) article saying that “if the educated varieties of the ‘Outer Circle’ assume the status of national standards without reference to the international norms of the Inner Circle, will they diverge too far to remain part of the international Standard English?” To put Greenbaum’s comment into context, Kachru (1985, 1988) used the term “Inner Circle” to refer to countries where English is used as a native language. “Outer Circle” countries are the countries where English is spoken as a L2. Countries where English is spoken as a foreign language belong to the “Expanding Circle”.

Still on the issue of the Inner, Outer and Expanding Circles, Kachru (1992) distinguishes between performance varieties and institutionalised varieties when discussing NNVE. He positions performance varieties like Japanese English and Iranian English in the Expanding Circle. Performance varieties have limited functions, which include tourism and commerce. From the discussion about functions of English\(^{20}\), it can be said that performance varieties

---

\(^{19}\) Sydney Greenbaum’s comment appears in a book that is a collection of papers and commentaries from a week-long colloquium in celebration of the 50th anniversary of the British Council. The commentary was included below Kachru’s article.

\(^{20}\) Cf. Section 2.1.2 for a discussion about functions of English.
have a lower function in the countries they are spoken. On the contrary, institutionalised varieties are found in the Outer Circle and have higher functions and their style and registers have been nativized and codified (Kachru, 1992).

Colonialism introduced English in many countries (Moag, 1992; Kachru, 1988, 1992; Wei, 2000; Jenkins, 2003; Mesthrie & Bhatt, 2008). After independence,

English has managed to stay, not only in formal and official functions; it has indigenised and grown local roots. It has begun to thrive and to produce innovative, regionally distinctive forms and uses of its own, in contact with indigenous languages and cultures and in the mouths of both native populations and the descendants of former immigrants, making even deeper inroads into local communities.

(Schneider, 2007: 1)

WE are usually labelled according to the country they are used, for example, Irish English, Indian English, Canadian English, South African English, Singapore English, Zimbabwean English etc. (Strevens, 1992; Richards, 1972; McArthur, 1998). McArthur remarks that one cannot provide a definite timeline for the evolution of modern English from its West Germanic origins. In the same way, a distinct timetable cannot be provided about when the new varieties of English that are emerging in different parts of the world will become accepted as varieties with their own reference material like grammar books and dictionaries. We now look at codification as an important process in variety formation.

2.3.2 Codification

For a language to gain official acceptance, it should have reference books such as grammar books, dictionaries and user guides (Schneider, 2003). The process of codification is important because speakers of the language will refer to the reference material for information about usage. According to McArthur (1987) some of NNVE are in the process of codification.

The next section deals with misconceptions concerning the users and uses of English.
2.3.3 Misconceptions about users and uses of English

Kachru (1992) draws attention to one of the misconceptions that L2 speakers of English mainly learn English in order to interact with L1 speakers. He argues that in some instances, this is true but in most countries in the Outer and Expanding Circle, indigenised varieties are used by L2 English speakers to interact among themselves. A discussion about the Inner, Outer and Expanding Circle is given in Section 2.4.3.

Kachru explains that, although in some instances English is learned for a better understanding of British and American cultures, the situation is different in the Outer Circle because local cultures and traditions are reflected in the Outer Circle English. Another fallacy is that “the native speakers of English as teachers, academic administrators and material developers provide a serious input in global teaching of English, in policy formation and in determining the channels for the spread of the language” (Kachru, 1992: 358). Kachru claims that the global diffusion of English is not spearheaded by the native speaker. Due to financial limitations, many countries, especially those in the developing world, rely on L2 speakers to teach English in schools. In addition, there are not enough native-speaker teachers in the world to teach English to all the non-native speakers currently in classrooms, not to mention that many native speaker teachers are not keen to relocate to other parts of the world. Therefore, theoretically, even if finances were not a problem, there would still not be enough people to perform the function of a native speaker as a teacher of English. This means that the input received by learners may not be the Standard English variety.

According to Kachru (1992) the diverse nature of English is sometimes taken as a sign of linguistic decay and as a sign of failure on the part of ESL teachers. This has led to any deviation from Standard English to be labelled as “errors” even though the functions of English are different in L2 contexts. As a theoretical point of departure these errors could form part of the interlanguage as part of the second language acquisition process, again outlining that these bilinguals will fall on various points of the continuum toward being proficient speakers of the language dependant on the in terms of the DyM and Kachru’s model.
In addition to misconceptions about users and uses of WE, the attitudes of people towards WE need to be taken into consideration. This is because, as mentioned by Schneider (2007), acceptance of a new variety is important for it to be considered as a norm (cf. Section 2.8.9).

**2.3.4 Attitudes towards world Englishes**

In order to better understand the history of WE, one needs to also consider the attitudes of people towards WE. These attitudes have implications for the acceptance of WE as explained above. Lack of acceptance has been one the reasons why L2 varieties of English are not recognised in many communities (Nelson, 1992; Strevens, 1992). Kachru (1992) is of the opinion that the NNVE have generally not been accepted by native speakers. The nativisation process, which leads to NNVE is usually considered to be a deficiency in acquiring English. According to Gough, (1996a) and Marungudzi (2016a), although there is variability in the way that English is spoken globally, there is reluctance to accept these emerging varieties due to various reasons. Some of the reasons are highlighted by McArthur (1998) as follows:

(a) **A pessimistic view**
According to McArthur (1998), pessimists are fearful that StE will disappear if variability is allowed in the way English is spoken worldwide. They are worried that English will suffer the same fate that Latin has experienced, that is, the decline in teaching and use of Latin. Therefore, pessimists advocate for the promotion and use of StE. Despite some of the L2 speakers of English using the localised varieties, most of them do not accept the existence of such varieties (Kachru, 1992). “Accent” is still being used as a measure of one’s mastery of English and “even when the non-native models of English are linguistically identifiable, geographically definable and functionally valuable, they are still not necessarily attitudinally acceptable” (Kachru, 1992: 66).

(b) **An optimistic view**
Acceptance of WE is a characteristic of optimists (McArthur, 1998). People who hold the optimistic view recognise that WE have emerged in different parts of the world where English is spoken as a L2 or an additional language. According to McArthur, unlike pessimists who view the decreased use of Latin as a disaster, optimists believe that Latin has produced a lot of Romance languages such as Spanish, Portuguese and French.
(c) A neutral and pragmatic view

Neutralists believe that languages do not stay the same but change all the time (McArthur, 1998). Therefore, “investigating the background of Latin, English and comparable languages might clarify our perceptions of what tends to happen to prestigious, widely used languages” (McArthur, 1998: 182).

Still on the subject of attitudes towards WE, Mesthrie and Bhatt (2008) draw attention to opposing views from two of the leading scholars of English in the global context. Quirk (1985) initially acknowledges the emergence of divergent varieties of English. Later on, he advocates that the teaching and learning of English should be based on Standard British or American English and that other varieties should not be tolerated. Quirk (1990) remarks that the goal of students should aim for the acquisition of the Standard variety of English.

Quirk also sees the native speakers as important in teaching English to non-native speakers. Despite Quirk’s assertion that teachers of English worldwide should be native speakers, the reality of the situation is the opposite. As he seems to acknowledge, many teachers in countries where English is taught as a L2 are not native speakers and “the English of the teachers themselves inevitably bears the stamp of locally acquired deviation from standard language” (Quirk, 1990: 8). Mesthrie and Bhatt (2008) add to discussion by saying that financial resources are limited in many countries where English is learned and taught as a L2. This makes the hiring and paying of native speakers of English a burden for authorities. The subject of L1 and L2 teachers of English is discussed further in Section 2.8.10.

Kachru (1991) does not agree with Quirk’s view that the success in mastering English should be measured by one’s ability to speak StE. He advocates for the acceptance that English has spread and diversified to communities around the world leading to the emergence of different varieties. According to Kachru (1991: 11), the contexts in which English is used worldwide differ and “the sociolinguistic realities of identity, creativity and linguistic and cultural contact” are evident in the WE.
Kachru also proposes actions that non-native speakers (NNS) of English can take in order for the attitudes towards NNVE to change. These include, not considering English as a colonial heritage, acceptance of literature produced by local writers and accepting the local variety and not considering the variety as deficient.

Although Standard American English and Standard BrE were the norm in countries such as Canada, Australia and New Zealand, these countries have now established their own standards (McArthur, 1998). McArthur seems to suggest that despite the controversy surrounding the acceptance of and attitudes towards WE, the same standardisation process that Canada, New Zealand and Australia went through are evident in countries such as Singapore, Nigeria and Malaysia.

Due to the importance of WE, with the focus on an educational function as well as contact with indigenous languages and cultures and the teaching and learning of English in countries where it is not a L1, there are bound to be differences in the way English is acquired. Owing to these above-mentioned factors it is important to outline WE in terms of SLA, with specific focus on the African continent and Zimbabwe as, as well as bilingualism and language proficiency in the following sections.

2.3.5 Second language acquisition and world Englishes

Since most learners first encounter English as a second or additional language in the school environment, (Kachru, 1992) it is important to discuss the link between SLA and WE. The majority of research into SLA uses “L2” as a cover term to refer to the second, third, or any language learned after the first one (Ellis, 2015). In this study, the term “L2” will be used to refer to any additional language acquired after the first language. Despite the fact that most WE emanate from SLA environments, Mesthrie and Bhatt (2008) note that research into SLA refers to NNVE as interlanguages. The term “interlanguage” was coined by Selinker (1972) to refer to learner languages. Kachru (1992: 358) explains that one of the fallacies in SLA is “that the international non-native varieties of English are essentially “interlanguages” striving to achieve “native-like character”. Kachru questions the applicability of the hypothesis to Outer Circle varieties of English and advocates for a re-evaluation of the assumption that NNVE are interlanguages. According to Sridhar (1992) the development and testing of SLA theories has not adequately utilised data from NNVE. Sridhar and Sridhar
(1992) explain that the context in which English is used throughout the world should be considered when discussing ESL in non-native contexts.

In addition to the context mentioned above, integrative and instrumental motivations are important for SLA. A learner may be motivated by the need to associate with people who speak a certain language (integrative motivations). Where a language is acquired as a tool for communication, the motivation is instrumental. Another dimension that ought to be considered is the acquisition versus learning distinction in SLA. Krashen (1982) uses the terms “acquisition” to explain the process where a person acquires a L2 subconsciously and “learning” when a L2 is learned via formal instruction. When the ability to speak a language develops gradually, due to language use in natural communicative environments, Yule, (2017) refers to this process as acquisition. Yule adds that acquisition usually occurs without formal instruction from teachers. In contrast, learning occurs consciously when a person gets formal instruction. The majority of people who speak a L2 learned it in a formal setting at school (Yule, 2017).

Sridhar and Sridhar (1992) assert that understanding the different motivations for SLA may help researchers to understand NNVE better, since some scholars regard NNVE as interlanguages due to the fact that the varieties are different from Standard varieties of English. In the school environment, a problem emanating from not taking the cultural and linguistic diversity of learners into consideration when studying SLA and NNVE is that in some instances, learners are misdiagnosed as having learning disabilities (Orosco & Hoover, 2009).

As noted in Section 1.1, age, acculturation, attitude and motivation, learning style, native language proficiency, community and family are some of the cognitive and environmental factors that influence SLA (Orosco and Hoover, 2009: 41). Due to the nature and scope of this thesis, it is important to outline that these factors influence SLA but it falls outside the scope of this thesis to discuss each of them in detail. For a detailed overview of each of these factors, see the following: (Gardner & Lambert, 1959; Cook, 2008; Lightbown & Spada, 2006; Ellis, 2008; Brown, 2014; Yule, 2017).
With more interest being shown in WE research over the years, models have been developed to account for L2 varieties of English. An outline of the models is provided in the next section.

2.4 Some models proposed for World Englishes

Some researchers have proposed different models to explain WE. Some of the models are described in the subsections below. These include the circle of world English, the circle model of English, the three circles of English and the language contact typology of world Englishes.

Onysko (2016) emphasises that a variety of models attempt to explain different features of WE using varying approaches. He adds that research on WE models has evolved from focusing on Standard versus non-standard usage of English to models that focus on how different varieties develop, the functions of English in different environments where it has spread globally and developing theories to account for the variations in WE.

2.4.1 McArthur- Circle of world English (1987)

The circle of world English is designed in the shape of a wheel, which consists of spokes, a hub and rims (McArthur, 1998). StE is at the centre and regional varieties like American English, African English and Irish English come up next. After the regional varieties, there are sub-varieties such as Singapore English, Aboriginal English, Black English Vernacular, etc. (McArthur, 1987). The circle of World English is shown below.
2.4.2 Görlach - Circle model of English

"Görlach's circle model of English whose wheel of English rotates around the hub called “International English" (Onysko, 2016: 198) is shown in the figure below.
From figure 4 above, it can be noted that International English is the main branch at the centre, with sub-branches increasing outwards.

2.4.3 Kachru- Three concentric circles of English (1985, 1988)

Kachru (1985, 1988) visualises the spread of English through his concentric circles of English. He took into consideration the fact that the different contexts in which English diffuses into leads to the accommodation and adaptation of English to suit the different contexts. Below is a diagram showing the three concentric circles of English.
Another factor considered by Kachru is how English is used in different sociocultural contexts from when it was used before spreading throughout the world. This is true especially in situations where English is spoken by L2 speakers who “have access to only a subset within the patterns and conventions of cultures which English represents” (Kachru, 1988: 3).

Three circles namely the Inner Circle, the Outer Circle and the Expanding Circle make up the concentric circles of world English. Countries such as Canada, Australia, UK, the USA and New Zealand where English is the L1 belong to the Inner Circle. Former British colonies such as Kenya, Zimbabwe, India and Nigeria are found in the Outer Circle. “The
major features of this circle are that (a) English is only one of two or more codes in the
linguistic repertoire of such bilinguals or multilinguals and (b) English has acquired an
important status in the language policies of most of such multilingual nations” (Kachru,
2006:142). Indonesia (colonised by the Dutch and never had English in widespread use)
and Taiwan (colonised by the Chinese) are examples of countries in the Expanding Circle.
According to Kachru, due to the spread of English into territories where other languages
and cultures already exist, there is bound to be language contact. Countries like Israel
and Japan, where English is used as a foreign language are classified under the Expanding
Circle.

The remarkable story about the spread of English has also brought plenty of problems.
“With this diffusion, naturally, comes scores of problems concerned with codification,
standardisation, nativisation, teaching and description and, of course, a multitude of
attitudes about recognition of various varieties and subvarieties” (Kachru, 2006: 244).

There is no clear-cut distinction between the Inner Circle since they share a number of
features (Kachru, 1985). In addition, countries do not stay in the same circle but they can
move to another circle. Kachru (1985) labels the varieties of English found in the Inner
Circle as norm-providing varieties. This is because they are native speaker varieties and
are normally used in teaching and learning. Norm-developing varieties exist in countries
belonging to the Outer Circle. In these countries, English has undergone the process of
nativisation, leading to the development of WE. Varieties in the Expanding Circle are
labelled norm-depended varieties.

Mesthrie and Bhatt (2008) credit Kachru’s three concentric circles of English with playing
an influential role in classifying English worldwide. Kachru is also credited with
promoting the use of the plural term “Englishes” to show the pluricentric nature of WE
(cf. Onysko, 2016). Kachru’s (1985, 1988) model is not without criticism. For example,
Quirk (1990) does not believe that there are institutionalised varieties of English in non-
native English-speaking countries. He further notes that the aim for learners should be
StE. Although acknowledging the merits of the concentric circles of English, Jenkins
(2003) criticises Kachru’s reliance on the genetics and geography in his classification of
English-speaking countries rather than how speakers use English. Another issue raised
by Jenkins is that there is an overlap between the Inner, Outer and Expanding Circle.
 Countries like Argentina and Belgium seem to be transitioning from EFL to ESL countries. Jenkins also criticises the model for its failure to acknowledge the differences that exist between individual speakers of English as a L1, ESL and EFL and the different varieties that exist within countries.

### 2.4.4 Onysko (2016) Language contact typology of world Englishes

Onysko (2016) proposed the language contact typology (LCT) to provide a theoretical explanation about the diverse nature of English globally. WE are modelled from the language contact perspective. In Onysko’s model, language contact is the basic tenet of WE.

![Figure 6. The Language Contact Typology of world Englishes (Onysko, 2016)](image)

“The notion of language contact can refer to the individual performance of a single speaker and to the collective performance of entire speech communities” (Onysko, 2016: 206).

According to the LCT, there are five main groups of WE. These include global Englishes (e.g. the influence of English on Japanese and the influence of English on European languages); learner Englishes (e.g. Chinese English and Italian English); English in
multilingual constellations (e.g. Irish English and Welsh English); English-based pidgins and creoles (e.g. Ghanaian Pidgin English and Aboriginal Pidgin English) and Koine Englishes (e.g. standardised varieties of British, American and Canadian English. Onysko advocates for a consideration of the historical conditions of the main contact environment adding that:

If the mutual historical conditions are considered, the typology becomes dynamic and allows for flexible categorisations and overlap among the categories and its members. In addition, the conditions of contact can be different among speakers of the same variety, which calls for internal differentiation in the categorisation.

(Onysko, 2016: 215-216)

As seen from the discussion above, the LCT makes language contact the central issue in the description of WE. This is a deviation from other models, which distinguish between native and NNVE (e.g. Kachru's three concentric circles of English and McArthur's circle of world English described above). The LCT is not hierarchical and no single variety is more important than the other.

2.5 The Electronic World Atlas of Varieties of English

Although the aforementioned models are used in the description of NE varieties, the eWAVE is an important resource for the description of morphosyntactic features of different varieties of English. As explained in Section 1.6, the electronic eWAVE that was recently updated to eWAVE 3.0, is an interactive database on morphosyntactic variation in spontaneous spoken English. It maps 235 features from a dozen domains of grammar in now 51 varieties of English (traditional dialects, high contact mother tongue Englishes and indigenised L2 Englishes) and 26 English-based Pidgins and Creoles in eight Anglophone world regions (Africa, Asia, Australia, British Isles, Caribbean, North America, Pacific and the South Atlantic) (www.ewave-atlas.org).

A number of research questions can be answered by utilising the eWAVE. Investigations can be made on determining the most common features, the number of varieties that share a certain feature globally, the prevalence of a feature within a variety, determining variations in grammar between varieties among others. The eWAVE database continues to expand as
new varieties of English are added to it as shown by the inclusion of Croker Island English. The following table explains features that will be examined in this study.

Table 1. Variations reported in literature (source: Kortmann, Lunkenheimer & Ehret's (2020)

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Description of feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Use of definite article where StE has indefinite article.</td>
<td>I had the toothache; ...when they’d get the cold.</td>
</tr>
<tr>
<td>61</td>
<td>Use of indefinite article where StE has definite article.</td>
<td>A sun was shining.</td>
</tr>
<tr>
<td>62</td>
<td>Use of zero article where StE has definite article.</td>
<td>Did you get Ø mileage-claim for that trip? Ø main reason for their performance...</td>
</tr>
<tr>
<td>63</td>
<td>Use of zero article where StE has indefinite article.</td>
<td>Then he thought, what about getting Ø girl [to marry] from India?</td>
</tr>
<tr>
<td>64</td>
<td>Use of definite article where StE favours zero ([article].</td>
<td>He was appointed sales representative at the Nestlé, Ghana Ltd.; Do they keep the goats? Poor people were starved with the hunger.</td>
</tr>
<tr>
<td>65</td>
<td>Use of indefinite article where StE favours zero [article].</td>
<td>About a three fields.</td>
</tr>
<tr>
<td>66</td>
<td>Use of indefinite article one/wan.</td>
<td>They seen one ['a'] green snake tangled round a tree; Di uman sel wan mango.</td>
</tr>
<tr>
<td>67</td>
<td>Use of demonstratives for definite article.</td>
<td>That door bin close 'the door closed'.</td>
</tr>
<tr>
<td>121</td>
<td>Double modals</td>
<td>I tell you what we might should do.</td>
</tr>
<tr>
<td>123</td>
<td>Present tense forms of modals used where StE has past tense forms.</td>
<td>I wish that people in the world will ['would'] get educated; [...] they can ['might'] be wild, but they're human beings.</td>
</tr>
<tr>
<td>127</td>
<td>Non-standard use of modals for politeness reasons.</td>
<td>Must I make you some tea? Shall I ...; May you please lend me a pen.</td>
</tr>
</tbody>
</table>

The features explained in the table above were used as the basis for the description of ZE. Since this study is informed by the DyM of postcolonial Englishes as its theoretical framework, the following section gives a comprehensive description of the model.
2.6 The dynamic model of postcolonial Englishes

Schneider (2003, 2007) developed the DyM, “a coherent framework to explain the emergence of and thus the relationships among New Englishes” (Schneider, 2003: 234). As indicated earlier in Section 1.2, the DyM attempts to describe the five different stages of development in L2 varieties of English. According to the DyM, NE evolves in stages which are characterised by certain features and different phases [e.g. Fiji (stage 2), Hong Kong (stage 3), Singapore (stage 4), Australia and New Zealand (stage 5).

Schneider (2003) believes that although there are differences in the way English is spoken around the world, it developed due to similar sociolinguistic and language contact situations. Social identity is the principal idea in the DyM.

The following points are noted in the DyM:

1. The spread of English from native speaking countries has resulted in the emergence of NE. These NE pass through similar stages namely foundation, exonormative stabilisation, nativisation, endonormative stabilisation and differentiation.

2. The settlers and indigenous people's languages come into contact and influence one another.

3. The significance of the different stages in the DyM is that group identities will be recreated.

Schneider (2003) emphasises the importance of studying the sociohistorical contexts of NE since the different contexts account for the variations found in NE. One commonality is that the NE emerged from contact environments (e.g. contact between settlers and indigenous people). The stages and features characteristic of each stage are outlined below:
2.6.1 **Stage 1: Foundation**

The foundation stage is the beginner stage and English is introduced into a certain country by settlers during colonisation and gains prominent use. Despite the regular use of English, there are other indigenous languages in the country, which are also spoken frequently leading to language contact.

According to Schneider, the settler community's arrival introduces English to a new country but there is limited contact between the settlers and locals. Schneider explains that indigenous names are some of the first to be borrowed by the settlers.

2.6.2 **Stage 2: Exonormative stabilisation**

With the colonial government in charge, English acquired a higher status and is now spoken in a new cultural and sociolinguistic environment. Although the settlers try to preserve the original English, they soon gradually adjust and “begin to adopt local vocabulary, at first predominantly for objects that the settlers encountered for the first time in the new territory” (Schneider, 2003: 245-246). Schneider notes that new names of plants and animals are either created or borrowed from local languages.

Another characteristic of stage 2 is the spread of bilingualism among indigenous people. Due to language contact and education, local people start using English in addition to their native language, which allows for multiple forms of bilingualism to occur. These forms of multilingualism will be dependent on the types of interaction, which occurs in the different stages outlined by this model and will thus always fall on a continuum which reflects the competence and performance of a speaker. For a more detailed discussion on bilingualism in the context of this study, cf. Section 2.8.5. According to the DyM, local people view English as a valuable resource for their economic and social advancement. Structural nativisation starts at stage 2 as postulated by Schneider (2003). He further notes that, due to language contact, there will be linguistic transfer on a structural and phonological level especially in spoken discourse. At the exonormative stage, innovation features manifest in the English of indigenous people. “An innovation is seen as an acceptable variant” (Bamgbose, 1998: 2).
The emergence of innovative features is a hallmark of stage 2 of the DyM. Schneider notes that at stage 2, differences between indigenous and Standard strands of English emerge. Although the settlers still identify with BrE, they begin to slowly adjust to the local environment leading to the modification of the English spoken in the new country. “The settler strand in its spoken form begins to move towards a local language form” (Schneider, 2003: 245). Local vocabulary for new objects encountered is among the first features to be adopted by the settlers. Schneider adds that there is increased demand for English learning at this stage because locals begin to appreciate the social, economic and political value of English. The third stage of the DyM is discussed next.

### 2.6.3 Stage 3: Nativisation

According to the DyM, in the nativisation stage, both the settlers and indigenous people realize the changes that are taking place in the languages. At this stage, there is increased interaction between the locals and the settlers, meaning that languages will be in contact more often (Schneider, 2003). Schneider notes that language contact (between English and indigenous languages) leads to nativisation. The DyM posits that there is transformation, both culturally and linguistically. Due to an increase in cross-cultural communication, language change occurs, but there might be a clash between conservative language users and those who embrace innovations.

Another point made by Schneider is that heavy lexical borrowing occurs at stage 3. Language use reflects local accents and “indigenous speakers … will nativise the language grammatically and structurally by using constructions peculiar to the given country” (Schneider, 2003: 248). In addition, various forms of linguistic features are slowly adopted by the indigenous populace, but with time, they are used regularly. Changes at lexico-grammatical level are another characteristic of stage 3 of the DyM. This is evidenced by the frequency and usage patterns of innovation features.
Nativisation is . . . the central phase of both cultural and linguistic transformation in which both parties involved realise that something fundamental has been changing for good: traditional realities, identities and socio-political alignments are discerned as no longer conforming to a changed reality and the potentially painful process of gradually replacing them with something different, a new identity reflecting a changed reality, combining the old and the new, is in full swing.

(Schneider, 2003: 247)

The attainment of political independence is one of the features of the nativisation stage according to Schneider. The DyM claims that most of the changes are noticeable in the IDG strand language users via cultural and linguistic assimilation.

Although a lot of borrowing takes place, there is lack of acceptance of the nativised variety but there is slow and gradual use of variant forms leading to the features being used increasingly. The fourth stage of the DyM is discussed in the next section.

2.6.4 Stage 4: Endonormative stabilisation

A hallmark of the endonormative stabilisation stage is that innovations are now accepted as the norm by the majority of people in a given country (Schneider, 2003). The acceptance of innovations is evident in their use in formal channels of communication. Due to the use of 2 different languages (e.g. a local language and English), language transfer occurs. According to Schneider, stage 4 results in the birth of a new language variety and the production of dictionaries, grammar books and user guides. Schneider also notes that at stage 4, due to linguistic self-confidence, innovations are adopted, accepted and used regularly by the majority of people in a speech community.

Endonormative stabilisation, sometimes occurs smoothly and gradually after political independence of a country or may be caused by political events (Schneider, 2003). Schneider draws attention to “EVENT X”, which he believes leads to STL strand speakers to rethink and re-examine their position in the countries that they are based. In the case of Australia, the lack of support during World War 2 is noted by Schneider as being Event X.

The DyM states that acceptance of the NE happens during the endonormative
stabilisation stage. New norms that are different from the STL strands are accepted and increasingly used. The next process for a new language variety noted by Schneider is codification. User guides, grammar books and dictionaries need to be made available for the new variety. The final stage of the DyM is discussed next.

**2.6.5 Stage 5: Differentiation**

Schneider (2003) asserts that stage 5 of the DyM is characterised by the solidification of the new variety. With the emergence of a new English variety, there is a sense of ownership of the new variety, which reflects the local cultures and identities. New dialects of the new variety also emerge at stage 5.

The DyM is suitable for use in describing ZE because the “model yields remarkable exploratory power in regards to linguistic features and variety status of ZE … in the sense that it is possible to place the variety in any of the five phases depending on its structural characteristics as well as its sociolinguistic history and milieu” (Marungudzi, 2016a: 3).

Schneider notes that the DyM can be used to describe varieties of English around the world. He also suggests that there is variability in the way different societies use English, leading to new dialects of the language. However, whether these varieties are accepted or not is another issue. This study seeks to determine whether there are any variations in the way ZE is used, (by looking at articles, modal verbs and some discourse markers), but does not seek to impose on whether the variety should be accepted or not.

One of the studies that used the DyM as its theoretical framework is Borlongan (2016) who studied Philippine English and noted that although there seems to be an unwillingness to accept the local nativised variety of English in the Philippines, the language seems to have reached stage 4 of the model.

Marungudzi (2016a) has also studied ZE within the DyM framework. He studied some morphosyntactic features and from the study noted that ZE seems to be at stage 3. By studying more features, this study further seeks to situate ZE within the DyM. From an analysis of the selected features, it will be ascertained whether ZE is at stage 3 or any other stage.

Despite having been accepted as a viable model for describing WE, we look at some of the
criticism that has been levelled against the DyM in the following section.

### 2.7 Criticism of the dynamic model

The DyM has been criticised by some researchers. Mesthrie and Bhatt (2008) argue that since the DyM focuses mainly on ESL countries, there is no need to include what they call “dominion” countries like Australia and New Zealand. “The conditions surrounding the installation of dominions seem quite different from that of colonies and protectorates” (Mesthrie & Bhatt, 2008: 35).

Another point of contention is the DyM’s perception of “identity” as being a “public” notion. Mesthrie and Bhatt (2008) argue that the DyM does not give due consideration to other important features of identity such as class and status.

The importance of “EVENT X” in the DyM is also questioned by Mesthrie and Bhatt (2008: 35) who argue that too much emphasis is placed on the significance of Event X on the use of indigenous varieties.

Mesthrie and Bhatt also question the chronological order by which countries move from stage to stage in the DyM. They believe that countries may bypass a stage (for example, a country can move from stage 3 to stage 5). “This would be a territory in which English became nativised and subsequently differentiated into subdialects, without there being a commonly accepted endonormative standard” (Mesthrie & Bhatt, 2008: 35).

Class distinctions are not found in the DyM. According to Mesthrie and Bhatt (2008), middle class varieties of English differ from other varieties like pidgin and basilect varieties. Another criticism comes from Onysko (2016) who comments that the DyM fails to accommodate the development of Englishes in the expanding circle. Despite this criticism, attempts have been made to extend the DyM and to integrate postcolonial and non-postcolonial Englishes (e.g. Buschfeld & Schneider, 2017; Buschfeld & Kautzsch, 2017).
Despite the criticism levelled against it, the DyM enables the analysis of NE in individual countries and an interesting conclusion can be reached by applying the DyM to different countries (Mesthrie & Bhatt, 2008; Onysko, 2016). The next section discusses key terms that are associated with WE.

2.8 Definition of concepts

This section outlines the main terms which are applicable to the current study. The terms help to shed more light on the subject of WE and are described as they are used in literature. An explanation on how the terms are used in this study is given.

2.8.1 World Englishes/ New Englishes/ Global Englishes

According to Schneider (2003) typically, the terminology of L2 varieties of English is not consistent. This is evident in the different terms used in literature. These terms are "New Englishes" (e.g. Gough, 1996a; Schneider, 2003, 2007), "World Englishes" (e.g. Sridhar, 1992; Kachru, 1992; De Klerk, 2003a) and "Global Englishes" (e.g. McArthur, 2001). Kachru (2006) and Schneider (2003) note that the plural term "Englishes" seems to be used to show that there are different varieties of English in various socio-cultural contexts. Black South African English (Kruger & Van Rooy, 2017), Hong Kong English (Lam, 2009a, 2009b), Nigerian English (Unuabonah, 2019), Xhosa English, which is a part of Black South African Englishes (De Klerk, 2005a, 2005b), Irish English (Murphy, 2015) are examples of WE.

Makoni (1992) argues against the description and documentation of NE stating that doing so asserts the inferior status of these varieties of English. Instead, he advocates for the promotion of StE. Makoni (1993: 105) sees the English spoken in Zimbabwe as an interlanguage or learner language “because it has not yet been ‘localised’ following the criteria: stability, native speaker norms and degree of compactness within the speech community”. Kachru (1992) on the other hand denies the interlanguage assertion and suggests that social, personal and cultural factors should be taken into consideration when describing NNVE. Kachru advocates for the acceptance of WE since he believes that English operates in different contexts worldwide, leading to variations.
Some scholars (e.g. Mesthrie & Bhatt, 2008; Graddol, 2003) suggest that the indigenous varieties have emerged due to the fact that there are few native speakers and teachers of English. As a result, the input that learners get is from L2 speakers, which may influence the way English is spoken. However, Makoni (1992) does not agree with the assertion that the unavailability of native speakers in countries where L2 is spoken leads to localisation. He argues that technology has mitigated the need for the physical presence of native speakers. In Makoni’s opinion, broadcasting services (e.g. radio and television) help to provide the native speaker norms, which will benefit locals. Although it is true that the new century has brought in a surge in the use of the Internet and other technologies, there are still some remote places in the L2 speaking countries. In these places, learners still rely on the traditional ways of learning. In addition, even in areas like towns and cities, access is still a problem because some of the poor communities have limited or no access to the Internet.

Makoni (1992) remarks that NE are a result of British colonialism. He further notes that NE are spoken by non-native speakers of English, therefore they are different from native speaker varieties such as American English and BrE. A discussion about L1 and L2 speakers as English teachers is given in Section 2.8.10. In addition to English being spoken as a L2, there is a growing number of L2 writers of English (Sridhar, 1992). Some of the writers from West Africa are Wole Soyinka, Ngugi Wa Thiong’o and Chinua Achebe. Examples of writers from Southern Africa include Tsitsi Dangarembga, Charles Mungoshi, Chenjerai Hove, Sindiwe Magona and Njabulo Ndebele.

Schneider (2011) comments that the teaching and learning of English has received mixed reviews in different countries where it is used as a L2. Critics denounce the promotion and use of English because they believe that it leads to the death of indigenous languages and cultures. Despite the criticism, there are others who accept the benefits of using English because they believe that using English will benefit them socially, politically and economically. That is the reason why there is an increase in the number of L2 speakers of English globally.

Chang (2008) observes that one of the interesting characteristics of English in the 20th and 21st centuries is that different cultures use English in varying contexts. According to Mesthrie and Bhatt (2008: 3), it has become a tradition for researchers to use the plural
“Englieshes” when referring to the new English varieties that are emerging in L2 contexts “to stress the diversity to be found in the language today and to stress that English no longer has one single base of authority, prestige and normativity”. Mesthrie and Bhatt further argue that one of the characteristic features of WE is that they are L2 varieties that are rarely used at home.

Researchers (e.g. Kachru, 1992; De Klerk, 2003a, 2003b, 2005a, 2005b; Schneider, 2003; 2007; Mesthrie & Bhatt, 2008; Van Rooy, 2011; Kortmann, Lunkenheimer & Ehret (2020); Van Rooy & Kruger, 2016; Hichey, 2017) have documented some of the elements of WE that make them different from native varieties. One of the contentious issues regarding WE is whether they should be used as the basis in L2 teaching and learning. The issue of attitudes towards WE is discussed in Section 2.3.4.

One of the key arguments raised by Kachru (1992) is that English is used differently in countries where it is not spoken as a L1. Therefore, innovations (cf. Section 2.8.7) and nativisation (cf. Section 2.8.8) should not be dismissed as errors. Another noteworthy observation highlighted by Kachru is that the personal, social and cultural aspects that operate in the environments in which the WE varieties are utilised should be taken into consideration in order to better understand the status of NNVE. In taking this view, Kachru seems to be diverging from “the purist school of thought which uses derogatory labels to describe NNVE as approximate systems, idiosyncratic language variations, transitional competence and inter-languages” (Kadenge, 2009: 158). WE is an umbrella term from which other terms which depict varieties found in different continents emerge as illustrated in the following section about African Englishes.

### 2.8.2 African English/ African Englieshes

Where WE encompass NNVE spoken throughout the world, the term “African English” or “African Englieshes” (AE) describes the varieties of English that are being used in Africa by non-native speakers of English (Kadenge, 2009). Kamwangamalu (2019) views English as an African language noting that lexical, semantic, phonological, syntactic and phonetic features have been and continue to be introduced into the Englieshes spoken in different African cultures. He sees culture as influencing the way English is spoken in different contexts.

© University of Pretoria
Achebe (1965: 30) notes that English spoken in Africa will be different from L1 varieties by stating that “I feel that the English language will be able to carry the weight of my African experience. But it will have to be a new English, still in full communion with its ancestral home, but altered to suit its new African surroundings”.

Although WE varieties were once frowned upon, they are slowly being accepted. Examples of these are Black South African English (BSAE) (e.g. Gough, 1996a; De Klerk, 2003a, 2003b; Mesthrie, 2006; Parkinson & Singh, 2007; Makalela, 2007, 2013), Xhosa English (e.g. De Klerk, 2005a, 2005b, De Klerk 2006a, 2006b), Nigerian English (e.g. Werner & Fuchs, 2017), Ugandan English (Nassentein, 2016), Kenyan English (e.g. Schmied, 2006) and White South African English (Wasserman & Van Rooy, 2014).

The complexity of the language policies of former colonial countries in Africa has led to the promotion of both indigenous languages and English as official languages (Kachru, 1992; Gough, 1996a; Thondhlana, 2002; Schmied, 2006, Ndhlovu & Siziba, 2018). As a result of indigenous languages being in contact with English, by-products of language contact like borrowing and code switching. Borrowing entails “the introduction of single words or short idiomatic phrases from one variety into another” (Gumperz, 1982: 66). Codeswitching “involves the alternate use of two languages or linguistic varieties within the same utterance or during the same conversation” (Hoffmann, 1991: 110). Codeswitching and borrowing will not be discussed further because they are not the focus of this study.

Researchers have proposed two causes of the variations that occur in the syntax and semantics of African English varieties “L1 or mother tongue interference and analogical deviations based on English” (Bokamba, 1992: 139). Bokamba asserts that the deviations from StE that seem to be similar across varieties of AE, are an indication of the different environments in which L2 English operates. Other researchers have documented the different characteristics of AE (e.g. Arua, 1998; Banjo, 2000; De Klerk, 2003b, 2005a; Schmied, 2006; Bokamba, 1992, 2019; Fuchs, Van Rooy & Gut, 2019; Marungudzi, 2016a; Werner & Fuchs, 2017; Xu, 2017) The terms “contextualisation” or “nativisation” are used in literature to express the view held by researchers that “English is adapted to local or regional linguistic conditions and thereby deviates systematically from the standard dialect” (Bokamba, 1992: 140). The concept of nativisation is explored in Section 2.8.8.
Bokamba (1992) argues that it is not only speech that has manifested diversity, but also writing. The variations are evident in the works of African writers, primarily novelists, who, through their writing, try to safeguard and convey African culture. To sum up the contributing factors for the emergence of variation in AE, Bokamba (1992) notes that language contact between English and the mother tongue of a society contributes to the emergence of new English varieties.

Nigerian English is another example of an AE and has received a lot of scholarly attention. To provide some examples, Bamgbose's (1992) research focused on the morphological, phonological, syntactic and lexical features of Standard Nigerian English. Another study by Gut (2008) reported that Nigerian English has a reduced vowel system when compared to Southern BrE, which has 23 vowels. Gut also noted that Nigerian English has syllable-timed rhythm whilst StE has stress-timed rhythm. Another difference noted is in the intonation in Nigerian English, which Gut reports to be simplified when compared to L1 varieties of English. Using written samples from the 18th century to date, Jowitt (2019) discussed the distinctive features of Nigerian English focusing on phonology, lexico-semantics and morpho-syntax.

Gut and Fuchs (2013) analysed the progressive aspect in Nigerian English and noted that there is stylistic variability in the way the progressive is used in Nigerian English when compared to BrE. Unuabonah, (2019) studied the stylistic variability in Nigerian English and reported that DMs in Nigerian English and BrE varied significantly in terms of frequency and style. Similarities were also noted in the frequency of elaborative DMs in both varieties of English. Banjo (1997) and Taiwo (2020) reported on the different morphosyntactic features of Nigerian English.21

Allo and Mesthrie (2004) suggest that having been acquired mostly at school, interference between English, other local languages and cultural contact are some of the reasons why Nigerian English has developed distinct features. Branching from African Englishes are different varieties, which are usually referred to by the country of origin. One such variety, BSAE, is discussed below.

21 Although the reviewed studies have highlighted the fact that Nigerian English has been studied extensively, Nigerian English will not be explained further because the scope of this study is ZE.
2.8.3 Black South African English

BSAE is a type of English spoken by the black population in South Africa. As mentioned above, of all the varieties of African English, BSAE has received the most attention by researchers (Van Rooy, 2013). “BSAE is defined as a variety of “South African” English, which is further specified as “Black” to indicate its origin in township education and its status as a language spoken by speakers who have already acquired at least one other language” (Coetzee-Van Rooy & Van Rooy, 2005: 1). BSAE consists of different varieties some of the studies that have shown innovation in BSAE include Gough (1996a), De Klerk (2003a, 2003b; 2005a, 2005b), Mesthrie (2006) Parkinson and Singh (2007), Van Rooy (2000, 2007, 2013), Wasserman and Van Rooy (2014).

According to Van Rooy (2004) and Mesthrie (2006), the dawn of independence in South Africa in 1994 has resulted in economic and political opportunities for the previously marginalised black South Africans. This has led to more black people using English and the diversity in the way English is spoken is evident in the local media, especially on televisions and radios. Schneider (2011) observes that BSAE is prevalent among politicians and celebrities. To give an example of the diversity that exists in the way English is spoken, if one watches proceedings in the South African parliament, there are diverse ways in which English is spoken by politicians.

The diffusion of English “reflects a new diversity of lifestyles, educational and cultural mixing, which sees English not only as the main language of a multilingual black elite, but even making inroads into some homes” (Van Rooy, 2004: 943). De Klerk (2006b) asserts that an indigenous African language is the L1 for many black people whilst English is spoken as a L2. Van Rooy observes that in some instances, English is now the L1 in black households. As mentioned in Section 2.8.10, most of these children will still learn English from non-native speaker parents and teachers. This may have implications on the variety of English that the children will speak.

Studies done on Black South African English

Some of the studies done on BSAE include Van Rooy and Kruger (2016) who used corpora to study the progressive aspect of BSAE. One of their findings is that an increase in the proficiency level of a speaker leads to changes in the use of the progressive form. Van
Rooy and Kruger (2016: 215) highlighted that “the three dimensions of the progressive construction investigated change at differential rates and have different outcomes. The differences provide new insights into the interaction between proficiency and normative processes that play a role in the conventionalisation of an innovative feature, or its demise as a learner error”.

According to Van Rooy and Kruger (2016) learners who are in the early stages of acquisition tend to use innovations more often than the more proficient speakers. Van Rooy and Kruger further state that innovations may be reinforced if there are no interventions from teachers and editors. Another point made is that there are higher chances for innovations to become conventional in situations where English is widely used and where there are locally produced texts. Kachru (1985) claims that although India is a non-native English-speaking country, English book production is third largest after the UK and the USA. Therefore, following Van Rooy and Kruger’s (2016) explanation, India is an example of a country where there are higher chances of innovations becoming conventional because it produces many English books.

Another study on BSAE was done by Gough (1996a). Gough reported variations in vowel phonology, the consonant systems, word stress and intonation. Some of the grammatical features of BSAE include the use of non-count as count nouns, omission of articles, extensive use of the resumptive pronouns and the use of can be able to as a modal verb. The use of idiosyncratic DMs is another feature of BSAE mentioned by Gough (1996a).

De Klerk (2003b) studied a corpus of Xhosa English (a type of BSAE) consisting of 540 000 words to test whether Xhosa English can be classified as a NNVE. The corpus was compiled using 27% unscripted monologue and 73% private spoken dialogue. From her analysis, De Klerk asserts that Xhosa English meets the criteria of being a variety of English due to the different features, which distinguish it from StE. Some of the features reported by De Klerk (2003b) include the use of the resumptive pronoun in relative clauses, the extended use of the progressive, conflation of demonstratives, overgeneralisation in the use of quantifiers, no distinction between mass and count nouns, omission and insertion of inappropriate articles, can be able to as a modal verb and maybe used to express conditional modality.
Focusing on DMs De Klerk’s (2005b) investigated the DM *actually* using a spoken corpus drawn from Xhosa first language speakers who use ESL. She concluded that Xhosa English speakers generally use the DM *actually* the same way as native speakers except when *actually* is used as a contemplative and to mark disagreement by Xhosa English speakers.

In another study of DMs in Xhosa English, De Klerk (2005a) explored the DM *well*. She reported that there is diversified use of *well* by Xhosa English speakers. In addition, De Klerk (2005a) noted that there was less frequency in the discourse marker *well*, compared to the frequency in native speaker discourse. In a synthesis on her work on Xhosa English, De Klerk (2006a) discussed the various syntactic features of Xhosa English, how intensity levels are expressed and DMs in Xhosa English with special focus on *well* and *actually* using a corpus of over half a million words.

Mesthrie (2006) analysed undeletions in BSAE and reported on the variability in occurrence of undeletions in BSAE. Examples include “the complementiser *that*, infinitival marker to, pronoun (including resumptive pronouns, left dislocation and dummy it), to be in small clauses and the occasional undeletions with idiomatic wh-constructions” (Mesthrie, 2006: 82). In a study of the morphological and syntactic features of BSAE, Mesthrie (2008) reported on the variability in the tense-aspect-modality systems, verb phrase negation, relativisation, complementation, word order and agreement. Another feature reported by Mesthrie that was also investigated in the current study is the use of articles Ø, *a/an* and *the*.

Makalela (2007) examined whether the L1 influences phonology, morphosyntax, pragmatics and discourse features of BSAE. Evidence from the investigation shows the influence of mother tongue on the production of different features studied such as localisation. In another study, Makalela (2013) investigated the use of BSAE on the radio using a corpus of 209 000 words. He reported on the four features with the highest frequencies namely retention of question word order, the use of *maybe* as a conditional modality, consecutive tense and the substitution of *that*-complementiser. Makalela (2013) suggests that these features are a result of the influence of African languages on English due to language contact. Considering the DyM, Makalela situated BSAE on stage 3 (nativisation) and stage 4 (endonormative stabilisation).
Van Rooy and Terblanche (2006) used the Tswana Learner English Corpus (TLEC) and the Louvain Corpus of Native English Student Essays (LOCNESS) to analyse similarities of students’ writings. Results of this study show some similarity in the way students write. Kruger and Van Rooy (2017) studied the progressive aspect when examining the role played by editors in making innovation features legitimate. They concluded that editing plays a role in endonormative stabilisation of innovative features. In another study, Van Rooy (2007) examined consonant clusters and resyllabification of consonants in BSAE. He noted the existence of differences between BSAE and native English varieties. An overview of some of the work done on English varieties in South Africa affirms that the majority of research focuses on BSAE (Van Rooy, 2013).

Although BSAE has received more attention than any other type of African English in southern Africa as shown by a review of the studies mentioned above, researchers have also investigated ZE as will be discussed next.

2.8.4 Some studies on Zimbabwean English

Although research on ZE is still scarce, studies have been carried out on different aspects of ZE. Mlambo (2009) explored whether ZE is a variety of English and concluded that although not yet institutionalised, there are a number of varieties of English in Zimbabwe. Furthermore, he noted that the English spoken in Zimbabwe reflects some structural qualities of Shona especially in phonology.

Another study is the explanatory and exploratory research, which was used by Kadenge (2009) to explore the effects that Shona phonology has on the way English vowels are pronounced by Shona speakers. Kadenge observes that there is reduction in the number of vowels from twenty-five vowels found in StE to five monophthongs. The speech of Shona L1 speakers of English contains five monophthongs namely, \( [a], [e], [i], [o] \) and \( [u] \), as compared to the twenty-five vowels of StE. “The findings of this research demonstrated that this is achieved mainly through simplification strategies such as monophthongisation of diphthongs, glide epenthesis, vowel length reduction and vowel substitution” (Kadenge, 2009: 169). Kadenge suggests that the phonological processes that manifest in the speech of Shona L1 speakers of English are evidence of the indigenisation and nativisation.
Kadenge (2012) explored the lexical and semantic features of ZE in creative writing, focusing on the novels *Bones* and *Ancestors* by Chenjerai Hove. Kadenge reported that in the novels, the author used English as a medium to express Shona cultural values, beliefs, philosophies and ways of life. Direct lexical from Shona to English is one area of indigenisation of English in Zimbabwe according to Kadenge (2012). Examples include names of indigenous foods such as *sadza* ('pap – maize meal'), names of indigenous fruit trees such as *muonde* ('fig tree') and *muhacha*, adult dances such as *muchongoyo* and *mbakumba* and traditional games such as *dudu muduri* and *mahumbwe* ('children's games' – generic). The abovementioned names are unique to Shona culture.

Kadenge (2012) reports on the complexity of kinship terms like “mother”, “father”, “brother”, “sister” and “son” and the broad semantic fields that they cover in ZE when compared to StE. He comments that the terms are used as honorifics in ZE to show respect even to non-relatives.

Literal translation in figurative expressions such as metaphors, idioms and proverbs is another feature of ZE (Kadenge 2012). Kadenge gives an example of literal translation in “a man must die trying”. In a fight, a man can be said to be dead only when flies perch on his intestines”, which is the literal translation for the Shona proverb *kufa kwemurume kubuda ura* ('a man must die trying').

ZE is a term that is not used universally and Kadenge (2009) advocates for its acceptance and recognition. Kadenge hopes that more studies on ZE will help shed light on how indigenous languages of Zimbabwe have influenced L2 English. It is hoped that this study will add to the already available but scarce research on ZE.

In a study of morphosyntactic features of ZE, Marungudzi (2016a) examined eight features and concluded that due to their higher frequencies, these features can be labelled as ZE features. The features belong to category B (neither pervasive nor extremely rare – cf. Section 5.5) and are shown in the table below together with the feature numbers and different categories that are reported in Kortmann and Lunkenheimer’s (2012) taxonomy.

---

22 Examples with no translations occur as borrowed words in StE.
Table 2. Eight features of ZE (source: Marungudzi, 2016a: 89)

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Name of feature</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td>Extension of <em>be + V-ing</em> to stative verbs.</td>
<td>Verb phrase: tense and aspect</td>
</tr>
<tr>
<td>174</td>
<td>Deletion of auxiliary <em>be</em> before progressive.</td>
<td>Agreement</td>
</tr>
<tr>
<td>175</td>
<td>Deletion of auxiliary <em>be</em> before <em>gonna</em>.</td>
<td>Agreement</td>
</tr>
<tr>
<td>194</td>
<td>Resumptive/shadow pronouns.</td>
<td>Relativization</td>
</tr>
<tr>
<td>209</td>
<td>Addition of <em>to</em> infinitive where StE has bare infinitive.</td>
<td>Complementation</td>
</tr>
<tr>
<td>222</td>
<td>Use of <em>too, too much, very much for very qualifier.</em></td>
<td>Adverbs and preposition</td>
</tr>
<tr>
<td>227</td>
<td>Inverted word order in indirect questions.</td>
<td>Discourse organisation and word order</td>
</tr>
<tr>
<td>234</td>
<td><em>Like</em> a focussing device.</td>
<td>Discourse organisation and word order</td>
</tr>
</tbody>
</table>

In addition to the 8 features, Marungudzi (2016a) also reported on features which were not frequent enough to be in category B and were not considered to be characteristic of ZE. They include the following:

1. *She/her* used for inanimate referents (Feature 1).
2. *Me* instead of *I* in coordinate subjects (Feature 7).
3. *Meself/myself* instead of *I* in coordinate subjects (Feature 8).
4. Different count or mass noun distinctions resulting in use of plural for StE singular (Feature 55).
5. Double determiners e.g. demonstrative/article + possessive, with possessive pronoun preposed or postposed (Feature 59).
6. Extension of progressive to habitual contexts (Feature 89).
8. Use of *like* as a quotative device (Feature 235).
According to Marungudzi (2016a), there are similarities and differences between ZE features and those reported in other AE varieties. From the analysis of the morphosyntactic features, Marungudzi reported that ZE is at stage 2 (exonormative stabilisation) and stage 3 (nativisation) of the DyM. He also noted that the reasons for variations in ZE include overgeneralisation, simplification strategies, L1 influence and the context in which English is acquired as a L2.

The studies highlighted above provide interesting accounts of ZE and show how ZE as a research area is developing.

Provided that ZE is a product of bilingualism, it is necessary to describe bilingualism within the context of WE.

### 2.8.5 Bilingualism

From the above discussion, it can be deduced that bilingualism is a key component of WE in general and ZE in particular. This is because L2 speakers of English use more than one language. “Bilingualism” is a term that has been used variably to describe the use of two languages by an individual. One definition given is that “the ideal bilingual switches from one language to the other according to appropriate changes in the speech situation” (Weinreich, 1963: 73). The definition seems to suggest what is referred to in literature as native-like accent in a L2. Native-like mastery of both languages is sometimes used as a benchmark on a bilingual continuum for one to be called a bilingual as deduced from Weinreich’s definition. But other definitions of bilingualism encompass different levels of linguistic competence in a L2. For example, according to Hoffmann, (1991: 24) “different bilinguals have distinct uses, as well as various levels of competence for each code”. Wardhaugh, (1992: 89) defines a code as a language or a language variety by noting that “the term … can be used to refer to any kind of system that two or more people employ for communication”. Still on the subject of bilingualism, Hoffman also distinguishes between perfect bilingualism (whereby a person has native-like knowledge of two languages) and balanced bilingualism (whereby proficiency is not measured according to native-like standards). Bilingual conversation depends on speakers involved, speakers’ individual linguistic proficiency and language choice (Nel, 2012). For the current study, a bilingual is classified according to Hoffman's (1991) classification where bilinguals’ linguistic
competence and performance may vary.

Bilingualism is a common feature in many countries (Wei, 2000; Thondhlana, 2002). It results from contact between languages and is one of the by-products of interaction between people of different linguistic backgrounds (Hoffmann, 1991). In addition, Hoffmann notes that a child's lack of progress in school was considered to be caused by the child's bilingual status in the past.

“The most salient feature of bilingualism is that it is a multi-faceted phenomenon” (Hoffmann, 1991:14). For that reason, there is no clear-cut definition of bilingualism found in literature, again pointing to different possible points of proficiency on a bilingual continuum. Hoffmann comments that another complication arising from the study of bilingualism is the interdisciplinary nature of the subject with sociologists, psychologists, linguists, anthropologists, educators, among others using diverse methodologies in the study of bilingualism.

Hoffmann (1991: 22-24) remarks that “the notion of relativism, . . . is a central one in the discussion of any type of bilingualism [...] different bilinguals have distinct uses, as well as various levels of competence, for each code”. That is, there is variability in the competence levels of different bilinguals. She uses the term “receptive bilingual” to refer to someone who can understand the speech and written form of a L2. On the other hand, the term “productive bilingual” refers to a person who is able to speak and read a L2. Montanari and Nicoladis (2016) agree that the diversity in the degree of bilingualism is highlighted by the different proficiency levels of bilinguals and comment that researchers now accept that bilingualism is a norm for most people throughout the world. Due to the complex nature of trying to characterise a bilingual, Hoffmann (1991) draws attention to the different aspects that can be used to profile a bilingual. The variable aspects are as follows:

1. How a language develops, how a L1 or L2 is maintained or lost.
2. The sequence in which the L1 and L2 are learned, considering whether they are acquired concurrently or if the L2 was learned afterwards.
3. How well one speaks the L1 and L2, taking into account, which language is mostly used by an individual.
4. The functions of the L1 and L2.
5. Attitudes towards speakers, the languages and whether bilingualism is viewed favourably.
7. Psychological, motivational and social forces.
8. The natural setting of the bilingual.
9. The extent to which an individual is familiar with L1 and L2 cultures.

Most learners become bilingual at school because their home language is not the same as the language used for teaching and learning (Genesee, 2016). In most former colonies, the former coloniser’s language is usually used in schools (Thondhlana, 2002). Hoffmann (1991) draws attention to cases where English-speaking children undertake French immersion courses in schools.

According to Hoffmann (1991), one of the facets of bilingual speech include interference, whereby elements from one language are transferred to another language. Another hallmark of bilingual speech is borrowing of lexical items from one language to another. Codemixing and code switching, whereby speakers use more than one language in a conversation (cf. Myers-Scotton, 1993) are also given by Hoffmann (1991) as aspects of bilingual speech.

Nel (2012) notes that it is crucial to be mindful of the distinction between competence and performance.

Competent speakers of a language have tacit knowledge of more than just grammaticality, i.e. what is a well-formed sentence in their language and what is not. In addition, they are able to judge the acceptability of a given well-formed sentence in a given social context.

(Myers-Scotton, 1993: 79)

From the definition above, a speaker may have knowledge about a language but the way he or she speaks or uses the language (performance) may be affected by other issues such as cognitive, affective and socio-cultural skills. In studies that collect different forms of production data such as this study, competence is implied and not directly measured. The performance of the participant will depend on the context in which it is used or sampled.
Competence and performance are thus foundational in the mind and use of the bilingual speaker and will ultimately in this context only focus on use and production (written and spoken) in terms of language performance of bilinguals and specifically situated in the DyM.

Romaine (1995) and Mackey (2000) draw attention to the four interrelated issues that are crucial to the description of bilingualism. These are, proficiency level (degree), what the languages are used for (function), how a person alternates between the languages (alternation) and whether one is able to keep the languages apart (interference).

According to Unsworth (2016) the way bilingualism develops varies from one individual to another. In addition, the L2 learning experiences are also different. There are factors that contribute to diversity in the speech of bilinguals. “These include exposure to input from non-native speakers, the variety of speakers providing input in a given language and the existence of multilingual utterances in the input” (Unsworth, 2016).

2.8.5.1 Types of bilinguals

In research about bilingualism, distinctions are made between people who speak the L2 with varying degrees of bilingualism. For example, Wei (2000) refers to a person who is taught a L2 as a secondary bilingual and notes that if someone acquires a L2 after the L1 has started developing, he or she is referred to as a successive bilingual. If a person learns two languages in a similar environment and at the same time, he or she is called a compound bilingual. On the other hand, if one learns the two languages in different environments, he or she is referred to as a coordinate bilingual. A dominant bilingual has greater command of one language over another whilst a functional bilingual can use both languages with similar degrees of competence (Wei, 2000).

The notion that bilingualism is widespread in different countries around the world is supported by Schneider (2011). He attributes the development of WE to language contact whereby different languages come into contact leading to the languages influencing each other.
In bilingual situations, the two languages are used in varying contexts and “whereas the first language may be a sign of solidarity or intimacy, English, in many bilingual situations, carries overtones of social distance, formality or officialdom” (Graddol, 2000: 12). According to Mackey (2000) the four factors of bilingualism that have a variable interrelationship are alternation, function, interference and degree. Of the four factors, Mackey provides a list of different areas that need to be looked at when measuring the degree of bilingualism in individuals. The four skills are listening, reading, speaking and writing. The levels to be measured are phonological or grammatical, lexical, semantic, stylistic and graphic.

2.8.5.2 How bilinguals are classified in this study

This study used bilinguals who have varying degrees of competence. Speakers who had at least 10 years of formal education and who were able to converse in English were asked to participate in the study. This is in line with Romaine’s (1995: 320) assertion that “bilingualism is not a unitary phenomenon. It is shaped in different ways depending on a variety of social and other factors, which must be taken into account when trying to assess the skills of bilingual speakers”. The proficiency levels of participants varied. A questionnaire and semi-structured interviews were utilised to help determine whether participants meet the qualifying criteria for this study. From the discussion about bilingualism, it can be noted that different factors need to be probed in order to reach a better judgement when assessing bilingual competence. Since we have established that speakers’ degrees of bilingualism vary, the discussion now turns to linguistic proficiency in the following section.

2.8.6 Linguistic proficiency

There is no consensus about the definition and measurement of linguistic proficiency (Brière, 1971; Farhady, 1982; Llurda, 2000). It is important to discuss the different definitions provided in literature in order to highlight the different viewpoints provided by researchers. One of the definitions is that “proficiency is frequently defined as the degree of competence or capability in a given language demonstrated by an individual at a given point in time independent of a specific textbook, chapter in a book, or pedagogical method” (Brière 1971: 385). Bialystok (1981) associates linguistic proficiency with the degree of mastery of
a language. This definition implies that there are different levels of abilities in using a language. Another definition is that:

Language proficiency refers to the degree to which an individual exhibits control over the use of a language for one, some, or all of its numerous and diverse aspects. These include the phonological, syntactic, lexical and semantic systems and discourse and stylistic rules for oral and written communication for different varieties of a given language in various domains and social circumstances.

(Burt & Dulay, 1978: 178)

Another definition is that linguistic proficiency “refers to a speaker’s sum total knowledge and facility in using the particular language in question” (Gathercole, 2016: 123). Gathercole concurs that proficiency varies between individuals because of factors such as the amount of exposure to a language and how often one uses the language in question. Gathercole also gives the amount and quality of input that someone is exposed to, as important determiners of proficiency in a language. Despite linguistic proficiency not being well defined, researchers usually test the capability of individuals in using a specific language (Farhady, 1982).

Brosnahan (1958 as cited by Allo & Mesthrie, 2004) came up with four proficiency levels. Brosnahan’s classification is based on how an individual uses English and the level of education. He associates obtaining higher educational qualifications with higher levels of proficiency. In level 1 are individuals who have had no formal schooling. People who have finished only primary education are designated to level 2. Level 3 is allocated to speakers who have completed secondary tuition only. Those individuals who have had university education belong to level 4, the highest level according to Brosnahan’s classification. However, Allo and Mesthrie (2004) draw attention to the inaccuracies that may arise when classifying individuals using levels of education. Different factors may contribute towards an individual’s proficiency levels, leading to some people who have completed higher educational qualifications, having lower proficiency levels and vice versa.

Some of the contributing factors towards the development of a bilingual’s proficiency include, how a language is learned, how it is used, whether there is language contact and socioeconomic factors (Gathercole, 2016).
According to Lowenberg (1992), the variability found in the way English is spoken is often ignored, with research into linguistic proficiency being biased towards native speaker proficiency levels. Lowenberg, like Kachru (1992) advocates for the consideration of the varying contexts in which English is spoken as a L2 in former colonised countries.

In the case of language proficiency, Cummins (1979, 1999) differentiates between basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP). BICS is associated with the language used in everyday conversations whilst CALP is academic language. Another point made by Cummins is that the interrelationship between BICS and CALP that is apparent in a L1 can also be shown in L2 learning situations. BICS are acquired earlier by L2 learners. CALP on the other hand takes longer to acquire and keeps on developing throughout the school years. Cummins further notes that there is a clear contrast in the way BICS and CALP are acquired and developed. “BICS refers to conversational fluency in a language while CALP refers to students’ ability to understand and express, in both oral and written modes, concepts and ideas that are relevant to success in school” (Cummins, 2008: 71).

Brière (1971) and Llurda (2000) also add to the discussion about linguistic proficiency by drawing attention to the challenging task of defining and measuring language proficiency as evidenced by different definitions that are provided in literature. Burt and Dulay (1978: 178) explain that “language proficiency refers to the degree to which an individual exhibits control over the use of a language for one, some, or all of its phonological, syntactic, lexical and semantic systems as well as discourse and stylistic rules for oral and written communication for different varieties of a given language”.

From the discussion above, two views about linguistic proficiency are highlighted. Researchers (e.g. Brière, 1971, Burt & Dulay, 1978; Llurda, 2000; Bialystok, 1981; Farhady, 1982; Gathercole, 2016) view linguistic proficiency as the degree of knowledge and ability to use a specific language on the bilingual continuum (where competence and performance form the foundational concepts of use in proficiency).

Linguistic proficiency is also viewed from a BICS and CALP point of view. Cummins’ (1979, 1999) view is that linguistic proficiency differs when language is used in different contexts, with everyday speech (BICS) developing faster than academic speech (CALP).
Cummins, an individual may be fluent in a language but may have difficulties in academic speech and writing.

Measuring the proficiency levels is one area that seems to be controversial because there is no clear-cut method. The researchers mentioned above are of the same view that measuring knowledge and ability of individuals is difficult. It is important to note that this study will use extralinguistic proficiency data to contextualise the corpus linguistic data in order to classify the bilingual use of participants according to the DyM. The measurement of proficiency levels falls outside the scope of this thesis. The next section deals with the reasons why linguistic proficiency is difficult to measure.

2.8.6.1 Why proficiency is difficult to measure

Brière (1971) is of the view that the difficulty in defining proficiency lies in the determination of whether the level of competence should be based on communicative competence or linguistic competence. Another setback is that it is difficult to develop a model that can be used to design a proficiency test because behaviour is measured when testing language.

According to Romaine (1995), discussions about linguistic proficiency are further complicated by the variability that exists between different speakers of the same language. Hoffmann (1991) draws attention to the complexity of language behaviour, which leads to different proficiency levels, which are difficult to measure. The researcher acknowledges that linguistic proficiency is difficult to measure but for the purpose of this study, a short explanation of how linguistic proficiency was determined is provided below.

2.8.6.2 How linguistic proficiency was ascertained/determined

Although linguistic proficiency was not measured in this study, biographical and language background questionnaires provided some information regarding the levels that participants considered themselves to be at. A questionnaire was used to solicit information about participants’ linguistic proficiency levels. Participants were asked questions about how well they believe that they can speak English as a L2 (questions 10
Questions about the biographical and language backgrounds of participants helped shed light on the proficiency levels of participants.

In addition, semi-structured interviews were conducted with participants. These helped to determine whether participants can converse in English. The interviews were recorded. Information gleaned from the interviews was used to select suitable participants for the study. Due to time and financial limitations, a thorough test of linguistic proficiency was not conducted for this study.

In the following section, innovation and errors are discussed. The discussion about linguistic proficiency is closely linked to innovation because it is the varying degrees of proficiency, which in turn may lead to innovation in WE.

### 2.8.7 Innovation or error

One of the important issues that lacks consensus in WE studies is how to distinguish between an innovation and an error, due to the fact that many WE varieties are not codified (Bamgbose, 1998; Van Rooy, 2011; Hamid & Baldauf, 2013; Kruger & Van Rooy, 2017). This is evident in the different proposals and justifications for identifying innovations and errors in WE provided by researchers. In order to gain a better understanding of innovations and errors, a discussion about the different definitions and different classification criteria, is provided next.

Kachru (1992) refers to innovations as systematic deviations in NE varieties and notes that innovations emerge due to the nativisation of English where it is used in contexts that are different from native speaker contexts and not due to deficient acquisition. Whereas an innovation is a result of nativisation in WE, “an error is simply a mistake or uneducated usage” Bamgbose (1998: 2). According to Gut (2011: 120), errors and innovations cannot be clearly distinguished but are “structures representing two end-points of a continuum”. Distinguishing between errors and innovations is a contentious but crucial issue because corrective feedback is an essential part of L2 teaching and learning and for the classification of nativised features.
According to Mesthrie (2004) even though prescriptivists are quick to judge the deviations from StE as errors, the context in which ESL operates must be taken into consideration. Mesthrie (2004: 808-809) further remarks that the English “must therefore become referentially adequate to describe local typography, fauna, customs and . . . has to blend in with the local linguistic ecology by being receptive to favoured turns of phrases, structural possibilities and habits of pronunciation”. This whole process is what Kachru (1992) refers to as nativisation (cf. Section 2.8.8 for a discussion on nativisation).

A mistake may be unacceptable by a native speaker since it does not belong to the linguistic norm of the English language, it cannot be justified with reference to the sociocultural context of a non-native variety: and it is not the result of the productive processes used in an institutionalised non-native variety of English.

(Kachru, 1992: 62)

Kachru advocates for a distinction to be made between what he calls a “deficient variety” and a “different variety” with the former being associated with performance deficiency and the latter referring to a variety that is not similar to the L1.

Although research into WE is increasing, distinguishing between an error and an innovation is difficult due to the lack of codification in WE (Hamid & Baldauf, 2013). Scholars have provided several criteria to be used in distinguishing between an error and an innovation. These include high frequency counts to show grammatical stabilisation (with different frequency count scales resulting in different category classifications, use of feature regularly and widely by different speakers to show acceptability in a speech community, authoritative and norm orientation, and attestation in both spoken and written registers (cf. Bamgbose, 1998; Gut, 2011; Van Rooy, 2011; Mesthrie & Bhatt, 2008; Van Rooy & Kruger, 2016). Hamid and Baldauf further note that, even though literature contains guidelines for the determination of errors or innovations, there are limitations to the criteria. Hamid and Baldauf (2013: 488) comment that “the spread of L2 forms and their acceptability are both long-term and relative processes. It is unclear how widely L2 features have to spread to be accepted as varietal features”. Kachru (1988) claims that a range of models exist for the description of English used in the Outer Circle. Furthermore, he advocates for the recognition of WE innovations and states that the context in which English is used as a L2
should be taken into consideration when describing WE. That way, a better distinction between errors and innovations can be made.

Van Rooy (2011) draws attention to the identity dimension, which is central to the DyM whereby errors are conventionalised after acceptance from acrolectal and native speakers. “Social forces, familiar to scholars of variationist sociolinguistics, operate on the feature pool, to lead to the (social) conventionalization and (individual) entrenchment in the grammar and/or lexicon” (Van Rooy, 2011: 192). Van Rooy re-examined the difference between errors which are popular in NE and learner Englishes and innovations whose occurrences are likely in NE than in Foreign language English contexts. He noted that the social context of use creates higher chances to use English in NE contexts than in Foreign language English contexts because in NE contexts, English is used more in education, the media and courts and NE speakers rarely communicate with native English speakers (Van Rooy, 2011).

Van Rooy (2011) suggests that innovations may start as errors and then eventually attain conventional status. He uses 3 features of AE namely the progressive aspect in South African English, “can be able to” in South African English and “enable + bare infinitive verb” in East African English to illustrate his point. Regarding the progressive aspect in South African English, Van Rooy points out that the feature should be regarded as an innovation in BSAE due to the acceptability and grammatical systematicity. “Can be able to” in South African English is regarded as a conventionalised innovation in BSAE as evidenced by its use in South African websites and by editors and native speakers of English in South Africa because it shows acceptability of the feature. Van Rooy (2011) cites evidence from a survey of East African English, which showed acceptance of “enable + bare infinitive verb” to show that the feature is an innovation rather than an error. The extension of the bare infinitive complement to the verbs “allow” and “force” are given by Van Rooy as evidence of errors due to their fewer frequencies in East African English.

Norm orientation and proficiency are regarded by Kruger and Van Rooy (2017) as important in identifying innovations in NE. To illustrate this, increased proficiency and normative intervention by editors and explicit feedback were analysed to determine how innovation features become conventionalised. From an analysis of five corpora with different proficiency levels, Kruger and Van Rooy reported that the extension of the
progressive to longer time spans could be regarded as an established feature of BSAE. In addition, the disappearance of potential innovations such as the quantitative overuse of the progressive and the omission of the auxiliary BE, was attributed to increased proficiency and normative influence. Frequency of use of a feature by a language user and how frequent a user is exposed to a variant are important factors to consider when determining whether a construction is an innovation or an error (Van Rooy & Kruger, 2016).

Van Rooy and Kruger (2016) are of the opinion that the acceptance of innovations depends on whether the innovations are not noticeable enough to be earmarked for normative correction. In addition, the innovations should be attested adequately in both written and spoken texts “to become entrenched in the grammatical representations of learners as they turn into advanced users of the New English” (Van Rooy & Kruger, 2016: 205).

Bamgbose (1998) suggested that internal criteria should be considered when discussing innovations. These include demographic and geographical factors, codification, authoritative and acceptability factors. Discussions about demography are centred on the number of users of a feature. If there are more people who use the feature, chances are high for the acceptance of an innovation. The geographical spread of an innovation and whether or not it is codified are important factors for the acceptance of an innovation. Authoritative factors relate to whether a feature is accepted and used by teachers, writers, publishers and in the media.

Acceptability is an important measure to determine whether an innovation has been acknowledged by a speech community (Bamgbose, 1998; Kruger & Van Rooy, 2017). Once an innovation has been accepted, it will be used regularly. Codification and acceptance of an innovation will distinguish an innovation from an error (Bamgbose, 1998). Kruger and Van Rooy (2017) note that the criteria for innovation is that a feature should be accepted by the user group. If a convention is not accepted, it will be confined to informal speech or it will be regarded as an error, which will disappear as one's proficiency improves. Kruger and Van Rooy further note that if a feature is accepted by gatekeepers such as editors, then it is considered an innovation. The corpus used in this study included editorials and newspaper samples. If the features studies were available in other text genres and in the edited texts, then they were considered as innovations.
Questions to ask when considering innovation status (Source: Bamgbose 1998:3)

- How many people use the innovation?
- How widely dispersed is it?
- Where is the usage sanctioned?
- What is the attitude of users and non-users to it?
- Since the demographic factors concern numbers, the question immediately arises: which speakers?

Gut’s (2011) research focused on cross-linguistic influence and how it manifests in WE and proposed a model for the emergence of structural innovations in WE. He investigated “whether the claim that the structural properties of New Englishes are based on learner errors that have become fossilised serves as a sufficient explanation of currently observed linguistic structures” (Gut, 2011: 103). One of the arguments made by Gut is that speakers’ and speaker communities’ norm orientations and attitudes are central to the categorisation of an innovation or an error.

Gut (2011) draws attention to the different methodologies that scholars have adopted in the study of structural innovations in NE as shown below:

- Comparing NE constructions with similar constructions in StE varieties.
- Comparing similar structures in different NE varieties.
- Comparing NE features with similar constructions in indigenous languages in a country.
- Comparing the differences between learner language and NE.

Gut (2011: 119) proposed a model for the development of structural innovations as shown below:

1. Some structures based on cross-linguistic influence form part of the language productions of individual speakers who are learning English. Depending on the stage of acquisition these will be direct forms of transfer rather than indirect ones.
2. Only some of these structures – mainly structural mixing and preference patterns – remain in the speakers’ language productions even when they have attained
proficiency in English. These are mostly structures that are widely accepted and associated with positive attitudes as markers of a particular variety of English.

Some of these structures are adopted by speakers of subsequent generations, some of whom might acquire English as a first language.

Gut (2011: 105) draws attention to the complexity of cross linguistic influence, which sometimes does not manifest as direct structural innovations but as “direct borrowing or loans, mixed structures, conceptual transfer, avoidance patterns, preference patterns, hypercorrection and a general facilitating effect in language learning”. According to Gut, for a structure to be labelled as an innovation, it needs to be attested in the language of fairly proficient speakers of a language.

Regarding the describing structural innovations in NE, Gut (2011) advocates for the consideration of different types of cross-linguistic influence and the factors that restrict their occurrence. With regard to whether a structure should be regarded as an error or an innovation, Gut (2011) recommends that it should be based on extralinguistic instead of linguistic grounds. He regards language policy and norm orientation of a speech community as important in distinguishing between an error and an innovation. Another interesting point made by Gut (2011: 120) is that errors and innovations should be regarded as two end-points of a continuum because “the same structural feature of a language can move along this cline depending on a complex of extralinguistic factors” such as acceptance, how a feature is spread demographically and geographically, whether a feature has been codified and use by authorities such as teachers and writers.

Gorlach (1995: 44) predicted the following differences between English as a native language (ENL) and ESL communities, which may in turn lead to innovations:

(a) Contact with world English may become attenuated in ESL countries after independence . . . which may in turn result in a certain degree of fossilisation of words, meanings and concepts.

(b) Reduced linguistic self-confidence and the lack of a norm for English can restrict innovation- an archaic register, which is looked on as correct because it is documented in the “best authors” may well appear to most users to be safer than experimenting with innovations not authorized by the model, say BrE.
(c) The restriction of English to certain domains is likely to make innovations uneven in distribution: deviances will be more frequent with less formal, less educated, less norm-conscious speakers of English outside domains in which correctness can be monitored. However, “innovations” in these varieties are likely to be looked on as mistakes if there is a chance of comparing them with native-speaker English.

(d) The dominance of indigenous languages in other domains, or their co-existence with English in the same functions makes transfers or calques likely or even inevitable. Whether these are properly considered to be innovations will depend on the status of the indigenous language, speakers’ attitudes to transfers and mixes and the frequency or stability of individual items. (Gorlach, 1995: 44).

Bamgbose (1998) notes that the major issue that often creates controversy is deciding when a given feature is an innovation or an error. Paradowski and Jonak (2012) are of the opinion that linguistic creativity shows the innovative nature of different cultures and communities. Buchstaller (2008) comments that investigations into WE have helped shed light on the creativity of different speakers as they use L2 English in different cultural contexts. Kachru (1992) advocates for researchers to consider the cultural context in which English is spoken as a L2 in order to come up with a better analysis of innovations. The formal and functional features of innovations have been the focal point of early research. Researchers (e.g. Kortmann, Lunkenheimer & Ehret’s, 2020) have documented features that are similar in different varieties of English.

Drawing on comments by Bamgbose (1998), Van Rooy (2011) and Kruger and Van Rooy (2017), the conceptualisation of an innovation for this study outlines that it should be consistently attested in different genres, by different participants and in both spoken and written texts. This will show grammatical stability and widespread use of a feature in a speech community. If a feature is attested in edited texts such as newspaper articles and editorials, then this study considered it to be an innovation because it shows acceptance by norm-providing authority (editors). Therefore, this study used both internal criteria and external norms to classify innovations and errors. Section 4.4.2.6 further describes how the distinction between an error and an innovation was made in this study.
The discussion above highlighted the difficulty in determining whether an observed feature is an innovation or an error because of lack of codification in WE. The different classification criteria suggested by researchers, though they have limitations, can be used to distinguish between innovations and errors.

Since innovation and errors have been discussed above, it is imperative to focus on the term “nativisation” in the next section because nativisation and innovation are intertwined.

2.8.8 Nativisation

Nativisation is linked to innovation because variations occur due to the nativisation process where WE adopt features that are different from StE (Schneider, 2004; Van Rooy, 2019). Van Rooy, 2011) suggests that innovations may start as errors and then eventually attain conventional status through the nativisation process whereby WE features are stabilised and conventionalised to become innovations.

Different terms are used to refer to nativisation in literature. In addition to “nativisation”, Gough (1996a) uses the term “indigenisation” to refer to the same phenomenon. Reference is also made on the localised varieties of English (e.g. Schneider, 2011). Kachru (1992) believes that nativisation is a result of different languages and cultures coming into contact and converging. Innovation and nativisation are interlinked because innovations are part of the nativisation process (Bamgbose, 1998).

In addition to English being a L2 in many former colonial countries, Schneider (2011) reports on the fact that in Africa and Asia, some children of ESL users are now growing up speaking English as their L1. In these cases, the input that these children receive comes from ESL speakers whose English may be nativised to reflect local cultures.

Schneider (2011) remarks that, although English was introduced to many countries as a result of colonialism, the end of colonialism did not signal the end of English usage. In fact, the number of people learning English continues to grow due to the financial and economic opportunities brought about by using English. As noted in Section 2.3.4, acceptance of nativised varieties of English is one of the controversial issues in WE. “The absence of codification makes morphological, syntactic and phonological innovations difficult to accept” (Bamgbose, 1998: 5).
According to Fishman (1992) the spread of English is not a unique phenomenon. Other languages (e.g. Latin, French and Spanish) have spread throughout the world. Fishman sees the nativisation process that English is going through as a sign that local people are incorporating their cultures in the way they use English.

In an examination on whether the substrate system of Bantu languages affects the nativisation of BSAE, Makalela (2007: 143-144) notes that, “the analysis of morpho-syntactic, phonological and discourse and pragmatic features showed that the rules of Bantu languages logic naturalise the variety, paving its way to becoming a distinct endonormative outer circle English”. Following the arguments above, one can conclude that scholars agree that nativisation is a result of ownership of English (e.g. Kachru, 1992; Gough, 1996a; Schneider, 2003, 2007; Makalela, 2007). It is clear from the discussion above that nativisation is closely associated with innovation. This is because an innovation is described as a feature that has been nativised in a speech community. It is also important to note that if a linguistic feature has not been nativised in a speech community, then it is considered as an error.

From the discussion in Section 2.8.7, it can be deduced that deviation from the norm when using a language can result in linguistic innovation. Therefore, it is important to turn our attention to the term "norm" in the next section.

2.8.9 Norm

The study of the global spread of English has produced different opinions among researchers. Some of the contentious issues that have been raised are about codification (which norm should be used?), deciding the acceptability of innovations, the factors determining a norm and “the issue surrounding the de-Englishisation of the cultural contexts of English in the institutionalized non-native varieties” (Kachru, 2006: 247). According to Lindsey (2019), although “Received Pronunciation” was used to describe the accent that was taught in schools in England, contemporary Britain is diverse and a range of accents are evident in academia, politics and the media. He further notes that due to the fact that the Southern part of Britain is powerful, wealthy, and influential, accents of the South are dominant in the mass media. Standard Southern British pronunciation is taught in schools in England, but it is different from the type of English
spoken in Scotland and Northern Ireland (Lindsey, 2019).

The situation is different for L2 English speakers. As discussed in Section 2.3.5, the setting and context for the acquisition of English as a L2 is different from that prevailing in countries where English is spoken as a L1. It then follows that in the case of L2 English, conceptual and methodological challenges exist when trying to identify norms (Lowenberg, 1992). Lowenberg attributes these challenges to the scarcity of research into WE. However, research into WE has been increasing since the start of the new millennium as evidenced by the increase in the number of journals that publish WE research (e.g. World Englishes, Asian Englishes and English World-Wide: A journal of varieties of English).

According to Kachru (1992) and Schneider (2011), with English having spread to different linguistic and cultural contexts, it is challenging to describe the norm for different WE speakers. Strevens (1992) emphasises that despite the non-existence of a norm-determining academy or authority, the teaching and learning of English grammar and vocabulary is the same globally.

Bamgbose (1998) gives Nigerian English as an example of a variety, which is not accepted even by the speakers themselves. Acceptance of Nigerian English is made difficult because of the belief that if the variety is accepted, then English will dissipate into a different language. Bamgbose notes that although Nigerian English is not accepted, this has not dissuaded linguists from documenting the features of Nigerian English. The financial benefits that come with English teaching and learning by native speakers of English is another angle pursued by Makoni (1992). If NNVE are accepted and codified, then there will not be a need to supply teaching resources that generate a lot of income for the native speaker countries like the USA and the United Kingdom. Schneider (2011) gives Australian English as an example of a variety whose innovation features were once not accepted. Presently, Australian English is accepted as a norm.

Researchers often use native speaker norms to identify innovations by consulting grammar books, dictionaries and also using native speaker intuition (Bamgbose, 1998). Bamgbose notes that there are usually several norms (e.g. feature norm, code norm and behavioural norm), which makes it difficult to identify innovations.
2.8.10 L1 or L2 speakers as teachers of English

The aforementioned discussion noted that in the context of NE, the norm is usually not provided by the native speakers but by indigenous educated ESL speakers (Van Rooy, 2011). In addition, teachers at school are responsible for norm-orientation (Kachru, 1991). Although it is practically impossible to have native speakers teaching English throughout the world as noted in Section 2.3.3 one of the debates in ESL teaching concerns a discussion about the best teachers between L1 English speakers or ESL speakers (Mesthrie & Bhatt, 2008). Quirk (1990) seems to favour L1 speakers as best suited to teach ESL because he believes that some L2 or non-native teachers may have poor command of English to adequately teach StE. Kachru (1991), Sridhar and Sridhar (1992) and Mesthrie and Bhatt’s (2008) opinions differ from Quirk’s. They assert that, in the majority of countries in Africa and Asia ESL is taught by non-native teachers due to lack of financial resources to recruit L1 speaker teachers resulting in learners receiving different kinds of input. This scenario is true for Zimbabwe where most teachers are L2 speakers of English. According to Sridhar and Sridhar (1992), the main goal of learning ESL in WE settings is for a speaker to converse with other ESL users.

In a study conducted by Brain (2005) on non-native or L2 speaker English teachers, L2 speaker teachers acknowledged the differences between L1 and L2 English teachers as speakers in terms of linguistic proficiency and teaching methods. This view is also shared by Moussu and Llurda (2008).

Although the discussion above has highlighted the differing views regarding who should teach L2 English, the reality is that, as mentioned in Section 2.3.3, there are not enough English native speakers to teach the increasing number of L2 English learners and the native speaker teachers are unwilling to relocate to other parts of the world.

Another important concept to consider in the next section is register because both the spoken and written registers were used in this study.
2.8.11 Register

Different definitions are provided by researchers in literature. Some of the definitions are as follows: Register refers to “particular varieties or styles of speaking and writing” (www.dictionary.cambridge.org). “There are some varieties of language which can be associated with neither groups nor individuals but with the occasions when they are used. These varieties are called registers” (Brook 1979: 81). Brook further states that when studying registers, one will be examining language use within context. “Register refers to stylistically defined language varieties associated with certain channels (such as spoken political speech)” (Schneider, 2011: 16). Brook (1979) refers to the study of registers as an investigation of language within the environment in which it is used. Another definition found in literature is that “a register is a linguistic repertoire that is associated, culture-internally, with particular social practices and with persons who engage in such practices” (Agha, 2004: 24). In this study, the term "register" is used to refer to spoken and written registers. The term “genre” is used to refer to different text types that are found in spoken and written registers (e.g. private dialogue, newspaper reportage, social letters, editorials, creative writing, public dialogue, etc.).

This section outlined important concepts that will play a vital role in answering the research questions and will guide the analysis of the data in subsequent chapters. The following section provides a conclusion of how the conceptualisation and models of WE along with the exposition on the DyM can lead to possible language change.

2.9 Conclusion

In order to contextualise ZE, this chapter reviewed literature on WE. From the discussion, it can be noted that the fear that diffusion of English and its use in different countries will lead to language change is contributing to the non-acceptance of WE. In addition, opinions differ on the impact of English’s spread and use worldwide. According to Graddol (2000), some critics (e.g. Quirk, 1990) argue that the use of different varieties of English especially by L2 speakers will destabilise the language into different fragments, leading to incomprehensibility of some English varieties (especially L2 varieties). Others believe that the adaptations that English has made in countries where it is spoken as a L2 are testament of the innovative nature of speakers of English in different social and cultural contexts. This
uniqueness, commentators argue, is manifested in the different ways in which English is spoken in countries like India, Nigeria and South Africa. This is the reason why Kachru (1992) refers to English as “the other tongue” in countries where it is spoken as a L2 or additional language.

As the number of people who use English as a L2 continues to rise, Kachru (2006) advocates for researchers to use new paradigms in order to gain new perspectives on linguistic creativity in different cultural contexts where English is used as a L2. This was seen from the different theoretical frameworks proposed by researchers to account for the diffusion of English into different contexts around the world. The research design for this study will include corpus linguistics. Therefore, the next chapter reviews literature on corpora.
Chapter 3

Corpora: An overview

3.1 Introduction

In this chapter a general review of literature on corpus linguistics is done since the study is corpus-based. Firstly, a description of corpora as well as the history of corpus linguistics is given. Thereafter, different types of corpora are highlighted. Considerations when designing a corpus are reviewed. Tools for corpus analysis are highlighted. To highlight the widespread use of corpora in the study of different AE varieties, some studies, specifically studies done on AE using corpora are discussed.

3.2 Corpora

Machine readability is a criterion used by researchers to define corpora as will be highlighted in the definitions provided in literature. Anthony (2013) refers to corpora as linguistic data, which need to be analysed using software tools in order to observe patterns and to describe linguistic phenomena. Kennedy (1998) considers a corpus as a basis for analysing and describing language. A corpus is “a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language under consideration” (McEnery & Wilson, 2001: 32). Being machine-readable means that a corpus can be manipulated using computer software. Another definition given by Biber, Conrad and Reppen, (1998: 12) is that “a corpus is a large and principled collection of natural texts”. One of the qualities of a corpus is that one can explore a wide variety of topics using a corpus and that it can be used to explore features of language that have not been observed before. In another definition of a corpus, Tognini-Bonelli (2010: 18) distinguishes between a text and a corpus by explaining that “given that a corpus is a collection of texts, the aim of corpus linguistics has rightly been seen as the analysis and description of language use as realised in text(s)”. The differences are highlighted in the table below.
Table 3. A qualitative comparison of a text versus a corpus (Source: Tognini-Bonelli, 2010)

<table>
<thead>
<tr>
<th>A text</th>
<th>A corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read whole</td>
<td>Read fragmented</td>
</tr>
<tr>
<td>Read horizontally</td>
<td>Read vertically</td>
</tr>
<tr>
<td>Read for content</td>
<td>Read for formal patterning</td>
</tr>
<tr>
<td>Read as a unique event</td>
<td>Read for repeated events</td>
</tr>
<tr>
<td>Read as an individual act of will</td>
<td>Read as a sample of social practice</td>
</tr>
<tr>
<td>Instance of parole</td>
<td>Gives insights into langue</td>
</tr>
<tr>
<td>Coherent communicative event</td>
<td>Not a coherent communicative event</td>
</tr>
</tbody>
</table>

From the different definitions given above, it can be concluded that researchers consider a corpus as being linguistic data which is machine readable, a representative of a language and can be used as a basis for analysing a language.

A better understanding of corpus linguistics can be gleaned by looking at how the subject has evolved over the years. This is done in the next subsection.

3.3 A historical overview of corpus linguistics

From a literature review, it can be noted that initial corpus-based research began in 1876 with the use of non-machine-readable text. The work of biblical and literary scholars like Roberto Busa in the 1950s can be considered as the background of gathering, indexing and concordance data (McCarthy & O’Keeffe, 2010).

The early 1960s saw the development of corpus linguistics with the compilation of the Survey of English Usage (McEnery & Wilson, 2001; Taljard, 2006). Nelson Francis and Henry Kucera created the first electronic corpus, the Brown corpus in the 1960s. Corpus linguistics did not develop much in the late 1960s because of a shift from the use of corpora as sources of data. According to McEnery and Wilson (2001: 6), Noam Chomsky was influential in the shift because he “changed the object of linguistic enquiry from abstract descriptions of language to theories that reflect a psychological reality, cognitively plausible models of language”. Although corpus linguistic research slowed down partly due to Chomsky’s criticism, it did not stop completely. McEnery and Wilson
illustrate this by noting the work done in different disciplines. For example, Alfonse Juilland did research in mechanolinguistics from 1956 to the early 1970s.

Another setback for early corpus linguistics as noted by McEnery and Wilson (2001) and Biber, Conrad and Reppen, (1998), was the labour intensiveness of compiling and processing corpora because technology was not as advanced as it is now. Technological advancement, particularly the availability of desktop computers, has led to the growth of corpus linguistics (Kennedy, 1998; McCarthy & O’Keeffe, 2010; Tognini-Bonelli, 2010). As the ability of computers to store large amounts of data increases, corpora are getting bigger and bigger. Even as the corpora are getting bigger and bigger, smaller corpora of less than one million words in size are being created for research as evidenced by this study, which utilised a corpus of ZE consisting of 356 007 words. Tognini-Bonelli (2010) draws attention to the scarcity of certain categories of data due to the difficulty in compiling them.

According to Schmied (1990) corpora are now being used as the basis of research in areas such as grammar and lexicography. This point is supported by McCarthy and O’Keeffe (2010) who point out that some of the areas of study where corpus linguistics has been utilised include pragmatics, discourse analysis, speech technology, sociolinguistics, forensic linguistics and language teaching and learning. They also predict that corpus linguistics will continue to benefit from technological advancement in the future.

3.4 Types of corpora

This section outlines that corpora can be grouped into different categories. These categories include; (i) general-, (ii) specialised-, (iii) written- and (iv) spoken-, as well as (v) tagged and untagged corpora, which will be discussed in more detail in the sections below.

3.4.1 General corpora

In a general corpus, texts belonging to different genres are collected systematically and general corpora are used as sources of data for linguistic research (Lindquist & Levin, 2018; Esimaje & Hunston, 2019). A general corpus is compiled without any particular research question in mind and can be used to answer any research question, or as many
as possible (Kennedy, 1998). General corpora are compiled for comparison of different languages or varieties of a language (Esimaje & Hunston, 2019). In addition, general corpora can be used as a resource for the creation of dictionaries, as was done with the Collins COBUILD advanced learners’ dictionary, which was published in 2018. Texts can be added to general corpora regularly, meaning that they can increase in size.

Examples of general corpora include the BNC, which has 100 million words and the Global Web-based English (GloWbE) corpus, which consists of 1.9 billion words from different countries. The next section discusses specialised corpora.

3.4.2 Specialised corpora

A corpus may contain texts that deal with a specific subject or genre (Koester, 2010). According to Flowerdew (2004) different criteria can be used to classify a corpus as being specialised. An example is when a researcher compiles a corpus of a variety of English such as the International corpus of learner English (https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html). A corpus can also be considered to be specialised if it deals with a specific subject matter such as economics or if it is compiled to include the same type of text like biology textbooks. A further discussion including text-based corpora is overviewed in the next section, namely written corpora.

3.4.3 Written corpora

The ease of compiling a written corpus is noted by Baker (2010) as the reason why there is more written than spoken corpora. An example of a written corpus is a corpus of student writing (Walsh, 2010). Written corpora can be specialised if the written texts included in the corpus deal with a specific subject matter (e.g. scientific texts). For example, the English Scientific Text corpus consists of journal articles in computer science, biology, electrical engineering, digital construction, microelectronics, linguistics, computational linguistics, informatics, mechanical engineering and is annotated (https://www.clarin.eu/resource-families/corpora-academic-texts). The English Scientific Text corpus consists of 35 million words and is available online. The next section outlines a different type of proficiency related corpus, namely spoken corpora.
3.4.4 Spoken Corpora

Spoken corpora are fewer and usually small due to the labour intensiveness of compiling them and the expenses involved (Kennedy, 1998; Adolphs & Knight, 2010). One needs to record speech and to transcribe it first before including the text in a corpus. Adolphs and Knight (2010) explain that despite the compilation of spoken corpora being labour intensive, the exercise is worthwhile because spoken data is rich and unscripted. An example is, when one is studying natural conversations, the fact that they are unscripted will reveal real language use. While the abovementioned corpora outlined different proficiency and genre types, the next section provides an overview of linguistically descriptive corpora.

3.4.5 Tagged and untagged corpora

A tagged corpus contains more linguistic information, which is added during annotation. Depending on the type of research to be conducted, an annotated corpus can be manipulated to hide tags if they are not needed (Anthony, 2013). Van Rooy and Schafer, (2003) draw attention to the value of tagged corpora, especially as a resource in natural language processing and in dictionary development. An example given is the Collins COBUILD dictionary, which was created based on the data found in the Collins corpus of 4.5 billion words (https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus). Considerations for corpus design are discussed next.

3.5 What to consider in corpus design

In this section, different issues affecting corpus design are discussed. One of the issues is the purpose of the corpus. Kennedy (1998) notes that the nature, design and size of a corpus to be compiled is dependent on the purpose of the corpus. Reppen (2010) and Baker (2010) state that if the research questions are well constructed, then the task of corpus building will be easy. Practical and ethical issues pertaining to corpus construction are some of the issues that need to be considered. Adolphs and Carter (2013) note that decisions need to be made about how to record, transcribe, annotate and analyse a corpus. According to Reppen (2010) when designing a corpus, the decisions that are made will have an impact on the corpus to be compiled. These decisions also include defining the
sample population and text size. Reppen adds that spoken corpora are more time-consuming to compile because one needs to record and then transcribe the text for it to be available electronically. Nelson (2010) suggests that one should create a corpus if there is no other existing corpus that can be used to answer proposed research questions\textsuperscript{23}. In some cases, where access to an existing corpus is difficult, a new corpus needs to be created. For the current research, part of Marungudzi’s (2016a) corpus was utilised in addition to the one that was compiled by the researcher. To the best of my knowledge, there is no other corpus of ZE that can be used for this study.

### 3.5.1 Representativeness of the corpus

One of the basic tenets of corpus design is representativeness (Reppen, 2010). Sometimes it is impossible to compile a corpus, which is completely representative of a language or a given population due to time limitations. One needs to get enough material that accurately represents a feature under study (Reppen, 2010; Adolphs & Carter, 2013). In the case of the current project, time and financial constraints meant that some of the samples that were supposed to be included in the corpus were not collected. However, the samples that are not included in my corpus are available in Marungudzi’s (2016a) corpus, which were added to the corpus compiled for the study. Koester (2010) and Biber, Conrad and Reppen, (1998) note that even though a corpus is small, it has to be representative. As an illustration, if one is investigating a specific subject, or genre, the collected samples should be accurate in representing the subject or genre. Due to practical limitations, representativeness may be difficult to achieve in a small corpus. “What is important is to ensure that the samples are collected from a range of fairly typical situations” (Koester, 2010: 69). This was done in the current study. Size is another important consideration when compiling a corpus, as will be highlighted below.

\textsuperscript{23} This study warranted a corpus to be compiled because, as mentioned in Section 1.1, I managed to get access to part of Marungudzi’s (2016a) corpus. Therefore, I had to compile a corpus to be combined with Marungudzi’s corpus in order to answer research questions posed.
3.5.2 Corpus size

Whilst scholars like Fillmore (1992) and Sinclair (2004) consider large corpora to be better, other scholars believe that the efficacy of corpora is not based on size but on the type of research to be carried out on it (e.g. Reppen, 2010; Nelson, 2010; Anthony, 2013). Therefore, a corpus should be big enough to be representative of the language to be investigated.

Although the question of size and representativeness affect the validity and reliability of the corpus, it has to be stressed again that any corpus, however big, can never be more than a miniscule sample of all the speech or writing produced or retrieved by all of the users of a major language on even a single day.

(Kennedy, 1998: 66)

Technology advancement, especially the availability of computers and the fact that corpus sizes are getting bigger has enabled researchers to study features that were not possible before due to limitations in the size of the corpus (Taljard, 2006; Taljard & De Schryver, 2016).

Practicality is another factor to be considered (Reppen, 2010). Sometimes time and financial limitations mean that the corpus size may be affected. Some of the important issues to consider when looking at the size of a corpus include the number of words per sample and the number of texts from various categories (Biber, Conrad & Reppen, 1998). Another point raised by Biber, Conrad and Reppen is that size alone cannot make a corpus representative but diversity is also key for representativeness. McEnery and Wilson (2001) predict that the number of words in a corpus will keep growing as technology keeps on advancing. However, although corpora are getting bigger, Anthony (2013) predicts that smaller specialised corpora will keep growing in number owing to small research projects being carried out. The number of corpora in different languages also continues to rise. Anthony gives an example of the corpus sizes as represented by figure 7 below.

---

24 As mentioned above, time and financial constraints led to the corpus being 356 007 words.
25 The corpus used in this study is an example of a small corpus.
From figure 7 above, it can be deduced that early corpora were smaller compared to the ones that are more recent. For example, the British National Corpus (BNC), which was compiled in 1994, consists of 100 million words. In contrast, the more recent Oxford English Corpus consists of 2 billion words and the Cambridge English Corpus contains 1.6 billion words. The Collins Corpus (COBUILD), which started with less than 100,000 words in 1987 now has over 4.5 billion words (https://www.collinsdictionary.com/cobuild/). Anthony (2013) credits this increase to technological advancement, which has made it possible to store and analyse bigger corpora. The discussion above shows that corpus compilation projects and the number of words in a corpus have continued to rise over the past 50 years due to technology advancement and the increased use of corpora in linguistic studies (Anthony, 2013). Some of the merits of a computer include accuracy in data processing and speed and the handling of large quantities of data (Kennedy 1998; McEnery & Wilson, 2001). A third consideration in corpus design is the concept of balance in text size, which is outlined in the next section.
3.5.3 Balance in text size

Balance can be achieved through sampling equal numbers of texts for each category (McEnery, Xiao & Tono, 2006; McEnery & Hardie, 2012). According to Esimaje and Hunston (2019), the number of words in each genre should be the same in order to achieve balance. Balance also involves not including too many similar genres at the expense of others. In addition, if a corpus is being compiled to include both spoken and written registers, then, there should be equal numbers of texts for each register. Esimaje and Hunston draw attention to the difficulty of achieving balance when collecting spoken texts because of the time-consuming nature of recording, transcribing and annotation.

Due to the fact that other factors such as size can affect the balance of a corpus, “it can be seen that achieving balance involves a great many subjective decisions” (Esimaje & Hunston, 2019: 15). Since balance may be difficult to achieve, normalised frequencies can be used to compensate for the shortcomings of an unbalanced corpus (Esimaje & Hunston, 2019). The following section provides a discussion about different corpus analysis tools that are used by researchers.

3.6 Corpus analysis tools

Technological advancement has led to the creation of powerful software tools than those that were available for research during the 1960s and 1970s (Anthony, 2013). Some of the computer software available since the 1980s include Wordcruncher and the Oxford Concordance Program. In the 1990s, more advanced software like Mike Scott’s 1996 WordSmith was developed. The new generation tools are able to perform complex data analysis tasks on corpora (Kennedy, 1998).

Currently, there are software tools that can handle billions of words in a corpus. An example is Sketch Engine (https://www.sketchengine.eu/). Although the software tools can handle huge amounts of data, they have their own disadvantages. For example, the tools require data to undergo cleaning, processing, reformatting, indexing and uploading to a server before one can start analysing the data (Anthony, 2013). There may be limits in accessing the server. Some software tools like Sketch Engine need one to register and to pay for accessing them. Another issue “is that fourth generation tools blur the boundaries between the corpus data and the tools used to observe it. Due to the way these
tools store the corpus data in an index form on an external server, users have no way to observe the raw data directly with their own eyes” (Anthony, 2013: 153).

Due to design variation in corpus analysis tools, this will have an effect on corpus analysis. To illustrate this point, Anthony (2013) gives an example where developers use different definitions of a word. The differences will in turn have an impact on word frequency counts, key words and collocations.

### 3.7 Some of the studies done on African English using corpora

When the ICE project was launched, Schmied (1990) highlighted the benefits of using corpora in NNVE such as using corpora for dictionary compilation, aiding in the study of nativisation, morphosyntax, grammar, lexicography and phonology in AE. Greenbaum (1996) also predicted the possibilities of using ICE corpora to carry out contrastive research between L1 and L2 varieties, between L1 varieties and between regional varieties. Another prediction made by Schmied (1990) was that corpora in AE would enable researchers to quantitatively compare L1 and L2 varieties of English to determine whether there is stylistic variation. In addition to the benefits, Schmied (1990) also drew attention to the challenges of compiling corpora of L2 varieties of English. One of the challenges is that it is difficult to obtain some text types in L2 varieties of English.

A lot of research has since been carried out since Schmied (1990) highlighted the benefits and challenges of compiling and using corpora to study AE varieties. Corpora of English as a L2 have been collected under the ICE project for different countries in Africa. These include the ICE-East Africa, ICE-South Africa, ICE-Ghana, ICE-Namibia, ICE-Nigeria, ICE-Uganda. The predictions made by Schmied (1990) and Greenbaum (1996) have been realised since a lot of research has been carried out using ICE corpora, especially in studies that focus on AE varieties. Some of the studies done on AE using corpora are highlighted below.

Werner and Fuchs (2017) examined the present perfect tense in Nigerian English. They based their analysis on the Nigerian portion of the International corpus of English (ICE). “Results indicate an overall reduced present perfect friendliness of Nigerian English and show internal variability in terms of present perfect frequencies in different variable
contexts” (Werner & Fuchs, 2017: 129). Werner and Fuchs suggest that this could be a result of Irish and American English’s influence.

A comparative study of the frequency and stylistic variability of stance markers was done between West African English varieties (Nigerian English and Ghanaian English) and BrE by Gut and Unuabonah (2019). Three corpora were analysed namely ICE-Nigeria, ICE Ghana and ICE-GB. Gut and Unuabonah noted that Nigerian English and Ghanaian English users show systematic differences in the way they use stance markers. In addition, the frequency of stance markers was reported to be higher in BrE and lower in Nigerian English and Ghanaian English.

Using the historical corpus of English in Ghana and the ICE-Ghana, Brato (2019) examined lexical expansion at the nativisation stage of the DyM in Ghanaian English. Results from this study show that semantic shift plays an important role in word-formation processes and the traditional processes namely derivation or compounding play a lesser role in word formation in Ghanaian English.

Online corpora have also been used in the study of AE. For example, the Nigerian component of the GloWbE corpus was utilised by Unuabonah and Daniel (2020) to examine bilingual interjections in Nigerian English. Unuabonah and Daniel (2020) report that the use of haba, kai, chei, chai and mtchew is evidence of innovation in Nigerian English because the words are borrowed from Hausa, an indigenous Nigerian language.

Another area of focus in the study of AE is learner language. Esimaje (2019) studied the Nigerian and Cameroonian English Learner Language corpus consisting of 442,939 words to determine how Nigerian and Cameroonian learners use lexical forms and tenses. The author reports on the difficulty of learning the lexical form of capitalisation and tense.

As has been discussed in detail in Sections 2.8.2 and 2.8.3, BSAE is one type of AE that has received extensive scholarly attention using corpora in different areas such as morphology and syntax (e.g. Mesthrie, 2006, 2008; Makalela, 2007, 2013; Kruger & Van Rooy 2017) and phonology (e.g. Gough, 1996a). In addition to the studies reported in Section 2.8.3, another study that examined BSAE was done by Partridge (2019) using a parallel corpus. He focused on the written register. The ditransitive construction, anti-deletion in terms of the finite that-complementiser and anti-deletion in terms of the non-
finite infinitive to-marker were analysed. Results from this study show an increased use of the prepositional dative allostruction\(^{26}\) by BSAE speakers when compared to white South African English speakers and BSAE speakers show an increase in the omission of the that-complementiser in the reportage and academic registers. With regards to the anti-deletion of the infinitive to-marker, no significant differences were noted between BSAE speakers and white South African English speakers.

As gleaned from the discussion above, AE research has benefited immensely from the use of corpora. Researchers have utilised corpus linguistics as a method in AE research, with different research topics being explored using corpora of different sizes. Given that this study follows the ICE corpus design loosely, we will discuss the ICE below.

### 3.8 The International Corpus of English

The ICE is a project in which teams around the world are compiling one million-word corpora each of different varieties of English (http://ice-corpora.net/ice/). The project started in 1990 in 26 countries and research teams focus on different varieties of English. Some of the countries involved in the ICE are Australia, Cameroon, Kenya, India, Ireland, New Zealand, USA, Fiji, Britain and South Africa.

A standard corpus design and annotation scheme is being followed by researchers to allow for components to be compared. Teams compile data from participants aged at least 18 years whose medium of learning at school is English. Participants are supposed to have completed at least secondary education. Spoken and written samples are included in the corpora following a prescribed corpus design. For annotation, textual mark-up, word-class tagging (using an ICE tagger) and syntactic parsing are utilised. Teams choose the level of annotation they want. The ICE-GB, a corpus that is a part of the ICE project, was used for comparison with the ZE corpus in order to investigate variations.

---

\(^{26}\) Allostructions are “contextually realised variants of an underspecified verb-particle construction. Various other authors have demonstrated the viability of the allostructural analysis by applying it to numerous other alternations, e.g. the dative alternation and the locative alternation in present-day English” (De Vaere, Kolkmann, & Belligh, 2020: 97).
The following section provides a synthesis of the different corpora types, important considerations for corpus design and the analysis tools needed.

3.9 Conclusion

A literature review on corpora was done in this chapter. A corpus was defined and then a brief historical overview was given. Types of corpora were discussed together with considerations when designing a corpus such as representativeness, size and balance. This chapter also explained the different corpus-analysis tools. Some of the studies done on African English were highlighted, to contextualise the discussion, since ZE is a type of African English. Since this study followed ICE guidelines (with some deviations) on corpus design and annotation, the ICE was described. The next chapter discusses the research design and methodology that was used in this study and its strengths and weaknesses.
Chapter 4

Methodology

4.1 Introduction

The previous chapter provided a general background exposition on the use of corpora in research by outlining the history of corpora, different corpora types and possible corpus designs as well as tools that can be used in the analysis of corpus data. This chapter will outline the specific research design and methodology for the current study, which is used to investigate variation in ZE by examining selected features that are common in other WE as explained on eWAVE (cf. Section 1.6) and outlined in the research questions of this study.

The unit of analysis (what is being studied) in this study is the use of articles, modal verbs and selected DMs (so, well and but) within two corpora, both in ZE and ICE-GB, in order to determine whether there is innovation in ZE through the theoretical perspective of the DyM. After obtaining ethical clearance from the ethics committee at the University of Pretoria (ethics clearance number: GW20181012HS), the researcher gathered data that would assist to answer the research questions posed in this study (cf. Section 1.7).

The research questions that are tailored to gain insights into patterns of use of articles, modal verbs and selected DMs to ultimately measure ZE against the five different stages of development in L2 varieties of English in the DyM, are mentioned below for ease of reference:

1.1 What are the occurrence frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?

1.2 How does the occurrence of articles, modal verbs and discourse markers in ZE compare to that of BrE?

1.3 Are the variations between the corpora statistically significant?
2.1 What are the function frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?

2.2 How do the functions of articles, modal verbs and discourse markers in ZE compare to those of BrE?

2.3 Are the variations between the corpora statistically significant?

3.1 Which features of innovation are present with regards to articles and modal verbs in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE?

4. What is the stage of adult use of these features in ZE when considering the dynamic model of postcolonial English?

The following section outlines the difference between the concepts of method and methodology as this differentiation plays a vital role in the conceptualisation of the data collection- and analysis strategies of this study.

**4.2. Differences between methodology and method in this study**

A distinction is made between methodology and method. For the purpose of this study, the term “method” is used to refer to the research design that was used, which is the corpus linguistics. The term “methodology” is used to refer to the different techniques used in the data collection process. The following section, section 4.3 outlines the research design and method, while the section thereafter, section 4.4, will outline the research methodology for this study.

**4.3 Research design: Corpus linguistics method**

As stated in Section 1.9, a corpus linguistics method was used for this study. Researchers (e.g. Kennedy, 1998; McEnery & Wilson, 2001; Anthony, 2013; Biber, Conrad & Reppen, 1998; Bowker & Pearson, 2002; Gries, 2009; Lindquist & Levin, 2018; Esimaje, Gut & Antia, 2019) consider corpus linguistics to be a research method and not a new branch of linguistics such as morphology or psycholinguistics. McEnery, Xiao and Tono (2006) point
to the use of corpus linguistics in the exploration of different research areas in linguistics as evidence that it is an approach or method.

The major elements of corpus linguistics that make it suitable for this study are the following: (cf. Biber, Conrad & Reppen, 1998).

(a) It involves the analysis of real language texts.
(b) The basis of analysis is a corpus, which is machine-readable.
(c) Corpus analysis software is used to aid in the analysis of corpora.
(d) Both qualitative and quantitative analysis methods are utilised in the interpretation of results.

Corpus linguistics as an approach or method is suitable for this study because it uses corpora as the basis for analysis. In addition, both quantitative- and qualitative research methods are used to complement each other (Biber, Conrad & Reppen, 1998). A corpus-based approach is used in this study. The starting point for corpus data analysis is firstly focussing on the quantitative data collection in terms of frequency count determination of the features in question (cf. subsection b in Section 4.4.2.6 for an exposition) after which the qualitative functions of the features are determined (cf. subsection d in Section 4.4.2.6 for an exposition). Both the data and conclusions drawn from both the qualitative and quantitative aspects of the corpus-based approach will play a vital role in determining the variations reported in Kortmann, Lunkenheimer and Ehret (2020). Kortmann, Lunkenheimer and Ehret (2020) outline a list of possible morphosyntactic variations from which to base the analysis and to do comparisons with other varieties of English, therefore, the corpus is used for testing existing hypotheses about variation in L2 English usage.

The following section provides a discussion on the differentiation between corpus-based- and corpus driven approaches in order to outline why the study employs a corpus-based approach.

4.3.1 Concepts of corpus-based and corpus driven approaches

In corpus linguistics the corpus-based and corpus-driven approaches are adopted by researchers to make use of corpus data. A corpus-driven approach considers that direct observation of corpora should be the starting point of analysis. According to Tognini-Bonelli (2001) and Baker (2010), a corpus-driven approach is inductive and does not depend on
pre-existing theories to derive examples of language use. In the corpus-driven approach, evidence from a corpus is used to obtain new insights into a language (Römer 2005; Gries, 2012). Römer (2005) notes that in a corpus-driven approach, data is observed first in order to develop new theories of language from the findings.

To give examples of the corpus-driven approach, Tognini-Bonelli (2001: 101) studied “the way in which word choices interact with each other in fairly close proximity”. An examination of the corpus revealed that co-selection is hidden and unconscious, which makes it crucial in communication and that a variety of textual meanings emanate from the co-selection of two or more words. Römer (2005) utilised a corpus-driven approach to analyse the use of the progressive in BrE and concluded that there are differences in the way progressives are used in schools when compared to the way they are used in speech.

The other approach, a corpus-based, regards corpora and pre-existing linguistic theories as complementing each other (McEnery, Xiao & Tono, 2006; Taljard, 2006; Anthony, 2013). Corpus-based approaches test and improve existing linguistic theories (Gries, 2012). Biber, Conrad and Reppen, (1998) comment that, with the corpus-based approach, previously unexplained aspects of language can be examined.

[A c]orpus-based approach . . . is often complemented by introspection and information gleaned from mother-tongue speakers who act as informants, especially in cases where the linguist is not a mother-tongue speaker of the language

(Taljard, 2006: 69). Another reason why the corpus-based approach is suitable for this study is that, from an analysis of corpora, researchers can determine whether linguistic features are used distinctively (Partridge, 2019). In addition, researchers can also check if a feature has been nativised in a certain language variety (Van Rooy, 2011). Comparisons can also be made between different varieties of a language within a country, region and even between L1 and L2 varieties of a language (De Klerk, 2003a; Van Rooy, 2013).

Consistency, robustness and reliability are some of the advantages of using a corpus-based approach (Biber, Conrad & Reppen, 1998; McEnery, Xiao & Tono, 2006). This is because of the empirical nature of the corpus-based approach, the use of both qualitative and quantitative methods and the use of computers in the data analysis process (De Klerk,
This study used the corpus-based approach, which is suitable for this study, because a computer was used to do quantitative corpus analysis. Specifically, Sketch Engine corpus analysis software tools were used to generate frequency lists and to search for Key Word In Context (KWIC) and then analysing the concordance lines for functions of articles, modal verbs and DMs (so, well and but).

To highlight how a corpus-based analysis was used in research, two examples are given here. Utilising a corpus-based analysis, Louw and Jordan (1993) compiled a frequency list of the nouns he, she, man, boy, woman and girl using a corpus of eight English books used in Zimbabwean secondary schools. Their aim was to avail the corpus as a resource to be used in areas such as teaching and learning, research, writing and curriculum planning. Using a corpus-based approach to investigate the native and NNVE to determine the extent of pluralisation of mass nouns, Schmidtke and Kuperman (2017) examined the GloWbE. The results suggest that NNS used pluralised mass nouns more often than native English speakers. In Sections 4.3.3 and 4.2.4, introspective and observational data as well as qualitative versus quantitative data are explored respectively as they formed part of the data analysis process in this study.

### 4.3.2 Introspective data versus observational data collection and analysis

Introspective data and observational data are the two types of data that researchers collect in order to study language (Gries 2013a). Introspective data is “very informally collected linguistic acceptability judgements (primarily by the analyst him/herself) as the primary source of data” (Gries, 2003: 94). Svartvik (1992) draws attention to the challenge of using introspective data and native speaker intuitions only, to describe linguistic competence and performance. Despite the abovementioned challenge, introspective data is still regarded as a useful form of data (McEnery & Wilson, 2001).

Observational data is another form of data that is regularly being used in the corpus linguistics method. In corpus linguistics, corpora as observational data enable researchers to utilise usage-based approaches to observe and analyse naturally occurring language in order to identify and quantify linguistic features (Gries, 2006b). This enables researchers to determine whether the features have been conventionalised in a language
variety. The ZE corpus and the ICE-GB are the two examples of observational data used in this study.

Introspective data is not sufficient in this study because the study examines how L2 speakers of English use articles, modal verbs and selected DMs (so, well and but). Using observational data will allow for a description of the features studied and to check whether the features are prevalent in ZE. In addition, observational data is suitable for a comparative analysis done between ZE and BrE.

The use of both introspective data and observational data is considered beneficial to language research (Fillmore, 1992; Croft, 1998; Gries, 2006a; McEnery, Xiao & Tono, 2006). To demonstrate the complementary nature of these two types of data, Fillmore (1992: 12) says that:

> The first [observation] is that I don’t think there can be any corpora, however large, that contain information about all the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I’ve had the chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

With the understanding that a corpus cannot contain all the information about a language, Fillmore (1992) asserts that introspection and natural language observation are all beneficial in the study of language and should complement each other.

In this study, the ZE corpus and the ICE-GB provided observational data that allowed the quantification of the features studied in terms of calculating the frequency with which the features occurred in the corpora. A usage-based approach was utilised to ascertain whether there is innovation in ZE with regards to articles, modal verbs and selected DMs. The ZE corpus consists of both spoken and written texts by L2 users of English in Zimbabwe. The inclusion of written text (e.g. editorials and newspaper reportage) allows the researcher to check and establish whether the features are acceptable to editors, who are regarded as norm-orientation authorities.
One of the advantages of using corpora is that the study of actual language use is the basis of claims made about a language (Biber, Conrad & Reppen, 1998). With observational data from the two corpora used in this study, “real life” language use was examined and quantified. Corpus data provided frequencies of occurrences of the studied features and it was possible to draw examples from the corpus to illustrate the features. The frequency counts cannot be obtained from intuition and anecdotal data (Van Rooy (2013). Therefore, a corpus-based analysis is suitable for this study.

To illustrate how frequency counts obtained from corpora can be used to determine whether a feature is conventionalised in a variety, Minow (2010) shows that, for instance, the prevalence of the unmarked verb forms in the past time context is 15%. This evidence from a corpus is valuable because it can support or contradict existing theories about BSAE.

The aforementioned advantages support the use of corpora in this study because the ZE corpus was used to ascertain whether the morphosyntactic features that show variation with articles and modal verbs and that these constructions and associated features could be nativised and conventionalised in ZE. Regarding DMs, comparisons were made between ZE and BrE using the ZE corpus and the ICE-GB to determine whether there are similarities or differences in the use of selected DMs (so, well and but). Results from the analysis of articles, modal verbs and DMs will be used to determine the stage of ZE in the DyM.

It is also important to note that the biographical and language background questionnaires (as extra linguistic proficiency data) was another source of data in this study. Results obtained from the questionnaires were used in the interpretation of the qualitative data. From this section it also becomes clear in order to determine at what stage of the DyM the construction and associate features falls, the study needs to look firstly at the quantitative data in terms of the frequencies with which these features occur and secondly qualitatively at the functions in which these features occur. For this reason, the following section outlines the differentiation between quantitative and qualitative data analysis and the specific application of the corpus-data of this study.
4.3.3 Quantitative versus qualitative data collection and analysis

A better understanding of the research problem is gained through the blending of quantitative and qualitative research designs in a corpus linguistics method, (Biber, Conrad & Reppen, 1998). Corpus linguistics research method is suitable for this study because “qualitative analysis can provide greater richness and precision, whereas quantitative analysis can provide statistically reliable and generalisable results” (McEnery & Wilson, 2001: 77). Since a corpus linguistics approach combines quantitative and qualitative methods, in this study, the quantitative approach was used to obtain statistical data such as frequency counts in chapters 5, 6 and 7. The extralinguistic proficiency data, i.e. biographical and language background questionnaires were used to complement the qualitative interpretation of the data analysis.

Section 4.3 described the research design, namely corpus linguistic method, which was used in this study. The discussion also showed that a corpus and usage-based approach was followed in this study. In order to show why a corpus-based approach was chosen, comparisons were made between corpus-based and corpus driven approaches. This study is suited for a corpus-based approach because corpora were used to check whether existing theories about the use of articles, modal verbs and selected DMs can be confirmed or disregarded. It was also shown that the quantitative method or approach was used to determine the frequencies and functions of the selected features. The qualitative method was used to explain the functions of the features in the corpora and the biographical and language background questionnaire results were used to aid the interpretation of the data.

In the next section, the research methodology employed in this study will be explained. The discussion will focus on instrument development, the pilot study, sampling of participants, data collection, transcription and corpus compilation.
4.4 Research methodology

In this section, the different techniques employed to collect and analyse data are discussed. The following stages and steps within the stages were followed during the research process:

4.4.1 Stage 1: Theoretical research

The first stage was theoretical in nature. In order to inform on the theoretical framework of this study, a comprehensive literature review was carried out until the end of the research project.

4.4.2 Stage 2: Empirical research: Data collection

After obtaining ethical clearance from the faculty of Humanities at the University of Pretoria, the researcher proceeded to do the empirical part of the study. Empirical research was done in different steps as explained below.

4.4.2.1 Step 1: Pilot study and instrument development

In this section, firstly, an explanation is given about how the pilot study was conducted. Secondly, the manner in which the instruments were developed will be discussed. Both these steps are included in this section because conducting the pilot study played a vital role in creating and adjusting the research instruments for the main study. A detailed exposition is given in the two sections below.

Before embarking on the pilot study, the researcher first requested and was granted permission from various institutions and organisations as follows.

a. From North-West University to use Marungudzi’s (2016a) corpus.

b. From the University of Zimbabwe to carry out research at the institution.

c. From the Herald and the Sunday Mail to access their newspapers online.

A pilot study was carried out to test whether the questionnaires and the semi-structured interview schedule solicited suitable data for the study. The pilot study was carried out with participants who were similar to the target population for this study. Potential participants for the pilot study were contacted and recruited by means of snowball sampling and
subsequent convenience and purposive sampling. This recruitment occurred either via
telephone or informal meetings conducted in English, either in participants’ homes or the
residence of the researcher. In the questionnaire, participants were asked about their
biographical information and information and linguistic backgrounds. During this sampling
participants were asked to answer self-evaluating questions about their English language
proficiency. This was done to determine their suitability to participate in the study. But, it
has to be acknowledged that there are limitations of assessing linguistic proficiency this way
because some participants may not disclose the truth. Private conversations and private
semi-scripted dialogues were recorded with 10 participants to test whether recorded
conversations were suitable for corpus compilation. For written samples, the researcher
requested and was granted permission from The Herald and The Sunday Mail newspapers
to access online newspapers to compile a corpus. The setting for the pilot study is discussed
next.

a. Pilot study setting

For the pilot study, data was gathered in different parts of Zimbabwe. Private semi-scripted
dialogues and private dialogues were recorded in different social situations, which
depended on the snowball, purposive and convenience sampling. The setting was in
participants’ private homes or any preferred location where participants chose, or the
residence of the researcher. Below is an explanation of how instruments were developed.

b. Development and use of the research instruments

Different instruments were developed for this study, including (i) a biographical and
language background questionnaire (used for extralinguistic proficiency data in order to aid
the discussion on the language backgrounds of participants), (ii) a semi structured interview
schedule and (iii) informed consent forms for written and spoken texts. The purpose of
the semi structured interviews was to guide the conversations and to obtain data and create
a corpus in order to look for the above-mentioned features. Specific questions were included
in the questionnaire in order to solicit for biographical information and linguistic
backgrounds of participants to inform their language choices– this information was used as
extralinguistic proficiency data and did not form part of the compiled corpus. Collecting

---

27 The questionnaire, semi-structured interview schedule, and consent forms are shown in appendixes A, B, C and D.
background information is beneficial because the details can be used to add contextualisation and provide qualitative and objective conclusions when interpreting the quantitative data (Koester, 2010). The proficiency levels of the possible participants were also informed from these questionnaires. The purpose of the questionnaire was to obtain biographical information and linguistic backgrounds of participants and to ensure that participants who were able to converse in English were selected. The questionnaire contained what Dörnyei (2010) refers to as factual and behavioural questions. In factual questions, the researcher asked participants about their demographic features. These include age, address, gender, occupation and education. Using behavioural questions, participants were also asked about their language history and language use. Some of the questions required participants to write their responses on the blank spaces provided. Other questions required participants to choose from the given responses by ticking a box. Dörnyei (2010) is of the notion that questionnaires are popular because they are easy to construct. However, he warns against the assumption that every question posed in a questionnaire will draw the right answer. Instead, he suggests that researchers take extra care in formulating the right questions that will elicit suitable answers because “a well-constructed questionnaire can reduce the bias of interviewer effects and thus increase the consistency and reliability of the results” (Dörnyei, 2010: 6).

It was important to ask questions about proficiency because this assisted in the selection of participants for the study. Allo and Mesthrie (2004) report on the four proficiency levels that were proposed by Brosnahan in 1958. The levels are: level 1, which is used by people with no formal education, level 2 is used by those who finished primary education, level 3 for those who completed secondary school and level 4 for university graduates. According to Allo and Mesthrie (2004), Brosnahan assumes that proficiency levels increase, as one gets more educated. Allo and Mesthrie (2004) comment that the above assumption is not flawless as other factors like more exposure to a language may contribute to higher proficiency levels even for less educated people. This study used speakers with varying linguistic proficiency levels. The variation is due to acquisition, which occurs mainly at schools. Secondly, a semi-structured interview schedule was created in which questions were included to solicit language, which would contain the specific features outlined above. These interviews were conducted in participants’ private homes or any location that participants preferred.
The corpus design for this study was loosely based on the ICE (cf. Section 3.8). Under the ICE project, different teams around the world have compiled corpora of different English varieties (one million words each). The teams follow the same corpus design and use the same corpus annotation scheme to allow for comparison between components (Kirk & Nelson, 2018).

The following subsection will provide an overview of the participants used in the pilot study in terms of the corpus linguistic data as well as the extralinguistic data.

c. Participants

Ten Shona L1 speakers who used English as their L2 and have had at least 10 years of formal education participated in the pilot study. Their ages were 18 years and above. Participants’ linguistic proficiency varied in this study. From these participants, the following data was obtained.

i. Participant information - the Extralinguistic proficiency data

Participants completed biographical and language background questionnaires, which were then used as extralinguistic data. The table below shows the language backgrounds of participants in the pilot study.
Table 4. Summary of Linguistic backgrounds for participants in pilot study

<table>
<thead>
<tr>
<th>Participants</th>
<th>Mother tongue</th>
<th>Other languages spoken</th>
<th>Age when English was learned</th>
<th>Where English is used</th>
<th>Hours per week spent speaking English</th>
<th>When English was learned</th>
<th>Where English was learned</th>
<th>How many years has English been used?</th>
<th>English proficiency</th>
<th>Years taught by English L1 speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;PP1&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Work</td>
<td>15</td>
<td>School</td>
<td>School</td>
<td>15</td>
<td>Average</td>
<td>1-5</td>
</tr>
<tr>
<td>&lt;PP2&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Work, friends</td>
<td>25</td>
<td>School</td>
<td>School</td>
<td>30</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP3&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>5</td>
<td>Home, school, work, friends</td>
<td>10</td>
<td>Before school</td>
<td>School</td>
<td>25</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP4&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Home, school, friends</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>13</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP5&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Friends</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>56</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP6&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>5</td>
<td>Home, work, friends</td>
<td>40</td>
<td>Before school</td>
<td>Home</td>
<td>23</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP7&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>5</td>
<td>Home, work, friends</td>
<td>15</td>
<td>Nursery school</td>
<td>Home</td>
<td>31</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP8&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>4</td>
<td>School</td>
<td>4</td>
<td>Nursery school</td>
<td>School</td>
<td>17</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP9&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Work, friends</td>
<td>30</td>
<td>School</td>
<td>Home</td>
<td>32</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;PP10&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>7</td>
<td>Work, strangers</td>
<td>5</td>
<td>School</td>
<td>Home</td>
<td>18</td>
<td>Good</td>
<td>0</td>
</tr>
</tbody>
</table>
i. Participant information - the Corpus linguistic data

Data was collected in Zimbabwe between March and December of 2019 using semi-structured interviews. Six private semi-scripted dialogues were recorded for up to 20 minutes. In addition, four private conversations were also recorded. From these, a corpus totalling 9000 words was compiled. Only two genres were collected in the pilot study because the aim was to test the open ended semi-structured interview schedule. The genres are shown in the table below.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Number of texts</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private semi-scripted dialogues</td>
<td>6</td>
<td>5 500</td>
</tr>
<tr>
<td>Private conversations</td>
<td>4</td>
<td>3 500</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>9 000</td>
</tr>
</tbody>
</table>

4.4.2.2 Step 2: Main study

In this section, the setting, research instruments, corpus and extralinguistic data are described. In addition, a summary of the participants is given.

a. Setting

For this study, data was gathered in different parts of Zimbabwe. Private semi-scripted dialogues and private dialogues were recorded in different social situations, which depended on the snowball, purposive and convenience sampling. The setting was in participants’ private homes or any preferred location where participants chose, or the residence of the researcher. Participants were drawn from both urban areas and rural areas. The socio-cultural and linguistic backgrounds of participants varied. One is likely to encounter people who share English communication in urban areas. The urban population in Zimbabwe consists of people from different linguistic backgrounds (Thondhlana, 2002). Online newspaper articles (editorials and newspaper reportage) were also used to compile a corpus. The researcher was granted permission to use the online resources from two newspaper organisations namely The Herald and The Sunday Mail. Private business letters
were more difficult to obtain. Some of the people that were approached refused to grant me access to their letters citing confidentiality issues. Others were sceptical about sharing their private business letters. That is why five samples were collected instead of the ten initially planned. In the end, a variety of registers such as editorials, newspaper reportage, private dialogues, private scripted monologues and private business letters were collected. The research instruments used in this study are described next.

b. Research instruments

Based on the feedback received from the pilot study, adjustments and alterations were made to data collection instruments. The research instruments used in the main study include a biographical and language background questionnaire, which also focussed on proficiency, semi structured interview schedule and informed consent forms for written and spoken texts. From the pilot study, the researcher realised that some of the questions were ambiguous and some participants had difficulties answering them. In this case, the researcher provided simple and clear questions. During the pilot study, the researcher came to the realisation that some crucial questions had been omitted and added the questions in the final version of the questionnaire. In the case of semi-structured interviews, questions that were not easy for participants to respond to were removed. In addition, new questions were added to the semi-structured interview schedule because some participants did not manage to reach the target number of texts (2000 words) needed for this study. From the pilot study, it was apparent that some participants in private dialogues were not able to reach the 2000 words that were needed. So, the number of minutes had to be adjusted in order to get more texts from participants. Observations were also made that some participants were hesitant to answer the personal details section that was on the first page. For the main study, the researcher placed the personal questions about names, phone numbers and addresses at the end of the questionnaire so that participants would answer them last and reiterated that participants’ information will be treated with utmost confidentiality. Since the research instruments have been explained above, it is important to give an overview of how participants were selected, as outlined in the following section.
c. Participants

A tailored mixed method sampling technique was used in this study. This included three non-probability sampling methods namely, snowball, convenience and purposive sampling. “Non-probability sampling consists of a number of strategies that try to achieve a trade-off, that is, a reasonably representative sample using resources that are within the means of a researcher” (Dörnyei, 2010: 60). The selection of participants in non-probability sampling is not all-inclusive of the population. Instead, participants are chosen depending on the nature and objectives of the study (Etikan, Musa & Alkassim, 2016).

Firstly, snowball sampling was utilised. “Snowball sampling uses a small pool of initial informants to nominate other participants who meet the eligibility criteria for a study” (Morgan, 2008: 815). Through snowball sampling, initial participants refer the researcher to potential participants to be sampled. Dörnyei (2010) defines a sample as a group of people that is being studied and the population as the group of people that are targeted by the research. The researcher made use of initial contacts with family and friends, who were then asked to help recruit potential participants (those meeting the eligibility criteria were selected to participate in the study).

Secondly, using a convenience sampling method, data was gathered in the Faculty of Arts at the University of Zimbabwe. The researcher had obtained permission to carry out research from the university. “A convenience sample can be defined as a sample in which research participants are selected based on their ease of availability” (Saumure & Given, 2008). In this study, respondents who were willing to take part in the study are the ones who participated.

Thirdly, purposive sampling was used. Since convenience sampling may exclude some samples of the population, purposive sampling was also employed because it is used to select participants with particular characteristics. Through purposive sampling, participants who had at least 10 years of formal education were selected. In purposive sampling, “people [...] are chosen for a particular purpose, for instance, we might choose people who we have decided are ‘typical’ of a group” (Leedy & Ormrod, 2014: 221).
The minimum requirement for participation in this study was that all participants were supposed to be Shona first language speakers who could communicate in English. With Kennedy's (1998) idea that records should be kept about the type of data collected, the researcher gathered information about dates when data was recorded and the biographical information of participants. These were used to aid data analysis.

For private conversations, 10 pairs of participants were chosen. Forty-five participants who were at least 18 years old were chosen for private semi-scripted dialogue. The age groups were purposely sampled in terms of three groups of 15 participants each, which included the age ranges from 18 years to 29 years, from 30 years to 49 years and from 50 years and above. The population pyramid of Zimbabwe shows a decrease in the number of people aged from 65 and above. That is why the third group was made up of participants aged 50 and above. The Zimbabwe population pyramid is shown below in figure 8.

![Population pyramid of Zimbabwe](https://www.populationpyramid.net/zimbabwe/2019/)

*Figure 8. Population pyramid of Zimbabwe*
(Source: [https://www.populationpyramid.net/zimbabwe/2019/](https://www.populationpyramid.net/zimbabwe/2019/))

The life expectancy of Zimbabwe is 60 years according to the World Health Organisation ([https://www.who.int/countries/zwe/en/](https://www.who.int/countries/zwe/en/)). Hence, the age groups were created in such a way that the last group would have a starting age of 50.

The study assumed that learners who had at least 10 years of formal education would have been exposed to English and could communicate in English. After 10 years in formal
education, most people will have obtained their Ordinary Level certificate and will be either
continuing to advanced level (Form 6)\textsuperscript{28} or going for vocational training and colleges. Semi-
structured interviews were conducted in English with participants. The level of linguistic
proficiency was expected to be varied. A detailed description of the corpus linguistic data
used in this study, together with the challenges encountered, are given in the next two
sections.

i. Participant information - the Extralinguistic proficiency
data: Private semi-scripted dialogue

The biographical and language background questionnaires were used to aid in the
analysis of features. The corpus was used to check for the presence or absence of a
feature. This was then complemented by checking questionnaires results to see whether
the language backgrounds of participants contributed to the presence or absence of a
feature.

Table 6 shows the linguistic backgrounds of participants in private semi-scripted
dialogue. Shona was the mother tongue of all participants and English was used by
everyone in the group. The majority of participants indicated that they started learning
English when they started school. An interesting observation is that all the participants
considered themselves to have used English continuously from the time they learned it,
as shown by the number of years that they said they have been using English. When asked
to rate their linguistic proficiencies, 12 participants noted that they considered their
proficiencies to be average. Fifteen participants indicated that they were good whilst 17
participants believed that they were very good. A single participant reported to be not so
good. Another interesting observation is that 22 out of 45 participants were taught by an
English L1 teacher at school. The number of years taught ranged from 1 to at least 16
years. Twenty-three participants were never taught by an L1 English speaker. A link
between where English is used and the hours spent speaking English per week is evident.
School and work were the most popular places where English was used. This is because
the majority of participants (39) indicated that they were either formally employed or
were pursuing tertiary education.

\textsuperscript{28} In Zimbabwe, Secondary education starts at form 1 to form 4. High school education starts at form 5 and ends
at form 6. After passing form 6, learners can go to university.
Table 6. Summary of Linguistic backgrounds for private semi-scripted dialogue participants

<table>
<thead>
<tr>
<th>Participants</th>
<th>Mother tongue</th>
<th>Other languages spoken</th>
<th>Age when English was learned</th>
<th>Where English is/was used</th>
<th>Hours per week spend speaking English</th>
<th>When English was learned</th>
<th>Where English was learned</th>
<th>How many years has English been used?</th>
<th>English proficiency</th>
<th>Years taught by English L1 speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$A$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Work</td>
<td>25</td>
<td>School</td>
<td>School</td>
<td>27</td>
<td>Good</td>
<td>1-5</td>
</tr>
<tr>
<td>&lt;$B$&gt;</td>
<td>Sh</td>
<td>Eng, Swa</td>
<td>6</td>
<td>Home, school, friends</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>17</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>3</td>
<td>Home, school, work, friends</td>
<td>42</td>
<td>School</td>
<td>School</td>
<td>25</td>
<td>Very good</td>
<td>6-10</td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>School</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>17</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>4</td>
<td>Home, school, work, friends</td>
<td>60</td>
<td>Before school</td>
<td>Home</td>
<td>30</td>
<td>Very good</td>
<td>1-5</td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>10</td>
<td>Work</td>
<td>2</td>
<td>School</td>
<td>School</td>
<td>20</td>
<td>Not good</td>
<td>1-5</td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>School</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>20</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Friends</td>
<td>24</td>
<td>School</td>
<td>School</td>
<td>16</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>11</td>
<td>Home, friends</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>60</td>
<td>Very good</td>
<td>6-10</td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>3</td>
<td>Work</td>
<td>40</td>
<td>Before school</td>
<td>Home</td>
<td>33</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>4</td>
<td>Home, school, work, friends</td>
<td>70</td>
<td>School</td>
<td>School</td>
<td>20</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>School, work, friends</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>35</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>Sh</td>
<td>Eng, Ndau, Nd</td>
<td>7</td>
<td>School</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>11</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>8</td>
<td>School, work, friends</td>
<td>7</td>
<td>School</td>
<td>School</td>
<td>38</td>
<td>Average</td>
<td>1-5</td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Home, school, work, friends</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>37</td>
<td>Very good</td>
<td>16+</td>
</tr>
<tr>
<td>Sh</td>
<td>Eng</td>
<td>Nd</td>
<td>Tso, Por</td>
<td>7</td>
<td>Home, school, work, friends</td>
<td>8</td>
<td>School</td>
<td>School</td>
<td>44</td>
<td>Very good</td>
</tr>
<tr>
<td>$&lt;$KK$&gt;$</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>7</td>
<td>Home, school, work, friends</td>
<td>7</td>
<td>School</td>
<td>School</td>
<td>20</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>$&lt;$LL$&gt;$</td>
<td>Sh</td>
<td>Eng, Kal</td>
<td>7</td>
<td>Home, school, work, friends</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>61</td>
<td>Very good</td>
<td>11-15</td>
</tr>
<tr>
<td>$&lt;$M$&gt;$</td>
<td>Sh</td>
<td>Eng, Nd, Shan</td>
<td>7</td>
<td>Work</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>45</td>
<td>Good</td>
<td>1-5</td>
</tr>
<tr>
<td>$&lt;$NN$&gt;$</td>
<td>Sh</td>
<td>Eng</td>
<td>4</td>
<td>Home, school, work, friends</td>
<td>40</td>
<td>Before school</td>
<td>Home</td>
<td>50</td>
<td>Very good</td>
<td>6-10</td>
</tr>
<tr>
<td>$&lt;$OO$&gt;$</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>6</td>
<td>School, work</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>28</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>$&lt;$PP$&gt;$</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Home, school, work, friends</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>18</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>$&lt;$QQ$&gt;$</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Work</td>
<td>14</td>
<td>Before school</td>
<td>Home</td>
<td>28</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>$&lt;$RR$&gt;$</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>7</td>
<td>Home, school, work</td>
<td>4</td>
<td>School</td>
<td>School</td>
<td>35</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>$&lt;$SS$&gt;$</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Home, work</td>
<td>7</td>
<td>School</td>
<td>School</td>
<td>30</td>
<td>Good</td>
<td>0</td>
</tr>
</tbody>
</table>
ii. Participant information - the Extralinguistic proficiency data: Private dialogue

Twenty participants took part in the private dialogues. Their linguistic backgrounds are shown below.

*Table 7. Summary of Linguistic backgrounds for private dialogue participants*

<table>
<thead>
<tr>
<th>Participants</th>
<th>Mother tongue</th>
<th>Other languages spoken</th>
<th>Age when English was learned</th>
<th>Where English is used</th>
<th>Hours per week spent speaking English</th>
<th>When English was learned</th>
<th>Where English was learned</th>
<th>How many years has English been used?</th>
<th>How many years taught by English L1 speaker</th>
<th>English proficiency</th>
<th>Years taught by English L1 speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$A$&gt; Sh</td>
<td>Eng, Latin</td>
<td>10</td>
<td>Work</td>
<td>10</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>Good</td>
<td>6-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$B$&gt; Sh</td>
<td>Eng, Nd</td>
<td>7</td>
<td>Work, friends</td>
<td>15</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>34 Good</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$C$&gt; Sh</td>
<td>Eng</td>
<td>6</td>
<td>Home, school, work, friends</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>23 Average</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$D$&gt; Sh</td>
<td>Eng</td>
<td>6</td>
<td>Home, school, friends</td>
<td>24</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>18 Average</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$E$&gt; Sh</td>
<td>Eng</td>
<td>7</td>
<td>Friends</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>School</td>
<td>44 Very good</td>
<td>1-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$F$&gt; Sh</td>
<td>Eng</td>
<td>4</td>
<td>Home, work, friends</td>
<td>112</td>
<td>Before school</td>
<td>Home</td>
<td>39</td>
<td>Very good</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$G$&gt; Sh</td>
<td>Eng</td>
<td>5</td>
<td>Home, work, friends</td>
<td>30</td>
<td>Before school</td>
<td>Home</td>
<td>31</td>
<td>Good</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$H$&gt; Sh</td>
<td>Eng</td>
<td>6</td>
<td>School</td>
<td>5</td>
<td>School</td>
<td>School</td>
<td>17</td>
<td>Average</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$I$&gt; Sh</td>
<td>Eng, Nd</td>
<td>6</td>
<td>Work, friends, home</td>
<td>40</td>
<td>Before school</td>
<td>Home</td>
<td>32</td>
<td>Good</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$J$&gt; Sh</td>
<td>Eng, Nd</td>
<td>5</td>
<td>Home, work, friends, strangers</td>
<td>20</td>
<td>Before school</td>
<td>Home</td>
<td>45</td>
<td>Good</td>
<td>1-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$K$&gt; Sh</td>
<td>Eng, Nd</td>
<td>7</td>
<td>Home, work, friends</td>
<td>35</td>
<td>School</td>
<td>School</td>
<td>36</td>
<td>Good</td>
<td>1-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Country</td>
<td>Native Language(s)</td>
<td>Value</td>
<td>Environment</td>
<td>Time</td>
<td>Location</td>
<td>Score</td>
<td>Quality</td>
<td>Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>--------------------</td>
<td>-------</td>
<td>-------------</td>
<td>------</td>
<td>----------</td>
<td>-------</td>
<td>---------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>5</td>
<td>Work, friends</td>
<td>20</td>
<td>Before school</td>
<td>Home</td>
<td>38</td>
<td>Good</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>School</td>
<td>14</td>
<td>School</td>
<td>School</td>
<td>20</td>
<td>Very good</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Work</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>35</td>
<td>Average</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>4</td>
<td>School, family, friends</td>
<td>14</td>
<td>Before school</td>
<td>Home</td>
<td>17</td>
<td>Very good</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>4</td>
<td>School, home</td>
<td></td>
<td>Before school</td>
<td>Home</td>
<td>18</td>
<td>Average</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Home</td>
<td>20</td>
<td>School</td>
<td>School</td>
<td>11</td>
<td>Good</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>&lt;$R$&gt;</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Home, school, work, friends</td>
<td>67</td>
<td>School</td>
<td>School</td>
<td>25</td>
<td>Average</td>
<td>16+</td>
<td></td>
</tr>
<tr>
<td>&lt;$S$&gt;</td>
<td>Sh</td>
<td>Eng, Kal</td>
<td>7</td>
<td>Home, school, work, friends</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>48</td>
<td>Very good</td>
<td>1-5</td>
<td></td>
</tr>
<tr>
<td>&lt;$T$&gt;</td>
<td>Sh</td>
<td>Eng, Swa</td>
<td>7</td>
<td>School, work, strangers</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>35</td>
<td>Average</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
From table 7 above, English was the common L2 in all the respondents. Seven participants noted that they spoke a third language with 4 participants saying they spoke Ndebele and the other 3 languages being Latin, Swahili and Kalanga. Seven participants also noted that they learned English at home before they started school. These participants' learned English at an earlier age ranging from 4 to 6 years. The majority of participants (14) responded that they were never taught by a teacher who spoke English as a L1. It is interesting to observe that, when considering where English was used, work, home, school and friends seem to be popular responses.

i. Participant information - the Extralinguistic proficiency data: Private business letters

Table 8 below shows the language backgrounds of participants who provided private business letters.

Table 8. Summary of Linguistic backgrounds for private business letter participants

<table>
<thead>
<tr>
<th>Private business letter</th>
<th>Mother tongue</th>
<th>Other languages spoken</th>
<th>Age when English was learned</th>
<th>Where English is used</th>
<th>Hours per week spent speaking English</th>
<th>When English was learned</th>
<th>Where English was learned</th>
<th>How many years has English been used?</th>
<th>English proficiency</th>
<th>Years taught by English L1 speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Work, school</td>
<td>15</td>
<td>School</td>
<td>School</td>
<td>17</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>Sh</td>
<td>Eng, Nd</td>
<td>8</td>
<td>Work</td>
<td>30</td>
<td>School</td>
<td>School</td>
<td>58</td>
<td>Good</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>Sh</td>
<td>Eng</td>
<td>6</td>
<td>Work</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>58</td>
<td>Average</td>
<td>0</td>
</tr>
<tr>
<td>P4</td>
<td>Sh</td>
<td>Eng, Ndau</td>
<td>6</td>
<td>Friends</td>
<td>10</td>
<td>School</td>
<td>School</td>
<td>23</td>
<td>Very good</td>
<td>0</td>
</tr>
<tr>
<td>P5</td>
<td>Sh</td>
<td>Eng</td>
<td>7</td>
<td>Work, strangers</td>
<td>40</td>
<td>School</td>
<td>School</td>
<td>44</td>
<td>Average</td>
<td>0</td>
</tr>
</tbody>
</table>

As with private semi-scripted dialogue and private dialogues, participants in this category had Shona as their L1 and used English as a L2. All participants started learning English at school and 4 participants indicated that they used English at work with the hours spent speaking English ranging from 10 to 40. All 5 respondents were never taught by a L1 speaker of English.
d. Corpus linguistic data collection

The researcher obtained permission to use part of Marungudzi’s (2016a) corpus from North-West University. I accessed 72 texts totalling 150 000 words. Their genres include public spoken dialogues (47 texts), editorials (1 text), newspaper reportage (7 texts), popular writing (1 text), business letters (3 texts), academic writing (1 text), social letters (2 texts), public scripted spoken monologues (2 texts) and creative writing (8 texts). The unavailability of part of Marungudzi’s (2016a) corpus led me to adjust data collection. Since Marungudzi’s (2016a) corpus was lacking in terms of the number of private semi-scripted dialogue and dialogues, this study intended to fill that gap by collecting speech samples from private semi-scripted dialogue and dialogue. The ICE protocol was utilised loosely (as done by Marungudzi 2016a) for data collection and transcription so that both corpora can be combined to form one corpus for ZE.

Consent forms were handed out for participants (for detailed information about the participants for each genre (cf. tables 6, 7 and 8 in subsection c of section 4.4.2.2 above) to sign indicating their willingness to take part in the study. After signing the consent forms, participants were then handed the biographical and language background questionnaire to fill in. Participants were requested to complete the questionnaire in my presence so that I would explain to them if there was a question they did not understand. This way, maximum participation was ensured from the participants. After completing the questionnaire, some participants would then take part in the semi-structured interviews immediately. Other participants opted to be interviewed on a different day. In this case, a suitable day and time was negotiated. Participants answered specifically constructed questions to test for the features under study (cf. Section 1.6 and 2.5 for a detailed description of the features). For private semi-scripted dialogue, interviews of between 15 to 35 minutes duration between the researcher and participants were recorded on a smartphone. Recordings were done at a location chosen by participants. This was either the participants’ homes or the researcher’s residence. Other participants opted to be interviewed in public places like parks or shopping centres.
The researcher did not interrupt participants once they started talking. Where necessary, the researcher asked follow-up questions for participants to elaborate on a topic or sentence in order to elicit as much data as possible. The part where the researcher spoke was not included in the transcription in private semi-scripted dialogue because the researcher was not participating in the study. The recordings include segments where the researcher spoke whilst asking semi-structured interview questions for participants to respond to. In the case of private conversations, some participants opted to record themselves and then hand over the recordings. This is good because it limits observer’s paradox\textsuperscript{29}.

Other participant conversations were recorded by the researcher. Naturally occurring private unscripted dialogue were recorded for between 15 minutes and 35 minutes for each pair of participants. Recordings of private conversations were done on a smartphone. De Vries \textit{et al.} (2014) note that smartphones are now being used for recording because they are readily available as they get cheaper. Recording was done in private settings (as outlined above) with families of mixed backgrounds, as well as at universities and professional but private contexts (for which permission was to be obtained). For natural conversations, the researcher managed to get the 10 samples. This is because most conversations in Zimbabwe take place in Shona, which is spoken by about 75\% of the population (Mberi, 2009). Opportunities to record English conversations arose when diaspora families visited Zimbabwe. Some diaspora families cannot speak indigenous languages of Zimbabwe, especially those who left Zimbabwe when they were young or those who were born, raised and attended school in the diaspora. In these cases, people were communicating in English during private conversations in order to accommodate those who could not converse in Shona. The recordings were stored on a password protected USB flash drive.

Additionally, after getting permission from the newspaper organisations, different genres of online newspaper articles were collected from May 2019 to December 2019. These included editorials and newspaper reportage. Each compiled text for newspaper reportage was from

\textsuperscript{29} In his description of the observer’s paradox, Labov (1972: 209) notes that “the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation”.

© University of Pretoria
the same writer. Samples of private business letters were collected for corpus compilation. Some participants provided samples of the private business letters that they had emailed. Tognini-Bonelli (2010) draws attention to the value of using new sources of data such as emails. Other private business letters were typed. There were deviations from ICE guidelines because some genres were difficult to get. For example, it was difficult to obtain more private conversations. This is because the majority of conversations in Zimbabwe occur in indigenous languages (Shona in the case of this study). Marungudzi (2016a) reports on encountering similar problems when collecting private conversations. Another project that had to deviate from ICE guidelines due to difficulties in obtaining some genres is ICE East Africa (Hudson-Ettle & Schmied, 1999). Their corpus contains 35 broadcast discussions where ICE guidelines suggest 20.

### e. Challenges during data collection and how they were dealt with

The objective was to stick to ICE guidelines in corpus compilation so that the study would be comparable to other ICE projects. But due to time and financial limitations, the researcher encountered difficulties in gathering data for the private business letters and private conversations, which led to the collection of fewer samples than planned. For private business letters, some participants withdrew their consent to participate in the study when asked to provide more samples. For this reason, 5 samples of private business letters were collected instead of 10 samples as initially planned. Gathering private unscripted dialogues was challenging because most private conversations occur in indigenous languages in Zimbabwe. For private semi-scripted dialogue, some participants had to be re-interviewed in order to obtain the 2000-word length set by ICE guidelines.

Hudson-Ettle and Schmied (1999) report on encountering similar challenges when they compiled the East African component of the ICE. They had challenges getting 2000-word texts for spoken dialogue, business and social letters. One of the decisions taken by Hudson-Ettle and Schmied was to compile a 20 000-word file from each area. Baker (2014: 110) highlights that “sometimes the job of building a perfectly balanced corpus proves to be difficult and so we have to make compromises, settling for what we can get rather than what

---

30 Marungudzi (2016a) encountered similar problems when collecting private conversations.
we would like to have”. A discussion about how the texts were transcribed and annotated is provided next.

### 4.4.2.3 Step 3: Corpus linguistic data transcription and annotation

The next step was orthographic transcription of recorded conversations and interviews. The parts where the researcher spoke were not included in the transcription of semi-structured interviews. “One of the paradoxes of spoken language capture is that the way in which the capture is represented is often by means of the written language” (Adolphs & Carter, 2013: 11). Transcription was done so as to compile an electronic corpus that was used during the data analysis process. “How we transcribe doesn’t just reflect our theories of language, it also shapes them, drawing our eyes to some phenomena while leaving others in shadow” (Du Bois, 1991: 71). Since annotation was based on the ICE mark-up manual (cf. Nelson, 2002a) the transcription did not take phonetics into account. Pauses lasting longer than three seconds were transcribed as (...) and brief pauses as (.). Non-speech sounds such as laughter was transcribed as {LG} and coughing or grunting was transcribed as {CG}. Filled pauses such as *uhm, mm, ah,* were transcribed as they sounded. Intonation gestures such as pointing and nodding were not transcribed because they would not add value to the data analysis process.

In line with Reppen’s (2010) assertion that researchers need to establish file naming conventions during transcription, the transcribed texts were labelled using participants pseudonyms to preserve their anonymity (e.g. <$A> to represent participant 1, <$B> to represent participant 2 etc.). The ICE conventions of corpus design and annotation were also followed partially. The corpus was annotated at textual mark-up level using the ICE mark-up manual for spoken texts (Nelson, 2002a) and written texts (Nelson, 2002b). Corpus annotation provided interpretive linguistic information (cf. McEnery, Xiao & Tono, 2006). Written texts were marked for headings, sentence and paragraph boundaries. Orthographically transcribed spoken texts were marked for pauses, overlapping, speaker turns, false starts and hesitations. The compiled corpus was uploaded to Sketch Engine

31 The list of transcription conventions used in this study is provided in appendix I.

32 Sketch Engine is available on [https://www.sketchengine.eu/](https://www.sketchengine.eu/). It analyses large samples of texts online.
where part-of-speech-tagging was done automatically when the researcher uploaded the corpus on Sketch Engine, a commercially available online text analysis tool that was used for corpus analysis. Sketch Engine used the English Penn Treebank part-of-speech tagset which contains modifications made by Sketch Engine. Tagging should be as accurate as possible for a corpus to be used more usefully (Chabata, 2000; Adolphs & Knight, 2010).

Annotation of the ZE corpus was done in order to enhance it. Although there is debate about the value of annotation (e.g. Sinclair 2004), Anthony (2013) notes that modern corpus analysis tools can manipulate corpora so as to show or hide annotations as one wishes. Therefore, even if a corpus is annotated, a researcher can hide all the annotations using modern software.

McEnery, Xiao and Tono (2006) draw attention to the fact that inconsistencies and errors that can occur in automatic annotation of a corpus. They suggest that both automatic annotation and the human analysts should complement each other in order to reduce the inconsistencies and errors. In this study, automatic annotation was done in Sketch Engine. According to McEnery, Xiao and Tono (2006), the accuracy rate for English part of speech tagging is usually 97%. Even though this is the case, a random sample was created in Sketch Engine to check whether for accuracy of annotation on the tags that are relevant to this study. In this study, the tags and concordance lines were checked for accuracy and edited by using the Sketch Engine manual annotation (Skema) tool. Skema can be used to assign new tags in the concordance lines and automatically counts and displays the number of concordance lines that belong to similar groups or categories when new tags are added to the corpus.

4.4.2.4 Step 4: Corpus linguistic data: The Zimbabwean English corpus

The first corpus employed in this study is the ZE corpus. In total, the ZE corpus consists of about 356,007 words. Part of Marungudzi’s (2016a) corpus (about 150,000 words) was also added. Time and financial constraints were the determining factors for the size of the corpus compiled for this study. Marungudzi (2016a) observed that, in general, English is experienced mostly as a written component than as a spoken component in Zimbabwe. Examples include formal writing at schools, colleges, universities and the print media (books, magazines, newspapers). The scarcity of naturally occurring
conversations is the reason why few private conversation samples were collected. Hudson-Ettle and Schmied (1999) encountered similar problems during the compilation of the East African part of the ICE corpus. In order to solve the problem of lack of sufficient English samples, Schmied (1990) suggest that it is possible to deviate from the ICE corpus guidelines and to arrange texts using a multidimensional approach\(^{33}\). This study adopted a multidimensional approach in categorising texts. For easy handling and storage, texts were grouped into spoken and written texts. Spoken texts were named according to their genres. To give an example, private semi-scripted dialogue 3 refers to the third spoken unscripted text. Similarly, for written texts, the naming convention; editorials 1 refers to the first sample of written editorials. The summary of texts is shown in table 9 below.

**Table 9. ZE corpus summary**

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>Number of texts</th>
<th>Number of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>Private semi-scripted dialogue</td>
<td>45</td>
<td>106 410</td>
</tr>
<tr>
<td></td>
<td>Private dialogue</td>
<td>10</td>
<td>25 551</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>2</td>
<td>4 193</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>47</td>
<td>98 760</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>8</td>
<td>17 890</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>1</td>
<td>2 759</td>
</tr>
<tr>
<td></td>
<td>Academic writing: Examination</td>
<td>1</td>
<td>2077</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>11</td>
<td>26 342</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>15</td>
<td>48 939</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>2</td>
<td>5 008</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>3</td>
<td>6 608</td>
</tr>
<tr>
<td></td>
<td>Private business letters</td>
<td>5</td>
<td>11 470</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>150</strong></td>
<td><strong>356 007</strong></td>
</tr>
</tbody>
</table>

From the table above a summary of the number of texts and percentages for spoken texts (69%) and written texts (31%) is shown in figure 9 below.

\(^{33}\) Schmied (1990: 264) explains that “as there is no satisfactory comprehensive theory of socio-stylistic variation for speech and writing available and as it has been shown that text type categories can be broken down and defined in several parameters, a multidimensional approach has to be adopted”. 

© University of Pretoria
Copyright issues limited the collection of more written texts. This is because obtaining permission from copyright holders was difficult considering the limited time and financial resources available for this study. For example, the researcher attempted to get more business letters from a public institution but was referred to different people for authorisation. In the end, the researcher could not proceed because of time constraints. The same applied for popular writing and academic writing texts. Copyright holders were sceptical of the content of their magazines being used in the corpus because they felt that it would hurt their sales. In the end, only one text in the popular writing genre and one text in the academic writing genre that were compiled by Marungudzi (2016a) were included in the corpus. In order to mitigate the effects of less written texts, the results obtained were normalised in order to compare results.

---

34 Cf. McEnery, Xiao and Tono (2006) for a detailed explanation of how copyright laws may hinder corpus compilation for written texts.
4.4.2.5 Step 5: Corpus linguistic data: The ICE-GB

The second corpus employed in this study is the ICE-GB. The ICE-GB is the British part of the ICE project and has one million words, consisting of 500 samples, with each sample containing about 2,000 words (www.ucl.ac.uk/english-usage/projects/ice-gb/). Release 2 of ICE-GB was used for this study. It is an enhanced, tagged and parsed corpus containing additional material and improved transcription. Since all samples in the ICE-GB came from BrE speakers, they were suitable to be used for comparisons with ZE in order to answer research questions about whether there is variability in DMs.

In this study, only samples from the ICE-GB, which occurred in both the ZE corpus and the ICE-GB were used for comparison purposes. These included unscripted monologues, private dialogue (direct conversations), public scripted monologue (broadcast news), public dialogue, creative writing, academic writing: examinations, editorials, newspaper reportage, social letters, business letters (both private and public) and popular writing. The registers, genres, types of texts, number of samples, as well as number of words are illustrated in the table below.

<table>
<thead>
<tr>
<th>Register</th>
<th>Text type</th>
<th>Number of samples in the ICE-GB</th>
<th>Words in the ICE-GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written</td>
<td>Private dialogue (direct conversations)</td>
<td>90</td>
<td>185 208</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue (broadcast news)</td>
<td>20</td>
<td>43 061</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>80</td>
<td>171 062</td>
</tr>
<tr>
<td></td>
<td>Creative writing</td>
<td>20</td>
<td>42 646</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>10</td>
<td>21 199</td>
</tr>
<tr>
<td></td>
<td>Academic writing: Examination</td>
<td>10</td>
<td>21 225</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>10</td>
<td>20 719</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>20</td>
<td>41 539</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>15</td>
<td>31 085</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>15</td>
<td>30 491</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>290</td>
<td>608 235</td>
</tr>
</tbody>
</table>
4.4.2.6 Step 6: Data analysis

Data was examined with the aim to uncover whether there is innovation in the use of articles, modals and selected DMs. The corpus linguistic data (from both the ZE corpus and the ICE-GB) were analysed in this study, while the extralinguistic proficiency data (biographical and language background questionnaires) were used to aid in the interpretation of the corpus linguistic data.

In this section, the tools that were used in the analysis of the corpora are described. Interpretation of corpus data was done from a functionalist usage-based perspective (cf. Botha, 2012, Partridge, 2019). Standard reference grammars of English (e.g. Quirk, *et al.*, 1985; Huddleston & Pullum, 2002) were used as guides for rules of English grammar. In addition, an overview of the statistical tools that were utilised to summarise and interpret the data are given. The empirical data obtained from corpora is given as an example of how language is used as the functions of the features in the corpus.

For comparison purposes, the following text categories shown in table 11 were used because they occurred in both corpora.

*Table 11. Word count for text categories in ZE corpus and ICE-GB*

<table>
<thead>
<tr>
<th>Register</th>
<th>Text type</th>
<th>Number of samples in ZE corpus</th>
<th>Number of samples in ICE-GB</th>
<th>Words in ZE corpus</th>
<th>Words in ICE-GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>Private dialogue (includes direct conversations from ICE-GB, private semi-scripted dialogue and private dialogues)</td>
<td>55</td>
<td>90</td>
<td>131 961</td>
<td>185 208</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>2</td>
<td>20</td>
<td>4 193</td>
<td>43 061</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>47</td>
<td>80</td>
<td>98 760</td>
<td>171 062</td>
</tr>
</tbody>
</table>
### Table of Written Corpora Data

<table>
<thead>
<tr>
<th>Written</th>
<th>Creative writing</th>
<th>Popular writing</th>
<th>Academic writing: Examination</th>
<th>Editorials</th>
<th>Newspaper reportage</th>
<th>Social letters</th>
<th>Business letters</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>15</td>
<td>2</td>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>290</td>
</tr>
<tr>
<td></td>
<td>17 890</td>
<td>2 759</td>
<td>2 077</td>
<td>26 342</td>
<td>48 939</td>
<td>5 008</td>
<td>18 078</td>
<td>356 007</td>
</tr>
<tr>
<td></td>
<td>42 646</td>
<td>21 199</td>
<td>21 225</td>
<td>20 719</td>
<td>41 539</td>
<td>31 085</td>
<td>30 491</td>
<td>608 235</td>
</tr>
</tbody>
</table>

#### a. Corpus analysis tools used

Two corpora, namely the ZE corpus and the ICE-GB were used as resources for data analysis to test whether there are systematic differences in the use of articles, modal verbs and selected DMs. The ICE-GB was used as the control corpus. “Comparison with a Standard variety of English and use of a native control corpus are convenient to identify potential areas of innovation and stabilisation of features” (Botha: 2013: 16). That is why the ICE-GB was used as a control corpus in order to determine whether there are variations in the features studied. This is because the features in Kortmann, Lunkenheimer and Ehret (2020) are reported to show variations with Standard English. Therefore, the ICE-GB represents a variety of Standard English (BrE).

Since a corpus linguistics approach or method was used in this study, data was analysed both quantitatively and qualitatively, meaning that quantitative and qualitative data analysing methods complemented each other. Firstly, a quantitative analysis of the use of articles, modal verbs and selected DMs (so, well and but) in the two corpora was undertaken by determining the frequency of the features (cf. subsection b of section 4.4.2.6 for a detailed overview).

In this study, results from the quantitative analysis allowed the researcher to identify what deserves to be qualitatively discussed (cf. Gries, 2013b: 14). Results from the biographical and language background questionnaires were used to aid interpretation in...
the qualitative data analysis. The quantitative data analysis process was aided by the use of a computer and Sketch Engine Tools software (www.sketchengine.eu/). Sketch Engine was used to analyse the ZE corpus to get information on frequencies and once a wordlist has been generated, the corpus was queried by using a selection of articles, modal verbs and DMs (so, well and but). Concordances were generated in order to examine the KWIC35. A concordance is “a list of all the occurrences of a particular search item in a corpus, presented within the context that they occur in; usually a few words to the left and right of the search term” (Baker, 2006: 71).

To give examples, the concordances of the and an are shown in the screenshots below.

![Figure 10. Screenshot showing KWIC concordances of “the” in Sketch Engine](image)

---

35 "KWIC is the red text highlighted in a concordance. The red text is the result that matches the search criteria" (www.sketchengine.eu/).
From these lists of concordances, the **frequencies** of each construction and feature was measured. The section below explains how statistics were used in the data analysis.

b. **Quantitative Analysis: Calculation of occurrence and function frequencies.**

In order to quantitatively answer research question 1.1 focusing on occurrence frequencies and research question 2.1 focussing on function frequencies, the occurrence and function frequencies of all constructions have to be calculated. The applicable research questions are repeated here for ease of reference:

1.1 What are the occurrence frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?

2.1 What are the function frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?
i. Calculation for occurrence frequency counts (research question 1.1).

The frequencies for articles, modal verbs and discourse markers were determined in three different ways. The frequency determination of articles is outlined first, followed by that of modal verbs and discourse markers.

- Articles

To determine the frequency of articles, it was necessary to find the tagset for determiner, together with the specific determiner (e.g. the, a, an) in the ZE corpus. This tagset was searched in the concordance tool. This made it possible to get results of the instances where the words were used as determiners. Thereafter, the articles the, a/an were counted. Regarding the zero article, the different tagsets of nouns were searched in order to identify the nouns. This was done in order to check for articles, which are said to be examples of central determiners, which precede noun phrases (NPs) or its modifiers as suggested by Quirk et al. (1985). The concordances to the left of nouns were checked to see which ones fit in the category of having the zero article and to see if the zero article was used or whether the definite or indefinite articles, or any other determiner was used instead. Nouns were chosen to check whether a determiner was used because, according to Quirk et al. (1985: 253), a noun is part of a NP and the determiner “function is typically realised by a set of closed-class items, or determiners, which occur before the noun acting as head of the noun phrase (or before its premodifiers)”. According to Botha, 2015), determiners include a/an/no, both, demonstratives, we/us/you, quantifiers/numerals, other, possessives, some, the/such, all any and certain/exact/specific (for more details on articles cf. Chapter 5).

Regarding the quantitative functional analyses of the, a/an and the zero article, their frequencies were more than 6 000 each. Due to feasibility issues, random sampling was done. Based on Sison and Glaz’s (1995) method of determining the sample size and confidence intervals, 400 instances of the definite article and 400 instances of the indefinite article were analysed for the frequency of functions. For the zero article, 400 NPs were analysed in each corpus. Sison and Glaz (1995) determined that a sample size of at least 384 is adequate if a researcher wants to achieve a confidence level of 95 per cent and a deviation
tolerance of 5 per cent. Researchers who analyse large amounts of concordance lines usually limit the number of samples they analyse due to practicality reasons (cf. Botha, 2012; Van Rooy & Terblanche, 2006; Wahid, 2013; Marungudzi, 2016a).

- **Modal verbs**

For modal verbs the tagset for modal verbs (MD) was searched in the concordance tool. Each modal verb was then counted for frequency. To check whether there is use of double modals, the researcher searched the concordances of all the nine modal verbs in Sketch Engine. The concordance lists were manually checked to determine whether feature 121 was attested. For features 123 and 127, the different functions of modal verbs were used to determine whether the features are attested in the ZE corpus. In the Wordlist tool in Sketch Engine, the tagset MD was searched and the concordance lines of each modal was checked to determine whether present tense forms of modal verbs were used where StE has past tense forms (feature 123) or whether there was non-standard use of modals for politeness reasons (feature 127).

- **Discourse markers**

The DMs (*so, well* and *but*), were searched each in the concordance tool. Further details about how the non-DM functions were excluded is provided in Section 7.6.1.3. A comparative analysis of the frequencies of three DMs (*so, well* and *but*) was done. Analysing all the DMs is outside the scope of this study. To get a better understanding of whether there are any statistically significant differences in the frequencies of the observed frequencies in article use, modals and DMs in the ZE corpus and the ICE-GB, a log likelihood test was used (cf. subsection c Section 4.4.2.6).

The frequencies and of the articles, modal verbs and selected DMs were also investigated in different genres and registers as noted in Section 1.6. In order to obtain comparable results, the absolute frequencies were normalised per 10 000 words. This was done because the ZE corpus and the ICE-GB are not equal in size. The ZE corpus contains fewer text types and fewer texts per category compared to the ICE-GB (cf. Sections 4.4.2.4 and 4.4.2.5).
ii. Calculation for function frequency counts. (Research question 2.1)

The different functions of articles, modal verbs and selected DMs (so, well and but) were quantified as explained below.

- **Articles**

Due to the high frequencies of the definite article in both corpora, it was necessary to do random sampling (cf. subsection b in Section 4.4.2.6 for a detailed explanation of how random sampling was done in this study). Four hundred concordance lines were analysed for the different functions of the definite article. For function frequency, the projected distribution of each function was calculated by dividing the frequency of each function by 400 and then multiplying the result by the total frequency of the article. The normalised frequency and the LL values were then calculated using the projected distribution numbers.

- **Modal verbs**

For modal verbs, the LL values, absolute and normalised frequencies of the different functions of modal verbs were given. This will help determine whether there are variations in the frequencies of functions. Regarding the functions of modal verbs, Collins (2009) and Wasserman’s (2014) classification of modals were used in this study. The modal verbs were grouped into three categories. The first category is necessity and obligation, the second category is possibility, permission and ability and the third category is prediction and volition. The frequency of each function was normalised per 10 000 words.

- **Discourse markers**

For DMs, the different functions of so, well and but were counted by checking the concordance lines where the DMs occurred in Sketch Engine. A qualitative analysis in terms of outlining the meanings of the different functions as semantic interpretations helped determine the functions. The identification of these functions in both the quantitative-qualitative analysis plays a vital role in determining the statistical significance in order to identify the stage of ZE in the DyM. The frequency of each function was normalised per
10 000 words and the LL values were calculated to determine whether the observed variations were statistically significant. For both the occurrence and function frequency counts, the analysis of the data was subdivided for type, per register and per genre.

c. Quantitative Analysis: Comparison of frequencies in corpora

In order to quantitatively answer research question 1.2 for occurrence frequencies and research question 2.2 for function frequencies as well as create data, on which to base the analysis for research questions 1.3, 2.3 and four of the study, the occurrence and function frequencies of all constructions and associated frequencies in the two corpora have to be statistically compared. In order to compare the two corpora, the frequencies were normalised per 10 000 words. For overall frequencies, this was done by dividing the frequency of each construction by the number of words in a corpus and then multiplying the result by 10 000 as shown below:

\[ \frac{\text{Total number of each article in a corpus}}{\text{Total number of words in a corpus}} \times 10000 \]

For registers, the frequencies were divided by the number of words in a register as shown below:

\[ \frac{\text{Total number of each modal verb in a register}}{\text{Total number of words in a register}} \times 10000 \]

Regarding genres, the frequencies were divided by the number of words in a genre and then multiplied by 10 000 as shown below:

\[ \frac{\text{Total number of each article in a genre}}{\text{Total number of words in a genre}} \times 10000 \]

The above calculations were used in chapters 5, 6 and 7 where normalisation was done. The statistical log likelihood test was used and a description of how it has been used for the different frequency types are outlined below. The specific research questions are provided for ease of reference below:
1.2 How does the occurrence of articles, modal verbs and discourse markers in ZE compare to that of BrE?

2.2 How do the functions of articles, modal verbs and discourse markers in ZE compare to those of BrE?

i. **Log likelihood test for occurrence frequency counts**

To check whether the results obtained were not due to coincidence, statistical tests were done. There are a number of statistical tests available for researchers. One of the tests is the chi-square test (Conover, 1999). Due to the fact that some of the values in the contingency tables were smaller than 5, the chi-square test was not used in this study. A log likelihood (LL) test was used in this study as is explained below.

In order to compare the occurrence frequencies of articles, modal verbs and DMs (so, well and but) to determine whether the observed results were statistically significant, an online log likelihood calculator created by Paul Rayson was used (cf. Rayson & Garside, 2000; Rayson et al., 2004). To interpret statistical significance, the critical values that are specified on the website [http://ucrel.lancs.ac.uk/llwizard.html](http://ucrel.lancs.ac.uk/llwizard.html) were used. The critical values are shown below.

- (a) 95th percentile; 5% level; \(p < 0.05\); critical value = 3.84
- (b) 99th percentile; 1% level; \(p < 0.01\); critical value = 6.63
- (c) 99.9th percentile; 0.1% level; \(p < 0.001\); critical value = 10.83
- (d) 99.99th percentile; 0.01% level; \(p < 0.0001\); critical value = 15.13

To determine whether the observed results are statistically significant, four critical values were used to report on the different levels of statistical significance. A critical value of 3.84 was regarded as significant where \(p < 0.05\). A critical value of 6.63 or higher was considered to be significant at the level of \(p < 0.01\), whilst a critical value of 10.83 or higher was regarded as statistically significant at the level of \(p < 0.001\). At the level of \(p < 0.0001\), a critical value of 15.13 or higher was considered to be statistically significant.

To summarise, one asterisk (*) is used where \(p < 0.05\), meaning that log likelihood > 3.84. Two asterisks (**) are used for \(p < 0.01\), meaning that log likelihood > 6.63, whilst three asterisks (***) are used for \(p < 0.001\), meaning that log likelihood > 10.83. Four asterisks
(***) are used for \( p < 0.0001 \), meaning that log likelihood > 15.13. This was done in all the chapters where the LL values were calculated. Where a comparison was found to be statistically significant it meant that there were variations which occurred not due to chance. It also helps to outline that if one naturally picks up a difference that this difference is large enough to show variation. This calculation thus subsequently played a vital role in answering research questions 1.3, 2.3, 3 and 4.

ii. **Log likelihood tests for function frequencies**

The log likelihood test was used to test whether there are significant differences in the use of articles, modal verbs and DMs (so, well and but) in ZE and BrE. Comparisons of the two varieties of English were based on frequencies of usage and on how identifiable the marked NPs were. “It must be emphasised that this markedness is not a discrete but continuous quality, recognising that even Inner Circle varieties may differ from each other on article usage” (Wahid, 2013: 30). The functions of articles were compared between ZE and BrE. Normalisation was done per 10 000 words as was done with the frequency counts. For both the occurrence and function frequency counts, the analysis of the data was subdivided for type, per register and per genre.

d. **Qualitative analysis of functions**

In order to qualitatively answer research question 1.1, 1.2, 2.1 and 2.2 and create data on which to base the analysis for research questions 1.3, 2.3 and four of the study, functions of the different constructions had to be described, explained and evaluated.

1.1 What are the occurrence frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?

1.2 How does the occurrence of articles, modal verbs and discourse markers in ZE compare to that of BrE?

1.3 Are the variations between the corpora statistically significant?

2.1 What are the function frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB?

2.2 How do the functions of articles, modal verbs and discourse markers in ZE compare to those of BrE?
2.3 Are the variations between the corpora statistically significant?

4. What is the stage of adult use of these features in ZE when considering the dynamic model of postcolonial English?

Investigating the functions of the definite article is crucial in the determination of whether there is variation in its use (Wahid, 2013). This statement is also true for the indefinite article, modal verbs and DMs. Each article was classified according to the different functions that are discussed in Section 5.3.

Regarding the qualitative analysis, **semantic interpretations** assisted in determining the functions of articles, modal verbs and DMs (cf. Wasserman, 2014). Examining the concordances of keywords is one of the criteria used to determine the functions of the definite and indefinite articles, modal verbs and DMs. Using the Skema tool in Sketch Engine, **concordance lines** of articles, modal verbs and DMs (so, well and but) were searched and annotated for their functions. Skema automatically counts and displays the number of concordance lines that belong to similar groups or categories.

- **Articles**

For articles the qualitative analysis was based on **semantic interpretations** of the concordance lines where the articles appeared and the different functions of articles provided in literature. For the zero article, the concordances and collocations of the noun head were examined to determine the different functions.

- **Modal verbs**

A functional analysis was done by checking the concordance lines where modal verbs occurred in the concordance tool in Sketch Engine. For modal verbs, bearing in mind that their meanings can overlap, classification was based on Collins (2009) and Wasserman's (2014) **semantic clusters**. **Deontic, epistemic and dynamic meanings** were used (cf. Section 6.6.1.4). Considering that researchers have reported on the difficulty of classifying modal verbs according to their meanings as either deontic, epistemic or dynamic (cf. Collins, 2009; Wasserman, 2014), the indeterminate group included ambiguous instances.
For DMs, a functional analysis was done by checking the concordance lines where DMs occurred in the concordance tool in Sketch Engine. **Semantic interpretations** of the concordance lines where the DMs occurred and the different functions of articles provided in literature were used to determine the frequency of functions of DMs (cf. Sections 7.4 and 7.6.2).

The following section outlines the identification of errors and innovation in order to answer research questions 3.1 and 3.2.

**e. Identification of errors and innovations / Attestation of innovation features according to Kortmann, Lunkenheimer and Ehret’s (2020)**

Distinguishing between error and innovation is crucial in this study because analyses of features are dependent on this distinction in order to answer research question 3.

3.1 Which features of innovation are present with regards to articles and modal verbs in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE?

According to Hundt and Mukherjee (2011), the distinction between an error and an innovation is difficult because there are no universal linguistic criteria for labelling features that deviate from L1 English norms. However, researchers have come up with different criteria, which, when combined together, can help make the distinction easier (cf. Section 2.8.7). In WE studies (such as the current study), the identification of errors cannot be done by using error-tag analysis because “what counts as an error in Standard English may not necessarily count as an error in institutionalised varieties” (Partridge, 2019: 142). Several methods were employed in this study to distinguish between an innovation and an error. Therefore, the different criteria used to distinguish between an error and an innovation are discussed below.
The first criterion is that of using frequency counts. Kortmann, Lunkenheimer and Ehret’s (2020) classification of innovation features were used as the basis for analysing whether the features studied were attested in the ZE corpus. In order to do this, it was necessary to use different Sketch Engine Tools to search for the specific features in order and to check concordance lines to determine the category of each feature. The different criteria used to identify innovations and errors for the features for articles (60, 62, 62, 63, 64, 65, 66 and 67) and modal verbs (121, 123 and 127) are given in section i and ii below. The discussion in this section will focus on the general criteria used to categorise the features into either A, B, C, or D.

Firstly, frequency counts of the selected features were used. For each feature, the corpus was queried for the frequency counts. Frequency counts are important because the frequencies and distributional patterns will reveal a feature’s structure, its meaning and function (Partridge, 2019). Another merit of frequency counts according to Biber et al. (1998) is that they help the researcher to observe patterns of language use. Frequency counts will also help in determining whether the features studied are innovations in ZE, which will be used as a basis to determine the stage of ZE in the DyM. Frequency counts were also used to measure grammatical stabilisation (Partridge, 2019). Grammatical stabilisation of a feature is evidence that it has been entrenched and is used widely and regularly in the grammars of individual speakers in a speech community (Van Rooy, 2011).

Although frequency counts are important, they are not without shortcomings. Partridge (2019: 142) draws attention to the challenge of using frequency counts to differentiate between errors and innovations by saying “frequency as a measure of conventionalisation raises the question of how often a specific linguistic form should occur in a corpus before it can be regarded as a feature”. Another limitation is that different distributional patterns may affect frequency counts as shown below.

A word \( w \) may occur 18-20 times in each of ten very different registers, or it may occur 190 times in only one of the ten registers. While these two results look the same in a frequency list of the complete corpus of ten registers, it is obvious that
these results would not be the same: they would not be the same for the corpus linguist who may be interested in register-dependent vocabulary differences and they would not be the same for the psycholinguist or language acquisition researcher who knows that learning processes in general exhibit a distributed learning or spacing effect.

(Gries, 2010: 198)

Despite their shortcomings, frequency counts enable the researcher to answer a variety of questions regarding how token frequency and type frequency interact and how autonomy slowly develops (Teubert, 2005). Due to the abovementioned limitation of frequency counts, it was necessary to take other classification criteria into consideration when categorising features to mitigate for the above-mentioned shortcomings as mentioned below.

One of the solutions to the abovementioned problem is to check whether a feature is accepted (cf. Section 2.8.7) as evidenced by its use by many members of a speech community (Minow, 2010; Gut, 2011; Van Rooy, 2011). Another way to check for acceptance is to determine whether linguistic gatekeepers such as editors accept a feature (as outlined by looking at different registers and genres) (Kruger & Van Rooy, 2017). With the aim to ascertain how stable the features are in ZE, authoritative measures and norm orientation were also gauged by checking if a feature was attested in texts that undergo editing (e.g. editorials and newspaper reportage) and formal texts. If a feature was used by linguistic gatekeepers such as writers, publishers, editors and in the formal media, it suggests that the feature is acceptable as an innovation and if a feature was not attested in edited texts and formal texts, then it was considered as an error (cf. Bamgbose, 1998; Kruger & Van Rooy, 2017). Botha (2015: 95) adds to the discussion by noting that “grammatical features attested in professional and academic writing are a more accurate indication of [innovation which leads to] stabilisation in the variety”. Another classification criterion is that for a feature to be regarded as an innovation, it had to be attested in both spoken and written texts and in different genres (cf. Bamgbose, 1998; Kruger & Van Rooy, 2017). To summarise the points made above, the following criteria was taken into consideration to identify innovations in this study:
1. Frequency counts: To show grammatical stabilisation (cf. Van Rooy 2011; Van Rooy & Kruger, 2016; Partridge, 2019). Kortmann, Lunkenheimer and Ehret’s (2020) criteria for the categorisation of features was used. If a feature occurred 1 to 15 times in the corpus and did not meet any other classification criteria mentioned below, then it was placed in category D where there is attested absence of the feature. If a feature was attested 16 to 30 times, it was classified as belonging to category C, meaning the feature exists, but is extremely rare. A feature that occurred 31 to 45 times was placed in category B, meaning the feature is neither pervasive nor extremely rare. Category A (feature is pervasive or obligatory), was assigned to a feature whose frequency count was 46 and above. However, other criteria that were also taken into consideration to determine the category of a feature are explained below.

2. For articles, calculations were done in order to account for both how many times an article was used in a “standard” manner and how many times it was used in a way that is different to the “standard”. The different calculations done are explained in Section 5.6.4. The results were given in percentages.


4. A feature should be attested in different genres. If a feature was attested in up to 3 genres only, it was classified as belonging to category D and if a feature was attested in 4 to 6 genres, then it was classified as belonging to category C. The occurrence of a feature in 7 to 8 genres meant that a feature belonged to category B. If a feature occurred in 9 to 10 genres, then it was classified as belonging to category A.

5. A feature should be attested in both spoken and written registers (cf. Van Rooy & Kruger, 2016).

If a feature did not meet any of the abovementioned criteria, then it was regarded as an error. For example, if a feature was only attested in informal speech, then it was regarded as an error (cf. Van Rooy & Kruger, 2016). If a feature met all the classification criteria but was attested with lower frequencies (e.g. once in a genre), then other criteria mentioned above were used to classify a feature. The extra linguistic biographical and language background questionnaire was used to check whether a feature was used by the same participant or if a feature was used by more participants. This would show widespread and regular use of a feature, meaning acceptability within a speech community. Due to the scope of this study, no statistical analyses were done on the age groups.

i. Identification of variations in articles

Master (2003) draws attention to the challenge of determining whether the zero article was omitted, deleted or inserted and suggests that examining NPs helps with identification of whether the zero article was omitted or replaced by the definite or indefinite article. Nouns and their collocations and concordances were searched in order to check for the presence or absence of articles.

The determination of errors or innovations in the use of the indefinite article was done by checking the StE grammar rules. One of the rules is that, “the indefinite article is only used before singular count nouns when the conditions for the definite article do not apply” (Sand, 2004: 284).

Quirk et al. (1985) list a number of rules governing the use of the definite article, indefinite article and the zero article in StE (cf. Section 5.2). To give an example, Quirk et al. say that the zero article is used with plural count nouns and noncount nouns, where the indefinite article does not occur. These rules were used as bases for checking whether there was omission, deletion or insertion of articles.

As an example, specific searches that were made for features 62, 63, 64 and 65 are discussed below.
1. Regarding the use of zero article where StE has definite article (feature 62), random sampling was done and nouns were checked for the specific contexts where the can be used or is obligatory to see if the zero article was used instead (cf. subsection b in Section 4.4.2.6)

2. For feature 63, which concerns the use of zero article where StE has indefinite article, random sampling was done and nouns were checked for specific contexts where a/an are obligatory to see whether there was omission (cf. subsection b in Section 4.4.2.6)

3. To check whether the definite article was used where StE favours zero (feature 64), the keyword the was searched in each genre. Afterwards, the KWIC was examined to see whether the was used in contexts, which would not be appropriate in StE.

4. To check if the indefinite article was used instead of the zero article (feature 65), the keywords a and an were searched in each genre. Thereafter, each keyword was examined in context (KWIC) to check if it was used in places where StE favours zero. The specific indefinite contexts where a/an can be used or are obligatory were examined to see whether the zero article was used instead. Again, a comparative statistical analysis was done.

ii. Identification of variations in modal verbs

To evaluate whether there is variation in the use of modal verbs as reported in literature (e.g. Bowerman, 2008; Mesthrie, 2008; Huber & Dako, 2008), qualitative functional analysis of the meanings of modal verbs was performed. The different meanings such as deontic, epistemic, dynamic and indeterminate were quantified. The log likelihood test was used to determine whether the observed variations were statistically significant. Further details regarding how each feature was examined are provided in Sections 6.6.4.1, 6.6.4.2 and 6.6.4.3.
iii. Identification of variations in discourse markers

A comparative analysis was done on the frequencies and functions of so, well and but in ZE and BrE. A log likelihood test was done to determine whether the observed variations were statistically significant (cf. subsection c in Section 4.4.2.6) in order to outline innovation. Due to a lack of innovation feature being present in the studied literature with regards to DMs none of the abovementioned criteria was used in the analysis as was done for article and modal verbs. A differentiation of whether there exists variation between the corpora in terms of the identification of errors and innovation can thus not occur for DMs and therefore variation will be identified according to statistical significances which occur in both occurrence- and function frequencies.

The following section provides information on how research questions 1, 2 and 3 are brought together to answer research question 4.

f. Stage of Zimbabwean English in the dynamic model

Depending on the results from the quantitative and qualitative analyses of articles, modal verbs and selected DMs, as well as the identification of errors and variations as outlined in research questions one to three and the section above, a determination was made regarding the stage at which ZE is at in the dynamic model in order to answer research question 4 (What is the stage of adult use of these features in ZE when considering the dynamic model of postcolonial English?) In addition, results from the log likelihood test were used to determine whether there are statistically significant variations in the use of articles, modal verbs and DMs (so, well and but), aiding in the determination (cf. research questions 1 and 2).

This section has detailed the research methodology that was employed in this study. The discussion focused on both theoretical and empirical research. The different stages of empirical research were outlined. The pilot study was important in this study because it enabled the researcher to adjust instruments before the main data collection process began. Extralinguistic proficiency data and corpus linguistic data were unpacked. In this section, the discussion also focused on how data was transcribed and annotated. These two processes were vital for compilation of the ZE corpus. Data analysis procedures were
highlighted. These included quantitative and qualitative analyses. The statistical test used (log likelihood test) was explained. The different criteria used to identify innovations and variations in articles, modal verbs and DMs (so, well and but) were also described.

An explanation about how ethical issues were handled is given in the next section.

4.5 Ethical considerations

Following guidelines given by Nelson (2010) and Adolphs and Carter (2013) about participants’ rights to privacy, participants were notified that participation in this study was voluntary and that the researcher undertook to guarantee confidentiality. An informed consent form was given to each participant from whom spoken language data was elicited. The form also explained the reasons for the study and asked for participants’ permission. Participants were also informed about their roles during data collection. The consent form also highlighted the use of the language background questionnaire and the semi structured interview process. All information regarding participation or withdrawal was included. The researcher explained that participants had the right to privacy and that their information will be handled confidentially. For written texts, a different consent form was given to participants who provided private business letters. As mentioned above in Section 4.4.2, the researcher requested and was granted permission from The Herald and The Sunday Mail in Zimbabwe by means of formal permission letters to access their online content. Steps were taken to ensure that no names of participants were used in this study. Instead, pseudonyms were used to identify participants, assuring anonymity in reporting and confidentiality throughout data collection and subsequent reporting. Participants were also notified that they had a right to voluntarily withdraw consent at any time during the study without incurring any consequences. If participants chose to withdraw from the study, data collected from them was destroyed. For example, when the researcher went back to participants to collect more samples of semi-structured interview in order to add to the first sample, some participants did not want to be interviewed again. So, their consent forms, questionnaires and recorded interviews were destroyed.
4.6 Assumptions

This study assumes that participants who have had at least ten years of formal education would have acquired some level of proficiency. Participants’ English proficiency levels are expected to be varied. In Zimbabwe, English is learned mainly when learners start formal education (grade 1). This is highlighted by Kadenge (2009: 157) who notes that “the mode of acquisition of English as a second language by first language speakers either through reading textbooks or from teachers who are themselves L2 English speakers has made the development of a distinctive Zimbabwean variety of English a reality”.

4.7 Conclusion

In this chapter, it has been reported that a corpus linguistics method, specifically a corpus-based approach was utilised in this study. This approach is appropriate because it enables the employment of statistical tools in the analysis of corpora. Both quantitative and qualitative methods were used to analyse results obtained from the corpus analysis. The features examined in this study were described from a usage-based perspective. The ZE corpus and the ICE-GB provided examples of “real life” language use by ZE speakers and BrE speakers.

An explanation was given about the methodology that was used in the data collection process including the steps taken during instrument development as well as providing a description of the research setting and the execution of the pilot and main study, transcription and compilation of a corpus. In this chapter, it was shown that data analysis involved quantitative statistical analyses, which aided in the identification of errors and innovations. A detailed description of how the statistical analysis was done was given. The discussion also encompassed how ethical issues were handled in the study.

The following three chapters focus on data presentation and analysis, starting with articles in chapter 5, then modal verbs in chapter 6. Data from the corpus about DMs (so, well and but) is presented and analysed in chapter 7.
Chapter 5

Articles

5.1 Introduction

This chapter provides an exposition on the nature of articles, an analysis of the data on articles and the subsequent discussion about the use of articles in ZE. The aim is to analyse the frequency and functions of articles in a corpus of ZE and BrE. In addition, an investigation is done on whether there is variation in the use of articles based on these abovementioned frequencies and functions. The identification of variation will be guided by Kortmann, Lunkenheimer and Ehret’s (2020) description of innovation features within the DyM.

Firstly, articles are discussed as they are used in StE and the different types of articles are provided in order to explain how different articles can occur in different contexts and functions and to provide the background on which article use in ZE and BrE are compared. Thereafter, a comparative analysis of the frequencies and functions of articles in ZE and the ICE-GB are examined by using the log likelihood test. In addition, possible variations on articles in L2 varieties of English are reviewed and then the ZE corpus is examined for these possible variations. A detailed description of the categorisation of features into Kortmann, Lunkenheimer and Ehret’s (2020) four categories and an explanation about the distinction of innovations and errors is given in subsection e in Section 4.4.2.6. The focus then moves to an analysis of the ZE corpus to determine how articles are used in ZE. This was done by analysing the corpus for features that are reported to show innovation in L2 varieties of English (cf. Sections 1.6 and 2.5). Results from biographical and language background questionnaires aided in the categorisation of features as explained in subsection e in Section 4.4.2.6.

The following section provides an overview about the use of articles in StE, as there is a need for a discussion about the nature of articles in StE together with an emphasis on the
different types and functions of articles, in order to use these contextual uses as the basis for the analysis of the function frequencies of articles in Section 5.6.1.4.

5.2 Types of English articles and their functions

According to Nel (2015), articles are considered to be in the category of determiners together with demonstratives, cardinal numbers, ordinal numbers, quantifiers, intensifiers, possessive pronouns and possessive nouns. This contextualisation of articles is important as the subsequent feature analysis of the use in articles in this study also include demonstratives in the noun phrase such as that man and numbers and quantifiers such as one/wan man. In order to better understand what articles are, two definitions found in literature are provided in this paragraph. On one hand, the Merriam-Webster dictionary defines article(s) as “any of a small set of words or affixes (such as a, an and the) used with nouns to limit or give definiteness to the application”. On the other hand, Minow (2010) puts English articles into three categories namely the definite article the, the indefinite article a/an and the category “no article”, which encompasses the zero article (Ø).

According to Quirk et al. (1985: 245) in a syntactically framed definition, articles are examples of central determiners, “which occur before a noun acting as head of the noun phrase (NP) (or before its premodifiers36)” as shown in the example below:

5.1 The girl.

5.2 The blonde girl in blue jeans.

(Quirk et al. 1985: 245)

In example 5.1 above, the definite article the, occurs before the noun girl, which is the head of the NP. In example 5.2, the occurs before the premodifier blonde, but still in front of the NP.

36 A premodifier is “the part of a noun group, adjective group or verb group, that comes before the most important word (the head), and adds information about it” (https://www.macmillandictionary.com/dictionary/british/premodifier).
An overarching distinction, which also occurs between definite and indefinite articles, is the difference between generic and non-generic and specific and non-specific (general). “Generic reference is used to denote the class or species generally” (Quirk, et al., 1985: 265). For example, in the sentence *a tiger can be dangerous*, there is generic reference because the sentence is not referring to a *particular tiger*. In the specific context, reference is made to a particular specimen or entity as shown in the sentence *a lion and two tigers are sleeping in the cage*.

In the description of what articles are in English and how they are used, the discussion is divided into two main parts. Firstly, a discussion on the use of three different article modification categories is provided and secondly the combination of these categories with different noun types. For reasons pertaining to the interpretation of nouns, it is important to differentiate between the different subclasses of nouns in English as the manner in which the noun can be pluralized as well as the article which can precede it, depend on its classification and type as outlined in figure 12 below. This study will focus on the inclusion of proper and common nouns and will differentiate between count and mass nouns specifically up to the third level of this figure below:

![Figure 12. The different subclasses of nouns in English](image)

© University of Pretoria
The following table illustrates the use of the abovementioned article modification categories with different noun types by creating a plural form and adding a determiner when the noun is used as the object in the English sentence *I like...*

Table 12. Use of articles with different noun types in English (Source: Nel, 2015: 56)

<table>
<thead>
<tr>
<th>Type of modification</th>
<th>Proper nouns</th>
<th>Common nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mass nouns</td>
</tr>
<tr>
<td>Proper nouns</td>
<td>Jack</td>
<td>Bread</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sand</td>
</tr>
<tr>
<td>Definite article</td>
<td>*the Jack</td>
<td>the bread the sand</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indefinite article</td>
<td>*a Jack</td>
<td>a bread</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*a sand</td>
</tr>
<tr>
<td>Indefinite quantitative</td>
<td>*some Jack</td>
<td>some bread some sand</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plural</td>
<td>*Jacks</td>
<td>*breads</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sands</td>
</tr>
</tbody>
</table>

In the table above, instances shown with an asterisk show improper use.

Expositions of each of the article types are given in the sections below. The first section outlines the definite article.

**5.2.1 Definite article**

“The definite article is used when referring to something, which can be identified uniquely in the contextual or general knowledge shared by speaker and hearer” (Quirk et al., 1985: 265). Quirk et al. (1985) outline that the sentence *the lamp on the table* shows that the speaker and hearer can identify *the lamp*. According to Master (1997), these English definite articles are used where there is clear or definite reference to the noun...
phrase. Richards and Schmidt (2010) further comment that in English, definiteness is conveyed where the reference of the NP is clear; while Hawkins (1978) asserts that placing a definite article before a referring predicate helps to focus the hearer's attention on a limited number of objects being referred to by the speaker. Therefore, the definite article is used for specific reference when contextualised within the categorisation that Quirk, et al. (1985: 265) gives. These articles can therefore have either general or specific reference. For example, in the sentence *the horse*, placing the definite article *the* before the noun *horse* refers to a particular and specific horse. Lastly, Abbott (2000) considers the definite article *the* in the following sentence to refer to a unique glass that the speaker is referring to.

5.3 *Please hand me the glass.*

(Abbott, 2000: 5)

From the sentence above, the speaker assumes that the addressee will be able to know which specific glass is referred to.

According to Abbott (2000) there are some instances, where there is no specific reference to something unique as shown in the following sentences:

5.4 *Yesterday the dog got into a fight with another dog.*

5.5 *Switch the light on.*

5.6 *The contestant gave the right answer and won a million dollars.*

In the above sentences, salience plays a major role where salience is defined as "the ease with which a linguistic item is perceived" (Richards & Schmidt, 2010: 506). In example 5.4, *the dog* referred to could be more salient in the context than any dog and could refer to the speaker's dog. The reference to *the light* in example 5.5 could be taken in the context that there is usually one light switch per room (Abbott, 2000). Lastly, example 5.6 assumes that *the right answer* could be the only answer.
Place names of plural form are also another example where the definite article is used (Quirk et al., 1985). Examples include the Bahamas and the Canary Islands. The definite article is also used to refer to public institutions (e.g. the Grand (Hotel), ships and planes (e.g. the Victory) and newspapers and periodicals (e.g. The Economist, The New York Times). In the example below, the is used to refer to the ocean.

5.7 The Atlantic (Ocean).

(Quirk et al. 1985: 288).

The following subsection outlines the nature of the second article type, namely the indefinite article.

5.2.2 Indefinite article

According to Quirk et al. (1985: 272), the indefinite article “is typically used when the referent has not been mentioned before and is assumed to be unfamiliar to the speaker or hearer”. The authors give the following example.

5.8 An intruder has stolen a vase. The intruder stole the vase from a locked case. The case was smashed open.

(Quirk et al. 1985: 272)

In the example above, an and a are used to introduce the referents intruder and vase. According to Minow, (2010) and Gabrielatos et al. (2010) there are two variants of the indefinite article namely a and an, with the difference between the two being that an is used before words that begin with a vowel (e.g. an apple). An exception to this rule is when an is used with some words that are spelled with the initial h (Palmer, Huddleston & Pullum, 2002). Examples include an heir, an honour and an hour. The indefinite article a is used before words starting with a consonant (e.g. a table is a useful article of furniture). According to Palmer, Huddleston and Pullum (2002: 1618), an exception to this rule is when a is used in front of words “pronounced with an initial consonant”. The authors give a unit and a eunuch as examples. It is also important to note that “the indefinite article a
is the most basic indicator of indefiniteness for singular\(^{37}\) count nouns” (Payne & Huddleston, 2002: 371). An exposition of different noun types in English is provided in Section 5.2.

Hewson (1972) refers to the indefinite article as an approach to the singular and particular\(^{38}\) as shown in the sentence; *A table is a useful article of furniture*. In the sentence, the first indefinite article *a* is used to refer to a single table but not a particular one. In StE, the indefinite article *a* occurs before consonant sounds (e.g. *a university*) whilst the indefinite article *an* occurs before vowel sounds, for example, *an apple* (Gabrielatos *et al.* 2010). Nel (2015) notes that as indefinite articles, *a* and *an* are not two types but instead belong to the same category. Their differences in use emanates from whether the first sound of the next word is a vowel or not. The article *an* is used in the former whilst *a* is used in the later.

In singular count nouns, the indefinite article *a/an* is used where the definite article can’t be used. “*A/an* is typically used when the referent has not been mentioned before and is assumed to be unfamiliar to the speaker or hearer . . . as shown in *an intruder has stolen a vase*” (Quirk *et al.* 1985: 272). In this example, both the subject position and the object position nouns are not known and therefore both uses occur as indefinite articles. Here again the differentiation between the use of *a* and *an* is also illustrated.

In their discussion of indefinite articles, Hawkins (1978) and Richards and Schmidt (2010) observe that they are used when referring to something general. An example given by Richards and Schmidt (2010) is that indefinite articles are used when referring to one example of a group or class (e.g. *Pass me a pencil please*). Another example is when reference is made to one example of a group or class (e.g. *A dog is a friendly animal*).

---

\(^{37}\) Singular refers to “a word form denoting one person, thing, or instance” (https://www.merriam-webster.com/dictionary/singular).

\(^{38}\) Particular refers to "a single person or thing" (https://www.merriamwebster.com/dictionary/particular).
The following subsection outlines the nature of the second article type, namely the zero article.

### 5.2.3 The zero article

In some instances, nouns are used without articles. This is referred to in literature as the zero article (cf. Master, 1997; Richards & Schmidt, 2010). Master (1992) reports that the zero article is the most frequent of the three types of articles in English and that classified plural count and singular noncount nouns are usually marked by the zero article. In example 5.9, the zero article is used before the noncount noun (*milk*).

#### 5.9 Ø Milk is good for you.

(Quirk *et al.* (1985: 274)

Master (1992: 1) also mentions that “the zero article is problematic because of the very practical difficulty of ascertaining when the zero article is correctly used and when the article or determiner slot has simply been left empty”. Other researchers (e.g. Botha, 2012; Master, 1992), have also reported on the same subject of the problematic nature of identifying where the zero article is used. Although it is difficult to identify the zero article, it is helpful to consider that

> When the zero article is selected, an amorphous\(^{39}\) environment embracing "massness", "generalness", "abstractness", "adjectiveness", "naminess", "vagueness" and "familiarity" can be said to have been activated [, T]he relevant environments [are] activated either by features on the noun, by sub categorisations of the verb or noun phrase, or by the context in which the NP occurs.

Master (1992: 16)

In order to clarify the aforementioned environments, Master (1992: 16) provides the following examples to show the linguistic environments in which the zero article occurs (given in brackets). These environments focus on the semantic interpretation in terms of

---

\(^{39}\) The *Merriam-Webster* dictionary defines amorphous as "having no definite form" (https://www.merriam-webster.com/dictionary/amorphous).
lexical features of the noun and the associated categories of classification and identification of binary features.

5.10  *There is Ø blood on the floor.* [-count]

5.11  *Mice like Ø cheese.* [+general]

5.12  *Ø Prison dehumanises people.* [+abstract]

5.13  *He was Ø man enough to accept his fate.* [+adjective]

5.14  *Get out of here, Ø dog.* [+name]

5.15  *Snoopy was selected Ø dog of the week.* [+vague]

5.16  *Hi, Ø dog.* [+familiar]

(Examples 5.10-5.16 taken from Master, 1992: 16)

In example 5.10, the zero article is used with a noncount noun (*blood*). In example 5.11, the sentence refers to cheese in general and not a specific *cheese*. Example 5.12 refers to the abstract sense of *prison* as an institution and not as a building where prisoners are kept. In example 5.13, the zero article is used where *man* is used as an adjective to show that he behaved like a man. The zero article is used before *dog* to show that a name was used for the dog in example 5.14. In example 5.15 the zero article is used where there is unclear or fuzzy meaning of *dog*. In example 5.16, the speaker is referring to a close friend.

In the category of the zero article, Master (1997) further distinguishes between the zero article (Ø1) and the null article (Ø2). Master (1997: 223) refers to the null article (Ø2) as the more definite when compared to the zero article and adds that the broad function of the null article is to label a one-member set, meaning that it appears “with singular count nouns in alternation with definite *the* and with singular proper nouns”. An example where the null article occurs before the noun *chairman* is given below:
5.17 Mr Jones was appointed (Ø)2 chairman.

In this study, the zero article will be treated as one entity (Ø) without subdivisions for the purpose of data analysis, since the features studied do not distinguish between the zero article and the null article (cf. Kortmann, Lunkenheimer & Ehret, 2020)

5.2.4 The contextual use of the three article types

To demonstrate the different environments where articles occur, Quirk et al. (1985) illustrate (in table 13 below) how articles are used in the context of count and noncount nouns, which are the two classes of common nouns. According to Quirk et al. (1985), count nouns are used to indicate individual quantifiable or countable units. Examples include book, bottle and chair. The authors define noncount nouns as nouns that indicate items that cannot be counted. Examples include beauty, light and sound.

Table 13. Uses of the articles with count and noncount nouns (Source: Quirk et al. 1985: 253)

<table>
<thead>
<tr>
<th>Noun type</th>
<th>Count</th>
<th>Noncount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>definite&lt;br&gt;The book&lt;br&gt;A book</td>
<td>The furniture&lt;br&gt;(Ø) furniture</td>
</tr>
<tr>
<td>Plural</td>
<td>definite&lt;br&gt;The books&lt;br&gt;(Ø) books</td>
<td>n/a</td>
</tr>
</tbody>
</table>

In the table above, Quirk et al. (1985) draw attention to the fact that it is possible for the definite article to occur with all three noun classes, with the singular count being the book, the plural count being the books and the singular noncount being the furniture.

Four binary distinctions in the article system are noted by Master (1992: 2) as shown in the examples below. The distinctions are given in brackets to show the feature that was used in the sentence provided.
(a) **Count [+count] a versus noncount [-count] Ø**

5.18  *I need a book.* [+count]

5.19  *I need Ø air.* [-count]

(Master, 1992: 2)

(b) **Singular [+singular] a/the versus plural [-singular] Ø/the**

5.20  *Give me a book. Give me the book.* [+singular]

5.21  *I need Ø books. I need the books.* [-singular]

(Master, 1992: 2)

(c) **Definite [+definite] the vs. indefinite [-definite] a/Ø**

5.22  *Give me the book.* [+definite]

5.23  *I need a book. I need Ø books.* [-definite]

(Master, 1992: 2)

(d) **Specific [+specific] a/Ø/the vs. non-specific [- specific] a/Ø/the**

5.24  *A tick entered my ear.* [+specific]

5.25  *Ø Ticks entered my ear.* [+specific]

5.26  *Ø Ticks carry diseases.* [-specific]

5.27  *The computer is changing our lives.* [-specific]

(Master, 1992: 2)

In examples in (a), there is a distinction between count and noncount. In example 5.18, the indefinite article *a* is used to refer to a countable entity (*book*), whilst in example 5.19, the zero article is used to refer to a non-countable noun (*air*). The examples given in (b) make a distinction between singular and plural. In example 5.20, the indefinite article *a* and the definite article (*the*) are used to refer to a single book, whilst in example 5.21, the zero article and the definite article are used to refer to more than one book. The indefinite and definite are distinguished in examples in (c). The definite article (*the*) refers to a particular book in example 5.22, whilst *a* and Ø are used to refer to any book or books in
example 5.23. Lastly, in examples in (d), a/an, the and Ø may appear in specific or non-specific (excluding some noun forms). The environment as reported by Master (1992) above is also seen as contributing to the distinction. Quirk et al. (1985) also highlight this viewpoint by noting that if the sentence: *The roses are very beautiful*, is said in a garden, the speaker will be referring to specific roses.

In the next section the functions of the three types of articles are outlined.

5.3 Functions of English articles

There are several functions or uses of the three articles that are analysed in this study. The functions of the definite article are explored first in Section 5.3.1. In Section 5.3.2, a description of the different functions of the indefinite article is given. The functions of the zero article are explained in Section 5.3.3. Examples are also given to illustrate each function.

5.3.1 Functions of the definite article

Several functions of the definite article are given by Hawkins (1978: 86-149) and Quirk et al. (1985). The following list provides an overview of these functions, which are subsequently explained in detail in the following subsections:

1. Reference
   a. Anaphoric reference
      i. Direct anaphoric reference
      ii. Indirect anaphoric reference
   b. Cataphoric reference
   c. Sporadic reference

2. Situation use
   a. Immediate situation use
   b. Larger situation use

3. The “logical” use of the

4. The use of the with reference to body parts

5. Generic use
The first function of the definite article is that of the reference, specifically anaphoric reference, which can be subdivided into direct- and indirect reference, followed by cataphoric- and sporadic reference, respectively. The second function is the situation use, which involves immediate situation use and large situation use. The third function is the “logical” use of the and the fourth function is the use of the with reference to body parts. Lastly, there is the generic use of the. Below is a description of each of these functions.

5.3.1.1a. i Anaphoric reference: direct

According to Quirk et al. (1985: 267) “anaphoric reference is used where the uniqueness of reference of some phrase is supplied by information given earlier in the discourse”. Quirk et al. (1985), further note that, where the same noun head has previously appeared in the text, a definite NP derives direct anaphoric interpretation. The authors give the following example:

5.28 Felicity bought a TV and a video recorder, but she returned the video recorder because it was defective.

(Quirk et al., 1985: 267)

In example 5.28, the object video recorder is introduced in the first part of the sentence by using the indefinite article a. Thereafter, the definite is used to refer to the video recorder that occurred before in the sentence as the second mention refers back to the first mention in which the object was introduced.

5.3.1.1a. ii Anaphoric reference: indirect

“Indirect anaphora arises when a reference becomes part of the hearer’s knowledge indirectly, not by direct mention” (Quirk et al., 1985: 267). For example, in the sentence below, the hearer infers that the bicycle has wheels from general knowledge:

5.29 John bought a new bicycle, but found that one of the wheels were defective.

(Quirk et al., 1985: 267).
5.3.1.1b Cataphoric reference

In cataphoric reference, what comes after the head noun instead of what comes before the head noun allows for the identification of the reference uniquely, as shown below.

5.30  The President of Mexico is to visit China. (Quirk et al., 1985, 268).

There is cataphoric use in the sentence above because the postmodifier of Mexico enables the hearer to identify the reference.

5.3.1.1c Sporadic reference

“The is sometimes used in reference to an institution of human society . . . which may be observed recurrently at various places and times” (Quirk et al., 1985: 269). The authors say that sporadic reference can be used when reference is made to aspects of mass communication such as the press, the radio and the television and modern transport and communication such as the mail, the post and the bus.

The following section focuses on situational use as a function of an article.

5.3.1.2 Situational use

This section focuses on the function of situation use, in which immediate- and larger situation use is outlined respectively.

5.3.1.2a Immediate situation use

With regard to the immediate situation use, reference of the is obtained from the extralinguistic environment. In this category, the referent will be in sight of the speaker and hearer as shown below:

---

40 The Macmillan dictionary defines a postmodifier as “the part of a noun group, adjective group, or verb group that comes after the most important word (the head) and adds information about it” (https://www.macmillandictionary.com/dictionary/british/postmodifier).
5.31  **The roses are very beautiful.** (said in a garden)  

(Quirk et al., 1985: 266).

In the example above, the hearer will derive the reference of *the* by looking at the immediate situation, in this case, the garden.

In another example of immediate situation use given by Quirk et al. (1985), the hearer will derive the referent from the domestic context in the sentence: *Have you fed the cat?* The hearer will infer that there is a cat in the immediate situation. In the immediate situation use category, inference is also used where the speaker and hearer cannot see the referent. In the sentence: *Beware of the dog*, the hearer will infer that there is a dog in the immediate situation (Hawkins, 1978: 112).

### 5.3.1.2b Larger situation use

Regarding the large situation, the speaker and hearer may share the same knowledge of the “larger” situation (Quirk et al., 1985). In the sentence below, the speaker and hearer will have knowledge of **the Prime Minister**.

5.32  **The Prime Minister.**  

(Quirk et al., 1985: 266).

Quirk et al. further note that similar knowledge may be shared worldwide (e.g. **the Pope**), or by people who live in the same country. For example, in the current context of South Africa, the mention of **the President** means that the hearer will know that the speaker is referring to President Ramaphosa.

Regarding the large situation use, “the speaker will be appealing to the hearer’s knowledge of entities, which exist in the non-immediate or larger situation of utterances” (Hawkins, 1978: 115). Hawkins further argues that the large situation can be of different sizes but will have the common context of utterance as the focal point. Two types are given in the large situation use. The first type is the specific knowledge in the larger situation. In the sentence below, the speaker and hearer will have knowledge of **the Gibbet**.
5.33  *The Gibbet* no longer stands.  

(Hawkins, 1978: 119)

The second type is the general knowledge in the larger situation. Hawkins (1978: 119) explains that “this refers to a general knowledge of the existence of certain types of objects in certain types of situations” (Hawkins 1978: 119). The sentence below shows that both the speaker and hearer have general knowledge of the existence of bridesmaids at the wedding:

5.34  *Have you seen the bridesmaids?* (said at a wedding)  

(Hawkins, 1978: 118)

The following explanation focuses on the logical use of *the*.

### 5.3.1.3 The “logical” use of *the*

The uniqueness of the meanings of certain words enable for logical use of *the*. Quirk *et al.* (1985) say that ordinals such as *first* and superlative adjectives such as *best* are associated with uniqueness and can show logical use of the definite article as shown below.

5.35  *When is the first flight to Chicago tomorrow?*

In the example above, the uniqueness of the ordinal *first* is that there can’t be two first flights to Chicago.

### 5.3.1.4 The use of the with reference to body parts

The use of the definite article to refer to body parts and following a preposition is illustrated below.

5.36  *They pulled her by the hair.*  

(Quirk *et al.*, 1985: 270)
5.3.1.5 Generic use

The “is often formal or literal in tone, indicating the class as represented by its typical specimen” (Quirk et al. 1985: 282). The following examples illustrate generic use of the definite article.

5.37  *A great deal of illness originates in the mind.*  

(ICE-GB: S1A-091 #346)

In the example above, *the* is used to refer to *the mind* in general of all humans and not to a specific *mind*.

The different functions of the definite article and examples from literature have been discussed in this section. These functions will be used as the basis for the determining the function frequencies of the definite article in Section 5.6.1.4.

The following section outlines the functions of the indefinite article.

**5.3.2 Functions of the indefinite article**

The different functions of the indefinite articles as shown in the table below.

*Table 14. Functions of indefinite article (Source: Botha, 2012: 264-271)*

<table>
<thead>
<tr>
<th>Function</th>
<th>Example</th>
<th>Explanation of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantification: Denoting amount, extent and duration.</td>
<td><em>Lucas Radebe left for Leeds United almost a decade ago.</em></td>
<td><em>a</em> is used to show that the duration was only ten years as the determiner and the noun forms a specific quantification of one times ten.</td>
</tr>
<tr>
<td>Reference to a specific thing.</td>
<td><em>Themba was given a gun and was told, that he should be the one that demanded the money.</em></td>
<td>In the sentence, <em>a</em> is used to refer to a particular gun that Themba was given and not just any gun.</td>
</tr>
<tr>
<td>Ascription (attribution of a quality to something).</td>
<td><em>This is quite a dodgy issue.</em></td>
<td>The issue is described as having the quality of being dubious.</td>
</tr>
<tr>
<td>Co-text bound non-particular interpretations based on linguistic markers of</td>
<td><em>God has given everyone a brain.</em></td>
<td><em>a</em> is used in the sentence which is not a fact as the conceptualisation of religion and</td>
</tr>
</tbody>
</table>
non-factual context. These are found in clauses where what is described is not a fact.

God is a subjective belief or notion.

Generic uses: Representing a class or type.

*A minor criminal can go to jail and come back from jail as a very dangerous.*

*A is used to show that minor criminal as a representative of a specific type of criminal as a class (minor vs. major).*

Idiomatic uses in fixed expressions: (Uses in words that must be used together).

*Make a difference*

In the fixed expression, *a* is used in the idiomatic expression *make a difference.*

### 5.3.3 Functions of the zero article

Master (1997) refers to the zero article as the most indefinite between the zero article and the null article. Several functions of the zero article are reported by Quirk *et al.* (1985) as shown in the table below.

<table>
<thead>
<tr>
<th>Functions of the zero article</th>
<th>Examples</th>
<th>Explanation of example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used with plural count nouns and with noncount nouns, where the indefinite article does not occur.</td>
<td>Ø Milk is good for you.</td>
<td>The zero article is used with the noncount noun milk as liquids occur in a state or quantity that is impossible to count.</td>
</tr>
<tr>
<td>Occurs in noun phrases in a copular relation: to imply that only one person holds the particular position mentioned.</td>
<td>Maureen is Ø captain of the team.</td>
<td>The zero article is used to show the unique role that Maureen has, which is captain of the team. This relation is syntactically expressed where the subject <em>Maureen</em> links to the subject complement <em>captain</em> by means of the verb <em>is</em>. The sentence <em>Maureen is the / a captain</em> has a completely different meaning.</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference*: Some “institutions” of human life and society.</td>
<td>Go to Ø prison. Be at Ø school.</td>
<td>The zero article is used to refer to <em>prison</em> and <em>school</em> as institutions.</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Means of transport and communication.</td>
<td>Travel by Ø car. Come by Ø boat.</td>
<td>Reference is made to the means of transport (<em>car</em> and <em>boat</em>).</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Times of day and night.</td>
<td>Ø Evening approached. At Ø sunrise.</td>
<td>The zero article is used when reference is made about the times of the day (<em>evening</em> and <em>sunrise</em>).</td>
</tr>
</tbody>
</table>
### Noun phrases with sporadic reference:

<table>
<thead>
<tr>
<th>Reference</th>
<th>Example</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasons.</td>
<td>Ø Winter is coming.</td>
<td>The zero article is used when referring to the winter season.</td>
</tr>
<tr>
<td>Meals.</td>
<td>Stay for Ø breakfast.</td>
<td>When meals (breakfast and lunch) are referred to, the zero articles occurs.</td>
</tr>
<tr>
<td></td>
<td>After Ø lunch.</td>
<td></td>
</tr>
<tr>
<td>Illnesses</td>
<td>Ø Chicken pox.</td>
<td>The zero article is used when referring to diseases (chicken pox and measles)</td>
</tr>
<tr>
<td></td>
<td>Ø Measles.</td>
<td></td>
</tr>
<tr>
<td>Parallel structures:</td>
<td>Ø Arm in arm.</td>
<td>In the sentence, the zero article is used where two nouns (arm and arm) are placed together.</td>
</tr>
<tr>
<td>Used in fixed phrases involving prepositions</td>
<td>Ø Face to face.</td>
<td>The zero article occurs before the noun face and the preposition to.</td>
</tr>
</tbody>
</table>

In the table above, sporadic reference of the zero article means that the zero article occurs in different noun phrases either occasionally or singly, or in irregular or random instances. The sporadic references in each of the functions described in the table above (e.g. Illnesses, meals, seasons). With regard to the functions of the zero article, the table above has highlighted the different uses and provided example sentences. There are “different categories of zero article usage which are frozen as part of idiomatic usage” in the table above. The different functions of the noun phrases with sporadic reference are shown in the table above where examples and explanations of examples are given.

The next section discusses the different studies that have reported on variations in the use of articles in L2 varieties of English.

### 5.4 Variations on articles reported in literature

In this section, some studies that have reported variations in article use in different varieties of English, are highlighted. This is done in order to show that research has been focusing on documenting the variations in L2 varieties of English. In addition, variation in the use of articles that are given by researchers in Kortmann, Lunkenheimer and Ehret’s (2020) are shown as they will be used in the analysis.
Variations or differences in the use of articles in L2 varieties of English has been reported in the literature. Researchers (e.g. Mesthrie & Bhatt, 2008; Wahid, 2009; Minow, 2010; Akinlotan, 2017) have reported the use of articles where StE does not have an article (article insertion) and the omission of articles where StE has articles (deletion). According to Sand (2004), there are variations in the use of definite and indefinite articles in Standard varieties of English such as British and American English. The choices made by speakers on the use of different articles in L2 varieties of English are “mainly pragmatically, but also situationally and stylistically determined” (Sand, 2004: 284). This means that the context, environment and style of writing may play a role in the choice of articles in L2 varieties of English. Sand goes on to say that it is expected that L2 varieties of English may show considerable differences, due to the different environments and contexts, in which L2 English varieties are used, there may be variations. The variations include article insertion, article deletion and omitting of articles where they are obligatory in StE or L1 varieties. The differences reported here are represented by the features under analysis in this study as shown in table 16. For instance, Sand (2004) did a comparative analysis of the frequencies of the and a/an between Kenyan English, Singapore English, Irish English, Jamaican English and Indian English on one hand and BrE and New Zealand English on the other hand. The analysis revealed that text type influences the use of articles in the varieties studied.

Another comparative study was done by Wahid (2009). The author studied the use of articles in Inner Circle English varieties such as BrE, American English, Australian English, New Zealand English and Outer Circle English varieties namely Singaporean English, Indian English, Philippine English and Kenyan English. He notes that Outer Circle varieties of English tend to use fewer definite articles when compared to Inner Circle varieties. Another difference mentioned by Wahid (2009) is in the use of articles in collocations. In these collocations, more unique clusters were reported to be shared by three varieties in the Inner circle whilst varieties in the Outer Circle share no clusters. In

---

41 In this section mentioning other studies that have reported variations is important because it informs on what other researchers have reported and helps in terms of comparisons with the current study.
another study, Wahid (2013) investigated whether there is variation in the use of the definite article in private dialogue, academic writing and reportage in Inner Circle and the Outer Circle varieties of English. Variations in the use of the definite article in Inner Circle and Outer Circle varieties of English were reported by Wahid. In addition, the use of the definite article *the* was influenced by text type. Results from the study show statistically significant variations in three registers across Inner and Outer Circle varieties.

Variations in the use of articles in L2 varieties of English has also been investigated by scholars. For example, in BSAE, variations in the use of articles include omission of articles as shown in the sentence: *He was Ø good man* (Gough, 1996a), insertion, whereby an article is used where it is not obligatory in StE (e.g. *He was in a pain*) or substitution, whereby an article is used in place of the one that is obligatory in StE. (Mesthrie & Bhatt, 2008: 50). An example of article substitution is *he smashed the vase in the rage* (Mesthrie & Bhatt, 2008: 50).

In their study of the use of determiners in English by first year BSAE speakers at the University of Natal, Greenbaum and Mbali (2002) reported that students tended to omit the definite article. The authors give the following example: *According to Ø equality clause*. In this sentence, the definite article is omitted. Another observation made is article substitution, whereby the indefinite article is used instead of the definite article as in the sentence: *If they could change a University policy*. In this sentence, Greenbaum and Mbali (2002: 242) note that the definite article is substituted by the indefinite article. Article insertion is another feature reported by the authors. In this case, an article is used where StE does not have an article as shown in the sentence: *130 students of the Indian origin*.

Akinlotan (2017) investigated the extent of variation in the use of articles in Nigerian English using the Nigerian component of the ICE (ICE-Nigeria). Results from his study show variability in article use. Akinlotan (2017) reports that indigenous Nigerian languages have distinct article systems, which influence the use of articles in Nigerian English thereby superseding the influence of genre.
In addition to the aforementioned studies, Minow (2010) investigated the rate of omission of the definite article and the indefinite article. This was done to determine whether different groups of participants omit articles at the same rate. Minow reported that the definite article is used in 92.21% of the instances whilst the indefinite article is used in 87.67% of the cases. Language background was also used to determine whether it affects supplance or omission of articles. After comparing Nguni and Sotho/Tswana speakers, (Minow, 2010) reported that, regarding the definite article, Nguni speakers used it more frequently than the Sotho/Tswana group. Regarding the indefinite article, Sotho/Tswana speakers used a/an more frequently than Nguni speakers.

Several examples of variation in the use of articles have been shown in the discussion above. In the table below, the variations in the use of articles (and examples) reported in Kortmann, Lunkenheimer and Ehret (2020) are outlined in order to show the features that will be examined in Section 5.6.4.

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Description of feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Use of definite article where StE has indefinite article.</td>
<td>Recently I have the golden opportunity to visit university.  I had the toothache; ...when they’d get the cold.  I am the vegetarian.</td>
</tr>
<tr>
<td>61</td>
<td>Use of indefinite article where StE has definite article.</td>
<td>A sun was shining.  She has a flu.  Teachers go to an extent of punishing students for speaking mother tongues in school.</td>
</tr>
<tr>
<td>62</td>
<td>Use of zero article where StE has definite article.</td>
<td>Did you get Ø mileage-claim for that trip? Ø main reason for their performance... Ø deadline is next month.</td>
</tr>
<tr>
<td>63</td>
<td>Use of zero article where StE has indefinite article.</td>
<td>He was Ø good man.  We decided to rent Ø apartment.  You want Ø banana?</td>
</tr>
<tr>
<td>64</td>
<td>Use of definite article where StE favours zero [article].</td>
<td>He was appointed sales representative at the Nestlé, Ghana Ltd.  He compares Kenyan English with the Standard English.</td>
</tr>
</tbody>
</table>
Given that the features explained above are used to examine data in this study, the following section shows how each of the features are categorised in Kortmann, Lunkenheimer and Ehret (2020).

### 5.5 Attestation and pervasiveness of features

In this section, the different categories of the features studied are reported as they appear in the eWAVE. Table 17 shows attestation (whether a feature exists in a variety or not) and pervasiveness (how widespread a feature is used, as shown by the different categories).

When considering L2 varieties of English, the levels of variations in the use of articles are reported with different features being attested to exist in some varieties and absent in other varieties in the eWAVE (Kortmann, Lunkenheimer & Ehret, 2020). The levels of variations are given as A, B, C and D. Below is an explanation of the levels:

- **A** - Feature is pervasive or obligatory.
- **B** - Feature is neither pervasive nor extremely rare.
- **C** - Feature exists, but is extremely rare.
- **D** - Attested absence of the feature.

Examples of the value scores for the different L2 varieties are illustrated in the table below.
<table>
<thead>
<tr>
<th>Feature number</th>
<th>Values and examples of varieties in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>60</strong>: Use of definite article where StE has indefinite article.</td>
<td>Pervasive or obligatory (A): Hong Kong English, Pure Fiji English, BSAE, Tanzanian English, Ugandan English, Pakistani English. Neither pervasive nor extremely rare (B): Kenyan English, Jamaican English. Exists, but is extremely rare (C): Cape Flats, English, Malaysian English, Nigerian English, Sri Lankan English. Attested absence of feature (D): None.</td>
</tr>
<tr>
<td><strong>61</strong>: Use of indefinite article where StE has definite article.</td>
<td>Pure Fiji English. Tanzanian English, Ugandan English, Pakistani English. BSAE, Indian English, Kenyan English. Nigerian English, Ghanaian English, Cape Flats English, Hong Kong English.</td>
</tr>
</tbody>
</table>
From table 17 above, it can be noted that the features that show variations in the use of articles are attested differently, because some features exist in a variety whilst others do not exist, as shown by the different categories (A, B, C and D). In Kortmann, Lunkenheimer and Ehret’s (2020), different researchers grouped features into different categories as explained below.

In Hong Kong English (Wong, 2020) and Pure Fiji English (Tent, 2020), the use of definite articles where StE has indefinite article (feature 60) is attested to be pervasive or obligatory with an A value. This means that there is widespread use of a feature in a variety. Feature 60 is reported to be neither pervasive nor extremely rare in some L2 varieties of English such as BSAE (Mesthrie, 2020a), Tanzanian English (Schmied, 2020), Ugandan English (Ssemuuma, 2020) and Pakistani English (Mahboob, 2020). In category C where the feature exists, but is extremely rare, is Kenyan English (Buregeya, 2020) and Jamaican English (Sand, 2020). The use of definite article where StE has indefinite article is attested to be absent in Cape Flats English (Mesthrie, Toefy & Bowerman, 2020), Malaysian English (Pillai & Greig, 2020), Nigerian English (Taiwo, 2020), Sri Lankan English (Meyler, 2020).

Table 17 also shows that, of the L2 varieties of English studied in the eWAVE, Pure Fiji English is the variety in which the use of indefinite article where StE has definite article (feature 61) is pervasive or obligatory. Examples where feature 61 is neither pervasive nor obligatory include Tanzanian English, Ugandan English and Pakistani English. Varieties where the use of indefinite article where StE has definite article is reported to exist but is extremely rare include BSAE, Indian English and Kenyan English. Attested absence of feature 61 is reported in Nigerian English, Sri Lankan English, Ghanaian English, Cape Flats English and Hong Kong English.

To give another example from table 17, the use of zero article where StE has definite article (feature 62) in Hong Kong English, Indian English, Kenyan English and Nigerian English is said to be pervasive or obligatory (Kortmann, Lunkenheimer & Ehret, 2020). In the neither pervasive nor extremely rare category, examples include Cape Flats English, Ghanaian English (Huber, 2020), Indian South African English (Mesthrie 2020b) and Sri
Lankan English. Feature 62 exists, but is extremely rare in BSAE and Acrolectal Fiji English (Hundt, Biewer & Zipp, 2020). Tanzanian English, Chikano English (Bayley, 2020) and Ugandan English have attested absence of feature 62.

The following section discusses the results from the data analysis.

5.6 Results

This section firstly outlines the quantitative data analysis in terms of determining the occurrence frequency counts and subsequently the function frequency counts for all three article types. During this quantitative analysis the absolute and normalised frequencies of the 3 articles are further divided according to registers and genres, in order to determine whether there are statistically significant variations between different genres and registers. The ZE corpus is further compared to the ICE-GB in order to determine whether there are variations in the use of articles between the corpora. A LL test is used to check whether the observed variations are statistically significant for each of the frequency count types and article types.

Secondly this section outlines the qualitative functional analysis by evaluating the different functions of the three articles. Thereafter, the ZE corpus is analysed to determine whether the innovation features are attested or not and to categorise the features using Kortmann, Lunkenheimer and Ehret’s (2020) four levels of variation in order to situate ZE in the DyM.

Reference grammars such as those of Quirk et al. (1985) and Huddleston and Pullum (2002) were used as resources and as starting points for determining if English grammar rules were followed or not. This conceptualisation allows for the deviations to subsequently be classified as innovations or actual errors. This classification is dependent on the criteria outlined in subsection e section 4.4.2.6. Several Sketch Engine tools were utilised in order to analyse the frequency and functions of articles and to determine whether there is variation and innovation in the use of articles in ZE. The
researcher used the wordlist tool to generate a list of words and their frequency and then looked at the absolute frequency of a, an and the. Due to Sketch Engine’s ability to show a word in the context it is used with its surrounding collocates, word sketch was used to determine collocations of the articles. This enabled an analysis of the functions of articles where they occurred because using word sketch made it easier to view how a word behaves. From the summaries provided in word sketch, categories of collocations of articles were analysed to determine the functions of articles. In addition, the concordance list, which shows examples of a word as it is used in context, was used to determine the use of articles in context by looking at the KWIC. This enabled an analysis of the functions of articles where they occurred in the corpus.

Using both quantitative and qualitative data analysis methods provides insights into article use in ZE because, as Partridge (2019) points out, quantitative information gleaned from a corpus can be described qualitatively. The quantitative analysis involved counting the frequencies of articles and the qualitative analysis involved the analysis of functions of articles in the contexts in which they appeared.

A LL test was done to determine whether the observed differences in frequencies are statistically significant and results were reported as follows (cf. subsection c Section 4.4.2.6). One asterisk (*) was used for p < 0.05, meaning that the LL > 3.84. Two asterisks (**) were used for p < 0.01, meaning that log likelihood > 6.63. Three asterisks (****) are used for p < 0.001, meaning that log likelihood > 10.83. Four asterisks were used for p < 0.0001, meaning that log likelihood > 15.13. The results are discussed below.

[42] According to Sketch Engine, wordlist is a generic name for various types of lists such as lists of words, lemmas, POS tags or other attributes with their frequency (hit counts, document counts or others).
[43] Absolute frequency refers to the number of occurrences of a word.
[44] Collocations are “the occurrence of two or more words within a short space of each other in a text” (Sinclair, 1991: 170).
5.6.1 Quantitative analysis of articles

5.6.1.1 Occurrence frequencies of all three article types (a/an, the and the zero article) for both corpora with Log likelihood comparison

Section 5.6.1 answers part of research questions 1.1, 1.2, 2.1 and 2.2 by focusing on the occurrence- and function frequencies of articles and determining whether there are variations in the use of articles. The research questions are shown for ease of reference below.

1.1 What are the occurrence frequencies of articles in the ZE corpus and the ICE-GB?
1.2 How does the occurrence of articles in ZE compare to that of BrE?
2.1 What are the function frequencies of articles in the ZE corpus and the ICE-GB?
2.2 How do the functions of articles in ZE compare to those of BrE?

Firstly, section 5.6.1 will focus on the absolute and normalised frequencies of the, a/an and the zero article. Thereafter, the absolute and normalised frequencies of the three articles are given for the spoken and written registers. After that, the absolute and normalised frequencies are given per genre. It is important to look at the frequencies because they will show whether there are any statistically significant variations in the use of articles in the ZE corpus and the ICE-GB, by calculating the LL values in all the three sections. LL values will show whether the observed values are statistically significant or not, meaning that, if the results are statistically significant, this shows variations in the use of articles between the two corpora. These possible variations will be used to determine the stage of ZE in the DyM.

The concordance tool in Sketch Engine generated a set of KWIC and the frequencies of these keywords, in this case the definite, indefinite and zero articles. As noted earlier in Section 4.4.2.3, the ZE corpus was automatically annotated for part-of-speech in Sketch Engine. This enabled the researcher to also search for articles and determine their frequency. The determination of frequency counts aids in the analysis of features in order to learn how language is used (Biber, Conrad & Reppen, 1998). This is because frequency
Counts were used as part of the classification criteria for determining whether a feature is an error or an innovation (cf. subsection e in Section 4.4.2.6).

For the definite and indefinite article, the tagset determiner (DT) and either the definite article the or the indefinite article a/an was searched in the wordlist tool in Sketch Engine. For the indefinite article, concordances of a/an were searched and results showed 6 159 entries of the article a/an were recorded in the ZE corpus. The absolute frequency of the as an article in the ZE corpus is 16 493. The absolute frequency was used in order to get figures that were then used to calculate the normalised frequency. In addition, the absolute frequencies were used to calculate the LL values in order to determine if there were statistically significant variations between the two corpora. In Sketch Engine, instances where the is not used as a determiner are excluded from the frequency count of the as a determiner. These include instances where the does not occur as an article but as newspaper titles namely The Herald, The Sunday Mail and The Standard. The convention used for searching for the definite article automatically excluded these examples occurring in sources from where data on newspaper reportage and editorials were sampled. In the ICE-GB, the tagset that represents articles is (ART). That is why this tagset was searched together with the specific article. For the zero article, the nouns were searched and concordance lines were checked to determine whether a determiner was used. The different tagsets of nouns were searched in order to identify the nouns. This was done in order to check for articles, which are said to be examples of central determiners, which precede noun phrases (NPs) or its modifiers as suggested by Quirk et al. (1985). The concordances to the left of the nouns were checked to see which ones fit in the category of having the zero article and to see if the zero article was used or whether the definite or indefinite articles, or any other determiner was used instead. (A detailed description about how the zero article was searched is given in subsection i in Section 4.4.2.6)

In order to compare the two corpora, the frequencies were normalised per 10 000 words. This was done by dividing the frequency of each article by the number of words in a corpus and then multiplying the result by 10 000 as shown below.
Total number of each article in a corpus  \( X \times 1000 \)

Total number of words in a corpus

In addition, to determine whether the observed differences are statistically significant, LL values were calculated with an online calculator. The LL values of 3.84 and above show that the observed results are statistically significant (cf. subsection c Section 4.4.2.6). Statistically significant results were needed in order to show that the variations are not due to chance as statistically significant comparisons outline that there is variation, which is not due to chance and that the differences between the two corpora are large enough to show variation. Results showing absolute frequencies and normalised frequencies of the articles *a/an, the* in the ZE corpus are shown in table 18.

Due to the fact that there were 79,176 nouns in the ZE corpus and 98,435 nouns in the ICE-GB, it was necessary to do random sampling (cf. Van Rooy & Terblanche, 2006; Botha, 2012; Wahid, 2013). This is because it was not feasible to check each noun to determine whether the zero article was used or not. Four hundred nouns were analysed in each corpus (cf. subsection i in Section 4.4.2.6). There were 89 instances of the zero article in 400 nouns in the ZE corpus and 118 instances of the zero article in 400 nouns in the ICE-GB. In the table below, the absolute frequency of the zero article is based on the projected distribution of the zero article. For the two corpora, the number of instances of the zero article was divided by 400, which is the number of nouns analysed in random sampling. The result was then multiplied by 100 to get a percentage of occurrence in 400 nouns. The projected distribution\(^\text{45}\) of the zero article was then calculated by dividing the frequency of the zero article in 400 nouns by 400 and then multiplying the result by the total number of nouns in a corpus. The normalised frequency and the LL values were then calculated using the projected distribution numbers. The absolute and normalised frequencies of the zero article that show projected distribution are indicated with a hash sign (#).

---

\(^{45}\) Projected distribution “describes possible values and likelihoods that a random variable can take within a given range” ([https://www.investopedia.com/terms/p/probabilitydistribution](https://www.investopedia.com/terms/p/probabilitydistribution)). In this study the range was too large in the initial search for noun phrases which then had to be randomly sampled to create a smaller sample size.
Table 18. Absolute and normalised frequencies of articles

<table>
<thead>
<tr>
<th>Articles</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>A/an</td>
<td>6 159</td>
<td>173</td>
<td>12 168</td>
</tr>
<tr>
<td>The</td>
<td>16 493</td>
<td>463.3</td>
<td>27 968</td>
</tr>
<tr>
<td>Zero</td>
<td>17 617#</td>
<td>495#</td>
<td>29 038#</td>
</tr>
<tr>
<td>Total</td>
<td>40 269</td>
<td>1 131.3</td>
<td>69 174</td>
</tr>
</tbody>
</table>

When considering research question 1 which seeks to determine the frequency of articles, the frequency of *a/an* per 10 000 words is higher in the ICE-GB (200.1 instances per 10 000 words) compared to ZE (173 instances per 10 000 words). The difference in frequencies is statistically significant because the LL value of 87.63 is higher than 15.13, which is a critical value where \( p < 0.0001 \). There are no differences in the frequencies of *the* in the ZE corpus and the ICE-GB because the normalised frequency numbers are 463.3 per 10 000 words for the ZE corpus and 459.8 per 10 000 words for the ICE-GB. The observed differences in the use of the definite article are not statistically significant because the LL value of 0.59 is less than the critical value of 3.84. Regarding the zero article, the normalised frequency per 10 000 words is higher in the ZE corpus (495), compared to the ICE-GB (477). The LL value of 14.07 is statistically significant where \( p < 0.001 \). This means that there are variations in the frequency of the zero article in the two corpora. Overall, the two corpora do not show statistically significant differences in the total frequency of the definite article, the indefinite article and the zero article. This is because the LL value of 0.74 is lower than the critical value of 3.84, where \( p < 0.05 \). The normalised frequencies per 10 000 words for the ICE-GB and the ZE corpus are 1 136.9 and 1 131.3 respectively. In order to get a better picture of how articles are distributed in the spoken and written registers, it was necessary to determine the absolute and normalised frequencies of articles in registers. The analysis was necessary in order to

---

46 Each time the Log likelihood test was done the difference between the two corpora was calculated, however this difference can have a higher value direction towards the ICE-GB or a higher value towards the ZE corpus. If the higher value occurs in the ZE corpus this direction is indicated with a plus sign in front of the LL value. In comparison to the higher value occurring in the ICE-GB, then the direction is indicated with a minus sign in front of the LL value.
determine whether there are statistically significant variations in the frequencies of articles in the spoken and written registers.

5.6.1.2 Occurrence frequencies of all three article types across registers for both corpora with Log likelihood comparison

As has been discussed in Section 2.8.11, in this study, the term “register” refers to spoken and written registers. The term “genre” refers to different text types that are found in spoken and written registers (e.g. private dialogue, newspaper reportage, social letters, editorials, creative writing, public dialogue, etc.). In this section the frequencies of all the definite article and the indefinite article across registers will be outlined. The frequencies of articles per genre are reported in Section 5.6.1.3. The data will be presented in a table format in which the distribution of articles is outlined in terms of different registers, namely, in the spoken and written registers along with a total of both registers combined. The first subsection outlines the frequency of the definite article the across registers, followed by the frequency of the indefinite article a/an across registers respectively. The occurrence frequency of the zero article is analysed last.

The normalised frequencies for registers were calculated by dividing the total number of each article in a register by the total number of words in a register and then multiplying by 10 000, as shown below.

\[
\frac{\text{Total number of each article in a register}}{\text{Total number of words in a register}} \times 10^{4}
\]

47 Due to the fact that random sampling was used to determine the frequency of the zero article in the two corpora, the frequency of the zero article in the spoken and written registers was not compared.
a. Occurrence frequencies of *the* across registers for both corpora with Log likelihood comparison

<table>
<thead>
<tr>
<th>Register</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>9 658</td>
<td>411.1</td>
<td>15 776</td>
</tr>
<tr>
<td>Written</td>
<td>6 835</td>
<td>564.4</td>
<td>12 192</td>
</tr>
<tr>
<td>Total</td>
<td>16 493</td>
<td>463.3</td>
<td>27 968</td>
</tr>
</tbody>
</table>

Interesting observations can be made from table 19 in terms of the frequency of articles in the spoken and written texts. When considering the frequency of the definite article *the*, the definite article occurs most frequently in the spoken register, in the ZE corpus (411.1 times per 10 000 words) compared to the occurrence of the definite article in the ICE-GB (395.1 per 10 000 words). This is a statistically significant difference because the LL value of 9.49 > 6.63 where p < 0.01. The normalised frequency of the in the written register, in the ICE-GB is higher (583.6) compared to the normalised frequency for the ZE corpus (564.4 per 10 000 words). This difference is statistically significant because the LL value of 4.90 is more than 3.84, where p < 0.05, which means that the variations observed in the frequency of the definite article in the written register are not due to chance and that the differences between the two corpora are large enough to show variation. Regarding the frequency of the definite article in the spoken and written registers, the normalised frequency for the ICE-GB is slightly higher in the ICE-GB (978.7 per 10 000 words) compared to the ZE corpus (975.5 per 10 000 words). The difference for the total frequency of the definite article is not statistically significant because the LL value of 0.59 is lower than the critical value of 3.84 where p < 0.05.

The following section outlines the frequency of the indefinite article *a/an* across registers.
b. Occurrence frequencies of *a/an* across registers for both corpora with Log likelihood comparison

The table below shows the frequency of *a/an* in the spoken and written registers in both the ZE corpus and the ICE-GB.

Table 20. Frequency of *a/an* across registers

<table>
<thead>
<tr>
<th>Register</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>4 038</td>
<td>171.9</td>
<td>7 416</td>
</tr>
<tr>
<td>Written</td>
<td>2 121</td>
<td>175.2</td>
<td>4 752</td>
</tr>
<tr>
<td>Total</td>
<td>6 159</td>
<td>173</td>
<td>12 168</td>
</tr>
</tbody>
</table>

Both normalised frequencies (per 10 000 words) of *a/an* in the spoken and written registers of the ICE-GB in the table above are higher than those in the ZE corpus. In the spoken register, the normalised frequency for the ICE-GB is 185.7 per 10 000 words compared to 171.9 per 10 000 words in the ZE corpus, whilst for the written register the normalised frequency for the ICE-GB is 227.5 per 10 000 words compared to 175.2 per 10 000 words in the ZE corpus. The significance of the observed differences is evident in the LL values for both spoken (15.75) and written registers (103.34), which are above the critical value of 15.13 where \( p < 0.0001 \). This means that there are variations in the frequency of the indefinite article in the spoken and written registers. From these results, it can be concluded that the frequency of the indefinite article is more in the ICE-GB compared to the ZE corpus. The difference for the total frequency of the indefinite article is statistically significant because the LL value of 87.63 is higher than the critical value of 15.13 where \( p < 0.0001 \).

c. Occurrence frequencies of the zero article across registers for both corpora with Log likelihood comparison

Regarding the absolute and normalised frequencies of the zero article in the spoken and written registers, the projected distribution of the absolute frequency was 17 617 for the ZE corpus and 29 038 for the ICE-GB (cf. Section 5.6.1.1). Due to feasibility issues, the projected distribution of the zero article in registers was not done because the sample size of the zero article was too small. In the ZE corpus, the random sample of 400 nouns yielded 89 zero
articles and in the ICE-GB, a random sample of 400 nouns yielded 118 zero articles. Therefore, based on Sison and Glaz’s (1995) method of determining the sample size and confidence intervals, sample sizes of 89 and 118 were too small to work with in order to determine the projected distribution in registers. Sison and Glaz (1995) determined that a sample size of at least 384 is adequate if a researcher wants to achieve a confidence level of 95% and a deviation tolerance of 5%.

In the next section, the absolute and normalised frequencies of the three articles are discussed for the ZE corpus and the ICE-GB. In addition, the LL values are reported to determine whether the observed variations are not due to chance.

5.6.1.3 Occurrence frequencies of all three article types across genres for both corpora with Log likelihood comparison

In order to better understand how the articles were used in different genres in the definite article and the indefinite article\textsuperscript{48}, the normalised frequencies and the LL values were calculated as will be shown next.

\begin{enumerate}
\item \textbf{Occurrence frequencies of the across genres for both corpora with Log likelihood comparison}
\end{enumerate}

To normalize the frequencies of the definite article \textit{the} per genre, the total number of each article in a genre was divided by the number of words in a genre and then multiplied by 10 000. Results are presented in the table below.

\footnotetext[48]{Due to the fact that random sampling was used to determine the frequency of the zero article in the two corpora, the frequency of the zero article in the different genres was not compared.}
Table 21. Frequency of the across genres

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>4 734</td>
<td>358.7</td>
<td>4 842</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>221</td>
<td>527.1</td>
<td>2 847</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>4 591</td>
<td>464.9</td>
<td>8 087</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>1 040</td>
<td>581.3</td>
<td>2 180</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>149</td>
<td>536.4</td>
<td>1 429</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>149</td>
<td>717.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>1 820</td>
<td>690.9</td>
<td>1 559</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>2 980</td>
<td>608.9</td>
<td>2 761</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>159</td>
<td>317.5</td>
<td>1 141</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>651</td>
<td>360.1</td>
<td>1 549</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>16 493</td>
<td>463.3</td>
<td>27 968</td>
</tr>
</tbody>
</table>

If the distribution of *the* in the different genres is considered, observations are that, the highest difference in normalised frequency per 10 000 words is recorded in private dialogue with a LL value of 238.38. The definite article *the* is attested more in private dialogue in the ZE corpus (358.7 per 10 000 words) whilst in the ICE-GB the number is 261.4, with a LL value of 238.38, a higher value which shows that the observed variations are statistically significant where $p < 0.0001$. Creative writing is another genre, which shows more frequency of *the* in ZE (581.3 per 10 000 words) than in BrE, because *the* is attested 511.2 times.

A different pattern emerges for business letters (508), newspaper reportage, public scripted monologue and popular writing. This is due to the fact that *the* is more frequent in the ICE-GB than in the ZE corpus. All the four genres have LL values that are above the critical value of 15.13 where $p < 0.0001$, which means that there are variations in the frequency of *the* in the four genres because the ICE-GB has more instances of *the* compared to the ZE corpus. Considering that the frequency of *the* in editorials is 752.4 per 10 000 words in the ICE-GB whilst in the ZE corpus the frequency per 10 000 words is 690.9, the difference is statistically significant at $p < 0.05$, because the LL (6.10) is higher than the critical value of 3.84. No statistically significant variations were recorded in social letters, academic writing and public dialogue, which means that there are no
variations in the use of *the* in the ZE corpus and the ICE-GB. The occurrence frequencies for the indefinite article in different genres are reported next, together with LL values.

### b. Occurrence frequencies of *a/an* across genres for both corpora with Log likelihood comparison

The number of occurrences of the articles in each genre were normalised by dividing the total number of occurrences of each article in a genre by the total number of words in the genre and then multiplied by 10 000 as shown below.

\[
\text{Total number of each article in a genre} \times 10000
\]

\[
\text{Total number of words in a genre}
\]

The frequencies of the indefinite article *a/an* are reported in the table below.

*Table 22. Frequency of *a/an* across genres*

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>2 436</td>
<td>184.6</td>
<td>3 188</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>57</td>
<td>135.9</td>
<td>942</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>1 540</td>
<td>155.9</td>
<td>3 286</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>430</td>
<td>240.4</td>
<td>982</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>61</td>
<td>221.1</td>
<td>562</td>
</tr>
<tr>
<td></td>
<td>Academic writing: Examination</td>
<td>43</td>
<td>207</td>
<td>495</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>470</td>
<td>178.4</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>900</td>
<td>183.9</td>
<td>1 092</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>54</td>
<td>107.8</td>
<td>611</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>168</td>
<td>92.9</td>
<td>560</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>6 159</td>
<td>173</td>
<td>12 168</td>
</tr>
</tbody>
</table>

Occurring 262.9 times per 10 000 words, newspaper reportage has the highest frequency, which is statistically significant, in BrE compared to ZE, which has 183.9 instances. The frequency of *a/an* per 10 000 words is also high in the ICE-GB in public scripted monologue (218.8), public dialogue (192.1), editorials (217.2), social letters (196.6) and business letters (183.7). In contrast, *a/an* is more frequent in the ZE corpus in private
dialogue (184.6 instances per 10 000 words) compared to the ICE-GB, which has 172.1. The LL value for this genre is above the critical value of 6.73 where p < 0.01. The LL values in popular writing, academic writing and creative writing are below the LL < 3.84, which is a critical value for p < 0.05.

c. Occurrence frequency of the zero article across genres for both corpora with Log likelihood comparison

The absolute and normalised frequencies of the zero article in the different genres were not calculated because the sample sizes that were obtained from random sampling of 400 nouns were too small. The projected distribution of the absolute frequency was 17 617 for the ZE corpus and 29 038 for the ICE-GB (cf. Section 5.6.1). In the ZE corpus, a random sample of 400 nouns yielded 89 zero articles and in the ICE-GB, a random sample of 400 nouns yielded 118 zero articles. Therefore, based on Sison and Glaz’s (1995) method of determining the sample size and confidence intervals, sample sizes of 89 and 118 were too small to work with in order to determine the projected distribution in different genres. Sison and Glaz (1995) determined that a sample size of at least 384 is adequate if a researcher wants to achieve a confidence level of 95% and a deviation tolerance of 5%. The function frequencies of the three articles are discussed next.

5.6.1.4 Function frequency count of all three articles for both corpora with Log likelihood comparison

In this section, the function frequencies of the definite, indefinite and zero article are reported. LL values are calculated to determine whether the observed variations are not due to chance. In addition, the function frequencies were normalised per 10 000 words, based on the projected distribution from 400 functions for each article. The hash sign (#) shows that the results were based on projected distribution.

a. Function frequencies of the definite article for both corpora with Log likelihood comparison

Due to the high frequencies of the definite article in both corpora, it was necessary to do random sampling (cf. subsection b in Section 4.4.2.6 for a detailed explanation of how random sampling was done in this study). Four hundred concordance lines were
analysed for the different functions of the definite article. The absolute and normalised frequencies were based on the projected distribution (cf. Section 5.6.1.1). The results are shown in the table below.

Table 23. Frequency of functions of the definite article

<table>
<thead>
<tr>
<th>Function</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency out of 400</td>
<td>Absolute frequency (projected distribution)</td>
<td>Normalised frequency (projected distribution)</td>
</tr>
<tr>
<td>Anaphoric reference: direct</td>
<td>72</td>
<td>2 969#</td>
<td>83.4#</td>
</tr>
<tr>
<td>Anaphoric referent: indirect</td>
<td>43</td>
<td>1 773#</td>
<td>49.8#</td>
</tr>
<tr>
<td>Immediate situation use</td>
<td>10</td>
<td>412#</td>
<td>11.6#</td>
</tr>
<tr>
<td>Larger situation use</td>
<td>87</td>
<td>3 587#</td>
<td>100.8#</td>
</tr>
<tr>
<td>Cataphoric reference</td>
<td>15</td>
<td>618#</td>
<td>17.4#</td>
</tr>
<tr>
<td>Sporadic reference</td>
<td>67</td>
<td>2 763#</td>
<td>77.6#</td>
</tr>
<tr>
<td>The “logical” use of the</td>
<td>58</td>
<td>2 391#</td>
<td>67.2#</td>
</tr>
<tr>
<td>The use of the with reference to body parts</td>
<td>14</td>
<td>577#</td>
<td>16.2#</td>
</tr>
<tr>
<td>Non-referential</td>
<td>34</td>
<td>1 402#</td>
<td>39.4#</td>
</tr>
<tr>
<td>Total</td>
<td>400</td>
<td>16 493</td>
<td>463.3</td>
</tr>
</tbody>
</table>

Regarding the frequency of different functions of the definite article, the non-referential function shows the biggest difference between the two corpora. The normalised frequency (per 10 000 words) of the function in the ZE corpus (occurring 39.4 times), is double compared to the normalised frequency in the ICE-GB (occurring 19.5 times). The observed variation is statistically significant because the LL value of 315.26 is higher than 15.13, which is a critical value where p < 0.0001, meaning that the variation is not due to chance. Other functions with higher normalised frequencies per 10 000 words in the ZE corpus compared to the ICE-GB include: cataphoric reference (occurring 17.4 times compared to 12.6 times in the ICE-GB), sporadic reference (occurring 77.6 times compared to 67.8 times in the ICE-GB) and the use of the with reference to body parts...
(occurring 16.2 times compared to 11.5 times in the ICE-GB). These variations are all statistically significant at \( p < 0.0001 \) because the LL values were higher than the critical value of 15.13.

For the anaphoric reference with direct and large situation use, as functions of the definite article, occurred more in the ICE-GB (95.4 and 112.7 times per 10 000 words respectively) compared to the ZE corpus (83.4 and 100.8 times per 10 000 words respectively). Other functions of the, which occurred more (per 10 000 words) in the ICE-GB compared to the ZE corpus, include the anaphoric referent with reference to: indirect use (occurring 52.9 times compared to 49.8 times in the ZE corpus), immediate situation use (occurring 14.9 times compared to 11.6 times in the ZE corpus) and the “logical” use of the (occurring 72.4 times compared to 67.2 times in the ZE corpus). These variations are statistically significant at different levels, meaning that there are variations in the frequencies of these functions in the two corpora.

**b. Function frequency of the indefinite article for both corpora with Log likelihood comparison**

For the frequencies of functions of *a/an*, a random sample of 400 concordance lines was analysed (cf. subsection b in Section 4.4.2.6 for a detailed explanation about random sampling). The absolute and normalised frequencies were based on the projected distribution (cf. Section 5.6.1.1).
When the frequencies of functions of the indefinite article are compared in the ZE corpus and the ICE-GB, the biggest difference observed is in the generic use of the indefinite article. The function is attested more in the ZE corpus, occurring 18.6 times per 10,000 words compared to 8.5 times in the ICE-GB. The LL value for generic function is the highest at 179.11, meaning that there are statistically significant variations in the frequency of the function. The other five functions of the indefinite article occurred more (per 10,000 words) in the ICE-GB compared to the ZE corpus as follows: Quantification (occurring 32.5 times compared to 25.5 times for the ZE corpus), reference to a specific thing (occurring 25.5 times compared to 20.3 times in the ZE corpus), ascription (occurring 46 times compared to 36.3 times in the ZE corpus), idiomatic uses in fixed expressions (occurring 9.5 times compared to 6.9 times in the ZE corpus) and co-text bound non-particular interpretations based on linguistic markers of non-factual context (occurring 78 times compared to 66.3 times in the ZE corpus). Since all the LL values are...
higher than 15.13, which is a critical value where $p < 0.0001$, this shows that there are variations in the frequencies of these functions in the two corpora.

c. Function frequency of the zero article for both corpora with Log likelihood comparison

Based on the qualitative analysis of the functions of the zero article, the frequency of each function was counted in the 89 instances of the zero article that were obtained from a random sample of 400 nouns. Due to the fact that there were 79 176 nouns in the ZE corpus and 98 435 nouns in the ICE-GB, it was necessary to do random sampling. (cf. Van Rooy & Terblanche, 2006; Botha, 2012; Wahid, 2013 for random sampling). In Sketch Engine, “a random sample is used to reduce the number of concordance lines while preserving the representativeness of the sample” (https://www.sketchengine.eu/). Data analysis yielded several functions of the zero article in both the ZE corpus and the ICE-GB, as shown in the table below.

<table>
<thead>
<tr>
<th>Function</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used with plural count nouns and with noncount nouns, where the indefinite article does not occur.</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>Occurs in noun phrases in a copular relation.</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Some “institutions” of human life and society.</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Means of transport and communication.</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Times of day and night.</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Seasons.</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Meals.</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Illnesses</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Parallel structures</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Fixed phrases involving prepositions</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>118</td>
</tr>
</tbody>
</table>
Although the table above shows the different functions of the zero article in each corpus, the frequencies of different functions of the zero article were not calculated because the sample of zero articles obtained from a random sample of 400 nouns was not adequate when considering Sison and Glaz’s (1995) determination that a sample size of at least 384 is adequate if a researcher wants to achieve a confidence level of 95% and a deviation tolerance of 5%.

5.6.2 Qualitative analysis of articles

This section answers part of research question 2.3 by focusing on the semantic interpretations of the function of articles and determining whether there are variations in the use of articles. The research question is shown below for ease of reference.

2.3 How do the functions of articles in ZE compare to those of BrE?

5.6.2.1 Functions of the definite article in both corpora

The different examples of the functions of articles are reported in this section. This is done to show that semantic interpretations helped determine the function frequencies that were done in Section 5.6.1.4. Results from the corpus show different functions of the as a definite article in ZE. The functions as outlined in Section 5.3.1 are discussed below.

a. Anaphoric reference: direct

Data analysis yielded 2,969 definite articles for direct anaphoric functions that are reported as frequencies in Section 5.6.1.4 below. The use of the definite article the as anaphor is shown in the sentence below where the is an anaphor for a known referent, which is aspirin.

5.38 Yes there is aspirin, which can be taken and in this particular regard the aspirin, is not taken for pain but the aspirin, is taken to eh ah reduce tendencies for of forming clots in the, in those arteries. (ZE corpus: Public dialogue 3)
Let us consider the sentences below from the ICE-GB:

5.39  Medical and dental academic staff may charge fees; only through their School or District, the fees being collected by the institution. (ICE-GB: W2D-008 #077)

The is used as a definite article in the fees to refer to a known referent fees that was introduced in the first part of the sentence. Another function of the definite article is discussed next.

b. Anaphoric reference: indirect

Indirect anaphoric reference is shown in the sentence below.

5.40  Also, when I was at college I read a scary book but I didn’t finish it because the last pages were missing. (ZE corpus: Private semi-scripted dialogue 28)

In the sentence above, once the topic of a scary book has been introduced, the speaker can go on and talk about the last pages.

Another example of indirect anaphoric reference is shown below.

5.41  The urge to say that everything in sight was like a fairy tale was of course rampant at the nuptials of the Prince of Wales and Lady Diana Spencer in July 1981. Tom Fleming, the BBC commentator for the fixture said the bride was like a fairy tale princess. (ICE-GB: W2B-010 #291)

In the example above, the topic of a wedding” is introduced by the phrase the nuptials.

There is a presumption of the bride.

c. Cataphoric reference

The cataphoric use of the is shown below.

5.42  We enjoyed because we visited the president of South Africa’s residence to take photos and also to see the view of Pretoria. (ZE corpus: Private semi-scripted dialogue 30)

There is cataphoric use in the sentence above because the postmodifier of South Africa enables the hearer to identify the reference.
When considering the ICE-GB, cataphoric reference is shown in the sentence below.

5.43 **The president of Mexico** organised it after a long campaign by hundreds of children’s groups. (ICE-GB: S2B-022 #145)

In the example above, the postmodifier of Mexico enables the hearer to pinpoint the reference.

**d. Sporadic reference**

There is sporadic reference in the two sentences below because the is used to refer to aspects of mass communication. In the first sentence, the aspect of mass communication being referred to by the is television and in the second example the aspect is radio.

5.44 *I liked soccer a lot coz that’s what we used to to see in the television.* (ZE corpus: Private semi-scripted dialogue 41)

5.45 *The Eagle units slipped into position and over the radio came words Attack, attack, attack*. (ICE-GB: W2C-011 #064)

**e. Situation use**

The two types of situation use are explained below.

**f. Immediate situation use**

With regard to the immediate situation use, reference of the is obtained from the extralinguistic environment. In this category, the referent will be in sight of the speaker and hearer as shown below:

5.46 *So these are the troublemakers.* [said when the troublemakers are in view of the hearer] (ZE corpus: Private dialogue 1)

5.47 *Just look at the beautiful scenery here.* [said in a park] (ICE-GB: S2A-016 #0370)

**g. Larger situation use**

When considering large situation use as a function of the definite article, an appeal is made by the speaker to the hearer’s understanding of entities in existence “in the non-immediate or larger situation of utterances” (Hawkins, 1978: 11). For example, in the sentence below, the speaker and hearer will have knowledge of the mine.
5.48 But then the mine collapsed. (ZE corpus: Private semi-scripted dialogue 6)

Another example is shown in the sentence below where the speaker and hearer will have knowledge of the jewellery.

5.49 And then I used to look at the jewellery. (ZE corpus: Private semi-scripted dialogue 7)

In the ICE-GB, examples that show large situation use of the are shown below:

5.50 How is the Prime Minister selected? (ICE-GB: S1B-011 #080)

5.51 The development which uh was accelerated first by Hitler’s concern to acquire the V Two weapon and then by the political rivalry to be first on the moon. (ICE-GB: S2B-048 #032)

h. The “logical” use of the

The logical use of the definite article is shown below.

5.52 Luke Steyn made history after becoming the first Zimbabwean athlete to compete at the writer Olympics. (ZE corpus: Public dialogue 1)

In the example above, the uniqueness of the ordinal first is that there can’t be two first Zimbabwean athlete to compete at the winter Olympics.

If we consider the ICE-GB, the example below shows the logical use of the.

5.53 And Indonesia is the largest island archipelago in the world with some some seventeen thousand islands. (ICE-GB: S2A-046 #016)

The superlative adjective largest shows the uniqueness of the referent, in this case, the largest island archipelago in the world.

i. The use of the with reference to body parts

The use of the definite article to refer to body parts and following a preposition is illustrated below for both the ZE corpus and the ICE-GB.

5.54 A woman was also shot in the leg, said the patron. (ZE corpus: Newspaper reportage 15)
5.55  *Michael despised her for er indisposition, roughly picking her up by the arm.*  
(ICE-GB: W2F-008 #13)

In example 5.54, *the* is used to refer to a body part (*leg*), whilst in example 5.55, *the* is used to refer to a body part (*arm*).

**j. Generic use**

The following examples illustrate generic use of the indefinite article.

5.56  *The most mean, darkest, unforgiving and formidable prison in the world is the mind.*  
(ZE corpus: Newspaper reportage 11)

5.57  *And she plays the piano.*  
(ICE-GB: S1A-091 #346)

**5.6.2.2 Functions of the indefinite article in both corpora**

An analysis of the 400 concordance lines of the indefinite article yielded different uses of indefinite article in both corpora used in this study as illustrated in the following examples:

**a. Quantification: Denoting amount, extent and duration**

As a quantification function, the indefinite article is used to denote amount, extent and duration. In the ZE corpus, quantification is shown in the following examples:

5.58  *His name had been on the City Council’s housing waiting list for more than a decade and he no longer deceived himself with the sorry fantasy that a house could be properly, allocated to him without him greasing the palms of one official or the other.*  
(ZE corpus: Creative writing 7)

In the sentence above, quantification is shown by the uses of *a decade*.

In the ICE-GB, the quantification function of the indefinite article is shown by the use of the indefinite article with *large number* in the example below.

5.59  *If the project is successful and a large number of ethnic minority families volunteer to adopt and foster, they may be asked to take in white children.*  
(ICE-GB: W2C-015 #053)
b. The indefinite article functions as a reference to a specific thing in the ZE corpus as shown in the sentence below.

5.60 You returned a dress to my shop yesterday saying it is too tight and you want a bigger size. (Private business letter 4)

In the sentence above, the indefinite article is used to refer to a specific dress that was returned to the shop.

The use of the indefinite article to refer to a specific thing (a car) in the ICE-GB is shown below.

5.61 And this is a car you’ve got at the moment?. (ICE-GB: S1B-074 #289)

c. The ascription or attribution function of the indefinite is exemplified in the two sentences below:

5.62 This is a labour issue. (ZE corpus: Public dialogue 23)

5.63 We need to proceed with the greatest care therefore, for embryo research is a complex issue which involves the whole spectrum of medical, scientific ethical and moral issues. (ICE-GB: S1B-060 #080)

d. Co-text bound non-particular interpretations based on linguistic markers of non-factual context is another function of the indefinite article as shown below.

5.64 I want to go to university and get my degree so that I can be a doctor and live a better life. (Private semi-scripted dialogue 8)

Another example of co-text bound non-particular interpretations based on linguistic markers of non-factual context is another function of the indefinite article in the ICE-GB is shown below.

5.65 They would have a better chance of holding their seats and keeping power under another leader. (ICE-GB: W2C-003 #067)

e. The generic uses of the indefinite article occur when “a/an picks out any representative member of the class“ (Quirk et al. 1985: 281).

The following examples illustrate the generic use of the indefinite article.

5.66 But also there are some services which even if someone stays in a rural area which are needed to done in towns. (ZE corpus: Private dialogue 4)
In the ICE-GB, the generic use of the indefinite article is illustrated below.

5.67  *This is payable once only per annum at every centre where a student enrolls for one or more classes.* (ICE-GB: S2B-044 #118)

*f. Idiomatic use in fixed expressions is another function of the indefinite article. According to Botha (2012), examples of fixed expressions include earn a living and as a result.*

5.68  *On Monday, nurses at council clinics, who also claimed incapacitation as a result of eroded salaries, went on strike.* (ZE corpus: Editorials 7)

5.69  *Well, you can’t earn a living if they’re going to keep cancelling.* (ICE-GB: S1A-083 #062)

In the examples above, the indefinite article is used in the fixed expressions as a result and earn a living.

The examples above have been used to illustrate the different functions of the indefinite article as they are used in the ZE corpus and the ICE-GB. From the qualitative analysis done above, the frequency of each function of the indefinite article is counted and reported in the next section for the 400 concordance lines analysed.

5.6.2.3 Functions of the zero article in both corpora

The table below shows examples of the different functions of the zero articles as they appear in the ZE corpus and the ICE-GB. The zero article occurs before the underlined words in the examples below.
<table>
<thead>
<tr>
<th>Functions</th>
<th>Examples in the ZE corpus</th>
<th>Examples in the ICE-GB (BrE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used with plural nouns and noncount nouns, where the indefinite article does not occur.</td>
<td>Perhaps his ears had been too deafened by Ø hunger. (Creative writing 1)</td>
<td>Ø Research must continue however as a matter of the highest priority into improving the quality of life for handicapped people. (S1B-060 #056)</td>
</tr>
<tr>
<td>Occurs in noun phrases in a copular relation.</td>
<td>The discussion was moderated by Mr Sangu Delle, Ø chairman of Golden Palm Investments in Ghana. (Newspaper reportage 3)</td>
<td>I didn’t realise she was Ø Chairman of the Conservatives. (S1A-068 #054)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Some &quot;institutions&quot; of human life and society.</td>
<td>But it took days for them to get to Ø hospital. (Private semi-scripted dialogue 22)</td>
<td>They were in Ø hospital for a week and I never even got a cold when I was there. (S1A-014 #184)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Means of transport and communication.</td>
<td>Another way is that you can send them by Ø bus. (Private business letter 2)</td>
<td>First of all by train and boat and finally a tedious two hours by Ø bus. (W2F-013 #005)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Times of day and night.</td>
<td>Heavy rains coming in, cold coming in and the rains were continuous. From Ø morning to Ø sunset, from Ø sunset to Ø sunrise. (Private semi-scripted dialogue 24)</td>
<td>The indigestion is worse at Ø night, is it? (S1A-052 #074)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Seasons.</td>
<td>During Ø summer it got so hot that even if you spat, the spittle wouldn’t hit the ground. (Newspaper reportage 12)</td>
<td>On top of it I will be wanting things for Ø winter. (S1B-025 #027)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Meals.</td>
<td>For Ø lunch, Coventry said she has a sandwich and salad before going back to the pool for three hours. (Newspaper reportage 15)</td>
<td>England’s search for their first elusive first-class win on tour ended 40 minutes after Ø lunch here today. (W2C-014 067)</td>
</tr>
<tr>
<td>Noun phrases with sporadic reference: Illnesses</td>
<td>You mentioned that this is linked to your oral health when you talk about Ø heart disease and Ø diabetes as well. (Public dialogue 36)</td>
<td>Uhm I haven’t had Ø flu yet actually. (S1A-051 #227)</td>
</tr>
<tr>
<td>Parallel structures</td>
<td>She said that she was working Ø hand in hand, eh – with different organizations to ensure that eh all these people that were affected would have something to benefit. (Public dialogue 1)</td>
<td>One is to allow the contradictions to appear Ø side by side. (W2A-007 #013)</td>
</tr>
<tr>
<td>Fixed phrases involving prepositions</td>
<td>I am planning to take Ø advantage of the indigenisation and economic empowerment to manufacture things that we use like tents and chairs it really pains me to have such overheads. (Popular writing 1)</td>
<td>The hotel is conveniently located literally Ø on top of London Victoria Station. (W1B-016 #008)</td>
</tr>
</tbody>
</table>
The next section focuses on how articles are utilised in ZE.

5.6.3 Summary of quantitative and qualitative results for articles

In Section 5.6.1 data analysis provides the occurrence- and function frequencies of articles in the ZE corpus and the ICE-GB. A comparative analysis was also done in terms of the log likelihood test to see if there is a significant variation of articles between the corpora, in order to ascertain if the constructions themselves show innovation. This summary thus aims to answer research questions 1.3 and 2.3:

Are the variations between the corpora statistically significant?

When considering whether the variations between the corpora are statistically significant, there were variations observed in the frequencies of the indefinite article because *a/an* occurred more in the ICE-GB compared to the ZE corpus. The LL value of 87.63 was higher than 15.13, which is a critical value where *p* < 0.0001, meaning that the variations are not due to chance. There are no statistically significant variations noted in the frequencies of *the* in the ZE corpus and the ICE-GB because the LL value of 0.59 is less than the critical value of 3.84. Regarding the zero article, the normalised frequency per 10 000 words is higher in the ZE corpus compared to the ICE-GB. The LL value of 14.07 is statistically significant where *p* < 0.001, meaning that there are variations in the frequency of the zero article in the two corpora.

It was necessary to analyse the absolute and normalised frequencies of articles in registers in order to determine whether there are statistically significant variations in the two corpora. For the definite article, observations were made that *the*, occurred most frequently in the spoken register, in the ZE corpus compared to the ICE-GB. This is a statistically significant difference because the LL value of 9.49 > 6.63 where *p* < 0.01. The normalised frequency of the in the written register, in the ICE-GB was higher compared to the ZE corpus. This difference is statistically significant because the LL value of 4.90 is more than 3.84, where *p* < 0.05, which means that the variations observed in the frequency of the definite article in the written register are not due to chance.
Statistically significant variations were recorded in the frequencies of *a/an* in the spoken and written registers. In both the spoken and written registers, the ICE-GB had higher normalised frequencies than those in the ZE corpus. The significance of the observed differences is evident in the LL values for both spoken (15.75) and written registers (103.34), which are above the critical value of 15.13 where \( p < 0.0001 \). This means that the variations observed are not due to chance. The frequencies of the zero article in the spoken and written registers were not calculated due to the smaller sample size (cf. subsection c in Section 5.6.1.2).

If the distribution of *the* in the different genres is considered. Observations are that the highest difference in normalised frequency was recorded in private dialogue with a LL value of 238.38. The definite article *the* was attested more in private dialogue in the ZE corpus than in the ICE-GB with a higher value, which shows that the observed variations are statistically significant where \( p < 0.0001 \). Creative writing is another genre, which showed more frequency of *the* in the ZE corpus than in ICE-GB. A different pattern emerged for business letters, newspaper reportage, public scripted monologue and popular writing due to the fact that *the* was more frequent in the ICE-GB than in the ZE corpus. All the four genres had LL values above the critical value of 15.13 where \( p < 0.0001 \), meaning that there are variations in the frequency of *the* in the four genres. Considering that the frequency of *the* in editorials was higher in the ICE-GB compared to the ZE corpus, the difference is statistically significant at \( p < 0.05 \), because the LL (6.10) is higher than the critical value of 3.84. No statistically significant variations were recorded in social letters, academic writing and public dialogue, which means that there are no variations in the use of *the* in the ZE corpus and the ICE-GB.

The indefinite article occurring in newspaper reportage had the highest frequency, which is statistically significant, in BrE compared to ZE. The frequency of *a/an* per 10 000 words was also high in the ICE-GB compared to the ZE corpus in public scripted monologue, public dialogue, editorials, social letters and business letters. In contrast, *a/an* was more frequent in the ZE corpus in private dialogue compared to the ICE-GB. The LL value for this genre was above the critical value of 6.73 where \( p < 0.01 \), meaning that the observed variations are statistically significant. The LL values in popular writing, academic writing
and creative writing were not statistically significant at the lowest level where the critical value for p < 0.05 is 3.84. Due to the smaller sample, the frequencies of the zero article were not reported.

In order to determine whether there are statistically significant variations in the frequencies of different functions, the LL values were calculated for the articles. Regarding the frequency of different functions of the definite article, the non-referential function shows the biggest difference between the two corpora. The normalised frequency of the function in the ZE corpus was double compared to the number in the ICE-GB (19.5). The observed variation is statistically significant because the LL value of 315.26 is higher than 15.13, which is a critical value where p < 0.0001, meaning that the variation is not due to chance. Other functions showing higher frequencies in the ZE corpus compared to the ICE-GB include cataphoric reference, sporadic reference and the use of the with reference to body parts. These variations are all statistically significant at p < 0.0001 because the LL values were higher than the critical value of 15.13. Anaphoric reference with direct and large situation use, as functions of the definite article, occurred more in the ICE-GB compared to the ZE corpus. Other functions of the, which occurred more in the ICE-GB compared to the ZE corpus include the anaphoric referent with indirect immediate situation use and the “logical” use of the. These variations are statistically significant at different levels, meaning that the variations in the frequencies of these functions in the two corpora are not due to chance.

Comparison of the frequencies of functions of the indefinite article in the ZE corpus and the ICE-GB showed the biggest difference in the generic use of the indefinite article. The function was attested more in the ZE corpus compared to the ICE-GB. The LL value for generic function was the highest at 179.11, meaning that there are statistically significant variations in the frequency of the function. The other five functions of the indefinite article namely quantification, reference to a specific thing, ascription, idiomatic uses in fixed expressions and co-text bound non-particular interpretations based on linguistic markers of non-factual context occurred more in the ICE-GB compared to the ZE corpus as follows. The LL values for all the five functions are higher than 15.13, which is a critical
value where \( p < 0.0001 \), meaning that there are variations in the frequencies of these functions in the two corpora.

The following section outlines how the innovation features occurred in the ZE corpus and whether they can be attested in Kortmann, Lunkenheimer and Ehret’s (2020) list of innovation features.

5.6.4 Attestation of innovation features according to Kortmann et al. (2020) for articles used in Zimbabwean English

This section presents the results on how articles are used in ZE. This is done by qualitatively analysing the concordance lines to check whether the features studied are present or absent in the ZE corpus and then quantitatively, by counting the frequencies of these features. The corpus was analysed to check whether the features (60, 61, 62, 63, 64, 65, 66 and 67 – cf. table 27 below) that are reported to show morphosyntactic variation can be attested in ZE. A detailed description of the categorisation of features into Kortmann, Lunkenheimer and Ehret’s (2020) four categories and an explanation about the distinction of innovations and errors is given in subsection e in Section 4.4.2.

Table 27. Feature numbers and their descriptions

<table>
<thead>
<tr>
<th>Feature number and description</th>
</tr>
</thead>
<tbody>
<tr>
<td>60: Use of definite article where StE has indefinite article.</td>
</tr>
<tr>
<td>61: Use of indefinite article where StE has definite article.</td>
</tr>
<tr>
<td>62: Use of zero article where StE has definite article.</td>
</tr>
<tr>
<td>63: Use of zero article where StE has indefinite article.</td>
</tr>
<tr>
<td>64: Use of definite article where StE favours zero [article].</td>
</tr>
<tr>
<td>65: Use of indefinite article where StE favours zero [article].</td>
</tr>
<tr>
<td>66: Indefinite article one/wan</td>
</tr>
<tr>
<td>67: Demonstratives for definite article.</td>
</tr>
</tbody>
</table>

For features 60 to 63 the absolute distribution frequencies are very low and therefore no summary tables with quantitative data is provided. For features 64 to 67 these summary tables are however included.
Data analysis in this study intends to determine whether there is variation in the use of articles in ZE as reported and attested in other varieties of English in Kortmann, Lunkenheimer and Ehret’s (2020) list of innovation features. Below is an analysis of the ZE corpus to ascertain whether the following features are characteristic of ZE. Data interpretation is aided by biographical and language background questionnaire results.

Van Rooy, (2013: 12) notes that, “if the frequency is negligible in the face of an overwhelming trend to select the nativelike variant (or some other variant), then there is little reason to regard such a feature as an established feature”.

This is what was done in the current study. Features that were restricted to a few participants or limited genres were regarded as errors and not innovations.

This analysis focuses on answering research question 3:

3.1 Which features of innovation are present with regards to articles in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE?

The discussion includes a categorisation of the features according to the explanations given by Kortmann, Lunkenheimer and Ehret’s (2020). The explanations are as follows:

A- Feature is pervasive or obligatory.
B- Feature is neither pervasive nor extremely rare.
C- Feature exists, but is extremely rare.
D- Attested absence of the feature.

5.6.4.1 Use of definite article where Standard English has indefinite article: Feature 60

Using the different uses of the definite article (cf. Section 5.3.1) that are reported by Quirk et al. (1985: 265-272), the concordance lines of the were manually checked to determine whether StE would have used an indefinite article. The calculations for feature 60 were done on the basis of all the 16 493 instances of the definite article. Although this process
was labour intensive, it was done for several days to limit the risk of fatigue and not to compromise the accuracy of the results.

Data analysis seems to show that the use of definite article where StE has indefinite article is not prevalent in the ZE corpus, occurring only twice. Although there are examples of sentences where feature 60 occurs, they occur only once in each of the two genres. Another observation made is that the instances where the definite article is used instead of the indefinite article are recorded in spoken texts as shown in the following examples:

5.70  *We are all aware ZIFA is broke and we have been appealing to the corporate world the government to come in and and chip in and give the helping hand in football.* (Public dialogue 15)

5.71  *Uh I can just say I was just a a man of or I can say the Jack of all trades. Handball.* (Private semi-scripted dialogue 3)

The percentage of instances where the definite article is used where StE has indefinite article was calculated. This was done in order to determine the prevalence of feature 60. The instances where the definite article was used where StE has indefinite article was divided by the total number of occurrences of the definite article in the corpus (16 493). The result was multiplied by 100 to get a percentage. Feature 60 occurs 0.012% in the ZE corpus, whilst the percentage of instances where the definite article was used in a standard way is 99.98%.

Due to the fact that there are limited occurrences in which the definite article is used where StE has indefinite article, the two occurrences can be regarded as errors. This is in line with the guidelines proposed by Kruger and Van Rooy (2017). The authors emphasise that if a convention is not accepted, it will be confined to informal speech or it will be regarded as an error, which will disappear as one’s proficiency improves. Therefore, feature 60 can be classified as belonging to category D in which there is attested absence of the feature.

When considering African varieties of English, ZE seems to show similarities to such as Cape Flats English (Mesthrie, Toefy & Bowerman, 2020) and Nigerian English (Taiwo,
2020). As a comparison, varieties such as BSAS, Tanzanian English and Ugandan English classify feature 60 as being neither pervasive nor extremely rare (category B).

5.6.4.2 Use of indefinite article where Standard English has definite article: Feature 61

The concordances of plural nouns were checked to establish whether the indefinite article was used. There were no instances where the plural nouns were used with the indefinite article in the concordance lines of the indefinite article. There seems to be use of indefinite article where StE has definite article in the private semi-scripted dialogue genre only.

There were 10 instances of the use of indefinite article where StE has definite article in the ZE corpus. Self-correction by speakers was not included in the analysis. An example of self-correction is given below:

5.72 He is being relegated to a bench to the bench once again in place of Joram Muchambo. (Public dialogue 40)

The sentences provided below illustrate the use of feature 61.

5.73 I am the fourth born in a family. (Private semi-scripted dialogue 7)

5.74 There are, there is, eh in Dubai that’s where there is a tallest building in the world. (Private semi-scripted dialogue 33)

The absolute frequency of use of indefinite article where StE has definite article was divided by the number of instances where the indefinite article occurred (6 159) and the result was multiplied by 100. Results show that feature 61 occurred 0.16% of the time. The percentage of instances where the indefinite article was supplied is 99.84%.

Using the different uses of the indefinite article (cf. Section 5.3.2) that are reported by Botha (2012, 264-271), all concordance lines of the 6 159 instances of a/an were manually checked to determine whether StE would have used the definite article. Even though this process was time-consuming, the analysis was carried out over several days in order to limit the risk of fatigue and not to compromise the accuracy of the results. Due to the fact that feature 61 occurred 0.16% times in the ZE corpus and was attested 10
times out of the entire corpus in the spoken register and in only 1 genre, the use of indefinite article where StE has definite article can be absent in ZE (Category D). The 10 instances can be regarded as errors. Regarding the use of indefinite article where StE has definite article, Nigerian English (Taiwo 2020), Ghanaian English (Huber, 2020) and Cape Flats English (Mesthrie, Toefy & Bowerman, 2020) are also in category D.

5.6.4.3 Use of zero article where Standard English has definite article: Feature 62

Since a searchable entity is required for corpus analysis of the zero article, I searched for different tagsets that contain nouns in the ZE corpus. Only a random sample of 400 concordance lines of nouns was analysed due to the high frequency of the nouns (17,617). The nouns were then checked manually to see if the articles were present or absent (cf. Sand, 2004; Botha, 2012). Searching for the nouns was an important step because it enabled the researcher to check concordance lines to see which NP did not contain a determiner. A detailed description of the procedure for identifying nouns in order to check whether a determiner was used and whether the zero article was used where StE has definite article is provided in subsection i in Section 4.4.2.6. In addition, the different uses of the zero article (cf. Section 5.3.3) were used as guidelines to determine if the zero article was used in place of the definite article.

There were 89 instances of the zero article in a random sample of 400 nouns in the ZE corpus. The 89 instances were checked to see if the zero article was used where StE uses the definite article. Feature 62 occurred 3 times out of the 89 occurrences of the zero article as shown in the examples below.

5.75  *Concerning money issues, I was given $1500 by Svodai at Ø start of this month, so I had to borrow $1000 from Costain.* (Social letters 1)

5.76  *Ø Problem is money.* (Social letters 1)

5.77  *So, there were some members who were said to be of Ø ruling party.* (Private semi-scripted dialogue 14)

For feature 62, in order to determine the percentages of when an article was used in a standard manner and the number of times an article was used in a non-standard manner,
the number of instances (3) was divided by the total number of the zero article (89) in a sample of 400 nouns and the result was multiplied by 100. Results showed that the percentage of occurrence of feature 62 is 3.37%. The zero article was used in a standard way 96.63% of the time. Feature 62 can be classified under category C (feature exists but is extremely rare) because the feature was attested in both spoken and written registers. When considering other L2 varieties of English, Mesthrie (2020a) draws the same conclusion in BSAE and Hundt, Biewer and Zipp (2020) classify feature 62 as belonging to the C category in Acrolectal Fiji English.

In addition to the abovementioned studies, Minow (2010) notes that article omission can be classified as extremely rare in BSAE, given that the rate of omission by half of the participants in the study was very low (less than 5%). De Klerk (2003a) reports that omission of determiners is one of the features that shows innovation in Xhosa English. She illustrates this in the following examples:

5.78 You are asking about Ø death penalty.
5.79 There are many people saying Ø death penalty should be brought back.
5.80 Ja, Ø death penalty is such a cruel thing.

5.6.4.4 Use of zero article where StE has indefinite article: Feature 63

The same procedure that was used in Section 5.6.4.3 was used for identifying the zero article. A random sample of 400 concordance lines containing nouns was examined because there are 79 176 nouns in the ZE corpus. A detailed description of the procedure for identifying nouns in order to check whether a determiner was used and whether the zero article was used where StE has an indefinite article is provided in subsection i in Section 4.4.2.6. In addition, the different uses of the zero article (cf. Section 5.3.3) were used as guidelines to determine if the zero article was used in place of the indefinite article.

The percentage of instances where the zero article was used where StE has indefinite article was calculated. This was done by dividing the absolute frequency of feature 63 by...
the total number of occurrences of the zero article in the corpus and then multiplying the result by 100. Out of the 89 occurrences of the zero article in 400 nouns, feature 63 was attested twice. The percentage of instances of the feature 63 is 2.25% and the zero article was used in a standard way 97.75% of the time. Although the use of zero article where StE has indefinite article was attested in the corpus, the feature exists but is extremely rare. Feature 63 occurs in both spoken and written registers in the ZE corpus. In addition, the feature meets other criteria that warrant it to be assigned to category C. These include use in both spoken and written registers and authoritative and norm orientation (as shown by the occurrence of the feature in editorials). Feature 63 occurred once in public dialogue and once in editorials. Some of the examples are shown below with the omitted indefinite article given in brackets.

5.81 So, this is what happened, but this is Ø (a) church and it’s an apostolic church and when we dealt with Gumbura’s case. (Public dialogue 14)

5.82 More than half of its land is Ø (a) desert and the climate and lack of water resources are naturally not in favour of agriculture. (Editorials 8)

If one considers Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features, feature 63 fits in the C category because the feature exists but is extremely rare. As a way of comparison, other L2 varieties of English in which feature 63 exists but is extremely rare include BSAE (Mesthrie, 2020a), Cameroon English and Ghanaian English. One of the features of BSAE is omission of articles (Gough, 1996a). An example sentence given is; He was Ø good man. Minow (2010) notes that article omission can be classified as extremely rare or fairly marginal in BSAE, because the rate of omission in her study was very low.

5.6.4.5 Use of definite article where Standard English favours zero – Feature 64

To check whether the definite article was used where StE favours the zero article, different contexts reported by Sand (2004: 290-291) were used as the basis for analysis. The concordance lines of the were checked for the contexts which are given below:
1. In collective nouns with generic meaning. For example: *The girls tend to fare better in these subjects* (Sand, 2004: 290).

2. Before nouns referring to institutions, such as church, university, college, school or jail, when a type of institution rather than a specific entity or body is referred to. For example: “And her eh sister’s going over to Scotland to the university” (Sand, 2004: 291).

3. In proper names of places or institutions, where the article is not part of the name. For example: *Thankfully, he was sent to the Singapore General Hospital early enough* (Sand, 2004: 291).

4. In temporal expressions when a specific reference is not established by a postmodification. For example: *I’m going over for a meeting on the Thursday* (Sand, 2004: 291).

5. In deverbal nouns in -ing, which may occur in contact varieties instead of verbal expressions. For example: *So the parking is becoming a problem because of you know there are so many buildings there are so many vehicles* (Sand, 2004: 191).

6. In deverbal nouns in -ing, which may occur in contact varieties instead of verbal expressions. For example: *So the parking is becoming a problem because of you know there are so many buildings there are so many vehicles* (Sand, 2004: 191).

Although there seems to be more spoken texts where feature 64 occurs than written texts, it is interesting to note that the use of definite article where StE favours zero [article] is not attested in editorials. Kruger and Van Rooy (2017) are of the opinion that if a feature is accepted by gatekeepers such as editors, then it is considered an innovation.

Using several criteria described in subsection e in Section 4.4.2.6, feature 64 was analysed to check if the use of the definite article where StE favours zero can be regarded as innovation. Regarding the frequency cut-off points and other guidelines used to draw the line between classification of categories. To summarise, if a feature was attested 1 to 15 times, it was placed in category D. If a feature was attested 16 to 30 times, it was classified...
as belonging to category C, meaning the feature exists, but is extremely rare. A feature that occurred 31 to 45 times was placed in category B, meaning the feature is neither pervasive nor extremely rare. Category A (feature is pervasive or obligatory), was assigned to a feature whose frequency count was 46 and above. Subsection e in Section 4.4.2.6 also describes the additional criteria used for classifying features into categories A, B, C and D.

Below is a table showing the distribution of feature 64 according to genres.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private dialogue</td>
<td>7</td>
</tr>
<tr>
<td>Private semi-scripted dialogue</td>
<td>4</td>
</tr>
<tr>
<td>Public dialogue</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
</tr>
</tbody>
</table>

Example sentences where feature 64 occur are shown below:

5.83 If if you can have those facilities more than WhatsApp. We can have the the Facebook, eh Twitter, Myspace and the like. (Private dialogue 5)

5.84 Like, after that, the government seek help from, donations from the citizens. And the citizens donated. Some were send to the uh to help those who were, who survived from the the Idai 49. (Private semi-scripted dialogue 3)

5.85 Eh ideally the college life should be quite uh amusing. (Private semi-scripted dialogue 45)

In example 5.85, when participant <$SS> is asked what she thinks about college life in general, she replied with the definite article the even though she did not go to college.

To assist in determining whether the definite article was used where StE favours the zero can be regarded as a feature, the absolute frequency of feature 64 (19) was divided by the total number of occurrences of the definite article (16 493). The result was multiplied

49 In example 5.84, Idai refers to a cyclone.
by 100. Feature 64 occurs 0.12% in the ZE corpus, whilst the percentage of instances where the zero article was used in a standard way is 99.88%.

Subsection e in Section 4.4.2.6 provides guidelines for categorising features into A, B, C and D. The fact that feature 64 is attested in the spoken register only and in 3 genres namely private dialogue, public dialogue and public scripted monologue means that the feature does not meet the guideline about being attested in both spoken and written registers (cf. Van Rooy & Kruger, 2016). In addition, the fact that feature 64 occurs in 3 genres means that it can be assigned to category D, where there is attested absence of the feature. Another reason for assigning feature 64 to category D is that the percentage of occurrence is 0.12. The 21 instances where feature 64 occurred can be regarded as errors.

Given that the age ranges of participants in private semi-scripted dialogue and private dialogues are between 18 and 75 years and that the educational backgrounds and occupations of participants are varied, this shows the diversity in sampling participants and assists in determining whether a feature is an innovation or an error.

Kortmann, Lunkenheimer and Ehret’s (2020) classified the use of definite article where StE favours zero [article] in Ugandan English, Ghanaian English, Malaysian English and Pakistani English into the D category. The same category can be allocated to ZE.

5.6.4.6 Use of indefinite article where StE favours zero:

Feature 65

Data analysis seems to show that the use of indefinite article where StE favours zero [article] is more prevalent in private semi-scripted dialogue occurring 10 times followed by 7 times in public dialogue. The genres, absolute frequency and percentages are indicated in table 29 below.
Table 29. Distribution of feature 65 across genres

<table>
<thead>
<tr>
<th>Genre</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private semi-scripted dialogue</td>
<td>10</td>
</tr>
<tr>
<td>Newspaper reportage</td>
<td>2</td>
</tr>
<tr>
<td>Public dialogue</td>
<td>5</td>
</tr>
<tr>
<td>Popular writing</td>
<td>1</td>
</tr>
<tr>
<td>Public monologue</td>
<td>1</td>
</tr>
<tr>
<td>Academic writing</td>
<td>1</td>
</tr>
<tr>
<td>Private business letters</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
</tr>
</tbody>
</table>

It is worth noting that both spoken and written registers are represented in the table above. This shows that feature 65 is not only found in either spoken or written texts. In addition, the use of indefinite article where StE favours zero is attested in 7 genres namely private semi-scripted dialogue, newspaper reportage, public dialogue, public monologue, popular writing, private business letters and academic writing.

To help determine the category for feature 65, the absolute frequency of the use of indefinite article where StE favours zero divided by the total number of occurrences of the indefinite article. The result was then multiplied by 100. The percentage of occurrence of feature 65 is 0.34%. The indefinite article was supplied 99.66% of the time in the ZE corpus. When considering Kortmann, Lunkenheimer and Ehret’s (2020) taxonomy of features, feature 65 can be described as belonging to category C. The reason being that the use of indefinite article where StE favours zero [article] is attested 21 times and the percentage of occurrence of the feature is lower (0.34%). After considering the different guidelines for categorisation of features, the use of indefinite article where StE favours zero is assigned to category C in Kortmann, Lunkenheimer and Ehret’s (2020) grouping.

The following examples illustrate the use of indefinite article where StE favours zero[article]:

© University of Pretoria
But eh o what Mr Katsinha was saying about the salary increment it remains a hearsay until he proves it because he he’s he’s purporting to have stolen documents and now he can't even prove anything. (Public dialogue 25)

Uh it is actually important for me because it provides a decent accommodation for my family. (Private semi-scripted dialogue 10)

I can say there was a massive destruction mostly in Manicaland south. (Private semi-scripted dialogue 28)

He has always been playing with a lot of experience and coming into this game he comes with a valuable experience. (Public dialogue 40)

We would like to thank you for doing a great work at our school. (Private business letter 1)

Uh in fact what actually happened was that, at my school where I am working, there was an overstaffing. (Private semi-scripted dialogue 21)

A comparison of ZE with other L2 varieties of English represented on eWAVE shows that ZE is in the same category as Kenyan English (Buregeya, 2020) and Indian South African English (Mesthrie, 2020b) in which the use of indefinite article where StE favours zero [article] exists, but is extremely rare.

5.6.4.7 Indefinite article one/wan: Feature 66

Bowerman (2008: 482) reports that one is used as a non-specific determiner “to pick out one of a set, without specifying exactly”. The author gives the following examples:

5.92 My one cat is sick. (One of my cats is sick.) (Bowerman, 2008: 482)

5.93 He’s broken his one leg. (One of his legs/ a leg is broken.) (Bowerman, 2008: 482)

In order to draw the line between quantitative use and article use of one, different contexts in which one has a quantitative function were considered and excluded. According to Quirk et al. (1985), the quantitative function of one is realised when one occurs with singular count nouns (e.g. one sister), Another context in which one serves as a quantitative is when one is used “as a slightly emphatic equivalent of a (e.g. one or two miles).
Concordance lines of *one* were extracted and examined in the contexts in which they occurred. The Skema manual concordance annotation tool was used to label instances where *one* occurred as a non-specific determiner.

The indefinite article *one* appeared 8 times in the corpus. The feature was attested in 2 genres namely Private semi-scripted dialogue (5 times) and public dialogue (3 times) as shown in the table below.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private semi-scripted dialogue</td>
<td>5</td>
</tr>
<tr>
<td>Public dialogue</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
</tr>
</tbody>
</table>

The indefinite article *one* is illustrated below:

5.94 *Uh but having said that, as a society which is all these farms that were around, we had several common problems. Uh one major one being uhm thieves.* (Private semi-scripted dialogue)

5.95 *Interestingly this is one team that many people regarded as minors.* (Public dialogue 15)

From the guidelines of categorising features (subsection e in Section 4.4.2.6). The percentage of instances where the indefinite *one* was used was calculated. This was done by dividing the absolute frequency of *one* as an indefinite article by the total number of instances of *one* in the corpus and then multiplying the result by 100. The percentage of the use of *one* as an indefinite article is 0.74%. *One* was used in a standard manner 99.26% of the time.

Due to the fact that the percentage of occurrence of the indefinite article *one* is lower and due to the fewer number of genres in which the feature occurred, the instances in which feature 66 occur can be regarded as errors. This is in line with Van Rooy’s (2013) recommendation that if a feature’s frequency is negligible, then it cannot be regarded as an established feature, feature 66 is not regarded as an established feature in ZE due to the fact
that the indefinite article *one* was restricted to 2 genres only. Feature 66 can thus be classified in the D category where there is attested absence of the feature. The indefinite article *one* is reported to be absent in Cape Flats English, Ugandan English, Nigerian English and Malaysian English in Kortmann, Lunkenheimer and Ehret’s (2020) taxonomy.

5.6.4.8 Demonstratives for definite article: Feature 67

Using the rules of reference grammars (e.g. Quirk *et al.*, 1985; Huddleston & Pullum, 2002), the instances where demonstratives occurred were checked to determine whether they were used instead of definite articles. For example, Huddleston and Pullum (2002), comment that demonstratives are used to mark the NP as definite.

The use of the demonstrative *that* for definite articles was prevalent in the ZE corpus. Demonstratives include *this, that, these, those*.

“Demonstratives in English inflect for number as *this/these* as well as *that/those* show. While this and that point out one entity, these and those point out more than one entity by changing from singular to plural as in: (a) *This* boy is my friend. (b) *These* boys are my friends. (a) *That* man is in love. (b) *Those* men are in love” (Ekah & Udosen (2018: 21).

Several criteria were taken into consideration to identify innovation (cf. subsection e in Section 4.4.2.6). Instances where *that* was not used as a determiner were not considered. For example, the use of *that* as a conjunction as in the sentence “the president actually said *that* we mustn’t go out” (ZE corpus: Private dialogue 2). The use of *that* as an adverb was excluded, as exemplified in the sentence “I never dreamt I will go *that* far” (ZE corpus: Private unscripted monologue 40).

Results from the data analysis show that the frequency count for feature 67 is ninety. Even though the frequency count warrants feature 67 to be assigned category A where the feature is pervasive or obligatory, other classification criteria need to be considered as well. The percentage of instances where the demonstrative *that* was used as a definite article was calculated. This was done in order to determine the prevalence of feature 67. The instances where *that* was used in place of the definite article was divided by the total
number of occurrences of *that* as a demonstrative in the corpus. The result was multiplied by 100 to get a percentage. Feature 67 occurs 5.63% in the ZE corpus, whilst the percentage of instances where *that* was used as a demonstrative is 94.37%.

Based on the percentage of occurrence, feature 67 fits in category B in Kortmann, Lunkenheimer and Ehret’s (2020) categorisation because the feature is neither pervasive nor extremely rare, occurring 90 times in the ZE corpus. The different genres in which demonstratives are used instead of the definite article together with the frequencies and percentages are shown in the table below.

**Table 31. Distribution of feature 67 across genres**

<table>
<thead>
<tr>
<th>Genre</th>
<th>Absolute frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private semi-scripted dialogue</td>
<td>53</td>
</tr>
<tr>
<td>Private dialogue</td>
<td>26</td>
</tr>
<tr>
<td>Creative writing</td>
<td>1</td>
</tr>
<tr>
<td>Newspaper reportage</td>
<td>5</td>
</tr>
<tr>
<td>Social letters</td>
<td>3</td>
</tr>
<tr>
<td>Private business letters</td>
<td>1</td>
</tr>
<tr>
<td>Business letters</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>90</strong></td>
</tr>
</tbody>
</table>

It can be noted from the table above that private semi-scripted dialogue had the highest frequency followed by private dialogues. Feature 67 occurred 5 times in newspaper reportage and 3 times in social letters. The genres with the least frequency are private business letters and creative writing. The low frequency can be attributed to the composition of the corpus. Although efforts were made to get as many texts per genre as possible, it was not achieved due to time and financial constraints.

Interesting data is yielded from the table above, which shows that feature 67 occurred in both spoken and written texts. This is evidence that the feature is used across 7 different genres and therefore can be categorised as prevalent in ZE. Another interesting observation is that the use of demonstratives for definite articles occurs in texts of
participants of different age groups and varied educational backgrounds as shown in the table below:

If we consider the results from biographical and language background questionnaires, the. This can be taken as evidence of universal use of the innovation feature by L2 language users as explained by Bamgbose (1998) and Kruger and Van Rooy (2017).

In order to illustrate the use of demonstratives for definite articles, several examples are given below.

5.96 What what what gave you that confidence and the guts to say you know what we’re going to do it? (Public dialogue 27)

5.97 Mushambo will take that goal kick <,> from the right<,> from the left side to the right side. (Public dialogue 40)

5.98 The falling of that water into the gorge. (Private semi-scripted dialogue)

5.99 They will tell you the past was better even if, in the middle of the worst drought since 1981, they haven’t tasted that dreadfetid mealie-meal we used to call "kenya" in 1992. (Newspaper reportage 12)

5.100 I have seen your email. Please send that book to the address below: House No. 412/1 Mkoba 13 Gweru I’m yet to find a room to lodge in Gweru. (Social letters 2)

5.101 But the real work had been done before, with that victory in Kinshasa, that draw in Brazzaville and that demolition of Liberia. (Newspaper reportage 7)

5.102 I am sure that is where I said there is a, in our constitution there’s a gap in our constitution that our constitution will actually provide for that right to defend yourself, the right to defend your property. (Public dialogue 33)

5.103 Some people who are connected to the government are allowed into the country with that virus. (Private dialogue 1)

5.104 Just that guts to approach these guys who work at eh Makombe. (Private semi-scripted dialogue)

50 The yellow mealie meal was popularly known as “kenya” in Zimbabwe and was imported due to the drought.
When looking at the examples above, it can be observed that ZE is in the same category as L2 varieties of English such as BSAE (Mesthrie, 2020a), which also puts feature 67 in category B. There is attested absence of feature 67 in other African varieties of English such as Cape Flats English (Mesthrie, Toefy & Bowerman, 2020), Nigerian English (Taiwo, 2020), Ugandan English (Ssempuuma, 2020) and Ghanaian English (Huber, 2020).

Compared to other AE varieties such as BSAE, the use of demonstratives is one area that is reported to show innovation. For example, Van de Walt and Van Rooy (2002) report that the number distinction for demonstratives is discarded in BSAE. In a study by Botha (2012), the author shows that there are more uses of the distal demonstratives (that, those) in the TLEC.

### 5.7 Conclusion

In the overall analysis of articles, it was found that there were no statistically significant differences observed between the normalised frequencies of the definite article in the two corpora. This is because the ZE corpus recorded 463.3 instances per 10 000 words whilst the ICE-GB recorded 459.8 instances per 10 000 words. A comparative analysis of the indefinite article was done on the ZE corpus and the ICE-GB. Data analysis showed that the normalised frequency of the ICE-GB (200.1 instances per 10 000 words) compared to 173 instances per 10 000 words for the ZE corpus. For the indefinite article, results from a LL test showed that the variations observed were statistically significant due to the fact that the LL value of 87.63 is higher than the critical value of 15.13, where $p < 0.0001$. When considering the zero article, there are variations in the use of the zero article due to the fact that statistically significant differences were observed in the normalised frequencies of the two corpora. The ZE corpus has 463.3 instances per 10 000 words whilst the ICE-GB has 459.8 instances. Regarding the total frequency of the definite article, the indefinite article and the zero article, there are no statistically significant variations observed. This is because the LL value of 0.74 is lower than the critical value of 3.84, where $p < 0.05$. 

© University of Pretoria
Looking back at the data, the **frequency of the definite article** in the **spoken register** was more in the ZE corpus (411.1 times per 10 000 words) whilst the number was lower in the ICE-GB (395.1 per 10 000 words). These observed variations are statistically significant because the LL value of 9.49 > 6.63 where p < 0.01. Regarding the **written register**, the LL value of 4.90 showed statistically significant variations because it is higher than 3.84 where p < 0.05. In the written register, *the* recorded a higher frequency per 10 000 words in the ICE-GB.

If we consider the **indefinite article** *a/an*, the normalised frequencies per 10 000 words were higher in both the **spoken and written register** in the ICE-GB (185.7 and 227.5 respectively) when compared to the ZE corpus (171.9 and 175.2 respectively). The variations observed are statistically significant because the LL value for the spoken register is 15.75 whilst the LL value for the written register is 87.63. Both numbers are higher than the critical value of 15.13 where p < 0.0001.

When considering the **definite article**, data analysis of the different genres showed that the genres with higher normalised frequencies per 10 000 words in the ZE corpus are private dialogue (358.7 compared to 261.4 for the ICE-GB) and creative writing (581.3 compared to 511.2 for the ICE-GB). The LL values for these genres are 238.38 and 11.46 respectively, meaning that the observed variations are statistically significant. A different pattern emerged when considering the fact that *the* was more frequent, per 10 000 words, in the ICE-GB. In business letters, the normalised frequency per 10 000 words in the ICE-GB was observed to be 508 (compared to 360.1 for the ZE corpus), whilst in popular writing, the number is 742 (compared to 536.4 for the ZE corpus. The observed results were statistically significant where p < 0.0001 because the LL > 15.13. Other statistically significant differences were noted, with the ICE-GB recording higher normalised frequencies per 10 000 words in public scripted monologue and newspaper reportage compared to the ZE corpus.

Regarding the normalised frequencies of the **indefinite article** *a/an*, the genre showing a higher normalised frequency that is statistically significant in the ZE corpus include private dialogue. When considering the normalised frequencies in the ICE-GB,
Statistically significant variations were recorded in the use of the indefinite article in public scripted monologue, public dialogue, editorials, social letters and business letters.

The absolute and normalised frequencies of the zero article in the different genres were not calculated because the sample sizes that were obtained from random sampling of 400 nouns were too small.

A qualitative analysis of 400 concordance lines, together with projected distribution calculations helped in determining the **frequency of different functions** or uses of the definite article and the indefinite article. It was observed that the non-referential function of the definite article showed the biggest difference between the two corpora, occurring double the number of times in the ZE corpus, compared to the ICE-GB. Cataphoric reference, sporadic reference and the use of *the* with reference to body parts showed higher normalised frequencies in the ZE corpus compared to the ICE-GB. Due to the fact that the LL values were higher than the critical value for \( p < 0.0001 \), the variations are statistically significant. The functions of the definite article that occurred more in the ICE-GB compared to the ZE corpus include indirect anaphoric reference, immediate situation use, the “logical” use of *the*, direct anaphoric reference, and large situation use. The variations in the frequencies of these functions are statistically significant at different levels, meaning in the two corpora are not due to chance.

The frequencies of functions of the indefinite article were compared in the ZE corpus and the ICE-GB. The generic use of the indefinite article showed the biggest difference because the function occurred more in the ZE corpus compared to the ICE-GB, with a higher LL value showing statistical significance at \( p < 0.0001 \). The other five functions of the indefinite article namely quantification, reference to a specific thing, ascription, idiomatic uses in fixed expressions and co-text bound non-particular interpretations based on linguistic markers of non-factual context were attested more in the ICE-GB compared to the ZE corpus. These functions had higher LL values, meaning that the variations observed were statistically significant at \( p < 0.0001 \). The frequencies of different functions of the zero article were not calculated because the sample of zero articles obtained from a random sample of 400 nouns was not adequate when
considering Sison and Glaz’s (1995) determination that a sample size of at least 384 is adequate if a researcher wants to achieve a confidence level of 95% and a deviation tolerance of 5%.

In terms of the features studied, the following table provides an overview of how the examined features were attested in the ZE corpus when considering Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features.

Table 32. Summary of features and their categorisations in ZE

<table>
<thead>
<tr>
<th>Feature</th>
<th>Category of feature in ZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>60: Use of definite article where StE has indefinite article.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>61: Use of indefinite article where StE has definite article.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>62: Use of zero article where StE has definite article.</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>63: Use of zero article where StE has indefinite article.</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>64: Use of zero article where StE has indefinite article.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>65: Use of indefinite article where StE favours zero [article].</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>66: Indefinite article one/wan.</td>
<td>(D) Attested absence of the feature.</td>
</tr>
<tr>
<td>67: Demonstratives for definite article.</td>
<td>(B) Feature is neither pervasive nor extremely rare.</td>
</tr>
</tbody>
</table>

Several criteria were used to assign features into different categories proposed by Kortmann, Lunkenheimer and Ehret (2020). The criteria include frequency counts, regular and widespread use of a feature, whether a feature was attested in different genres, occurrence of a feature in both the spoken and written register and authoritative and norm orientation. A detailed description of the guidelines is given in subsection e in Section 4.4.2.6. The various calculations of non-standard features were done on the following basis that are summarised below:

(a) Feature 60 - The instances where the definite article was used where StE has indefinite article (2) was divided by the total number of occurrences of the definite article in the corpus (16 493). The result was multiplied by 100 to get a percentage.
(b) Feature 61 - The absolute frequency of use of indefinite article where StE has
definite article (10) was divided by the number of instances where the indefinite
article occurred (6159) and the result was multiplied by 100.

(c) Feature 62 - the number of times an article was used in a non-standard manner,
the number of instances (3) was divided by the total number of the zero article
(89) in a sample of 400 nouns and the result was multiplied by 100.

(d) Feature 63 - the number of times an article was used in a non-standard manner,
the number of instances (2) was divided by the total number of the zero article
(89) in a sample of 400 nouns and the result was multiplied by 100. The
percentage of instances where the indefinite article was supplied is 2.25%.

(e) Feature 64 - The absolute frequency of feature 64 (19) was divided by the total
number of occurrences of the definite article (16 493). The result was multiplied
by 100. Feature 64 occurs 0.12%.

(f) Feature 65 - The absolute frequency of the use of indefinite article where StE
favours zero (21) was divided by the total number of occurrences of the indefinite
article (6159). The result was then multiplied by 100. The percentage of
occurrence of feature 65 is 0.34%.

(g) Feature 66 - This was done by dividing the absolute frequency of one as an
indefinite article (8) by the total number of instances of one in the corpus and then
multiplying the result by 100. The percentage of the use of one as an indefinite
article is 0.74%.

(h) Feature 67 - The instances where that was used in place of the definite article was
divided by the total number of occurrences of that as a demonstrative in the
corpus. The result was multiplied by 100 to get a percentage. Feature 67 occurs
5.63% in the ZE corpus, whilst the percentage of instances where that was used
as a demonstrative is 94.37%.

Results showed attested absence of features 60, 61, 64 and 66. The examples observed
were either too few or occurred in fewer genres and in only one register. This warranted
the examples to be considered as errors. Features 62, 63 and 65 were assigned to
category C in Kortmann, Lunkenheimer and Ehret's (2020) categorisation of features.
This is because the features met some of the guidelines described in subsection e in
Section 4.4.2.6. However, the frequency counts were few and the percentages of occurrences of the features were lower, warranting the features to be assigned to category C. Feature 67 had the highest frequency in the ZE corpus and met most of the criteria for it to be assigned into category B.

When considering whether there is innovation in the use of articles in ZE, data analysis shows that there is innovation in some of the features, which occurred in a variety of genres and were recorded in the B category. It can be concluded that there is innovation as shown by the use of demonstratives for definite articles. The classification of features 62, 63 and 65 into category C shows that new norms may still be in the process of being formed. The emergence of innovative features is a hallmark of stage 2 of the DyM. Schneider mentions that, at stage 2, differences between indigenous and Standard strands of English emerge. A determination of the stage at which ZE is in the DyM is done in Section 8.2.4.

The next chapter discusses modal verbs and analyses how they are used in ZE.
Chapter 6

Modal verbs

6.1 Introduction

Variations in the use of modal verbs in L2 varieties of English have been reported in literature (e.g. Kortmann, Lunkenheimer & Ehret, 2020). The variations include the use of double modals (feature 121), present tense forms of modals used where StE has past tense forms (feature 123) and non-standard use of modals for politeness reasons (feature 127). The study of modal verbs in different varieties of English has been enhanced by corpus linguistics because corpora can be used as observational data to analyse modal verbs. For example, comparisons can be made between corpora to determine whether there are variations in frequencies of modal verbs (e.g. Collins, 2009; Wasserman, 2014; Wasserman & Van Rooy, 2014). Although modal verbs in English have been studied in different contexts, a lot needs to be learned about modal verbs in the ZE context where English is used as a L2. A discussion about the studies done on the use of modals in L2 varieties of English is provided in Section 6.4.

In order to contextualise the investigation, the chapter will firstly outline how modal verbs are used in English. This will link with section 6.3 on the functions of modal verbs, in which modal verbs are discussed according to their different meanings. Section 6.4 unpacks some of the studies done on modal verbs in L2 varieties of English. In Section 6.5, the three features examined in this study are explained, together with a summary of how the features are attested in different L2 varieties of English. Results from corpus analysis are then reported in Section 6.6, focusing on the frequency, functions and the attestation and prevalence determination of innovation features in ZE in order to determine whether there are variations and innovation in ZE.

Section 6.6 will provide an analysis of the data in terms of the following sub groupings: Section 6.6.1.1 discusses the overall occurrence frequencies of nine modal verbs in the
ZE corpus and the ICE-GB while also comparing the occurrence frequencies between the corpora by using the log likelihood test. The LL values are also reported in order to determine whether the variations observed are statistically significant. The LL values, absolute and normalised occurrence frequencies of the nine modal verbs in the spoken and written registers are discussed in Section 6.6.1.2. The function frequencies of functions of modal verbs are given in Section 6.6.1.4. Specifically, section 6.6.1.4 is subdivided according to the different categories of functions reported by Collins (2009) in order to account for the LL values, absolute and normalised frequencies of the different functions of modal verbs. In Section 6.6.2, the different functions of modal verbs are discussed. Section 6.6.4 also uses this subdivision (6.6.4.1, 6.6.4.2 and 6.6.4.3) in order to discuss each of the three features studied and to determine the feature’s category in Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features. The chapter is concluded in Section 6.7. These subgroupings occur in order to facilitate 4 main ideas in order to answer research questions, by focussing on modal verbs. For research question one, the frequencies and functions of modal verbs in the ZE corpus are discussed. For research question two, ZE corpus is compared to the ICE-GB in order to determine whether there is variation with modal verbs as outlined in Section 1.6. For research question three, the analysis focuses on whether there is innovation in ZE by looking at three features that are reported to show variation in in Kortmann, Lunkenheimer and Ehret (2020).

A discussion about how modal verbs are used in StE is given next, together with their functions. Thereafter, possible variations in the use of modals that have been reported in literature are discussed.

### 6.2 Modal verbs in English

Richards and Schmidt (2010: 369) describe English modal verbs as “any of the auxiliary verbs, which indicate attitudes of the speaker/ writer towards the state or event expressed by another verb, i.e. which indicate different types of modality”. Quirk et al. (1985: 219) are of the opinion that “modality may be defined as the manner in which the meaning of a clause is qualified so as to reflect the speaker’s judgement of the likelihood
of the proposition it expresses being true”. Modal verbs are also known as modals (Kakzhanova, 2013) and modal auxiliaries (Hykes, 2000) in literature. In this study, the term “modal verbs” is used whilst other terms are used as they appear in literature. According to Kukucz (2009) modal verbs contribute to the meaning of the verb phrase.

There is no consensus on what verbs should be included in the list of modal verbs (Salkie, 2009, Kakzhanova, 2013). Some scholars list nine modal verbs such as can, could, may, might, must, shall, should, will and would (e.g. Hykes, 2000) whilst others list 10 modal verbs namely can, could, may, might, must, shall, should, will, would and the tenth modal have (got) to (e.g. Richards & Schmidt, 2010). Another list consists of 11 modal verbs namely, can, could, may, might, must, shall, should, will, would as well as ought(to) and need (e.g. Coates & Leech, 1980). It is interesting to note that although scholars differ regarding what constitutes modal verbs, there are some modal verbs, which appear in all lists provided by researchers with different views as can be seen from the lists above. Quirk et al. (1985: 137) distinguish between ‘central’ modals (can, could, will, would, may, might, shall, should, must) and ‘marginal’ modals (dare, need, ought to, used to). Following Quirk et al. (1985), as well as Hykes (2000) and Gustilo’s (2011) lists, nine modal verbs (can, could, may, might, must, shall, should, will and would) were examined for variations in this study. The researcher chose nine modal verbs for analysis because more researchers seem to list them as modal verbs (e.g. Quirk et al., 1985; Hykes, 2000; Gustilo, 2011). Collins (2009) adds that will can be contracted to ‘ll and Gustilo (2011) notes that the modal verb would can be reduced to ‘d. In this study, the contracted forms were grouped together with the uncontracted forms for data analysis.

Characteristics that differentiate modal verbs from lexical verbs are shown below (Halliday, 1976 as cited in Hykes, 2000: 7; Huddleston & Pullum, 2002: 107; Collins, 2009):

1. Modals have no finite forms, (to, -ing, -ed, etc.): e.g. maying, to may.

2. Negative and interrogative forms are made without expansion. This means that, when making negative forms (to reply or deny something) and interrogative forms (asking a question), it is not necessary to include a "to do" or "to be" forms
when creating questions or making negative statements with modal verbs. For example: "He will not." “Can she go?” rather than “*Is she can go?” and “*He doesn’t will.” (Hykes, 2000: 7).

3. Huddleston and Pullum (2002: 107) note that modal auxiliary verbs are the verb class, which does not display the usual person-number agreement with the subject in the present tense (e.g. There must be a school on the island. / *There hope to be a school on the island).

4. Negative forms of modal verbs can be contracted (e.g. I can’t write this vs. I cannot write this).

5. Modals can be used as "code verbs" in ellipsis51 (Collins, 2009). “This means that when two equivalent clauses using a modal verb occur together, the second clause does not need to repeat the accompanying lexical verb, as in, "I can go. So can she." In this case the second modal verb can represents the lexical verb go” Hykes, 2000: 7).

6. Modals do not require a third person singular -s (e.g. *52She cans).

7. Modals are not found in imperative clauses.53

8. Modals do not combine with each other and do not co-occur in a clause; except in some nonstandard dialects; (e.g. She might could go).

9. Inversion of subject and auxiliary (e.g. Will he study? *Likes he study? (Collins, 2009: 12).

10. Emphasis: “emphatic polarity involving the use of contrastive stress” (Collins, 2009: 12). For example; They think he won’t study, but he will.

11. “The preterits54 of the modal auxiliaries- could, might, would, should- can be used with the modal remoteness meaning without the grammatical limitations that

---

51 “Ellipsis is the non-expression of a word or phrase that is, nevertheless, expected to occupy a place in the syntactic structure of a sentence” (McShane, 2005: 3).
52 The asterisk outlines that an utterance is ungrammatical.
53 “The imperative clause is a basic sentence type along with the declarative and interrogative. Each type differs in its communicative function (orders, statements, questions) and often displays salient morphosyntactic differences” (Alcázar & Saltarelli, 2014: 2).
54 Richards and Schmidt (2010) define preterits as the simple past tense.
apply in the case of other verbs, where it is found only in a small set of subordinate constructions” (Huddleston & Pullum (2002: 107). [e.g. Could you move it?]

The next section discusses the different functions of modal verbs, together with examples.

6.3 Functions of modal verbs

It is important to explore the different functions of modal verbs because this will inform the analysis of the modal verbs in ZE and BrE. In addition, examples of the functions of modal verbs in the ZE corpus and the ICE-GB (section 6.6.2) and the frequencies of functions of the nine modal verbs examined in this study (section 6.6.1.4) are based on the discussion in this section.

The functions of modal verbs can be described in terms of modality. In literature, the terms “meanings of modal verbs” and “functions of modal verbs” are used to refer to the same concept. Some researchers refer to meanings of modal verbs (e.g. Hinkel, 1995; Salkie, 2009; Collins, 2009; Richards & Schmidt, 2010; Myllyniemi, 2015) whilst other researchers refer to the functions of modal verbs (e.g. Yasumasa, 2010; Gustilo, 2011; Sadia & Ghani, 2019). For example, Hinkel (1995) notes that must and should denote obligation and necessity and Quirk et al. (1985) consider can and could to indicate ability, permission and possibility. This shows the polysemous and overlapping functions served by the nine modal verbs examined in this study, as highlighted in the table below.
Table 33. Modal verbs and their functions (Sources: Quirk et al. 1985; Swan, 1995; Collins, 2009)

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>Function(s)</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Can        | Ability, Permission, Possibility | *I can* read Italian, but I can’t speak it.  
Can I ask you something?  
Anybody who wants to *can* join the club. |
| Could      | Ability, Possibility, Permission | She *could* read when she was four.  
It was a place where anything *could* happen.  
*He could* borrow my car if he asked. |
| May        | Possibility, Permission | You *may* be right.  
You *may* use my desk. |
| Might      | Ability, Possibility, Permission | Frantic now, not thinking how he *might* control his rage, he ran from room to room.  
*Peter might* phone.  
*Might* I ask whether you are using a typewriter? |
| Must       | Necessity, Obligation | There *must* be some mistake.  
*We must* all share our skills and knowledge. |
| Shall      | Prediction, Volition | When *shall* we know the results of the election?  
*WE shall* uphold the wishes of the people. |
| Should     | Obligation, Necessity | People *should* drive more carefully.  
Oh then maybe you *should* go.  
*He should* be arriving any minute. |
| Will/’ll   | Prediction, Volition | *Oil will* float on water.  
*I’ll* write as soon as I can. |
| Would/’d   | Prediction, Volition | *I was told I would* feel better after this medicine.  
*She would* keep interrupting me. |

Regarding modality, Collins (2009) and Wasserman (2014) classify modals according to their meanings, namely epistemic, deontic and dynamic modality. “Epistemic modality is prototypically concerned with the speaker’s attitude towards the factuality of the situation, the speaker’s judgement of the likelihood that the proposition on which the
utterance is based is true, located along a scale ranging from weak possibility (it may be the case) to strong necessity” (Collins, 2009: 21).

The sentences below make determinations about the possibility and necessity, the likelihood of the accuracy of the statements (Palmer, 1990). From examples 6.1 and 6.2 below, the epistemic modal must is making judgements about the probability of John being there now.

6.1 John may be there now.
6.2 John must be there now

(Palmer, 1990: 5)

From examples 6.1 and 6.2 the epistemic modals may and must are making judgements about the probability of John being there now.

Obligation and permission are the meanings of deontic modality and “with deontic modality the conditioning factors are external to the relevant individual” (Palmer, 2001:9). Examples given are as follows:

6.3 John may/can come in now. [permission]
6.4 John must come in now. [obligation]

(Palmer, 1990: 9)

Palmer explains that, in the sentences above, the action of John coming is influenced by him getting permission or by him being obliged to come. Deontic modality is “concerned with influencing actions, states, or events” (Palmer, 1990: 6).

According to Wasserman (2014: 138), the “deontic meanings of obligation and permission are aligned with social functions: the possibility or necessity of acts performed” as shown in the sentence; I must go home now (Mum says so.), whilst epistemic meaning is linked to the expression of beliefs and knowledge as in the sentence: They might be in the cinema (They talked about going this afternoon).

The semantic meanings of modal verbs such as ‘necessity’, ‘obligation’, ‘permission’ and ‘possibility’ are associated with modal verbs (Collins, 2009). These meanings can overlap,
such that one modal verb can have two or more meanings (Quirk et al., 1985; Collins, 2009; Wasserman, 2014). An example of the overlap in meaning is provided by Quirk et al. (1985), who categorise modal verbs according to the functions they serve by showing that they can serve more than one function as shown below:

a) 'Permission', 'ability', 'possibility': *Can, might, could, may.*

b) 'Necessity', 'obligation': *Should, must.*

c) 'Prediction', 'volition': *Shall, would, will.*

The next section will explore some of the studies done on modal verbs in L2 varieties of English, with the goal of highlighting the different focus areas that have been explored.

### 6.4 Studies of modals in L2 and New Englishes

One of the areas that has garnered a lot of attention in new varieties of English has been to establish how modal verbs are used in different varieties of English and to determine whether there is any variation in the use of modal verbs (e.g. Hinkel, 1995; Coates & Leech, 1980; Gustilo, 2011). Some studies have shown that modal verbs are used in different contexts in L2 varieties of English (Hinkel, 1995). Hinkel comments that one of the reasons for the variation in use is that modal verb usage in L2 varieties of English, shows the pragmatic framework norms that are available in speakers’ L1 environments (Hinkel, 1995). Some of the studies on modal verbs in L2 varieties of English are highlighted below.

Within the South African context, Van Rooy and Wasserman (2014) did a comparative study of modal verbs in BSAE (an L2 English variety) and WSAE (a L1 English variety) to ascertain whether there is convergence in the use of modals and quasi-modals. The frequencies of modals and quasi-modals in BSAE are reported to be increasing whilst the frequencies of modals and quasi-modals in WSAE are reported to be decreasing. Van Rooy and Wasserman (2014) also noted similarities in the use of *must* in BSAE and WSAE. A semantic comparison done by the authors revealed that there is a small change in the

---

55 Quasi-modals are “periphrastic modal forms, a somewhat loosely-defined grouping, formally distinguishable from, but semantically similar to the modal auxiliaries” (Collins, 2009: 15).
use of *must* and *should* in BSAE, signalling that there is no convergence between BSAE and WSAE.

Gustilo (2011) examined the semantic functions of nine modal verbs (*will*'ll, *can*, *would*'d, *should*, *may*, *could*, *must*, *might* and *shall*) in five text categories in newspapers in the Philippine component of ICE. This was done in order to determine whether there is variation in the use of modal verbs in Philippine English. Gustilo (2011) concluded that, except for *shall* and *would*, Philippine English speakers use modal verbs in the same manner that L1 speakers of English do.

Hinkel (1995) compared the use of modal verbs by native and non-native speakers of English in essays. Findings from the study suggest that the use of modals of obligation and necessity (*must, have to, should, ought to and need to*) in native and non-native writing is influenced by culture and context.

Ali’s (2013) corpus-based study on how modal verbs (*can, could*) are used in Pakistani English established that there is innovation in the use of *can* and *could* by Pakistani English users. The investigation was a comparative one between Pakistani English fiction and BrE fiction. When considering the modal verb *could*, Ali (2013) observed that in Pakistan English, *could* expresses surprise whilst in BrE *could* is used to show politeness.

From the aforementioned studies, research into the use of modals in L2 varieties of English shows some variations in the use of some modals; such as *can* and *could* in Pakistani English (Ali, 2013). Research also shows similarities in the way modal verbs are used in L1 and L2 varieties of English; for example, *will*'ll, *can, should, may, could, must, might* and *might*).

In the next section, a description of the three features examined in this study, that are reported to show variation in Kortmann, Lunkenheimer and Ehret’s (2020) list of features. The discussion also reports on the categories assigned to the features in different varieties of English.
6.5 Features that show variation in varieties of English

Kortmann, Lunkenheimer and Ehret (2020) compiled a list of features that show variations in the use of modal verbs in different varieties of English (cf. Sections 1.6 and 2.5). The list and examples of features that show variation were provided by different scholars.

In order to establish whether there is variation in the use of modal verbs in ZE, the analysis was guided by the possible variations (cf. Sections 1.6 and 2.5) reported by different researchers in Kortmann, Lunkenheimer and Ehret (2020). As was done with articles, the features with the highest occurrence are classified as pervasive or obligatory (level A). Features that are neither pervasive nor extremely rare are classified as belonging to level B. Features that exist, but are extremely rare and features that are attested to be absent are classified into levels C and D respectively.

For this study, 3 features, features 121, 123 and 127, were investigated in order to determine whether there is variation. The table below highlights the variations in the use of modal verbs (and examples), that were examined in this study.

*Table 34. Variations with modal verbs reported in Kortmann, Lunkenheimer & Ehret (2020)*

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Description of feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>double modals</td>
<td><em>We might could do that.</em></td>
</tr>
</tbody>
</table>
| 123            | Present tense forms of modals used where StE has past tense forms | *They can (might) be wild, but they're human beings.*  
*I wish that people in the world *will (would) get educated.*  
*I knew the car *will (would) be there.* |
| 127            | Non-standard use of modals for politeness reasons | *I would be visiting your place tomorrow.*  
*Must I give you some water?*  
*I wouldn't (don't) know.*  
*This work *may (will) be finished tomorrow* |
When considering the 4 categories proposed by Kortmann, Lunkenheimer and Ehret (2020) for the features in table 34, different L2 varieties are reported to belong to different categories. This is shown in the table below.

*Table 35. Summary of how features are attested in different L2 varieties*

<table>
<thead>
<tr>
<th>Feature number</th>
<th>Values and examples of varieties in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Pervasive or obligatory (A)</strong></td>
</tr>
<tr>
<td>121</td>
<td>Chikano English</td>
</tr>
<tr>
<td>123</td>
<td>Hong Kong English</td>
</tr>
<tr>
<td>127</td>
<td>BSAE, Hong Kong English</td>
</tr>
</tbody>
</table>

The table above shows that feature 121 is reported to be neither pervasive nor extremely rare in Chikano English, an L2 variety of English (Bayley, 2020). In the eWAVE, the use of double modals is attested to be absent in African L2 varieties of English such as BSAE (Mesthrie, 2020a), Ugandan English (Ssempuuma, 2020), Kenyan English (Buregeya, 2020), Ghanaian English (Huber, 2020) and Nigerian English (Taiwo, 2020). Data analysis in Section 6.6.4.1 will ascertain whether ZE conforms to StE or if double modals are used in ZE. One of the ways in which new varieties of English show variation is the use of present tense forms of modals where StE has past tense. This feature is believed to be attested in some L2 varieties of English. Mesthrie (2020a) notes that in BSAE, feature 123 is neither pervasive nor extremely rare. Another variety where present tense forms of the modals are used instead of past tense forms is attested is Indian South African.
Feature 123 is reported to be neither pervasive nor extremely rare. In Sri Lankan English, a L2 variety of English, feature 123 is also neither pervasive nor extremely rare (Meyler, 2020).

In addition to the abovementioned studies, other studies have also reported on the use of present tense forms of modals where StE has past tense forms. The varieties studied include Ugandan English (Ssempuuma, 2020), Kenyan English (Buregeya, 2020), Hong Kong English (Wong, 2020) and Pakistani English (Mahboob, 2020).

Researchers have explored whether there is non-standard use of modals for politeness reasons in different varieties of English as shown on the eWAVE. Some of the L2 varieties of English in which feature 127 is attested are given below together with examples.

Sharma (2020) notes that feature 127 is attested to be neither pervasive nor extremely rare and adds that Indian English speakers use modals in a non-standard way for politeness reasons. According to Mesthrie (2020a), one feature that shows innovation in BSAE is the non-standard use of modals for politeness reasons because the feature is pervasive or obligatory. Hong Kong English is another variety, which shows feature 127 as being pervasive or obligatory, meaning that there is non-standard use of modals for politeness reasons in Hong Kong English as shown in the next sentence (Wong, 2020).

In Indian South African English (Mesthrie, 2020b), Pakistani English (Mahboob, 2020) and Ugandan English (Ssempuuma, 2020), feature 127 is attested to be neither pervasive nor extremely rare. Other L2 varieties of English, which show that feature 127 exists but is extremely rare include Cape Flats English (Mesthrie, Toefy & Bowerman, 2020), Jamaican English (Sand, 2020) and Sri Lankan English (Meyler, 2020).

Results from the data analysis are reported next.

### 6.6 Results

This section is divided into four sub-sections. Section 6.6.1.1 reports on the quantitative data in terms of the occurrence- and function frequencies (with reference to the overall absolute and normalised frequencies) of modal verbs. The absolute and normalised
frequency of the modal verbs in the spoken and written registers are discussed in Section 6.6.1.2. In section 6.6.1.3, the absolute and normalised frequencies in genres are reported. Example sentences of the different functions of modal verbs from the ZE corpus and the ICE-GB are given in Section 6.6.2. The frequencies of these functions are reported in Section 6.6.1.4.

In order to examine the frequency and functions of modal verbs in ZE, a concordance list was created to show the collocations of the 9 modal verbs examined in this study. All modal verbs appearing in the corpus were checked for frequency in the wordlist tool. Word sketch tools in Sketch Engine were used to determine collocations and concordance lists. Using the concordance tool enabled the researcher to get a glimpse of how a word behaves in the context in which it is used. This enabled the researcher to check the functions of modal verbs in the contexts in which they were used. As was done in chapter 5, one asterisk (*) was used for \( p < 0.05 \), where the LL > 3.84. Two asterisks (**) were used to show that the LL > 6.63, where \( p < 0.01 \). Three asterisks (****) were used for \( p < 0.001 \), where the LL > 10.83 and four asterisks (****) were used for \( p < 0.0001 \), where LL > 15.13.

The concordance lists were also checked to determine whether the use of modal verbs shows variation in ZE.

Corpus linguistics methodology, which involves the use of both quantitative and qualitative functional analyses was utilised. The approach is particularly appropriate because this study describes how specific linguistic features are used and examines whether there is innovation in the use of specific features in ZE so as to determine the stage of ZE in the DyM of postcolonial English.

### 6.6.1 Quantitative data analysis of modal verbs

Section 6.6.1 answers part of research questions 1.1, 1.2, 2.1 and 2.2 by focusing on the occurrence- and function frequencies of modal verbs and determining whether there are variations in the use of modal verbs. The research questions are shown for ease of reference below.
1.1 What are the occurrence frequencies of modal verbs in the ZE corpus and the ICE-GB?

1.2 How does the occurrence of modal verbs in ZE compare to that of BrE?

2.1 What are the function frequencies of modal verbs in the ZE corpus and the ICE-GB?

2.2 How do the functions of modal verbs in ZE compare to those of BrE?

### 6.6.1.1 Occurrence frequencies of the nine modal verbs

As a way of comparing between ZE and BrE, the number of occurrences of the modal verbs was normalised by dividing the total number of occurrences in a corpus by the total number of words in the corpus and then multiplied by 10,000 as shown below.

\[
\text{Total number of each modal verb in a corpus} \times 10\,000
\]

\[
\text{Total number of words in a corpus}
\]

Following Collins (2009) and Gustilo’s (2011) assertion that the modal verb *will* can be contracted to *’ll* and *would* can be reduced to *’d*, respectively, the reduced forms *’ll* and *’d* were analysed as part of their unreduced forms as shown in the table below.

#### Table 36. Summary and frequencies of modal verbs per 10,000 words

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10,000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Can</td>
<td>1 360</td>
<td>38.2</td>
<td>1529</td>
</tr>
<tr>
<td>Could</td>
<td>406</td>
<td>11.4</td>
<td>946</td>
</tr>
<tr>
<td>May</td>
<td>138</td>
<td>3.9</td>
<td>426</td>
</tr>
<tr>
<td>Might</td>
<td>108</td>
<td>3</td>
<td>495</td>
</tr>
<tr>
<td>Must</td>
<td>200</td>
<td>5.6</td>
<td>361</td>
</tr>
<tr>
<td>Shall</td>
<td>80</td>
<td>2.2</td>
<td>115</td>
</tr>
<tr>
<td>Should</td>
<td>396</td>
<td>11.1</td>
<td>679</td>
</tr>
<tr>
<td>Will/’ll</td>
<td>1 638</td>
<td>46</td>
<td>2275</td>
</tr>
<tr>
<td>Would/’d</td>
<td>744</td>
<td>20.9</td>
<td>2849</td>
</tr>
<tr>
<td>Total</td>
<td>5 070</td>
<td>142.3</td>
<td>9 675</td>
</tr>
</tbody>
</table>
Out of the nine modal verbs examined in this study, *would'd* was the most frequent modal verb, occurring 46.8 times per 10 000 words in the ICE-GB compared to 20.9 times per 10 000 words in the ZE corpus. This difference is statistically significant because the LL value of 442.96 is higher than the critical value of 15.13, where $p < 0.0001$. This means that there are variations in the frequency of *would'd* in the two corpora. *Can* and *will'll* were more frequent in the ZE corpus (with 38.2 and 46 instances per 10 000 words respectively) compared to 25.1 and 37.4 instances per 10 000 words respectively in the ICE-GB. This shows that there are variations in the frequency of *can* and *will'll*, with the ZE corpus recording more instances of the two modals. Variations were also observed because *could*, *may* and *might* were more frequent in the ICE-GB with normalised frequencies of 15.6, 7 and 8.1 respectively. In contrast, the occurrences of *could*, *may* and *might* in the ZE corpus are 11.4 times, 3.9 times and 3 times per 10 000 words respectively. The three modal verbs have LL values that are above the critical value of 15.13 where $p < 0.0001$. No statistically significant variations were recorded in the frequencies per 10 000 words of *must*, *shall* and *should* because their LL values were below the critical value of 3.84 where $p < 0.05$, meaning that there are no variations in the frequencies of *must*, *shall* and *should* between the two corpora.

### 6.6.1.2 Occurrence frequencies of all nine modal verbs across registers for both corpora with Log likelihood comparison

The normalised frequencies for registers were calculated by dividing the total number of each modal verb in a register by the total number of words in a register and then multiplying by 10000, as shown below.

\[
\text{Total number of each modal verb in a register} \times 10000
\]

Total number of words in a register

In order to make a comparative analysis of the frequencies of modal verbs in the ZE corpus and the ICE-GB, the tagset for modal verbs (MD) was searched in the wordlist tool in Sketch Engine. For the ICE-GB, the tagset AUX was searched. An analysis was done to determine the frequency of modal verbs per genre as illustrated in the table below.
Table 37. Frequency of modal verbs in spoken and written registers

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>ZE corpus register</th>
<th>ICE-GB registers</th>
<th>Log likelihood</th>
<th>Log likelihood for spoken register</th>
<th>Log likelihood for written register</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spoken absolute frequency</td>
<td>Written absolute frequency</td>
<td>Total</td>
<td>Spoken normalised frequency/10 000 words</td>
<td>Written normalised frequency/10 000 words</td>
</tr>
<tr>
<td>Can</td>
<td>1074</td>
<td>286</td>
<td>1360</td>
<td>45.7</td>
<td>23.6</td>
</tr>
<tr>
<td>Could</td>
<td>273</td>
<td>133</td>
<td>406</td>
<td>11.6</td>
<td>11</td>
</tr>
<tr>
<td>May</td>
<td>103</td>
<td>35</td>
<td>138</td>
<td>4.4</td>
<td>2.9</td>
</tr>
<tr>
<td>Might</td>
<td>77</td>
<td>31</td>
<td>108</td>
<td>3.3</td>
<td>2.6</td>
</tr>
<tr>
<td>Must</td>
<td>109</td>
<td>91</td>
<td>200</td>
<td>4.6</td>
<td>7.5</td>
</tr>
<tr>
<td>Shall</td>
<td>20</td>
<td>60</td>
<td>80</td>
<td>0.9</td>
<td>5</td>
</tr>
<tr>
<td>Should</td>
<td>220</td>
<td>176</td>
<td>396</td>
<td>9.4</td>
<td>14.5</td>
</tr>
<tr>
<td>Will/’ll</td>
<td>1032</td>
<td>606</td>
<td>1 638</td>
<td>43.9</td>
<td>50</td>
</tr>
<tr>
<td>Would/’d</td>
<td>498</td>
<td>246</td>
<td>744</td>
<td>21.2</td>
<td>20.3</td>
</tr>
</tbody>
</table>
Statistically significant differences can be observed in the uses of *can* and *will* in the spoken register, with both modal verbs registering LL values that are above the critical value of 15.13 where \( p < 0.0001 \). *Can* was attested 45.7 times per 10 000 words whilst *will* was attested 43.9 times per 10 000 words in the ZE corpus. In contrast, *can* and *will* occurred less frequently in the ICE-GB (27 times and 32.5 times per 10 000 words respectively). If we consider the LL values of *can* and *will* in the spoken register, the numbers are 147.13 and 51.64 respectively, which is well above the critical value of 15.13 where \( p < 0.0001 \). This shows that there are variations in the frequency of *can* and *will* in the spoken register. Considering the modal verbs *could, may, might, shall and would*, they all are attested more frequently per 10 000 words in the spoken register, the ICE-GB compared to the ZE corpus. Statistically significant differences are also evident in the written register, with *would* having the highest LL value of 122.94 and occurring 42.9 times per 10 000 words compared to 20.3 times per 10 000 words in the ZE corpus. This means that there are variations in the frequency of *could, might, shall and would* in the spoken register. Other modal verbs that show statistically significant variations in the written register include *could, may and might*. These three modal verbs occurred more in the ICE-GB, with *could* occurring 16.5 times per 10 000 words with a LL value of 16.88, *may* occurring 9.4 times per 10 000 words with a LL value of 52.90, whilst *might* was attested 9.5 times per 10 000 words with a LL value of 62.25. The LL values are higher than the critical value, which is 15.13 where \( p < 0.0001 \). In comparison, *could, may* and *might* appeared less frequently per 10 000 words in the written register in the ZE corpus. *Could* was attested 11 times per 10 000 words whilst the numbers for *may* and *might* are 2.9 and 2.6 respectively, meaning that there are variations in the frequencies of these modal verbs, with *could, may and might* occurring more in the written register in the ICE-GB compared to the ZE corpus.

This section has shown that there are statistically significant variations in the normalised frequencies of the modal verbs *can* and *will* in the spoken register. It was observed that the two modal verbs were attested more in the ZE corpus compared to the ICE-GB. Results have also shown that *could, might, shall and would* were more frequently in the spoken register, in the ICE-GB compared to the ZE corpus. Regarding the written register,
this section has shown that there are statistically significant variations in the frequency of *would, could, may* and *might* in the ICE-GB compared to the ZE corpus.

### 6.6.1.3 Occurrence frequencies of all the nine modal verbs across genres for both corpora with Log likelihood comparison

In this section, the absolute and normalised frequencies of the nine modal verbs in different genres were examined. In addition, the LL values were calculated in order to determine whether the observed variations were statistically significant, meaning that they were not due to chance. The number of occurrences of the modal verbs in a genre were normalised by dividing the total number of occurrences of each modal verb in a genre by the total number of words in the genre and then multiplied by 10 000 as shown below.

\[
\text{Total number of each modal verb in a genre} \times 10^4
\]

\[
\frac{\text{Total number of words in a genre}}{\text{Total number of each modal verb in a genre}} \times 10^4
\]

**a. Occurrence frequencies of *can* across genres for both corpora with Log likelihood comparison**

This section will provide the analysis of the occurrence frequencies of the modal verb *can* for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:
Table 38. Frequency of can across genres

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>702</td>
<td>53.2</td>
<td>562</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>11</td>
<td>26.2</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>361</td>
<td>36.6</td>
<td>455</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>23</td>
<td>12.9</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>2</td>
<td>7.2</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>8</td>
<td>38.5</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>58</td>
<td>22.0</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>67</td>
<td>13.7</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>41</td>
<td>81.9</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>87</td>
<td>48.1</td>
<td>71</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1360</td>
<td>38.2</td>
<td>1529</td>
</tr>
</tbody>
</table>

The table above shows that can was more frequent in the ZE corpus in private dialogue (53.2 instances per 10 000 words compared to 30.3 in the ICE-GB), public dialogue (36.6 instances per 10 000 words compared to 26.6 in the ICE-GB), social letters (81.9 instances per 10 000 words compared to 27.3 in the ICE-GB) and business letters (48.1 instances per 10 000 words compared to 23.3 in the ICE-GB). The observed variations are statistically significant because the LL values in the four genres were above 15.13, which is a critical value where $p < 0.0001$. The ICE-GB recorded higher instances of can in the popular writing genre (46.2 per 10 000 words) compared to the ZE corpus (7.2 per 10 000 words). The observed variations are statistically significant because the LL of 13.02 > 10.83, which is a critical value for $p < 0.001$. There were no statistically significant variations between the two corpora in public scripted monologue, creative writing and academic writing, editorials and newspaper reportage because the LL values were lower than 3.84, which is a critical value for $p < 0.05$. © University of Pretoria
b. Occurrence frequencies of *could* across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb *could* for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>194</td>
<td>14.7</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>2</td>
<td>4.8</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>75</td>
<td>7.6</td>
<td>272</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>44</td>
<td>24.6</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>4</td>
<td>14.5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>7</td>
<td>33.7</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>26</td>
<td>9.9</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>45</td>
<td>9.2</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>5</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>4</td>
<td>2.2</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>406</td>
<td>11.4</td>
<td>946</td>
</tr>
</tbody>
</table>

Comparison of the ZE corpus and the ICE-GB revealed that there were statistically significant variations in the normalised frequencies of *could* in some of the genres examined. The normalised frequency (per 10 000 words) of *could* was higher in the ICE-GB in public scripted monologue (17 compared to 4.8 in the ZE corpus), public dialogue (15.9 compared to 7.6 in the ZE corpus), editorials (20.3 compared to 9.9 in the ZE corpus), newspaper reportage (17.6 compared 9.2 in the ZE corpus) and business letters (21.3 compared to 2.2 in the ZE corpus). These variations are statistically significant because the LL values were higher than the lowest LL value of 3.84, which is a critical value where $p < 0.05$. Regarding academic writing, *could* was more prevalent in the ZE corpus, with 33.7 instances per 10 000 words compared to 5.2 for the ICE-GB. No
statistically significant variations were reported in private dialogue, creative writing, popular writing and social letters because the LL values < 3.84 where p < 0.05.

c. Occurrence frequencies of *may* across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb *may* for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>36</td>
<td>2.7</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>2</td>
<td>4.8</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>67</td>
<td>6.8</td>
<td>136</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>5</td>
<td>2.8</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>8</td>
<td>38.5</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>2</td>
<td>0.8</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>9</td>
<td>1.8</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>6</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>3</td>
<td>1.7</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>138</td>
<td>3.9</td>
<td>426</td>
</tr>
</tbody>
</table>

In the table above, statistically significant variations were recorded in four genres. The LL value for editorials (35.07) showed that *may* was more frequent in the ICE-GB, with 14 instances per 10 000 words compared to 0.8 for the ZE corpus. When considering the newspaper reportage genre, the LL value of 30.09 showed that *may* was more frequent in the ICE-GB, with 10.4 instances per 10 000 words compared to 1.8 instances in the ZE corpus. The normalised frequencies in popular writing and business letters were higher in the ICE-GB (8.5 and 9.5 respectively) compared to the ZE corpus (0 and 1.7
respectively). The LL values in private dialogue, public scripted monologue, public
dialogue, creative writing, academic writing and social letters were lower than the critical
value of 3.84, where \( p < 0.05 \), meaning that there were no statistically significant
variations in these genres.

d. Occurrence frequencies of \textit{might} across genres for both corpora with
Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb
\textit{might} for both the ZE corpus and the ICE-GB. The frequency results in each corpus are
then compared with the Log Likelihood test. For a summary of these statistics please see
the table below:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
\textbf{Register} & \textbf{Genre} & \textbf{ZE corpus} & & \textbf{ICE-GB} & \\
 & & \textbf{Abs. frequency} & \textbf{Norm. frequency/10 000 words} & \textbf{Abs. frequency} & \textbf{Norm. frequency/10 000 words} & \textbf{Log likelihood} \\
\hline
\multirow{3}{*}{Spoken} & Private dialogue & 35 & 2.7 & 154 & 8.3 & -45.95**** \\
& Public scripted & 1 & 2.4 & 16 & 3.7 & -0.21 \\
& monologue & & & 126 & 7.4 & -11.11*** \\
& Public dialogue & 41 & 4.2 & & & \\
\hline
\multirow{8}{*}{Written} & Creative writing & 5 & 2.8 & 77 & 18.1 & -28.48**** \\
& Popular writing & 1 & 3.6 & 5 & 2.4 & +0.14 \\
& Academic writing: & 2 & 9.6 & 7 & 3.3 & +1.44 \\
& Examination & & & & & \\
& Editorials & 7 & 2.7 & 38 & 18.3 & -31.57**** \\
& Newspaper & 10 & 2 & 29 & 7 & -13.04*** \\
& reportage & & & & & \\
& Social letters & 6 & 12 & 23 & 7.4 & +1.00 \\
& Business letters & 0 & 0 & 20 & 6.6 & -18.62**** \\
\hline
\textbf{Total} & & 108 & 3 & 495 & 8.1 & -104.53**** \\
\hline
\end{tabular}
\end{table}

For the modal verb \textit{might}, statistically significant variations were recorded in six out of
the ten genres analysed. In the six genres, \textit{might} was more frequent in the ICE-GB
compared to the ZE corpus as will be shown by the normalised frequencies per 10 000
words: In private dialogue, the ICE-GB had 8.3 instances compared to 2.7 for the ZE
corpus, in public dialogue, the ICE-GB had 7.4 instances compared to 4.2 in the ZE corpus
and in creative writing, the ICE-GB had 18.1 instances compared to 2.8 in the ZE corpus.

© University of Pretoria
The other three genres are editorials, with 18.3 instances compared to 2.7 instances in the ZE corpus, newspaper reportage, with 7 instances compared to 2 in the ZE corpus and business letters, with 6.6 instances compared to 0 in the ZE corpus. No statistically significant variations were reported in four genres namely public scripted monologue, popular writing, academic writing and social letters due to the fact that the LL > 3.84 where p < 0.05.

e. Occurrence frequencies of must across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb must for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>28</td>
<td>2.1</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>Public scripted</td>
<td>17</td>
<td>40.5</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>monologue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>63</td>
<td>6.4</td>
<td>89</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>5</td>
<td>2.8</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>0</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>44</td>
<td>16.7</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>36</td>
<td>7.4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>7</td>
<td>3.9</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>200</td>
<td>5.6</td>
<td>361</td>
</tr>
</tbody>
</table>

When the normalised frequency per 10 000 words in the ZE corpus and the ICE-GB were compared, it was observed that must was more prevalent in the ICE-GB in private dialogue, with 4.7 instances compared to 2.1 in the ZE corpus, social letters, with 8.4
instances compared to 0 in the ZE corpus, popular writing, with 14.6 compared to 0 in the ZE corpus and creative writing, with 10.6 instances compared to 2.8 in the ZE corpus. Must was attested more in the ZE corpus in public scripted monologue, with 40.5 instances compared to 5.3 in the ICE-GB and in newspaper reportage, with 7.4 instances compared to 2.6 in the ICE-GB. No statistically significant variations were recorded in public dialogue, academic writing, editorials and business letters.

f. Occurrence frequencies of shall across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb shall for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

Table 43. Frequency of shall across genres

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>3</td>
<td>0.2</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Public scripted</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>monologue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>17</td>
<td>1.7</td>
<td>30</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>1</td>
<td>0.6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>2</td>
<td>0.8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Newspaper</td>
<td>4</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>reportage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>4</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>49</td>
<td>27.1</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>80</td>
<td>2.2</td>
<td>115</td>
</tr>
</tbody>
</table>

In the table above, shall was attested more in the ZE corpus in newspaper reportage, with 0.8 instances per 10 000 words compared to 0 in the ICE-GB and in business letters, with 27.1 instances per 10 000 words compared to 5.6 in the ICE-GB. These variations are statistically significant because, for newspaper reportage, the LL value of 4.92 is higher.
than the critical value of 3.84, where \( p < 0.05 \) and for business letters, the LL value of 37.37 is higher than 15.13 where \( p < 0.0001 \). Regarding the private dialogue genre, the normalised frequency of *shall* was higher in the ICE-GB (2.5 compared to 0.2 in the ZE corpus). This variation is statistically significant because the LL value of 33.13 > 15.13, which is a critical value for \( p < 0.0001 \). There were no statistically significant variations in the frequency of *shall* in seven genres namely public scripted monologue, public dialogue, creative writing, popular writing, academic writing, editorials and social letters because the LL values were lower than 3.84 where \( p < 0.05 \).

### g. Occurrence frequencies of *should* across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb *should* for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

**Table 44. Frequency of *should* across genres**

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Absolute frequency</strong></td>
<td><strong>Normalised frequency/10 000 words</strong></td>
<td><strong>Absolute frequency</strong></td>
<td><strong>Normalised frequency/10 000 words</strong></td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>123</td>
<td>9.3</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>6</td>
<td>14.3</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>91</td>
<td>9.2</td>
<td>181</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total</strong></td>
<td><strong>396</strong></td>
<td><strong>679</strong></td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>7</td>
<td>3.9</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>2</td>
<td>7.2</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>12</td>
<td>57.8</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>95</td>
<td>36.1</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>52</td>
<td>10.6</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>1</td>
<td>2.0</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>7</td>
<td>3.9</td>
<td>72</td>
</tr>
</tbody>
</table>

Regarding *should*, the normalised frequency per 10 000 words were higher in the ICE-GB compared to the ZE corpus in popular writing (35.9 compared to 7.2) and business letters.
(23.6 compared to 3.9). Shall was attested more in the ZE corpus in academic writing, with 57.8 instances per 10 000 words compared to 3.3 in the ICE-GB) and editorials, with 36.1 instances per 10 000 words compared to 24.6 in the ICE-GB. There were no statistically significant variations in the frequency of should in private dialogue, public scripted monologue, public dialogue, creative writing, newspaper reportage and social letters. This is because the LL values were lower than the critical value of 3.84, where \( p < 0.05 \).

h. Occurrence frequencies of will across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb will for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

*Table 45. Frequency of will/’ll across genres*

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>616</td>
<td>46.7</td>
<td>540</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>18</td>
<td>42.9</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>398</td>
<td>40.3</td>
<td>551</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>24</td>
<td>13.4</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>6</td>
<td>21.7</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>Academic writing: Examination</td>
<td>4</td>
<td>19.3</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>94</td>
<td>35.7</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>174</td>
<td>35.6</td>
<td>269</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>61</td>
<td>121.8</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>243</td>
<td>134.4</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1 638</td>
<td>46</td>
<td>2 275</td>
</tr>
</tbody>
</table>

Out of the ten genres of the modal verb will/’ll analysed, there were statistically significant variations in eight genres. The normalised frequency per 10 000 words of will/’ll was higher in the ZE corpus in the following genres: private dialogue, with 46.7
compared to 29.2 in the ICE-GB, public dialogue, with 40.3 compared to 32.2 in the ICE-GB, social letters, with 121.8 compared to 30.6 in the ICE-GB and business letters, with 134.4 compared to 69.5 in the ICE-GB. The normalised frequency of will/’ll per 10 000 words was higher in the ICE-GB in the following genres: creative writing, with 23 compared to 13.4 in the ZE corpus, popular writing, with 52.4 compared to 21.7 in the ZE corpus, editorials, with 74.8 compared to 35.7 in the ZE corpus and newspaper reportage, with 64.8 compared to 35.6 in the ZE corpus. Due to the fact that the LL values in public scripted monologue and academic writing were lower than the critical value of 3.84, where p < 0.05, the variations recorded were not regarded as statistically significant.

i. Occurrence frequencies of would across genres for both corpora with Log likelihood comparison

This section will provide the analysis of the occurrence frequencies of the modal verb would for both the ZE corpus and the ICE-GB. The frequency results in each corpus are then compared with the Log Likelihood test. For a summary of these statistics please see the table below:

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus Absolute frequency</th>
<th>Normalised frequency/10 000 words</th>
<th>ICE-GB Absolute frequency</th>
<th>Normalised frequency/10 000 words</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>311</td>
<td>23.6</td>
<td>823</td>
<td>44.4</td>
<td>-98.61****</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>2</td>
<td>4.8</td>
<td>183</td>
<td>42.5</td>
<td>-21.61****</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>185</td>
<td>18.7</td>
<td>946</td>
<td>55.3</td>
<td>-26.30****</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>48</td>
<td>26.8</td>
<td>274</td>
<td>64.2</td>
<td>-37.81****</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>7</td>
<td>25.4</td>
<td>10</td>
<td>4.7</td>
<td>+9.67**</td>
</tr>
<tr>
<td></td>
<td>Academic writing: Examination</td>
<td>6</td>
<td>28.9</td>
<td>30</td>
<td>14.1</td>
<td>+2.17</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>42</td>
<td>15.9</td>
<td>89</td>
<td>43</td>
<td>-30.41****</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>93</td>
<td>19</td>
<td>167</td>
<td>40.2</td>
<td>-35.23****</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>6</td>
<td>12</td>
<td>124</td>
<td>39.9</td>
<td>-12.12***</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>44</td>
<td>24.3</td>
<td>203</td>
<td>66.6</td>
<td>-44.52****</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>744</td>
<td>20.9</td>
<td>2849</td>
<td>46.8</td>
<td>-442.96****</td>
</tr>
</tbody>
</table>
The table above shows that there were statistically significant variations in the frequency of *would*/*d* in nine genres. The normalised frequency of *would*/*d* per 10 000 words was higher in the ZE corpus in popular writing, which had 25.4 instances compared to 4.7 instances in the ICE-GB. The normalised frequency of *would*/*d* per 10 000 words was higher in the ICE-GB in the following genres: Private dialogue had 44.4 instances compared to 23.6 instances in the ZE corpus, public scripted monologue had 42.5 instances compared to 4.8 instances in the ZE corpus, public dialogue had 55.3 instances compared to 18.7 instances in the ZE corpus, editorials had 43 instances compared to 15.9 instances in the ZE corpus. The other genres with higher normalised frequencies of *would*/*d* per 10 000 words include newspaper reportage, with 40.2 instances compared to 19 instances in the ZE corpus, social letters, with 39.9 instances compared to 12 instances in the ZE corpus and business letters, with 66.6 instances compared to 24.3 instances in the ZE corpus. No statistically significant variations were recorded in the academic writing genre because the LL values were lower than the critical value of 3.84, where \( p < 0.05 \).

### 6.6.1.4 Function frequency of modal verbs

In this section, the LL values, absolute and normalised frequencies of the different functions of modal verbs are given. This will help determine whether there are variations in the frequencies of functions. This section will have three subsections in which modal verbs are subdivided by their different functions. In sub-section 6.6.1.4a, the modals of obligation and necessity (*should* and *must*) are examined in terms of their frequencies in the ZE corpus and the ICE-GB. The LL values, absolute and normalised frequencies of the functions of modals possibility, permission and ability (*may, can, might* and *could*) are provided in sub-section 6.6.1.4b. The frequencies of functions of modals of prediction and volition (*will, would* and *shall*) are given in subsection 6.6.1.4c.

Regarding the functions of modal verbs, Collins (2009) and Wasserman’s (2014) classification of modals are used in this study. The modal verbs are grouped into three categories. The first category is necessity and obligation, the second category is possibility, permission and ability and the third category is prediction and volition.
Tables 45, 46 and 47 below illustrate the different functions of modal verbs as attested in the ZE corpus and the ICE-GB.

a. Function frequencies of modal verbs of necessity and obligation (should and must) for both corpora with Log likelihood comparison

The table below shows the frequencies of different functions of modals of obligation and necessity (should and must) in both the ZE corpus and the ICE-GB. This was done in order to determine whether there are variations in the frequencies of the functions in the ZE corpus and the ICE-GB.

Table 47. Frequencies of functions of should and must

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>Meaning</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Should</td>
<td>Deontic</td>
<td>281</td>
<td>7.9</td>
<td>469</td>
</tr>
<tr>
<td></td>
<td>Epistemic</td>
<td>37</td>
<td>1</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Subjunctive</td>
<td>28</td>
<td>0.8</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Should = would</td>
<td>8</td>
<td>0.2</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>42</td>
<td>1.2</td>
<td>66</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>396</td>
<td>11.1</td>
<td>679</td>
</tr>
<tr>
<td>Must</td>
<td>Deontic</td>
<td>118</td>
<td>3.3</td>
<td>209</td>
</tr>
<tr>
<td></td>
<td>Epistemic</td>
<td>64</td>
<td>1.8</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>11</td>
<td>0.3</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>7</td>
<td>0.2</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>200</td>
<td>5.6</td>
<td>361</td>
</tr>
</tbody>
</table>

Interestingly, the modals of necessity and obligation namely should and must do not show any statistically significant differences regarding their functions in ZE and BrE. Regarding research questions 1.3 and 2.3, which seeks to determine whether there are variations between ZE and BrE, this result means that there are no variations in the frequency of functions of should and must. This is evidenced by all the LL values, which are below the critical value of 3.84 where p < 0.05. To give an example, the deontic function of the modal verb should is attested 7.9 times per 10 000 words in the ZE corpus whilst the same function of should was attested 7.7 times per 10 000 words. This shows that the normalised frequencies are similar for the deontic function of should, therefore, there are no variations. Similar observations can be made for dynamic function of the modal verb must. The normalised frequency for the ZE corpus is 0.3 per 10 000 words, whilst that for
the ICE-GB is 0.4 per 10,000 words with a LL value of 0.31, which is well below the critical value of 3.84 where \( p < 0.05 \). This result means that there are no variations between ZE and BrE, in for dynamic function of the modal verb *must*.

b. **Function frequency of modal verbs of possibility, permission and ability (may, can, might and could) for both corpora with Log likelihood comparison**

Regarding the modals of possibility, permission and ability, the table below shows the frequencies of different meanings of *may*, *can*, *might* and *could*.

*Table 48. Frequencies of functions of modal verbs of possibility, permission and ability (may, can, might and could)*

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>Meaning</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td><strong>May</strong></td>
<td>Epistemic</td>
<td>89</td>
<td>2.5</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>18</td>
<td>0.5</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>23</td>
<td>0.6</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>8</td>
<td>0.2</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>138</td>
<td>3.8</td>
<td>426</td>
</tr>
<tr>
<td><strong>Can</strong></td>
<td>Epistemic</td>
<td>11</td>
<td>0.3</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>142</td>
<td>4</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>101</td>
<td>30.9</td>
<td>1238</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>106</td>
<td>3</td>
<td>123</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>1360</td>
<td>38</td>
<td>1529</td>
</tr>
<tr>
<td><strong>Might</strong></td>
<td>Epistemic</td>
<td>83</td>
<td>2.3</td>
<td>381</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>17</td>
<td>0.5</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>8</td>
<td>0.2</td>
<td>27</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>108</td>
<td>3</td>
<td>495</td>
</tr>
<tr>
<td><strong>Could</strong></td>
<td>Epistemic</td>
<td>53</td>
<td>1.5</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>12</td>
<td>0.3</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>312</td>
<td>8.8</td>
<td>732</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>29</td>
<td>0.8</td>
<td>52</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>406</td>
<td>11.4</td>
<td>946</td>
</tr>
</tbody>
</table>

Statistically significant differences can be observed in the epistemic functions of *may*. The LL value of 51.25 is higher than the critical value 15.13 where \( p < 0.0001 \). The epistemic function is more frequent in BrE (5.5 instances per 10,000 words) compared to ZE (2.5 instances per 10,000 words). This means that there are variations in the frequency of the epistemic functions of *may*. Regarding the modal verb *can*, there are variations because
the dynamic function was attested more in the ZE corpus, occurring 30.9 times per 10 000 words compared to 20.4 instances per 10 000 words in the ICE-GB. This variation is statistically significant because the LL value of 100.44 is above 15.13, which is the critical value for $p < 0.0001$. The deontic function of *can* was also attested more in the ZE corpus (4 times per 10 000 words) compared to the ICE-GB (2.5 times per 10 000 words), with a LL value of 16.22, meaning that there are variations in the frequency of the deontic function of *can*. The epistemic function occurred 6.3 times per 10 000 words in the ICE-GB whilst the number is 2.3 times per 10 000 words. This difference shows that there are variations because the LL value of 80.64 is higher than the critical value of 15.13 where $p < 0.0001$. The dynamic function of *might* was more frequent in the ICE-GB, occurring 1.4 times per 10 000 words compared to 0.5 times for the ZE corpus, with a statistically significant LL value of 19.74. This shows that there are variations in the frequency of the dynamic function of *might*, with the ICE-GB recording more functions than the ZE corpus. BrE speakers use the dynamic function of *could* more frequently (12 times per 10 000 words) compared to ZE speakers (8.8 times per 10 000 words), with a LL value of 22.89, which is higher than the critical value of 15.13 where $p < 0.0001$. There are no statistically significant differences in the deontic and indeterminate functions of *could*, meaning that there are no variations in the frequencies of these functions in the two corpora.

c. Function frequency of modal verbs of prediction and volition (*will*, *would* and *shall*) for both corpora with Log likelihood comparison

In this section, the LL values, absolute and normalised frequencies of the three modal verbs of prediction and volition (*will*, *would* and *shall*) are provided. In table 49 below, comparisons are made between the ZE corpus and the ICE-GB to determine whether there are variations.
### Table 49. Frequencies of functions of will, would and shall

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>Meaning</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
</tr>
<tr>
<td>Will</td>
<td>Epistemic</td>
<td>1 016</td>
<td>28.5</td>
<td>1 361</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>33</td>
<td>0.9</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>507</td>
<td>14.2</td>
<td>728</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>82</td>
<td>2.3</td>
<td>118</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1 638</td>
<td>46</td>
<td>2 275</td>
</tr>
<tr>
<td>Would</td>
<td>Epistemic</td>
<td>483</td>
<td>13.6</td>
<td>1 833</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>171</td>
<td>4.8</td>
<td>652</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>90</td>
<td>2.5</td>
<td>364</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>744</td>
<td>20.9</td>
<td>2 849</td>
</tr>
<tr>
<td>Shall</td>
<td>Epistemic</td>
<td>32</td>
<td>0.9</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Deontic</td>
<td>2</td>
<td>0.1</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Dynamic</td>
<td>45</td>
<td>1.3</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Indeterminate</td>
<td>1</td>
<td>0.02</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>80</td>
<td>2.32</td>
<td>115</td>
</tr>
</tbody>
</table>

Considering the epistemic functions of the modal verb *will*, the ZE corpus recorded more instances per 10 000 words (28.5) compared to the ICE-GB, which has a normalised frequency of 22.4 instances per 10 000 words. The observed differences are statistically significant because the LL value is 33.94, which is higher than the critical value of 15.13 where p < 0.0001. This means that there are variations in the frequency of the epistemic function, with ZE recording a higher frequency compared to BrE. With a normalised frequency of 14.2 instances per 10 000 words in the ZE corpus, the dynamic function of *will* occurs more in the ZE corpus compared to the ICE-GB (12 instances per 10 000 words). This observed difference is statistically significant because the LL value of 8.93 is higher than the critical value 6.63, where p < 0.001, meaning that there are variations in the frequency of the dynamic function of *will*. No statistically significant differences were recorded for the deontic and indeterminate functions of the modal verb *will*, meaning that there are no variations in the frequency of these functions.

All the functions of *would* were more frequent in the ICE-GB compared to the ZE corpus. The epistemic function of *would* occurred 30.1 times per 10 000 words in the ICE-GB whilst in the ZE corpus the number is 13.6 occurrences per 10 000 words. This variation in frequencies is statistically significant because the LL value of 280.02 is way higher than...
the critical value of 15.13 where \( p < 0.0001 \). With a statistically significant LL value of 100.53, the dynamic function of *would* occurred more frequently (10.7 times per 10 000 words) in the ICE-GB when compared to the ZE corpus (4.8 times per 10 000 words). The indeterminate function of *would* was more prevalent in the ICE-GB (6 times per 10 000 words) than in the ZE corpus (2.5 times per 10 000 words). Due to the fact that the LL value of 62.66 is above the critical value of 15.13 where \( p < 0.0001 \), the difference is statistically significant. Therefore, it can be concluded that there are variations in the frequencies of the epistemic function, the dynamic function and the indeterminate functions of *would*.

The epistemic function of *shall* was more prevalent in the ZE corpus (0.9 times per 10 000 words) than in the ICE-GB (0.2 times per 10 000 words). This difference in frequencies is statistically significant because the LL value of 25.00 is above the critical value of 15.13 where \( p < 0.0001 \). This means that there are variations in the frequency of the epistemic function of *shall*, with the function being attested more in the ZE corpus. In the ICE-GB, the deontic function of *shall* is more frequent, occurring 0.9 times per 10 000 words compared to 0.1 times per 10 000 words in the ZE corpus. Given that the LL value for deontic *shall* is 35.65, which is above the critical value of 15.13, where \( p < 0.0001 \), the observed variations in frequencies are statistically significant. Therefore, there are variations in the frequency of the deontic function of *shall*. There are no statistically significant variations in the frequencies of the indeterminate functions of the modal verb *shall*, meaning that there are no variations.

### 6.6.2 Qualitative analysis of the modal verbs

Before the frequencies of functions of modal verbs are given, it is imperative to highlight some examples of these functions. The tables in this section show examples of the functions of the different modal verbs that were analysed in this study. Examples are given from both the ZE corpus and the ICE-GB. These examples will show that the two corpora were qualitatively analysed in order to determine the functions of modal verbs.
Table 50. Functions of modal verbs in the ZE corpus

<table>
<thead>
<tr>
<th>Modal verb</th>
<th>Function(s)</th>
<th>Examples</th>
<th>Explanation of example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can</td>
<td>Possibility</td>
<td><em>This grassroots approach can make a big difference.</em> (Editorials 5)</td>
<td>There is a chance of the grassroots approach making a difference.</td>
</tr>
<tr>
<td></td>
<td>Ability</td>
<td>With 50 people attending, I <em>can</em> supply 5 tables. (Private business letters 3)</td>
<td>The writer expresses his capability of supplying 5 tables.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td><em>I have included photos of some of the stuff I sell. You can look at them and tell me what you think.</em> (Private business letters 5)</td>
<td>The recipient is allowed to look at the photos in order to make an opinion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Could</td>
<td>Possibility</td>
<td><em>For example, I sometimes thought I <em>could</em> become a doctor</em> (Private semi-scripted dialogue 3)</td>
<td>There was a probability of the speaker becoming a doctor.</td>
</tr>
<tr>
<td></td>
<td>Ability</td>
<td><em>Ih since I liked athletics, I <em>could</em> run following the dog or the hares, the bucks that we could meet in the forest.</em> (Private semi-scripted dialogue 17)</td>
<td>The speaker was able to run following the dog or the hares, the bucks that we could meet in the forest.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td>*Minister Mavima explained that learners <em>could “skip” A’ Level and proceed straight to institutions of Higher and Tertiary Education, with the choice being incumbent upon the learner, parent or guardian.</em> (Newspaper reportage 13)</td>
<td>Learners were authorised to skip ‘A’ level and proceed straight to institutions of Higher and Tertiary Education,</td>
</tr>
<tr>
<td>May</td>
<td>Possibility</td>
<td><em>One <em>may</em> have a meeting from Zimbabwe with a business partner in Kuwait through skype without needing to go to Kuwait.</em> (Academic writing: Examination 1)</td>
<td>It is feasible for someone to have a meeting from Zimbabwe with a business partner in Kuwait through skype without needing to go to Kuwait.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td><em>Yes you <em>may</em> go ahead doctor</em> (Public dialogue 19)</td>
<td>The doctor is allowed to speak.</td>
</tr>
<tr>
<td>Might</td>
<td>Possibility</td>
<td><em>We <em>might</em> graduate on the 28th.</em> (Social letters 2)</td>
<td>There is a prospect of graduating on the 28th.</td>
</tr>
<tr>
<td>Must</td>
<td>Necessity</td>
<td><em>I <em>must</em> say it is very good to hear.</em> (Private business letter 4)</td>
<td>The writer gives an objective conclusion that it is very good to hear.</td>
</tr>
<tr>
<td></td>
<td>Obligation</td>
<td><em>It’s a brutal road, but we <em>must</em> travel together to make the load lighter.</em> (Editorials 7)</td>
<td>It’s a brutal road, but it is our responsibility to travel together to make the load lighter.</td>
</tr>
<tr>
<td>Will/’ll</td>
<td>Prediction</td>
<td><em>I will come and visit you over the weekend.</em> (Private business letters 4)</td>
<td>The writer is anticipating that he is going to visit over the weekend.</td>
</tr>
<tr>
<td></td>
<td>Volition</td>
<td><em>“I <em>will</em> tell you one day,” Ambuya said, yawning.</em> (Creative writing 5)</td>
<td>Ambuya intends to tell one day.</td>
</tr>
</tbody>
</table>
Would/’d  Prediction  Bukhosi would have appreciated it. (Creative writing 1)  Recently, Ambassador Nichols said the economy would recover, but cautioned it won’t happen “overnight”. (Newspaper reportage 10)  If I were to meet Canadian rocker Bryan Adams, I would ask for his permission to rewrite some of the lyrics from his hit song, “Summer of ’69 . . .” (Newspaper reportage 6)  Ambassador Nichols forecast the recovery of the economy, but cautioned it won’t happen “overnight”.  If I were to meet Canadian rocker Bryan Adams, I intend to ask for his permission to rewrite some of the lyrics from his hit song, “Summer of ’69.

Should  Obligation  Zimbabweans should have faith in local expertise. (Editorials 5)  Zimbabweans have a duty to have faith in local expertise.

Necessity  You said I should tell you next time I go there. (Private business letters 5)  It is essential for the writer to inform the reader when he goes.

Shall  Prediction  All players in the print media shall also charge for their products in foreign currency. (Business letters 1)  The writer is forecasting that all players in the print media are also going to charge for their products in foreign currency.  We shall talk about how you will pay for my destroyed house later. (Creative writing 6)  We are going to talk about how you will pay for my destroyed house later.

Volition

The table above has highlighted some of the examples of different functions of modal verbs, together with explanations of the examples. Based on the functions described above, it can be noted that ZE seems to follow StE norms in the use of the nine modal verbs examined.

The following table provides an overview of the examples of functions in the ICE-GB

<table>
<thead>
<tr>
<th>Modal</th>
<th>Function(s)</th>
<th>Examples</th>
<th>Explanation of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can</td>
<td>Possibility</td>
<td>Everybody can make informal arrangements. (S1B-078 #246)</td>
<td>There is a chance of everybody making informal arrangements.</td>
</tr>
<tr>
<td></td>
<td>Ability</td>
<td>I can read books. (S1A-084 #059)</td>
<td>I am able to read books.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td>Yes you can refer to your notes whenever you need to. (S1B-068 #013)</td>
<td>Yes, you are allowed to refer to your notes whenever you need to.</td>
</tr>
</tbody>
</table>

© University of Pretoria
<table>
<thead>
<tr>
<th>Could</th>
<th>Possibility</th>
<th>We could do that. (S1A-039 #170)</th>
<th>There is a probability of doing that.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ability</td>
<td>I could see these eyes sort of glaring at me. (S1B-049 #134)</td>
<td>I was able to see these eyes sort of glaring at me.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td>Could I ask my honourable friend to offer some reassurance? (S1B-059 #034)</td>
<td>The speaker is asking for permission to offer some reassurance.</td>
</tr>
<tr>
<td>May</td>
<td>Possibility</td>
<td>Something may happen. (S1A-094 #210)</td>
<td>There is a chance of something happening.</td>
</tr>
<tr>
<td></td>
<td>Permission</td>
<td>May I sit down for a minute? (W2F-018 #170)</td>
<td>The speaker is seeking permission to sit down for a minute.</td>
</tr>
<tr>
<td>Might</td>
<td>Possibility</td>
<td>We might die and then find ourselves going straight down. (S1A-084 #155)</td>
<td>There is a prospect of dying</td>
</tr>
<tr>
<td></td>
<td>Necessity</td>
<td>I must say I have driven once. (S1A-097 #155)</td>
<td>It is necessary to say I have drive once.</td>
</tr>
<tr>
<td></td>
<td>Obligation</td>
<td>You must pay these if you are normally self-employed. (W2D-004 #050)</td>
<td>You have a duty to pay these if you are normally self-employed.</td>
</tr>
<tr>
<td>Will/’ll</td>
<td>Prediction</td>
<td>I don’t think it will be. (S1A—77 #011)</td>
<td>The speaker is anticipating that it will not be.</td>
</tr>
<tr>
<td></td>
<td>Volition</td>
<td>“I will”, said Patience gratefully. (W2F-007 #144)</td>
<td>Patience is saying that she intends to do it.</td>
</tr>
<tr>
<td>Would/’d</td>
<td>Prediction</td>
<td>Well I would have been if I did the last year. (S1A-048 #017)</td>
<td>The speaker is forecasting what she would have been</td>
</tr>
<tr>
<td></td>
<td>Volition</td>
<td>We decided that we would work together. (S1A-001 #048)</td>
<td>We decided that we intend to work together.</td>
</tr>
<tr>
<td>Should</td>
<td>Obligation</td>
<td>You should keep your ancestral paintings. (S1A 007 #154)</td>
<td>You have a duty to keep your ancestral paintings</td>
</tr>
<tr>
<td></td>
<td>Necessity</td>
<td>If your vehicle is stolen, you should report it immediately to the police. (W2D-010 #063)</td>
<td>If your vehicle is stolen, it is essential that you report to the police immediately.</td>
</tr>
<tr>
<td>Shall</td>
<td>Prediction</td>
<td>I shall have a fever by tonight, blood poisoning soon after. (W2F-015 #131)</td>
<td>The speaker is forecast having a fever by tonight, blood poisoning soon after</td>
</tr>
<tr>
<td></td>
<td>Volition</td>
<td>Ok Ian, I shall try to get to know you. (S1A-017 #267)</td>
<td>Ok Ian, I intend to try to get to know you.</td>
</tr>
</tbody>
</table>
The table above has highlighted some of the examples of different functions of modal verbs, together with explanations of the examples.

Since some examples have been provided in tables 43 and 44 above, the next section will provide the frequencies of functions of modal verbs in the ZE corpus and the ICE-GB.

6.6.3 Summary of quantitative and qualitative results for modal verbs

In Section 6.6.1 data analysis provides the occurrence- and function frequencies of modal verbs in the ZE corpus and the ICE-GB. A comparative analysis was also done in terms of the log likelihood test to see if there is a significant variation of modal verbs between the corpora, in order to ascertain if the instructions themselves show innovation. This summary thus aims to answer research questions 1.3 and 2.3:

Are the variations between the corpora statistically significant?

In order to determine whether there are statistically significant variations in frequencies of modal verbs in the ZE corpus and the ICE-GB, the log likelihood was calculated in genres, registers and in the overall frequencies of the modal verbs.

Data analysis revealed that the overall variations observed in the occurrence frequencies of *can* and *will/’ll* were statistically significant, with the ZE corpus recording higher normalised frequencies compared to the ICE-GB. Statistically significant variations were observed because the ZE had higher normalised frequencies of *could, may, might and would* compared to the ICE-GB. The LL values were all higher than 15.13, which is the critical value where $p < 0.0001$, meaning that the variations were not due to chance. Three modal verbs namely *must, shall* and *should* did not show any statistically significant variations.

When the two corpora’s registers were compared, observations were that, in the spoken register, *can* and *will/’ll* were attested more in the ZE corpus compared to the ICE-GB and the variation is statistically significant. The LL values were higher in the spoken register, where the ICE-GB had higher normalised frequencies for *could, may, might, shall and...*
would. In the spoken register, no statistically significant variations were noted in the frequency of must and should. In the written register, statistically significant variations were noted in the frequencies of could, may and might and would/'d, with the ICE-GB recording more frequencies compared to the ZE corpus. In the ZE corpus, shall was attested more than in the ICE-GB and the LL value was higher than the critical value where p < 0.0001. There were no variations in the frequencies of can, must, should and will/'ll in the written register.

The LL values of each genre were also calculated in order to determine statistical significance. Can was more frequent in the ZE corpus in private dialogue, public dialogue, social letters and business letters compared to the ICE-GB. In the ICE-GB, can was more frequent in the popular writing genre compared to the ZE corpus. The observed variations are statistically significant because the LL value was above the critical value for p < 0.001. No statistically significant variations were noted between the two corpora in public scripted monologue, creative writing, academic writing, editorials and newspaper reportage.

Could was more frequent in the ICE-GB in public scripted monologue, public dialogue, editorials, newspaper reportage and business letters. These variations are statistically significant because the LL values were higher than the lowest LL value of 3.84, which is a critical value where p < 0.05. Could was more prevalent in the ZE corpus in academic writing compared to the ICE-GB. No statistically significant variations were reported in private dialogue, creative writing, popular writing and social letters, meaning that there were no variations in the frequencies in these genres.

The LL value for editorials showed that may was more frequent in the ICE-GB compared to the ZE corpus. When considering the newspaper reportage genre, may was more frequent in the ICE-GB compared to the ZE corpus. The normalised frequencies in popular writing and business letters were higher in the ICE-GB compared to the ZE corpus. Due to the lower LL values in private dialogue, public scripted monologue, public dialogue,
creative writing, academic writing and social letters, it was observed that there were no statistically significant variations in these genres.

*Might* was more frequent in the ICE-GB compared to the ZE corpus in private dialogue, public dialogue, creative writing, editorials, newspaper reportage and business letters. In public scripted monologue, popular writing, academic writing and social letters, no statistically significant variations were reported because the LL were lower than 3.84, which is a critical value for \( p < 0.05 \).

It was observed that *must* was attested more in the ICE-GB compared to the ZE corpus in private dialogue, social letters, popular writing and creative writing. *Must* was more prevalent in the ZE corpus compared to the ICE-GB in public scripted monologue and in newspaper reportage. No statistically significant variations were recorded in public dialogue, academic writing, editorials and business letters.

In the ZE corpus, *shall* was attested more in newspaper reportage and in business letters compared to the ICE-GB. These variations are statistically significant because the LL values were higher than the critical values set in at different levels of significance in this study. Regarding the private dialogue genre, the normalised frequency of *shall* was higher in the ICE-GB compared to the ZE corpus, meaning that the variations observed were statistically significant because the LL value was higher than 15.13, which is a critical value for \( p < 0.0001 \). Due to the lower LL values, no statistically significant variations were noted in the frequency of *shall* in seven genres namely public scripted monologue, public dialogue, creative writing, popular writing, academic writing, editorials and social letters, meaning that there were no variations between the two corpora.

Statistically significant variations were recorded in the frequency of *should*, where the normalised frequencies were higher in the ICE-GB compared to the ZE corpus in popular writing and business letters. *Should* was attested more in the ZE corpus in academic writing and editorials in the ICE-GB. No statistically significant variations were in the frequency of *should* in private dialogue, public scripted monologue, public dialogue, creative writing, newspaper reportage and social letters because the LL values were lower than 3.84, a critical value used for \( p < 0.05 \).
There were statistically significant variations in eight genres of *will/’ll*. The frequencies were higher in the ZE corpus in private dialogue, public dialogue, social letters and business letters, compared to the ICE-GB. The normalised frequency of *will/’ll* per 10 000 words was higher in the ICE-GB compared to the ZE corpus in creative writing, popular writing, editorials and newspaper reportage. Due to the lower LL values reported in public scripted monologue and academic writing the variations recorded were not regarded as statistically significant.

Data analysis revealed statistically significant variations in the frequency of *would/’d* in nine genres. The frequency of *would/’d* was higher in the ZE corpus in popular writing compared to the ICE-GB. In the ICE-GB the statistically significant variations were noted in private dialogue, public scripted monologue, public dialogue, newspaper reportage, social letters, business letters and editorials compared to the ZE corpus. No statistically significant variations were recorded in the academic writing genre because the LL values were lower than the critical value of 3.84, where $p < 0.05$.

The ZE corpus and the ICE-GB were also compared for function frequencies. *Should* and *must*, which are the modals of necessity and obligation, did not show any statistically significant differences regarding their functions in ZE and BrE. This result means that there are no variations in the frequency of functions of *should* and *must* because all the LL values were below the critical value of 3.84 where $p < 0.05$.

Statistically significant differences can be observed in the epistemic functions of *may* because the function was more frequent in BrE compared to ZE, meaning that there are variations in the frequency of the epistemic functions of *may*. There were variations in the frequency of the dynamic function, which was attested more in the ZE corpus compared to the ICE-GB. This variation is statistically significant because the LL value of 100.44 is above 15.13, which is the critical value for $p < 0.0001$. The deontic function of *can* was also attested more in the ZE corpus compared to the ICE-GB, meaning that there are variations in the frequency of the deontic function of *can* between the two corpora. The epistemic function occurred more in the ICE-GB compared to the ZE corpus, showing that there are variations between the two corpora because the LL value of 80.64 is higher than the critical value of 15.13 where $p < 0.0001$. 
The dynamic function of *might* was more frequent in the ICE-GB, compared to the ZE corpus. This shows that there are variations in the frequency of the dynamic function of *might*. BrE speakers used the dynamic function of *could* more frequently compared to ZE speakers because the LL value of 22.89, was higher than the critical value of 15.13 where $p < 0.0001$. There were no statistically significant differences in the deontic and indeterminate functions of *could*, meaning that there are no variations in the frequencies of these functions in the two corpora.

The ZE corpus recorded more instances of the epistemic functions of the modal verb *will* compared to the ICE-GB and the observed differences are statistically significant because the LL value is higher than the critical value of 15.13 where $p < 0.0001$. The dynamic function of *will* occurred more in the ZE corpus compared to the ICE-GB, a statistically significant variation because the LL value of 8.93 is higher than the critical value 6.63, where $p < 0.001$. This means that in the two corpora, there are variations in the frequency of the dynamic function of *will*. No statistically significant differences were recorded for the deontic and indeterminate functions of the modal verb *will*, meaning that there are no variations in the frequency of these functions. The LL values of all the functions of *would* (epistemic, dynamic and indeterminate) were higher, meaning that the functions were more frequent in the ICE-GB compared to the ZE corpus and showing statistically significant variations.

The epistemic function of *shall* was more prevalent in the ZE corpus compared to the ICE-GB. This difference in frequencies is statistically significant because the LL value is above the critical value of 15.13 where $p < 0.0001$, meaning that the observed variations are not due to chance. In the ICE-GB, the deontic function of *shall* was more frequent, compared to the ZE corpus. Due to the fact that the LL value for deontic shall was above the critical value of 15.13, where $p < 0.0001$, the observed variations in frequencies are statistically significant. No statistically significant variations were noted in the frequencies of the indeterminate functions of the modal verb *shall*, meaning that there are no variations between the two corpora.
The following section outlines how the innovation features occurred in the ZE corpus and whether they can be attested in Kortmann, Lunkenheimer and Ehret’s (2020) list of innovation features.

6.6.4 Attestation of innovation features according to Kortmann et al. (2020) for modal verbs used in Zimbabwean English

Up to now, the results section has been focusing on the absolute and normalised occurrence-and function frequencies of modal verbs, the functions of modal verbs and their frequency in the ZE corpus and the ICE-GB and on whether there are statistically significant variations between the two corpora which could indicate innovation. This section ties in with section 6.5 because it aims to determine whether the three innovation features that are reported to show innovation in Kortmann, Lunkenheimer and Ehret’s (2020) list of features. The ZE corpus is also analysed in order to determine the categories of the three features in ZE. This will help to answer research question three, which seeks to determine whether there is innovation in ZE, when considering the three features (121, 123 and 127) as outlined in the table below.

<table>
<thead>
<tr>
<th>Feature number and description</th>
</tr>
</thead>
<tbody>
<tr>
<td>121: Double modals</td>
</tr>
<tr>
<td>123: Present tense forms of modals used where StE has past tense forms</td>
</tr>
<tr>
<td>127: Non-standard use of modals for politeness reasons</td>
</tr>
</tbody>
</table>

This analysis focuses on answering research question 3:

3.1 Which features of innovation are present with regards to modal verbs in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE?

Considering that the classification criteria for an innovation included grammatical stabilisation, acceptability and widespread use of a feature by different participants, in different genres and authoritative norm orientation, it was necessary to check the
extralinguistic data (biographical and language background questionnaires) of participants.

The three features that show variation in their use are examined in Sections 6.6.4.1, 6.6.4.2 and 6.6.4.3. A usage-based functional approach is employed in the analysis, but it is also informed by the way modal verbs are categorised in English reference grammar books (e.g. Quirk, et al, 1985; Huddleston & Pullum, 2002). Using Sketch Engine software tools, the corpus was analysed to check whether the three features (121, 123 and 127) that are reported to show morphosyntactic variation can be attested in ZE. The discussion includes a categorisation of the features according to the explanations given by Kortmann, Lunkenheimer and Ehret's (2020). The explanations are as follows:

A- Feature is pervasive or obligatory.
B- Feature is neither pervasive nor extremely rare.
C- Feature exists, but is extremely rare.
D- Attested absence of the feature.

6.6.4.1 Double modals: Feature 121

According to Lebedeva & Orlova (2019) and Hasty (2012), double modals occur in the different L1 varieties of English in southern United States, northern England, Northern Ireland and Scotland (Fennell, 1993; Battistella, 1995; Hasty, 2012; Lebedeva & Orlova, 2019). In literature, the use of double modals has also been reported in L2 varieties of English such as BSAE (Van Rooy, 2011) and Pure Fiji English (Tent, 2020). Lebedeva & Orlova (2019: 76) give the following examples of double modal use, where the epistemic and dynamic meanings are used.

6.5 It's an important cause because it's relatable to us girls,' Richardson said. 'Any girl could be diagnosed with breast cancer right now and if we teenagers get involved in raising money, we might could find a cure.'

(Lebedeva & Orlova, 2019: 76)

The authors note that *might* indicates epistemic possibility whilst *could* conveys the dynamic meaning which shows that ability. In saying we *might could*, there is an expression of uncertainty regarding the possibility of finding a breast cancer cure. Hasty
(2012: 1717) notes that there are different examples of double modals found in literature (e.g. *might should, must could, might would, may can, might can, may could, might will, may will and may should*). The author notes that *might could* is the most frequently used double modal and gives an example of the use of double modals in questions as shown below.

6.6 *Might could you go to the store for me?*  
(Hasty, 2012: 1718)

Double modals are also used in negation, where negation can occur after the second modal as shown below.

6.7 *I might could not go to the store.*  
(Hasty, 2012: 1718)

According to Hasty (2012: 1718), negation can also occur between two modals (e.g. *I might not could go to the store*).

To check whether there is use of double modals, the researcher searched the concordances of all the nine modal verbs in Sketch Engine. The concordance lists were manually checked to determine whether feature 121 was attested.

In the ZE corpus, observations were made about insertion of a grammatical morpheme *be able to*, which is a quasi-modal, after the modal *can*. Insertion is a type of anti-deletion (Mesthrie, 2006). The term “anti-deletion” was coined by Mesthrie (2006) to refer to instances of insertion, undeletions and non-deletion in BSAE. Undeletion “restores an element that is assumed to be deleted or to have an empty node in generative analyses of English” (Mesthrie, 2006: 80). The author notes that insertion refers to instances where grammatical morphemes are added and non-deletion shows the presence of a feature of standard English that is deleted in some L1 dialects of English” (Mesthrie, 2006: 80).

There is interesting evidence of the use of *can* together with *be able to* in the ZE corpus. This can be regarded as use of double modals because, according to Van Rooy (2011: 198), “the construction ‘can be able to’ combines the extrinsic possibility meaning with the
ability meaning”. This means that there is use of double modals because a modal and a quasi-modal are used in succession.

In order to check whether the phrase can be able to was used in the corpus, the phrase can be able to was searched in the concordance function and concordance lines of were checked for the phrase. The concordance list contained nine examples of can be able to in the ZE corpus. All the nine instances are listed below, with the phrase can be able to shown in bold.

6.8 <$H> <$#> Ih you have to pass your ordinary level and then you have to go to ‘A’ level. And you have to have like ih more points so that you can be able to qualify to go to university and be a a and do your degree so that you can become a a doctor. (Private semi-scripted dialogue 8)

6.9 <$B> We’ve got people that uh can be able to help us with that. (Public dialogue: 32)

6.10 <$B> They must be dealt with efficiently and effectively so that we can be able to recover the loot. (Newspaper reportage 4)

6.11 We are hoping going forward we can be able to produce the calibre of players who can remain focused and be able to cross the finish line. (Newspaper reportage 14)

6.12 Eh sometimes there will be no water and it’s tough to work ih in an environment where there is no water especially at a clinic since most of the work we do need ih one to to wash hands after every procedure so that you can be able to do sterile procedures. (Private semi-scripted dialogue 18)

6.13 Which is my desire so that people eh students can be able to access the internet and research their material. (Private dialogue 6)

6.14 Does the funding also include uhm training a training aspect so that maybe the locals can be able to do that by themselves? (Public dialogue 19)

6.15 Uhm in the in the in the interest of serving our listeners I know we’ve uh we’ve we’ve appealed uhm on this particular program that whenever we’re discussing a certain chapter let’s stick to that chapter so that we can be able to cover uh more depth (Public dialogue 33)

6.16 But next week . . . we will be able to uh hopefully appeal to everyone and everyone can be able to listen and understand the constitution. (Public dialogue 33)
Can be able to is attested nine times in four different genres namely newspaper reportage, public dialogue, private dialogue and private semi-scripted dialogue. The normalised frequency of can be able to in the ZE corpus is 25 instances per 1 million words. This is evidence that the use of can be able to is not an error but is an emerging feature of ZE. To compare the use of can be able to in other L2 varieties of English such as BSAE, Van Rooy (2011) noted that the frequency of can be able to in the Xhosa English corpus as 42 instances per one million words and noted that this is regarded as an established feature of BSAE.

Of the nine instances, the profiles of three participants are available. The other six instances occur in the data in which there are no language backgrounds (four from Marungudzi, 2016a and two from newspaper reportage). Participant <$OO> was a 34-year-old engineering technician, participant <$HH> was a 53-year-old high school teacher and participant <$Q> was a 30-year-old banking officer. All the three participants reported that they learned English at school and were formally employed. They reported that they spoke English at least 24 hours per week. All the three participants hold undergraduate degrees. Schmied (1991: 48) proposed a lectal continuum for the different ESL varieties in Africa. The author groups L2 English speakers into basilect speakers, mesolect speakers and acrolect speakers. According to Schmied (1991), basilect speakers have only had primary education and usually work as shop assistants, clerks or taxi drivers. Mesolect speakers have had secondary education and their occupations include junior civil servants and senior nurses. Acrolect speakers are speakers who obtained at least a university degree and work as newspaper editors and lawyers. If the three types of speakers suggested by Schmied (1991) are considered all three participants can be regarded as acrolectal speakers, which is the highest level. This is because the three participants all have higher education qualifications.

It is also interesting to note that can be able to is attested twice in formal texts (newspaper reportage). Although double modals are not attested in all genres, they are attested in both spoken and written texts. Therefore, the use of double modals partially meets the criteria for an innovation, save for the fewer number of occurrences. It is because of this that feature 121 can be categorised as existing, but is extremely rare (category C).
Therefore, the use of double modals is not established in ZE. The results are similar to those reported on L2 varieties of English like Pure Fiji English (Tent, 2020). In comparison, Van Rooy (2011) regards the use of can be able to as an established feature of BSAE due to its high frequency.

6.6.4.2 Present tense forms of modals used where StE has past tense forms: Feature 123

According to Walter and Cook (1978) and Shinobu (2018), the forms of modal verbs include present and past tense forms. The authors note that in contexts where discourse refers to the past, past tense forms of modal verbs have past meanings. Bybee (1995) gives the following sentence as an example of present tense use of might.

6.17 I think it unlikely actually, but he might do it today.  
(Bybee, 1995: 504)

In the example above, the present tense modal might is used to show the possibility of doing it today. Another example of present tense use of a modal is shown below.

6.18 You can smoke in the lounge.  
(Walter & Cook, 1978: 7)

In the example above, can is used to show that the hearer is permitted to smoke in the lounge. An example of the use of the past tense modal is given below.

6.19 George Bush couldn’t run a laundromat.  
(Bybee, 1995: 504)

The example above could be interpreted as meaning that in the past, George Bush attempted to run a laundromat and failed. Another example showing past tense use of a modal is shown below.

6.20 It could have been a shark.  
(Walter & Cook, 1978: 10)

In example 6.20 above, there is reference to the past, as shown by the use of should have.
“The past tense forms of modal verbs do not simply refer to the past time. Rather, what past form modals usually signify is peripheral in events of the present and future time” (Shinobu, 2018: 79). For example, could you sign this now, Doctor? Due to the fact that there is reference to the present time (as shown by the use of now), could is used to refer to the present time.

Statistical results in Section 6.6.1.1 indicate that two present tense forms, can and will were considerably more frequent in ZE (38.2 and 46 instances respectively) compared to BrE (25.1 and 37.4 instances respectively). This warranted checking all the concordance lines of can and will to determine whether present tense forms were used where StE has past tense forms. Examples of feature 123 are given below.

6.21  <$L> <#> Cyclone Idai was very bad. It destroyed houses, schools, and bridges. A lot of people did not think that the cyclone will (would) be bad<$L> <#> So, when it came, people were relaxed. (Private semi-scripted dialogue 12)

6.22  <$NN> <#> Well, there are several uh things I could put. <#> But also it’s kind of a difficult question. <#> How would you weigh the important or the things you have been proud of? <#> But I would pick uh completing my studies, getting my P h D. <#> I think to me it’s a huge achievement. I’m proud of myself. <#> I never dreamt I will (would) go that far. (Private semi-scripted dialogue 38)

6.23  <$H> The assistance that they get was from the from the government. <#> The government helped in rescuing people using helicopters to take people to safe places. <#> And Strive Masiyiwa also helped with money and so that people can (could) buy clothes. (Private semi-scripted dialogue 8).

6.24  Workers did not have anything to say about the finished goods. <#> All the decisions were in the hands of the management team. <#> Henry Fayol proposed that management can (could) be put into four main functions. (Academic writing: Examination)

6.25  I hope that you had a merry Christmas. <#> Happy new year. <#> Thank you for replying to my email. <#> You asked if I can (could) print your school logo on sports bottles. <#> I will do that. (Private business letters 2)
In the Wordlist tool in Sketch Engine, the tagset for modal (MD) was searched and the concordance lines of each modal was checked to determine whether present tense forms of modal verbs were used where StE has past tense forms. An example sentence is *they can* (might) *be wild, but they’re human beings* (Kortmann, Lunkenheimer & Ehret, 2020). By checking the context in which the modal verbs occurred in the concordance lines, it was established that the use of present tense forms where StE has past tense forms was attested (35) times in the ZE corpus. The following criteria was taken into consideration to identify innovations (the classification criteria are explained in detail in subsection e in Section 4.4.2.6).

1. Frequency counts: To show grammatical stabilisation (cf. Van Rooy 2011; Van Rooy & Kruger, 2016; Partridge, 2019).
2. Use of a feature regularly and widely by different speakers to show acceptability in a speech community (cf. Bamgbose, 1998; Gut, 2011; Van Rooy, 2011; Mesthrie & Bhatt, 2008).
3. A feature should be attested in different genres.
4. Feature should be attested in both spoken and written registers (cf. Van Rooy & Kruger, 2016).

If a feature did not meet any of the abovementioned classification criteria, then it was regarded as an error. For example, if a feature was only attested in informal speech, then it was regarded as an error (cf. Van Rooy & Kruger, 2016). If a feature met all the classification criteria but was attested with lower frequencies (e.g. once in a genre), then other criteria mentioned above were used to classify a feature. The biographical and language background questionnaire was used to check whether a feature was used by the same participant or if a feature was used by more participants. This would show widespread and regular use of a feature, meaning acceptability within a speech community.
Data analysis showed that feature 123 was attested 35 times in the ZE corpus. The use of present tense forms where StE has past tense forms was attested in three genres namely academic writing: Examination, private business letters, and private semi-structured dialogue. Due to the fact that feature 123 occurred 33 times in the spoken register and twice in the written register, the feature can be placed in category C despite the high frequency count of 35, which allows for the feature to be in category B. This is because other classification criteria such as authoritative norm orientation and widespread use in the written register were not met.

Feature 123 can be assigned to category C because the feature exists but is extremely rare in the ZE corpus. Therefore, present tense forms of modals used where StE has past tense forms, is not an innovation feature in ZE but can be described as an emerging feature. To compare ZE with other L2 varieties of English, Buregeya (2020) makes similar observations that the use of present tense forms of modal verbs where StE has past tense is absent in Kenyan English. The feature is also categorised as C in Ghanaian English (Huber, 2020), and Tanzanian English (Bayley, 2020).

6.6.4.3 Non-standard use of modals for politeness reasons: Feature 127

Since feature 127 refers to non-standard use of modals for politeness reasons, an overview of “politeness” will be done. According to Brown & Levinson, (1987: 55) “patterns of message construction, or ‘ways of putting things,’ or simply language usage, are part of the very stuff that social relationships are made of or [...] crucial parts of the expressions of social relations.”

The authors believe that context is an important aspect of language and regard politeness as a special element, which is widespread and well established in society. The actions and interactions of society and language signal the interrelationship between the two, with “rationality” and “face” being two elements that all language users share (Brown & Levinson, 1987). Wasserman (2014) uses the term “negative face”, which concerns the need to be unobstructed by others and “positive face”, which concerns the desire to be accepted and regarded as desirable to describe the two types of face. Although there is universality in the need for sensitivity in interactions, face is considered as:
Something that is emotionally invested and that can be lost, maintained, or enhanced and must be constantly attended to in interaction. In general, people cooperate (and assume each other’s cooperation) in maintaining face in interaction, such cooperation being based on the mutual vulnerability of face.

(Brown & Levinson, 1987: 61)

In literature, one of the features that show variation is the use of the illocutionary force\textsuperscript{56} \textit{must}, which shows strong obligation to show considerably less social impact. In this use, \textit{must} becomes less face threatening in some contexts. According to Bowerman (2008: 477) \textit{must} is often used in place of polite \textit{should/shall} as in the sentence: \textit{You must turn left at the robots}. There were no statistically significant differences in the meanings of modals of obligation \textit{must} and \textit{should} in ZE and BrE. Therefore, no further analysis was done on these modal verbs.

Another example of non-standard use of modals for politeness reasons mentioned by Bowerman (2008: 477) is the use of “two directive softeners: the voluntative modal \textit{will} and the negative \textit{won’t}” to soften a request as shown in the sentence: \textit{Won’t you do me a favour?}

Huber and Dako (2008) report on the rarity of polite requests (e.g. \textit{could I/you, might I, would it be possible}) in Ghana English when compared to BrE. The authors further state that, what BrE speakers regard as orders are perceived as polite requests when used with in Ghanaian English as shown below:

6.26 \textit{I want} to borrow your book, please.

(Huber & Dako, 2008: 371)

Regarding politeness, Makalela (2004: 362) reports that BSAE speakers use “downgraders (detensifiers), to play down the force of impact an utterance has on the addressee”. The author gives the following sentence to show that downgraders are a sign of politeness in Bantu languages:

---

\textsuperscript{56} Crystal (2008: 446) regards illocutionary force as “the intentions of speakers while speaking”. The author lists commanding, promising, requestion etc. as examples of illocutionary acts.
Myself, I think I can say divorce is not a solution.

(Makalela, 2004: 362)

In the sentence above, I think and I can say are used as double downgraders to downplay the effects of bringing up an issue, which may seem offensive or disagreeable. To check whether double or multiple downgraders were used in the corpora, concordance lines were searched for the contexts in which modal verbs were used.

With regard to non-standard use of modals for politeness reasons, the use of the strong obligation must in less face-threatening contexts was not attested in the ZE corpus. The different meanings of must analysed in the ZE corpus and the ICE-GB do not show any statistically significant differences. This means that there are no variations in the use of must between the two corpora. In ZE, some examples that show politeness are given below:

6.28 Plz may you accompany us to Ps Tete in Chivi in Dec and then to my relatives. (Social letters 2)

6.29 Plz may you assist her to register English for both June and Nov. (Social letters 2)

6.30 So I will like to repeat this one. (Public dialogue 35)

6.31 So, I will like to see people with safe water. (Private semi-scripted dialogue 15)

6.32 <$BB> First of all I will like to if I have I I am a millionaire, I will help my family members who are poor.

The use of may in examples 6.25 and 6.26 expresses volition. This is because the speaker wishes or hopes to be accompanied to Chivi and to his relatives. In the two examples, examples, may is used to soften a request. In examples 6.25, 6.26, 6.27, 6.28 and 6.29, will is used in a non-standard way for politeness reasons because in StE, would is preferred.

Considering the frequency counts, feature 127 is attested 8 times. All three instances of non-standard use of modals for politeness reasons are found in a sample from one speaker (Social letters 2). In addition, two out of three instances in private semi-scripted dialogue are from one speaker (participant <$SS$>). This shows that more than half of the instances come from two participants. Therefore, there is no widespread use by different speakers. Regarding the criteria that feature 127 should be attested in different genres,
all the genres where examples are found are informal genres. This can mean that the non-standard use of modals for politeness reasons is confined to the informal genres. Feature 127 is not attested in editorials, published material and newspaper reportage, meaning that there is no authoritative and norm orientation as suggested by Bamgbose (1998) and Van Rooy and Kruger (2016). Given the aforementioned reasons, feature 127 belongs to category D where there is attested absence of the feature in ZE. The 8 instances can be regarded as errors. As a comparison with other African L2 varieties of English, feature 127 is reported to be absent in Ghanaian English (Huber, 2020) and Pure Fiji English (Tent, 2020).

6.7 Conclusion

A comparison of the ZE corpus data and the ICE-GB data revealed that the frequencies of must, shall and should per 10 000 words did not show any statistically significant differences. Statistically significant variations were recorded in the frequencies of can and will/ll, with the ZE corpus recording more occurrences per 10 000 words for both modals compared to the ICE-GB. The LL values for the frequencies of could, may and might were all above the critical value of 15.13 where p < 0.0001, with the ICE-GB recording more frequencies per 10 000 words than in the ZE corpus. Another important observation made in this chapter is that modal verbs are more frequent in spoken texts than in written texts. In the ZE corpus, can and will occurred more frequently compared to the ICE-GB. Could, might, shall and would, were more prevalent per 10 000 words in the ICE-GB compared to the ZE corpus. The highest LL values that were above the critical value of 15.13 were recorded in the use of would, could, may and might in the written register.

A qualitative functional analysis of the meanings of modal verbs was performed in order to inform on the frequencies of the different meanings. No statistically significant differences were recorded in the use of obligation and necessity modals must and should. Regarding the modals of possibility, permission and ability, some variations were observed in the frequencies per 10 000 words. For example, it was recorded that the epistemic function of may is attested more in the ICE-GB. The dynamic function of can
was more prevalent in the ZE corpus. BrE speakers use the dynamic function of *could* more frequently compared to ZE speakers as evidenced by the fact that the LL > 15.13 where p< 0.0001. When it comes to modals of prediction and volition, the epistemic *will* was more frequent in the ZE corpus whilst the epistemic *would* and the deontic *shall* were more frequent in the ICE-GB. Other functions of the modals did not show statistically significant differences because the LL values were lower than the minimum critical value of 3.84. These include the dynamic and indeterminate functions of *shall* and the deontic and indeterminate functions of *will*.

When considering the three features explored in this chapter, results indicate that ZE demonstrates one feature that is, to a lesser degree, characteristic of L2 varieties of English, but in the other 2 features analysed, ZE demonstrates similarities with L1 English norms.

The table below shows a summary of the features studied in this chapter along with the categorisation attributed to it according to the abovementioned analyses.

<table>
<thead>
<tr>
<th>Feature and description</th>
<th>Category of feature in ZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>121: Double modals</td>
<td>(C) Feature exists, but is extremely rare.</td>
</tr>
<tr>
<td>123: Present tense forms of modals used where StE has past tense forms</td>
<td>(C) Feature exists, but is extremely rare.</td>
</tr>
<tr>
<td>127: Non-standard use of modals for politeness reasons</td>
<td>(D) Attested absence of feature</td>
</tr>
</tbody>
</table>

To summarise, features 121 and 123 were categorised as belonging to category C, where the feature exists, but is extremely rare. Findings obtained from analysing the ZE corpus together with results from questionnaires show that ZE users tend to follow L1 speaker norms when considering feature 127 because there is attested absence of the feature.

The use of double modals was attested nine times, in 4 genres in the ZE corpus. This was evidenced by the use of *can be able to*. In the ICE-GB, there are no instances of *can be able to*. Due to the fewer frequency counts of feature 121, it was categorised as belonging to
group C where the feature exists, but is extremely rare. There is evidence of use of present tense forms of modals where StE has past tense forms in the ZE corpus as evidenced by the frequency count of 35. Therefore, feature 123 belongs to category C. Feature 127, which shows the non-standard use of modals for politeness reasons, exists, but is absent in the ZE corpus.

The aforementioned discussion highlights the diversity in the use of modal verbs between ZE speakers and BrE speakers. While some of the frequencies of modals per 10,000 words in different registers and genres showed statistically significant variations, others did not.

The next chapter discusses discourse markers and analyses how they are used in ZE.
Chapter 7

Discourse markers

7.1 Introduction

The use of DMs in L2 varieties of English has become one of the focus areas by researchers (e.g. Müller, 2005; Lam, 2009a, 2009b; De Klerk, 2005a, 2005b; Unuabonah, 2019). Some scholars have observed that there are variations in the use of DMs by L2 users (e.g. Müller, 2005; De Klerk, 2005a, 2005b) In this chapter, the frequencies and functions of the DMs so, well and but in ZE are investigated. The study also examines whether there is variability in the use of DMs by ZE speakers. Comparisons are made between the ZE corpus and the ICE-GB. The ICE-GB was chosen because it represents L1 English usage (BrE), which is a type of StE if we consider the definition of StE given in the Merriam-Webster dictionary (cf. Section 1.7). Observing the frequency and functions of DMs in different text types in the ZE corpus and making comparisons to the ICE-GB will provide a better understanding of whether there are variations in the use of DMs between ZE, a L2 variety of English and BrE, a L1 variety of English. BrE (an example of StE), as represented in the ICE-GB was used for comparison to ZE.

In order to contextualise the discussion about DMs, an exploration of DMs and how they are used in StE is done in Section 7.2. Different characteristics of DMs are discussed in Section 7.3. Section 7.3 links with section 7.4, where functions of DMs are unpacked. The discussion about functions of DMs is important because the ZE corpus and the ICE-GB will be explored for examples of each function (sections 7.6.2.1, 7.6.2.2 and 7.6.2.3) and the frequencies of these functions will be reported in Section 7.6.1.4. Section 7.4 is subdivided into three sections. Section 7.4.1 explores the functions of so, section 7.4.2 focuses on the functions of well and section 7.4.3 discusses the functions of but. Since this study examines the use of DMs in L2 varieties of English, section 7.5 reports on some of the studies done on DMs in L2 varieties of English. The following subgroupings will be used in Section 7.6: The quantitative
data analysis of DMs is done in Section 7.6.1. This section is further divided into sub-sections. Section 7.6.1.2 provides the overall frequency of ten DMs in ZE and the ICE-GB and the absolute and normalised frequencies of *so*, *well* and *but* across different registers. In Section 7.6.1.3, the LL values, absolute and normalised frequencies of *so*, *well* and *but* across different genres are given. The frequencies of the functions of the three DMs are then given in Sections 7.6.1.4. A qualitative functional analysis is done in different sub-sections in Section 7.6.2 in order to determine the frequency of each function of the three DMs (*so*, *well* and *but*). Examples of the functions of *so*, *well* and *but* are given for the two corpora in Sections 7.6.2.1, 7.6.2.2 and 7.6.2.3 respectively. The chapter concludes in Section 7.7.

Bearing in mind that discourse and pragmatic features of a new variety may be nativised as stated by Makalela (2007), a discussion about DMs is provided next.

### 7.2 Discourse markers

There is no consensus on the definition of DMs (Maschler, 2000). This is evident in the several definitions provided by researchers. To illustrate this, some of the definitions are given below.

According to Schiffrin (1987: 31), DMs are “sequentially dependent elements which bracket units of talk”. Fraser (1990: 383) defines DMs as “expressions such as *now*, *well*, *so*, *however* and *then*, which signal a sequential relationship between the current basic message and the previous discourse”. “The term 'discourse marker' is used to describe a class of seemingly meaningless or empty elements that occur predominantly in spoken language (Jucker, 1997: 91). Another definition provided by Aijmer (2002: 2) is that DMs “seem to be dispensable elements functioning as signposts in the communication facilitating the hearer’s interpretation of the utterance on the basis of various contextual clues”. In another definition, Fraser (1999: 950) defines “discourse markers as a pragmatic class, lexical expressions drawn from the syntactic classes of conjunctions, adverbials and prepositional phrases”.

© University of Pretoria
When discussing DMs, it is important to be mindful of the different theoretical frameworks that have been proposed to analyse and account for DMs such as the relevance theory and the coherence theory. To better understand the differences between the two frameworks, a review of the distinctions made by researchers is done below.

When looking at the coherence theory and the relevance theory it can be noted that:

The coherence approach suggests that the best way to account for discourse interpretation is to look at coherence relations between topics in discourse. By contrast, the relevance approach argues that the recognition of coherence relations between discourse topics is neither necessary nor sufficient condition for a successful discourse.

(Algouzi, 2014)

Algouzi (2014) distinguishes between the coherence theory and the relevance theory by noting that the former focuses on textual functions whilst the latter is centred on cognitive processes. The mention of both these theories are relevant with regards to discourse markers as DMs, in terms of coherence, firstly provide contextual links between sentences or larger paragraphs in discourses and secondly, in terms of relevance, as each DM outlines a different meaning and contextual link that needs to be appropriate in the specific sentence, paragraph or discourse, as a micro and macro structure in the discourse.

In her study, Blakemore (2002: 5) used the relevance framework, to discuss “the cognitive processes underlying successful linguistic communication” and argued that DMs ought to be examined according to their contribution to the cognitive process.

Fraser (1993, 2009) states that DMs have a procedural meaning, which aids discourse coherence. He groups DMs into three classes namely elaborative, contrastive and inferential DMs. As their name suggests, elaborative DMs show how information in the first sentence is elaborated in the second sentence. Examples given by Fraser include and, above all, after all, also, alternatively, analogously, besides, by the same token,
correspondingly, equally. The second class is contrastive DMs, which show direct or indirect contrast between the first sentence and the second sentence. In this class are DMs but, well, alternatively, although, contrary to expectations, conversely, despite etc. In the third class are inferential DMs. According to Fraser (2009: 301) an inferential DM “signals that sentence 1 (S1) provides a basis for inferring sentence 2 (S2)”. So, all things considered, as a conclusion, then, therefore, thus are examples of inferential DMs.

The aforementioned distinctions have shown that scholars who utilise the coherence theory concentrate on indexical functions of DMs whilst those who use the relevance theory are mainly concerned with the contribution made by DMs in text processing and comprehension (cf. Aijmer, 2002). Given the indexical nature of DMs, in this study, the coherence framework is used as a basis for analysis because DMs are examined in the contexts in which they appear to check their relationship with other words.

Besides different definitions given by scholars, another issue that lacks unanimity involves the list of DMs (De Klerk, 2005a; Fraser, 2009, Fox Tree, 2010). Researchers include different items in their lists of DMs. For example, one of the ground-breaking works on DMs was done by Schiffrin (1987). Using the coherence theory as her framework for analysing DMs, oh, well and, but, or, so, because, now, then, I mean and y'know, she highlighted that:

My discourse model has both non-linguistic structures (exchange and actions) and linguistic structures (ideational). Speaker and hearer are related to each other and to their utterances, in a participation framework. Their knowledge and meta knowledge about ideas is organised and managed in an information state. Local coherence in discourse is thus defined as the outcome of joint efforts from interactants to integrate knowledge, meaning, saying and doing.

(Schiffrin 1987: 29)

It should be pointed out that from the list of DMs studied by Schiffrin, the DMs well, but and so are discussed in this study. This study will use Schiffrin’s (1987) definition of coherence because it analyses contextual links between sentences or larger paragraphs in discourse.
Detachability is characteristic of DMs according to Fraser (1993) and Norrick (2001) who explain that a DM is removable and can be omitted without affecting the grammar and content meaning of a sentence. To show that DMs can be removed from the sentence without affecting its grammaticality, Fraser (1993) gives the following sentences in which DMs are underlined:

7.1 A: *I like him.*
   B: **So**, you think you’ll ask him out then.

7.2 *John can’t go. And* Mary can’t go either.

7.3 A: *Did you like it?*
   B: **Well**, not really.

7.4 **But** when do you think he will really get here?

(Fraser, 1993: 1)

From the examples given above, the DMs *so and, well and but* can be removed from the sentences but the sentences will remain meaningful. Fraser (1993) also notes that, although removing a DM from a sentence will not make a sentence ungrammatical, it removes clues about the commitment made by the speaker with regards to the relationship between previous and current discourse. Another point made by Fraser is that DMs are drawn from a variety of grammatical categories such as verbs (e.g. *look, listen*), adverbs (e.g. *now, then*), literal phrases (e.g. *as a result, to repeat*), idioms (e.g. *by and large*), interjections (e.g. *well*), subordinate conjunctions (e.g. *however*).

From the analysis above, it can be deduced that researchers utilise different criteria when identifying DMs. Different terminology is used in literature to refer to DMs. “Discourse markers” is one of the terms used (e.g. Jucker, 1997; Schourup, 1999; Müller, 2005; De Klerk, 2005a, 2005b; Hussein, 2005; Blakemore, 2002; Ying, 2006; Fox Tree, 2010; Lee, 2017; Sarfo-Kantankah & Yussif, 2019). “Discourse connectives” is a term used by Blakemore (2002). Another term mentioned by scholars is discourse particles (Aijmer, 2002; Mesthrie & Bhatt, 2008; Lam, 2009a, 2009b). “Pragmatic markers” is a term utilised by Brinton (2010) and Oladipupo and Unuabonah (2020).
The aforementioned terms provide some insights into the reason why scholars include different items in their lists of DMs and provide different lists of DMs (e.g. Schiffrin, 1987; Fraser, 1993; Aijmer, 2002). Although there are a host of terms used for DMs as shown above, researchers seem to agree on the important contribution made by DMs to the pragmatic meaning of utterances” (Müller, 2005). The following section looks at different characteristics of DMs.

7.3 Characteristics of discourse markers

Researchers suggest that DMs have different characteristics. The following are characteristic of DMs are noted by Schourup, (1999) and Fraser (2006):

(i) DMs are optional in the sense that removing them from a sentence will not affect the grammaticality of a sentence.
(ii) DMs are made up of items from different grammatical categories.
(iii) DMs are free morphemes.
(iv) Occur at discourse-segment initial.
(v) Signal a specific message.
(vi) DMs are classified in terms of their semantic or pragmatic functions.
(vii) DMs connect discourse units.

In addition to the above list, some scholars suggest that DMs have a procedural core meaning (e.g. Fraser, 1993, 1999) whilst others consider DMs to be characteristic of speech (e.g. Brinton, 2010; Jucker, 1997; Lam, 2009a). Jucker (1997) states that although DMs are mainly found in spoken discourse, they are also found in written language. Another feature of English DMs is that the majority of DMs seem to occur in sentence initial position with some occurring in the middle and at the end of a sentence as shown in examples below:

7.5 Sentence initial: TS: Have you spoken to him?  
   AN: **Well**, not really [laugh].
   (De Klerk, 2005a: 1195)
7.6 Sentence middle: MV: *It’s your identity man.*
KM: *Hey!*
MV: *No, well it’s your belief anyway.*

(De Klerk, 2005a: 1196)

7.7 Sentence final: *<Ir>: Uhu uhu em any films that you’ve seen recently <\Ir>*
*<Ie>: Erm well I hate films so <\Ie>*
*<Ir> you hate [films <\Ir>*
*<Ie> [yeah I hate films <\Ie>*

(Buysse, 2012: 1769)

Fraser (1988) notes that DMs belong to the pragmatic category and can indicate basic communication intent, can be used as commentary and can provide parallel message. In addition, pragmatic meaning signals the different messages that a speaker wishes to relay. Below is a description of the functions of DMs.

### 7.4 Functions of discourse markers

When considering the functions of DMs, some researchers claim that DMs are multifunctional and that some of the functions of DMs include “contributing to local coherence of adjacent phrases, assisting in turn-taking or repair, or contributing to social solidarity” (Fox Tree, 2010: 270). Fox Tree further notes that DMs have been examined from different perspectives such as L2 acquisition of DMs and child acquisition of DMs.

In an investigation of DMs in native and non-native English, Müller (2005: 9) lists the following functions of DMs:

1. To initiate discourse.
2. To mark a boundary in discourse (shift/partial shift in topic).
3. To preface a response or a reaction.
4. To serve as a filler or delaying tactic.
5. To aid the speaker in holding the floor.
6. To affect an interaction or sharing between speaker and hearer.
7. To bracket the discourse either cataphorically or anaphorically.
8. To mark either foregrounded or backgrounded information.
Aijmer (2002) adds to the discussion by observing that although DMs have a core function, they are categorically multifunctional. On the issue of core meaning, De Klerk (2005a) remarks that researchers seem to agree that DMs have a general core meaning. Examples include *well, actually* (dissonance), *anyway* (reorientation), *similarly* (parallelism). The author is of the opinion that, despite the consensus on the DMs having a core meaning, there is still debate about what the core meanings are. The different functions of each DM examined in this study are listed next.

### 7.4.1 Functions of *so*

Different functions of the DM *so* are reported in literature (e.g. Lam, 2009b; Algouzi, 2021; Vickov & Jakupčević, 2017). One of the comprehensive studies of *so* as a DM was done by Müller (2005). She asserts that the DM *so* has several functions both at textual and interactional level. The functions given by Muller are also observed by Vickov and Jakupčević, (2017). The functions are shown in the below table together with example sentences.

*Table 54. Functions of *so* (Sources: Müller 2005; Vickov & Jakupčević, 2017)*

<table>
<thead>
<tr>
<th>Level</th>
<th>Function</th>
<th>Example sentences</th>
<th>Explanation of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>Marking result or consequence</td>
<td><em>Some guy was short on his bill, So they beat him up and kicked him out.</em></td>
<td><em>So</em> shows that the guy was beaten and kicked out because he did not pay his bill.</td>
</tr>
<tr>
<td></td>
<td>Main idea unit marker</td>
<td><em>He orders, So basically he orders like some beans.</em></td>
<td>The main idea about ordering is marked by the use of <em>so.</em></td>
</tr>
<tr>
<td></td>
<td>Summarizing/ rewording / giving an example</td>
<td><em>And think of your own examples, so don’t use examples from the book.</em></td>
<td><em>So</em> marks rewording of the first part of the sentence to clarify that the hearer is not supposed to use examples from the book.</td>
</tr>
<tr>
<td></td>
<td>Sequential <em>so</em></td>
<td><em>It’s raining outside. So they go around the corner.</em></td>
<td><em>So</em> shows the sequence of events from realising that it was raining to going around the corner.</td>
</tr>
<tr>
<td>Interactional</td>
<td>Boundary marker</td>
<td><strong>So</strong> what happened when you went out was that they arrived in America.</td>
<td>So is the boundary marker between types of talk. In this case, the hearer is being told that when she went out, the characters in the movie (that they were watching) arrived in America.</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Speech act marker – question or request</td>
<td>B: It’s that’s what happened in the B part, A: Mhm. B: . . . <strong>so</strong> what did you think of it?</td>
<td>The start of a question is marked by the use of <strong>so</strong>.</td>
<td></td>
</tr>
<tr>
<td>Speech act marker – opinion</td>
<td>Charlie an’ the lady walking inside of the marriage license door, <strong>So</strong> I assume they got married.</td>
<td>The speaker uses <strong>so</strong> to give an opinion about marriage.</td>
<td></td>
</tr>
<tr>
<td>Marking implied result</td>
<td>It’s ten cent short and <strong>so</strong> ... all the waiters beat him up.</td>
<td>The speaker uses <strong>so</strong> to show that due to the bill being ten cents short, he was beaten up.</td>
<td></td>
</tr>
<tr>
<td>Marker of a transition relevance place</td>
<td><strong>So</strong> now we talk about what was . . . what did we like?</td>
<td><strong>So</strong> is used to move to another topic.</td>
<td></td>
</tr>
</tbody>
</table>

The table above has integrated the different functions of **so** as provided by Müller (2005) and Vickov and Jakupčević (2017). Given the indexical nature of **so**, like all the other DMs, the possible number of functions is not limited (cf. Aijmer, 2002). The functions provided in the table above were used as bases for analysing the functions and the frequency of functions of **so** in the ZE corpus and the ICE-GB in Sections 7.6.2.1 and in subsection a in Section 7.6.1.4

### 7.4.2 Functions of well

As noted by Aijmer (2002), DMs are indexical in nature, making the number of functions that can be observed unlimited. This is true of the DM well. Some of the functions of **well** are provided by Jucker (1997: 91), who notes that **well** has different uses in English namely; “as a frame it introduces a new topic or prefaces direct reported speech; as a
qualifier it prefaces a reply, which is only a partial answer to a question; as a face-threat mitigator it prefaces a disagreement; and as a pause filler it bridges interactional silence”.

Touching on the subject of functions of the DM well, Kirk (2018) is of the opinion that well mostly occurs in turn-initial position in conversations with the function being response to what was said before and to introduce what comes later. “When well introduces the response, it signals that the answer is not going to be the expected one but a modified one, possibly even a divergent one. At the same time, well expresses the co-operation expected in the exchange and reduces any face-threat by doing so” (Kirk, 2018: 142). As she did with the functions of so as a DM, Müller (2005) suggested different functions of well in a comparative study of the use of different DMs in native and NNVE as shown below.

<table>
<thead>
<tr>
<th>Level</th>
<th>Function</th>
<th>Examples</th>
<th>Explanation of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>Searching for the right phrase</td>
<td>[And then] he just ends up on the streets of well I guess New York?</td>
<td>The speaker is using well to search for the right phrase to use, in this case, New York.</td>
</tr>
<tr>
<td></td>
<td>Rephrasing/correcting</td>
<td>And then they go outside, they say that- - well he says that tomorrow he’ll engage them in some sort of a … job.</td>
<td>The speaker uses well to rephrase the sentence.</td>
</tr>
<tr>
<td></td>
<td>Quotative well</td>
<td>And he’s like … well let’s play poker my way.</td>
<td>Well is used to quote direct speech “well let’s play poker my way”.</td>
</tr>
<tr>
<td></td>
<td>Move to the main story</td>
<td>Well um, what about… the acting. What do you think of the manor (sic!) character?</td>
<td>The move to the main story is signalled by the use of well.</td>
</tr>
<tr>
<td></td>
<td>Introducing the next scene</td>
<td>Well then … Chaplin’s thinking oh you know I’d better … find my money. And then this beggar comes in off the street.</td>
<td>The function of well in the sentence is to introduce the next scene.</td>
</tr>
<tr>
<td></td>
<td>Conclusive well</td>
<td>But um the girl finds the money in her pocket and is s- so happy and um um wants to thank um this tramp, so that um the crew member-- … well. now knows that this is no pickpocket, … well.</td>
<td>Well is used at the end of a sentence, to conclude.</td>
</tr>
<tr>
<td>Interactional</td>
<td>Indirect answer</td>
<td>\textit{Well} the minor characters didn’t do anything.</td>
<td>The speaker starts a reply with \textit{well} and gives an indirect answer to the question.</td>
</tr>
<tr>
<td>Direct answer</td>
<td>\textit{Well} I liked it cause it was just very comical.</td>
<td>The speaker starts a reply with \textit{well} and gives a direct answer to the question.</td>
<td></td>
</tr>
<tr>
<td>Response to self-raised expectations</td>
<td>\textit{Well} you gotta like Charlie Chaplin movies in order to like this movie.</td>
<td>The speaker is treating what he said before as something that the hearer needs to respond to.</td>
<td></td>
</tr>
<tr>
<td>Contributing an opinion</td>
<td>\textit{Well} also like when .. when they rocked back and forth, they could be only be moving the camera and not the people.</td>
<td>\textit{Well} is used to add a contribution to an opinion.</td>
<td></td>
</tr>
<tr>
<td>Continuing an opinion/answer</td>
<td>B: Mhm they [just did it maybe once or twice], A: [. . . right mhm]. B: Mhm. A: And \textit{well} he cuts .. this very very .. good.</td>
<td>The speaker uses \textit{well} to continue giving an opinion.</td>
<td></td>
</tr>
<tr>
<td>Evaluating a previous statement</td>
<td>B: I don’t really like .. um .. silent films. A: I don’t either. ...but-- .. it was OK for .. as .. a silent film goes, B: Yeah. \textit{Well} I’ve actually never seen one.</td>
<td>The speaker starts a statement that he does not like silent films. He then evaluates the statement and then uses \textit{well} to say that he has never seen a silent film.</td>
<td></td>
</tr>
</tbody>
</table>

The aforementioned discussions about the functions of \textit{well} as a DM shows that researchers list different functions. It was shown that due to the indexical nature of DMs, providing a clear-cut number of functions of \textit{well} is not feasible. For the purpose of this study, the analyses of the functions of \textit{well} in the ZE corpus and the ICE-GB were based on the functions outlined by Müller (2005). The functions provided in the table above were used as bases for analysing the functions and the frequency of functions of \textit{well} in the ZE corpus and the ICE-GB in Section 7.6.2.2, and subsection b in Section 7.6.1.4.
7.4.3 Functions of but

In literature, the functions of but are given as contrastive and sequentially contrastive as will be discussed below. Fraser (1999) uses the following example to show but as a DM.

7.8 Sue left very late. But she arrived on time.

(Fraser, 1999: 931)

Fraser (1993) observes that but functions as a contrastive DM, which indicates a sense of dissonance and provides the following examples:

7.9 Son (whining): I can’t do it. Father: But I know that you CAN do it.

(Fraser, 1993: 14)

7.10 Job interviewer: The position has been filled. But do come in anyway and talk for a minute.

(Fraser, 1993: 14)

7.11 Witness: I didn’t think I should talk about it. Attorney: But what did you actually say?

(Fraser, 1993: 14)

Bell (1998) asserts that but is sequentially contrastive when it cancels what is expected in the next part of the discourse. The purpose of but as a sequential DM is to show a return to the main topic or point as shown in the example below:

7.12 Suddenly, his telephone is ringing with producers interested in his next project. But what most delights him is that Americans will see his film.

(Bell 1998: 530)

The functions provided above were used as bases for analysing the functions and the frequency of functions of but in the ZE corpus and the ICE-GB in Sections 7.6.2.3 and subsection c in Section 7.6.1.4. With the above assertion in mind, this study set out to examine whether there are variations in the way DMs are used in ZE by comparing the use of DMs in the ZE corpus and ICE-GB. This was done by exploring the frequency and functions of so, well and but. This section has unpacked the different functions of so, well
Examples from literature were also used. The next section discusses the use of DMs in L2 varieties of English.

### 7.5 Discourse markers in L2 English

One of the areas explored by researchers in L2 varieties of English is whether there are variations in the use of DMs by L2 speakers of English. Results from some studies show variability in the use of DMs in L2 varieties of English whilst other studies report that there is no variability. Below are some of the studies that focus on the use of DMs in L2 English.

As noted in Section 1.6.3, the study of DMs in L2 varieties of English has been scarce until recently (cf. Buysse, 2010, 2012; Diskin, 2017). Mesthrie and Bhatt (2008) note that some studies show innovation in the use of DMs in WE varieties. Another point raised by Mesthrie and Bhatt (2008) is that WE not only show structural differences from Standard varieties but there are considerable differences between English varieties in terms of linguistic structure. The authors attribute social, cultural and historical diversity to the differences reflected in the way English is used in different communities. To have a better understanding about DMs in WE and to situate this study within context, some of the studies that have examined DMs in L2 varieties of English are summarised below.

De Klerk (2005a, 2005b) focused on Xhosa English. De Klerk’s (2005b) investigation of the DM “actually” used a spoken corpus drawn from Xhosa first language speakers who use ESL. She noted that Xhosa English speakers generally use the DM actually the same way as native speakers except when actually is used as a contemplative and to mark disagreement by Xhosa English speakers. In another study of DMs in Xhosa English, De Klerk (2005a) explored the DM well. She reported that there is diversified use of well by Xhosa English speakers. In addition, De Klerk (2005a) noted that there was less frequency in the DM well compared to the frequency in native speaker discourse.

How text type influences the use of DMs is the focal point of Lam (2009b). The author studied the DM so in different contexts in Hong Kong English in order to determine the
different functions of DMs. One of the findings of this study is that the frequency and functions of DMs in Hong Kong English vary according to text type.

In a comparative study between Saudi English and BrE, Algouzi (2014) noted that BrE speakers used the DMs so and like more frequently than Saudi English speakers. In addition, the DM you know was used more frequently by Saudi English speakers than by BrE speakers.

Gough (1996a) reports on one of the characteristic features of BSAE as the idiosyncratic use of DMs. He gives an example of in fact, which indicates emphasis in L1 conversations but shows topic change or as a topic-initiation marker as shown in the sentence below:

7.13 A. Hello doctor.
   B: Hi, Vuyisile.
   (pause)
   A: In fact, I want to talk to you about my essay.

   (Gough, 1996a: 66)

In the example above, Gough attributes the use of idiosyncratic in fact to mother tongue influence (in this case Xhosa).

Unuabonah, (2019) studied the stylistic variability in Nigerian English. Results from the study showed that DMs in Nigerian English and BrE varied significantly in terms of frequency and stylistic variability. Similarities were also noted in the frequency of elaborative DMs in both varieties of English.

In a corpus-driven study, Trillo (2002) explored pragmatic fossilisation of DMs in non-native English speakers and noted the differences in the use of DMs between the two groups.

To show that there may be innovations in the way DMs are used in WE, Mesthrie and Bhatt put it succinctly when they comment that:
An innovation structure in a particular New English may serve a new function or add a certain nuance not generally found in other varieties. Or what appears to be the same structure as in, say, StE, may serve a new function.

(Mesthrie & Bhatt, 2008: 131)

Müller (2005) studied the use of DMs by German EFL speakers and American native speakers in a movie-telling experiment and noted that when comparing native and non-native data, the former has more functions of DMs and the latter has fewer.

In order to ascertain whether there are differences in the use of DMs by non-native speakers of English, Fuller (2003) analysed the use of DMs *oh, well, y’know, like and I mean* in native and non-native speaker interviews and conversations. Findings from her study show that there are low occurrences of DMs in non-native speaker data. Similarly, Jabeen, Rai and Arif (2011) reported that non-native speakers utilise fewer DMs compared to native speakers in a comparative study of the DMs *I mean, you know, I think, kind of, sort of, well, you see and so* in Pakistani English and BrE.

Some researchers note that there are similarities in the use of DMs by L1 and L2 users of English. For example, using corpora of political discourse, Dalili and Dastjerdi (2013) explored whether there are differences between native and non-native English usage of DMs. They observed that there are no notable differences in the relative frequencies of the DMs studied.

The examples of studies given above illustrate a growing list of studies of DMs in L2 English.

As can be gleaned from the aforementioned studies, similarities and differences exist in the use of DMs in L2 English compared to L1 English. The following section will provide the results of DMs research in this study.
7.6 Results

7.6.1 Quantitative data analysis of discourse markers

Section 7.6.1 answers part of research questions 1.1, 1.2, 2.1 and 2.2 by focusing on the occurrence- and function frequencies of discourse markers and determining whether there are variations in the use of discourse markers. The research questions are shown for ease of reference below.

1.1 What are the occurrence frequencies of discourse markers in the ZE corpus and the ICE-GB?
1.2 How does the occurrence of discourse markers in ZE compare to that of BrE?

2.1 What are the function frequencies of discourse markers in the ZE corpus and the ICE-GB?
2.2 How do the functions of discourse markers in ZE compare to those of BrE?

7.6.1.1 Occurrence frequencies of the discourse markers

This section explores the frequencies and functions of DMs so, well and but. The three DMs were chosen because, in the case of ZE, there is no study that has done a comparative analysis of DMs between ZE and a L1 variety of English. This study is a small contribution towards a better understanding of whether there is variation in the frequency and use of DMs by doing a comparative analysis of the frequencies and functions of the three DMs in order to answer the research questions posed. The frequencies of all DMs were not searched because the ZE corpus is not tagged for DMs. In addition, a comparison of functions identified by other researchers on L2 English was made. The corpora were queried to determine whether there is variability in the use of DMs in the ZE corpus when compared to the ICE-GB. The DMs so, well and but were manually checked and instances

57 Analysing the frequencies of all DMs in the two corpora was outside the scope of this study.
where the words did not function as DMs were not considered for analysis (c.f. Unuabonah, 2019).

For comparison purposes, only genres which occurred in both the ZE corpus and the ICE-GB were used (cf. Section 4.4.2.6). In order to achieve consistency and for comparison purposes, the frequencies of *so*, *well* and *but* were normalised per 10 000 words. Since the ZE corpus and the ICE-GB are of different sizes, the number of occurrences of the DMs *so*, *well* and *but* was normalised to occurrences per 10 000 words. Normalisation “involves extrapolating raw frequencies from the different-sized corpora, which are being compared so that they can be expressed by a common factor such as a thousand or a million words” (Evison, 2010: 126). The different calculations that were done to normalise the frequencies are given in each section where normalisation was done. The log likelihood test was used to determine whether the observed variations were statistically significant (cf. subsection c Section 4.4.2.6). One asterisk (*) is used for $p < 0.05$, where $LL > 3.84$. Two asterisks (**) are used for $p < 0.01$, where $LL > 6.63$. Three asterisks (*** are used for $p < 0.001$, where $LL > 10.83$. Four asterisks (****) are used for $p < 0.0001$, where $LL > 15.13$.

The frequency of the DMs in the ZE corpus and the ICE-GB was compared. This process was aided by the use of concordance, wordlist and word sketch tools in Sketch Engine corpus analysis software. Word sketch processes words adjacent to the search word and the results are arranged in a clear format that can be easily understood. Some of the information displayed include a word’s frequency, collocates and grammatical relations (www.sketchengine.eu/).

The next section will analyse the occurrence frequencies of the DMs in the corpus.
7.6.1.2 Occurrence frequencies of three discourse markers (so, well and but) for both corpora with Log likelihood comparison

In the ICE-GB, the corpus is tagged for DMs with the tagset discourse marker (DISMK). Therefore, when the three words were searched, results showed only instances where the so, well and but occurred as DMs. The ZE corpus is not tagged for DMs. Instances where so, well and but did not function as DMs were not included for analysis. This is explained further in Section 7.6.1.3. The summary of absolute frequency and normalised frequency is provided in table 56 below. Due to the fact that the two corpora compared differ in size, it was necessary to normalise the frequencies. Fuchs, Van Rooy and Gut (2019) note that if corpora are of different sizes, the frequencies of a particular phenomenon studied need to be normalised in order to reach better conclusions. As a way of comparing between ZE and BrE, the number of occurrences of the DMs so, well and but was normalised by dividing the total number of occurrences in a corpus by the total number of words in the corpus and then multiplied by 10 000 as shown below.

\[
\text{Total number of each DM in a corpus} \times 10 000
\]

\[
\text{Total number of words in a corpus}
\]

Results for the normalised values for each DM studied are shown in the table below.

<table>
<thead>
<tr>
<th>Discourse marker</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
</tr>
<tr>
<td>So</td>
<td>1 588</td>
<td>44.6</td>
</tr>
<tr>
<td>Well</td>
<td>231</td>
<td>6.5</td>
</tr>
<tr>
<td>But</td>
<td>890</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>2 709</td>
<td>76.1</td>
</tr>
</tbody>
</table>

When ZE and BrE are compared, the normalised frequencies show some differences with so occurring in the ZE corpus almost twice (44.6 per 10 000 words) the rate in the ICE-GB (29 per 10 000 words). The differences observed are statistically significant because the log LL value of 156.95 is more than the critical value 15.13, where \( p < 0.0001 \). This
means that there are variations in the frequency of *so* in the two corpora. *Well* is more prevalent in the ICE-GB than in the ZE corpus as shown by the normalised frequency of 32 per 10,000 words compared to 6.5 per 10,000 words for the ZE corpus, meaning that there are variations in the frequency of *well*, with the ZE corpus recording a higher frequency. This result is statistically significant because the LL value of 782.81 is more than the critical value 15.13 where *p* < 0.0001. There are variations in the frequency of *but*, with the ICE-GB recording more instances per 10,000 words (29) compared to the ZE corpus (25). The variation in the frequency of *but* between the two corpora is statistically significant because the LL > 15.13, where *p* < 0.0001.

a. **Occurrence frequencies of *so* across registers for both corpora with Log likelihood comparison**

To get a better picture about the distribution of *so* in the spoken and written registers, an illustration is made in the table below. The normalised frequencies for registers were calculated by dividing the total number of each DM in a register by the total number of words in a register and then multiplying by 10000 as shown below.

<table>
<thead>
<tr>
<th>Register</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>1484</td>
<td>63.2</td>
<td>1556</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>8.6</td>
<td>196</td>
</tr>
<tr>
<td>Written</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1588</td>
<td>44.6</td>
<td>1752</td>
</tr>
</tbody>
</table>

Although the ZE corpus and the ICE-GB are different in size, normalising the frequencies helps with comparisons between the two corpora. When comparisons are made on the written register, no notable variations can be observed in the frequency per 10,000 words because in ZE, *so* occurs 8.6 instances per 10,000 words whilst in BrE *so* occurs 9.4 instances. The picture is a bit different in the spoken register because the normalised
frequency of so in ZE is 63.2 instances per 10 000 words, which is higher than the 39 instances in BrE.

In order to better understand whether there are statistically significant differences in the frequency of the DM so, the LL value was calculated. In the spoken register, the LL value is 174.99. This value is higher than the critical value margin of 15.13 where \( p < 0.0001 \). Since the LL value is higher, we can therefore conclude that there are statistically significant differences in the overall frequency of the DM so in the spoken register in the ZE corpus and the ICE-GB. The written register does not show statistically significant differences because the LL < 3.84, which is the critical value for \( p < 0.05 \).

b. Occurrence frequencies of well across registers for both corpora with Log likelihood comparison

A better picture regarding the frequency of well in the spoken and written registers can be gleaned from the table below.

<table>
<thead>
<tr>
<th>Register</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>206</td>
<td>8.8</td>
<td>1 888</td>
</tr>
<tr>
<td>Written</td>
<td>25</td>
<td>2.1</td>
<td>61</td>
</tr>
<tr>
<td>Total</td>
<td>231</td>
<td>6.5</td>
<td>1 949</td>
</tr>
</tbody>
</table>

Results above imply that there are no notable variations in the normalised frequency in the written register in ZE and BrE (2.1 and 2.9 instances per 10 000 words respectively). There are variations in the spoken register, which shows that BrE has more instances of well per 10 000 words (47.3 compared to 8.8 in ZE). As was observed with the LL value for so, the LL value for well in the spoken register is 809.71, which is far much higher than the critical value margin of 15.13 where \( p < 0.0001 \). This is evidence of statistically significant variations.
c. Occurrence frequencies of but across registers for both corpora with Log likelihood comparison

The frequency of but in spoken and written registers is shown in table 59 below.

<table>
<thead>
<tr>
<th>Register</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>793</td>
<td>33.8</td>
<td>1 487</td>
</tr>
<tr>
<td>Written</td>
<td>97</td>
<td>8</td>
<td>302</td>
</tr>
<tr>
<td>Total</td>
<td>890</td>
<td>25</td>
<td>1 789</td>
</tr>
</tbody>
</table>

The results presented in the table above are interesting because they show notable differences in the normalised frequency per 10 000 words in both registers. The normalised frequency in the spoken register in the ZE corpus is 33.8 whilst in the ICE-GB it’s 37.2. In the written register, the normalised frequency per 10 000 words is 8 whilst in ICE-GB the figure is 14.5. The LL value of but in the spoken register is statistically significant at 28.04, considering that the critical value for p < 0.0001 is 15.13. The spoken register also shows statistically significant variations since the LL > 3.84, where p < 0.05.

7.6.1.3 Occurrence frequencies of discourse markers (so, well and but) across genres for both corpora with Log likelihood comparison

In this section, the absolute and normalised frequencies of so, well and but in different genres are reported for the ZE corpus and the ICE-GB. The LL values are also calculated in order to determine whether the reported variations are statistically significant. The number of occurrences of the DMs so, well and but in a genre were normalised by dividing the total number of occurrences of each DM in a genre by the total number of words in the genre and then multiplied by 10 000 as shown below.

\[
\text{Total number of each DM in a genre} \times 10 000
\]

\[
\text{Total number of words in a genre}
\]
a. **Occurrence frequencies of so across genres for both corpora with Log likelihood comparison**

Considering that this study focused on the use of *so* as a DM, propositional meanings of *so* such as a substitute form, a conjunction of purpose and a modifying adverb, (cf. Müller, 2005; Lam, 2009b) respectively shown in the examples below were not discussed because they were not part of the study. This was done using the Skema manual concordance annotation tool in Sketch Engine. All the concordance lines of *so* were manually annotated with the tagset “Not DM” to indicate instances where *so* was not used as a DM. Annotations are saved immediately to be accessed from the dashboard. The examples given below (7.14, 7.15 and 7.16) illustrate some of the uses of *so* that were not discussed.

7.14 Mm *yeah I think so*. (ZE: Private semi-scripted dialogue 3)

7.15 *We produce results so that when they are ready to distribute funds for the next round, Zimbabwe is in a favourable position*. (ZE: Newspaper reportage 1)

7.16 *And her kids were so happy*. (ZE: Private semi-scripted dialogue 43)

The discussion was centred on instances where the propositional content of a sentence does not change when *so* is left out or is excluded as illustrated below:

7.17 *So I thought we can just go straight to section fifty-one the right to human dignity*. (Public dialogue 31)

7.18 *So approximately half of these deaths can be prevented*. (ZE: Public dialogue 38)

7.19 *So I will try to do everything asap, just expect the parcel*. (ZE: Social letters 2)

In order to obtain normalised frequencies of each DMs per genre, the total number of occurrences in the genre were divided by the total number of words in a genre and then multiplied by 10 000 as shown below.

\[
\text{Total number of each DM} \times 10000
\]

\[
\text{Total number of words in a genre}
\]
The table below illustrates the absolute and normalised frequency, the frequency per 10 000 words in different genres and the LL values.

Table 60. Frequency of so across genres

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus tokens</th>
<th>ICE-GB tokens</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>1136</td>
<td>86.1</td>
<td>902</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologue</td>
<td>14</td>
<td>33.4</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>334</td>
<td>33.8</td>
<td>615</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>22</td>
<td>12.3</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>1</td>
<td>3.6</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Exam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>8</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>23</td>
<td>4.7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>29</td>
<td>57.9</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>21</td>
<td>11.6</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1 588</td>
<td>44.6</td>
<td>1 752</td>
</tr>
</tbody>
</table>

If the 10 genres analysed are considered, statistically significant differences can be observed in private dialogue, public scripted monologue, creative writing and social letters in the two corpora. The highest LL value (164.51, where p < 0.0001), is found in private dialogue where so occurs more frequently in the ZE corpus (86.1 times per 10 000 words) compared to the ICE-GB (48.7 times). There are no statistically significant variations in the frequencies in public dialogue, popular writing, newspaper reportage, academic writing and editorials. This is because the LL values in these genres are below the critical value of 3.84, where p < 0.05.

b. Occurrence frequencies of well across genres for both corpora with Log likelihood comparison

In this study, instances of well that did not serve the DM purposes were not included. As mentioned in Section 7.6.2.3, the ICE-GB is annotated for DMs. Therefore, there was no need for manual annotation. In the ZE corpus, this was done using the Skema tool by manually checking and annotating the corpus for all instances where well was not used.
as a DM and excluding the instances that, according to Müller (2005) show “in addition”, adverbial and conjunction use of well are illustrated in examples 7.20, 7.21 and 7.22 respectively.

7.20 We would like to thank them for a job well done. (ZE corpus: Editorials 5)

7.21 They performed very well. (ZE corpus: Private semi-scripted dialogue 26)

7.22 We would like to thank all parents who have send their children back as well as those who have enrolled their children with us. (ZE corpus: Private business letters 2)

Examples 7.20, 7.21 and 7.22 above illustrate some of the functions of well that were excluded because they were not DMs. Well was used 251 times in the ZE corpus, of which 231 times were DMs. In comparison, well was used 2 877 times of which 2 102 times were DMs. The frequencies across genres are indicated in table 61.

Table 61. Frequency of well across genres

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/ 10 000 words</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>98</td>
<td>7.4</td>
<td>1287</td>
</tr>
<tr>
<td></td>
<td>Public scripted</td>
<td>4</td>
<td>9.5</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>monologue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>104</td>
<td>10.5</td>
<td>557</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>7</td>
<td>3.9</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>6</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>2</td>
<td>1.1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>231</td>
<td>6.5</td>
<td>1949</td>
</tr>
</tbody>
</table>

It can be noted that the normalised frequency of the DM well shows some variations when considering ZE and BrE. This is evident in private dialogue and public dialogue, which show well as being more prevalent in BrE than in ZE (69.5 and 32.6 instances per 10 000 words compared to 7.4 and 10.5 respectively). A significant difference can be noted in the
private dialogue genre where the LL value of well (848.58) is way higher than the critical value for p < 0.0001, which is 15.13. The LL value for public dialogue is also higher (141.37) than the critical value for p < 0.0001. The newspaper reportage genre shows statistically significant differences, with ZE recording 2 instances per 10 000 words compared to 0.5 instances per 10 000 words for the ICE-GB. There are no statistically significant differences in public scripted monologues, creative writing, business letters and social letters. Another observation worth noting is that both corpora do not have any instances of the DM well in popular writing, academic writing and editorials. This finding can be linked to the fact that some scholars consider DMs to be mainly found in the spoken discourse (e.g. Brinton, 2010; Jucker, 1997; Lam, 2009a).

a. Occurrence frequencies of but across genres for both corpora with Log likelihood comparison

An analysis of both corpora showed the frequencies of but across different genres as manifested in the table below.

<table>
<thead>
<tr>
<th>Register</th>
<th>Genre</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
</tr>
<tr>
<td>Spoken</td>
<td>Private dialogue</td>
<td>640</td>
<td>48.5</td>
<td>751</td>
</tr>
<tr>
<td></td>
<td>Public scripted monologues</td>
<td>4</td>
<td>9.5</td>
<td>148</td>
</tr>
<tr>
<td></td>
<td>Public dialogue</td>
<td>149</td>
<td>15.1</td>
<td>751</td>
</tr>
<tr>
<td>Written</td>
<td>Creative writing</td>
<td>30</td>
<td>16.8</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Popular writing</td>
<td>6</td>
<td>21.7</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Academic writing:</td>
<td>25</td>
<td>9.5</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Examination</td>
<td>32</td>
<td>6.5</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Editorials</td>
<td>4</td>
<td>8</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Newspaper reportage</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Social letters</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Business letters</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>890</td>
<td>25</td>
<td>1 789</td>
<td>29</td>
</tr>
</tbody>
</table>
The table above reveals that there are statistically significant variations in the frequency of the DM *but* in six out of 10 genres analysed. In the ZE corpus, the normalised frequency for private dialogue is 48.5 per 10 000 words. In contrast, the normalised frequency in ICE-GB is 31.7. This indicates that *but* as a DM is more frequent in the ZE corpus than in the ICE-GB. Public scripted monologues show that the normalised frequency per 10 000 words in ZE is less (i.e. 9.5) than in ICE-GB (34.4). When considering public dialogues, the normalised frequency per 10 000 words is 15.1 and in ICE-GB the figure is 43.9. This indicates that *but* occurred more in ICE-GB than in ZE. Similar patterns can be observed in editorials and newspaper reportage where the normalised frequencies are higher in ICE-GB (29.4 and 22.1 respectively) than in the ZE corpus (9.5 and 6.5 respectively).

Considering that in this study, a critical value of 3.84 or higher was considered to be significant at the level of $p < 0.05$, the LL values of *but* across different text genres are higher than 3.84 in six of the ten genres. Statistically significant variations are noted in public scripted monologue (9.89) and popular writing (7.95) where $p < 0.01$. Private dialogue, public dialogue, editorials and newspaper reportage all recorded LL values that are higher than the critical value of 15.13 where $p < 0.0001$ as shown by the four asterisks (***) after each LL value.

### 7.6.1.4 Function frequency of discourse markers

In this section, the frequencies of functions of *so*, *well* and *but* are reported. The section consists of three sub-sections. The frequencies of functions of *so* *well* and *but* *are* reported in subsections a, b and c respectively.
a. Function frequency of so for both corpora with Log likelihood comparison

The table below shows a comparison of the frequencies of the different functions of so.

**Table 63. Frequencies of functions of so**

<table>
<thead>
<tr>
<th>Level</th>
<th>Function</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Textual</td>
<td>Marking result or consequence</td>
<td>934</td>
<td>26.2</td>
<td>793</td>
</tr>
<tr>
<td></td>
<td>Main idea unit marker</td>
<td>162</td>
<td>4.6</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>Summarizing / rewording / giving an example</td>
<td>132</td>
<td>3.7</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>Sequential so</td>
<td>20</td>
<td>0.6</td>
<td>272</td>
</tr>
<tr>
<td></td>
<td>Boundary marker</td>
<td>56</td>
<td>1.6</td>
<td>88</td>
</tr>
<tr>
<td>Interaction</td>
<td>Speech act marker – question or request</td>
<td>189</td>
<td>5.3</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>Speech act marker – opinion</td>
<td>52</td>
<td>1.5</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Marking implied result</td>
<td>12</td>
<td>0.3</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Marker of a transition relevance place</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Unclassified</td>
<td>Unclassified = Unfinished sentences, repetitions</td>
<td>30</td>
<td>0.8</td>
<td>34</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1 588</td>
<td>44.6</td>
<td>1 752</td>
</tr>
</tbody>
</table>

As gleaned from the table above, the function of the DM so that has the highest statistical significance is marking result or consequence with a LL value of 209.45. This function is attested more in ZE with a normalised frequency of 26.2 per 10 000 words compared to BrE with 13. The second most statistically significant difference is observed in the sequential so function, which has a LL value of 144.68. Main idea unit marker, summarising/rewording/giving example and speech act marker - question or request are also above the critical value for $p < 0.0001$, which is 15.13. Interestingly, ZE speakers use so to mark result or consequence, to summarise, as a main idea unit marker and as a speech act marker for a question or a request more than BrE speakers.
It has been reported in Section 7.6.2 that the challenging task of classifying functions of DMs is caused by the multifunctional nature of DMs. In this study, there are 16 unclassified instances where so did not fit into the functional categories. In the ICE-GB, the number of unclassified instances is 44 Some instances of so in spoken texts could not be classified because they occurred in contexts where speech was inaudible or interrupted. With regard to written texts, some words were deleted or incomplete. Repetitions of so were also put in the unclassified instance group if they did not have a different function from that of the adjoining so.

b. Function frequency of well for both corpora with Log likelihood comparison

Below is a table showing the different functions of well and how they are attested in the two corpora.

<table>
<thead>
<tr>
<th>Level</th>
<th>Function</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textual</td>
<td>Searching for the right phrase</td>
<td>55 1.5</td>
<td>611 10</td>
<td>-293.01****</td>
</tr>
<tr>
<td></td>
<td>Rephrasing/correcting</td>
<td>16 0.4</td>
<td>231 3.8</td>
<td>-126.25****</td>
</tr>
<tr>
<td></td>
<td>Quotative well</td>
<td>5 0.1</td>
<td>34 0.6</td>
<td>-11.43***</td>
</tr>
<tr>
<td></td>
<td>Move to the main story</td>
<td>13 0.4</td>
<td>129 2.1</td>
<td>-57.85****</td>
</tr>
<tr>
<td></td>
<td>Introducing the next scene</td>
<td>18 0.5</td>
<td>33 0.5</td>
<td>-0.06</td>
</tr>
<tr>
<td></td>
<td>Conclusive well</td>
<td>4 0.1</td>
<td>12 0.2</td>
<td>-1.04</td>
</tr>
<tr>
<td>Interaction</td>
<td>Indirect answer</td>
<td>23 0.6</td>
<td>126 2.1</td>
<td>-33.75****</td>
</tr>
<tr>
<td></td>
<td>Direct answer</td>
<td>57 1.6</td>
<td>549 9</td>
<td>-241.59****</td>
</tr>
<tr>
<td></td>
<td>Response to self-raised</td>
<td>0 0</td>
<td>0 0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>expectations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contributing an opinion</td>
<td>8 0.2</td>
<td>43 0.7</td>
<td>-11.26***</td>
</tr>
<tr>
<td></td>
<td>Continuing an opinion/answer</td>
<td>13 0.4</td>
<td>117 1.9</td>
<td>-49.21****</td>
</tr>
<tr>
<td></td>
<td>Evaluating a previous statement</td>
<td>3 0.1</td>
<td>20 0.3</td>
<td>-6.60*</td>
</tr>
<tr>
<td>Unclassified</td>
<td>Unclassified = Unfinished sentences e.g. Repetitions</td>
<td>16 0.4</td>
<td>44 0.7</td>
<td>-2.84</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>231 6.5</td>
<td>1949 32</td>
<td>-782.81****</td>
</tr>
</tbody>
</table>
Statistically significant differences in the frequencies of functions of *well* can be seen in the table above. BrE shows more normalised frequencies of *well*, with searching for the right phrase having the highest normalised frequency of 10 per 10,000 words compared to 1.5 for ZE. Direct answer is the second highest statistically significant function of *well* in BrE with a normalised frequency of 9 whilst ZE has 1.6. Other functions that show statistically significant differences at $p < 0.0001$ between the two corpora are; rephrasing/ correcting (126.25), move to main story (57.85), indirect answer (33.75) and continuing an opinion or answer (49.21) as shown by the **** after the LL number in the table.

c. Function frequency of *but* for both corpora with Log likelihood comparison

<table>
<thead>
<tr>
<th>Function</th>
<th>ZE corpus</th>
<th>ICE-GB</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute frequency</td>
<td>Normalised frequency/10 000 words</td>
<td>Absolute frequency</td>
</tr>
<tr>
<td>Contrastive</td>
<td>689</td>
<td>19.4</td>
<td>1 172</td>
</tr>
<tr>
<td>Sequentially contrastive</td>
<td>184</td>
<td>5.2</td>
<td>581</td>
</tr>
<tr>
<td>Unclassified</td>
<td>17</td>
<td>0.5</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>890</td>
<td>25</td>
<td>1 789</td>
</tr>
</tbody>
</table>

Of the two functions of *but*, the sequentially contrastive functions show statistically significant differences between the two corpora, with ICE-GB showing a normalised frequency of 9.6 per 10,000 words compared to 5.2 for the ZE corpus. The LL value for sequentially contrastive *but* is 58.03, which is above the critical value for $p < 0.0001$, which is 15.13. There are no statistically significant differences in the contrastive function of *but* because the LL value of 0.01 is less than the critical value of 3.84 where $p < 0.05$.

The frequencies of functions of *so, well* and *but* have been discussed in this section. This was done by reporting on the normalised frequencies per 10,000 words. The LL values which show whether the observed variations are statistically significant were calculated. A summary of the quantitative and qualitative data is provided next.
7.6.2 Qualitative analysis of discourse markers

This section is divided into three sub-sections. In subsections 7.6.2.1, 7.6.2.2 and 7.6.2.3, examples (from the ZE corpus and the ICE-GB) of the different functions of so, well and but are given respectively.

The task of assigning functions of DMs is not well defined and uncomplicated due to the fact that “natural language is not a sharp instrument with absolute or rigid boundaries, but is blurred at the edges” (Sinclair & Mauranen, 2006: 61). Another impediment to functional analysis of DMs is the fact that DMs are multi-functional in nature, making categorisation of functions difficult (Lam, 2009b). However, searching for specific DMs as KWIC and examining the contexts in which they appear and the prosodic clues helps with determining the functions of DMs (Holmes, 1984). That is why some prosodic features such as filled and unfilled pauses and non-speech sounds such as laughing and coughing were transcribed. In this study, the functions of so, well and but were analysed qualitatively by checking the concordance lines to see the contexts in which the DMs were used (KWIC). After the DMs were assigned functional categories, the frequency counts of each functional category were reported quantitatively. A statistical test was used to check whether the observed differences were statistically significant.

In order to explore the patterns of usage of DMs so, well and but in ZE, the corpus was queried using Sketch Engine software tools. A comparison was done between the ZE corpus and the ICE-GB. The concordance lists of the three DMs were generated to determine the contexts in which the DMs occurred.

The research investigated whether there are variations in the functions of DMs in ESL in Zimbabwe and BrE. The goal was to establish whether variations can be attested in the ZE corpus and the ICE-GB. The functions reported by Müller (2005) were used as a guideline. Results from the data analysis are reported next.
7.6.2.1 Functions of so

The functions of DMs were analysed qualitatively by checking concordance lines and using the Skema tool to assign functional labels to each occurrence of DMs.

a. Textual level functions

i. Marking result or consequence

An analysis of the corpora shows that both ZE and BrE use the DM so to mark result or consequence. Let us consider two examples below from the ICE-GB:

7.23 Uhm uh Chris has done that and we’ve done that when we’ve moved or he’s moved on. I mean he’s done different things so like tons of psychology books went because now he’s moving on or teaching something else. So he passed them on to somebody else. (ICE-GB: S1A-013 #216-220)

7.24 The principal is now in Australia but her Secretary’s going to be away too so I’ll be running between two offices. (ICE-GB: W1B-001 #185)

From example 7.23 above, so marks he passed them on to somebody else as a result of Chris moving on or teaching something else. In example 7.24, so indicates running between two offices as a consequence of the principal being in Australia and her Secretary being away.

The marking of result or consequence is also evident in the ZE corpus as shown in the examples below.

7.25 I failed to secure a Res, so I’m squatting with my former Res-mate in New Complex 5 and I will contribute towards the payment of the accommodation fee of $8000. (ZE corpus: Social letters 1)

7.26 <$NN>: I didn’t throw a party. I I was uh away. I didn’t have any close relatives to celebrate with. <#>Uh so there was no party. (ZE corpus: Private semi-scripted dialogue 40)

In example 7.25 it can be deduced that so marks I’m squatting with my former Res-mate in New Complex 5 as a result of not securing res. So marks there being no party as a result of being away and not having any close relatives to celebrate with in extract 7.26.
ii. Main idea unit marker

The main idea unit marker is illustrated below.

7.27 (A): Is that an irritation when you have a vague feeling you've lent a book to somebody and you can't quite figure it out.
   (A): It's not there.
   (E): If it's a paperback no.
   (E): If it's a hardback.
   (E): And on occasion I have lent hardbacks and not got them back.
   (E) **So** I just don't lend hardbacks to anyone now. (ICE-GB: S1A-013 #092-097)

In the above example, participant (E) first answers the question; *is that an irritation when you have a vague feeling you've lent a book to somebody and you can't quite figure it out?* with explanations that he doesn’t lend paperbacks and hardbacks. The answer to the question that is the main idea comes in the last sentence with the DM *so* marking the main idea *I don’t lend hardbacks to anyone now.*

Another illustration for main idea unit marker in the ICE-GB is shown below:

7.28  (B): I do appreciate you seeing me so quickly though.
   (A): That’s ok.
   (A): About five minutes a day use the mouthwash.
   (A): Just repeat that.
   (B): Oh.
   (B): Uhm.
   (A): Equal spacing *so* basically after each meal and uhm uh an extra time or two in the evening.
   (A): But certainly before you go to bed. (ICE-GB: S1A-087 #278-285)

Participant (A) repeats that participant (B) has to use mouthwash five minutes a day. In the example above, *so* is used to mark the main idea in the conversation.

ZE seems to have instances where the DM *so* is used as a main idea unit marker as illustrated below.

7.29  <$HH$>: Yeah, there were unusually heavy heavy heavy rains that led to to flooding. Unfortunately, eh areas like ours are not easily affected by, are not greatly affected by these these cyclones because my understanding is that cyclones actually have their eh most effect in
areas where there is rugged relief. So basically, when cyclones occur, it means we are receiving unusually heavy rainfall. (ZE corpus: Private semi-scripted dialogue 34)

7.30 <$OO>: Yes, hah there is, I would say a lot. But I will talk about, in the Bible I really love Moses. You know, the way God pulled him from from where he was. The way he was, the setup, just the setup for Moses being a leader was done. . . He wasn’t afraid. He would challenge God especially when they came to river Jordan and he will say, God, you are the one who send me to take these people away from Israel, now you, from Egypt. Now you want me to to just let them die here. And then God answers him and says no, just call every miracle that God did to Moses, it was like a challenge. I’m not the one who came here to take the people. You are the one. You know what you will do. So basically that’s why I like Moses in the Bible. (ZE corpus: Private semi-scripted dialogue 41)

In example 7.29 so functions as a main idea unit marker because participant <$HH> repeats her explanation that cyclones cause heavy rains. Participant <$OO> explains in detail how Moses was chosen by God to lead the Israelites from Egypt and also how Moses would challenge God. Participant <$OO>. So is introduced in this case, to mark the main idea about the reasons for liking Moses in the Bible in example 7.30.

iii. Summarizing/ rewording / giving an example

Let us consider the following examples from the ICE-GB.

7.31 I’ve read for work and for study but I’m coming round to it now with reading to my children, so like I’m picking up like I picked up. (ICE-GB: S1A-013 #222)

7.32 I could do what I like in that respect so there was no restriction there uhm which was very unusual, very uncommon. (ICE-GB: S1A-072 #144)

In the two examples above from the ICE-GB, so is used to put what has been said before in different words. In example 7.31 the idea of coming around is reworded with so showing the rewording to I’m picking up and I picked up. So in example 7.32 shows rewording to so there was no restriction from I could do what I like.

7.33 <$G>: I just know the Lord is my shepherd, I shall not want. So you need to trust in God coz she is our saviour. (ZE corpus: Private semi-scripted dialogue 8)
Example 7.33 indicates *so* as a marker of summary. Participant <$G>$ starts by stating the verse and then summarises it immediately afterwards with *so* used as a marker of the summary.

**iv. Sequential *so***

*So* is introduced to show the sequence of events from the storm that destroyed part of the tower to someone building it as a sort of thanksgiving in the example below from the ICE-GB.

7.34 (A): *There’d been some subsidence or some terrible storm that’d uhm uh demolished part of the tower.*

(A): *The original.*

(A): *So then he built this on as a sort of thanksgiving you might say.* (ICE-GB: S1A-094 #275-277)

The sequential function of *so* is evident in the example below because *so* is introduced to show the sequence of events from realising that the keeper was offline to kicking the ball towards the goal.

7.35 *When the ball suddenly came to me I realised the keeper was well off his line. *So* I just gave it a big kick back towards the goal.* (ICE-GB: W2C-014 #010)

If we analyse the ZE corpus, the sequential function of *so* is attested in the corpus.

7.36 <$F>$: *So*, Chombo came and I was uh junior councillor. *So*, you would give a speech, right. *So*, when you give a speech you will just uh, you were expected to give a slogan at the end. *So* unfortunately, I was representing children in parliament. *So*, I was not obliged to give a slogan. *So*, when I sat down, his guys were like uh no councillor, you left something, a slogan. *So* then I stood up and I tried to do the slogan of which I did not know how to. (ZE corpus: Private semi-scripted dialogue 6)

Participant <$F>$ first explains that when he was a junior councillor he was required to give a speech and do a slogan but didn’t. Participant <$F>$ further explains that after being reminded to do so, he tried to do the slogan. In this case, *so* sequentially marks the act of doing the slogan.
7.37 <$RR>$: I could not be there. They were now following me wherever I am. They could just be at my residence before I was there, you see.
<$RR>$: Trying to capture me, you see. So then that’s when I decided, no I will go to work at Gutu. That was very scary because they wanted, in fact they wanted to kill me. (ZE corpus: Private semi-scripted dialogue 44)

Participant <$RR>$ first mentions the reasons why he could not be there. The next scene is introduced by so to show the sequence of events that followed.

v. Boundary marker

According to Müller (2005), so as a boundary marker is prevalent between instructions and the beginning of utterances. In the ICE-GB, the following examples show the use of so as a boundary marker.

7.38 (B): Well yes I mean the idea is I’m not interested really in uh museums more generally.
    (B): I mean I wouldn’t be interested in more sort of.
    (A): Ok so it’s a limited number of museums.
    (B): Very limited. (ICE-GB-S1A-066 #063-066)

In example 7.38 above, so is the boundary marker between types of talk.

ZE has examples where so was used as the very first word as a boundary marker as illustrated below.

7.39 <$NN>$: So, at the farm we we sort of do everything. Uhm I would say anything in a a rural setting home can do is what we do. (Private semi-scripted dialogue 40)

7.40 <$Q>$: So, Tomuda can you tell me, can you tell me how, I mean can you tell me how you hoping or how are you hoping in your life about his coronavirus? (ZE corpus: Private dialogue 9)

b. Interactional functions

i. Speech act marker – question or request

Regarding the function of speech act marker for a question or a request, Müller (2005) categorised all instances where a question or a request is preceded by so. In both the ZE
corpus and the ICE-GB, some of the instances where so is a speech act marker indicating a question or a request are shown below.

7.41 So why am I going? (ICE-GB: W1B-010 #067)

7.42 Well Liz is quite keen to go there, so why not do it? (ICE-GB: S1A-061 #139)

7.43 <$B>: Mm. So what can be done? Is there anything that’s been done? (ZE corpus: Public dialogue 20)

7.44 <$A>: That sounds so exciting. So what is your criteria? How are you picking these models, how are you picking these fashion designers? (ZE corpus: Public dialogue 20)

ii. Speech act marker – opinion

Müller (2005) describes the use of so as a speech act marker of opinion and notes that sometimes the words I assume, I think, I guess, etc. can be used to show that it’s an opinion. Starting with the ICE-GB, the following examples indicate so as a marker of opinion where it is used.

7.45 And the way I’m going about it is uhm I use a lot of turps and rub the colour in into the grey areas, so I suppose that’s one of the reasons why it appears luminous. (ICE-GB: SB1-008 #113).

7.46 You say that you’ll be looking for a “more permanent dwelling”, so I guess you’re planning a longish stay in Britain. (ICE-GB: W1B-015 #018)

In the two examples above, so is used as a speech act marker of opinion. Example 7.45 shows opinion about the colour appearing luminous and example 7.46 indicates opinion about plans to stay in Britain longer.

The speech act marker for opinion is attested in ZE as illustrated in the examples below.

7.47 <$L>: Ok. From your narration, thank you. From your narration I think you said you are a teacher by profession and you are also a family member. Eh so I assume that you’ve got children who go to school. (ZE corpus: Private dialogue 6)
7.48 <$NN>: Well, I have a lot of relatives who are actually struggling. **So** I think I will start from there. That’s the charity I will do.  (ZE corpus: Private semi-scripted dialogue 40)

In example 7.47 above, *so* prefaces an opinion in which participant <$L> assumes that the other participant has children who go to school. In example 7.48, *so* is a speech act marker used to share an opinion about not being trusted.

### iii. Marking implied result

The function of *so* as a marker of implied result in the ICE-GB is illustrated below:

7.49  (F): Well actually the thing is it’s not even particularly far.
      (F): It’s just the travel.
      (A): Yes.
      (A): *The difficulty is the travel.*
      (A): You know it’s all right if you’ve got a car but
      (D): *I mean,*
      (D): . . . *you have to go to Cambridge and get the village bus out.*
      (D): And it took over three and a half hours to get there and over three and a half hours to get back and uh probably four all told.
      (D): *Involves getting the tube.*
      (D): . . . *so.*
      (E): *Not worth it is it?*
      (C): *No.*
      (F): Certainly wasn’t worth it. S1A-019 #366-377)

In example 7.49 above, speaker D is explaining how difficult and time consuming it is to travel using public transport and ends with the DM *so*. Here, the DM *so* indicates that speaker D wants other speakers to deduce the implications of the long journey. In this case, speaker E responds by asking whether it is not worth it to which speaker C and speaker F respond that it’s not worth it.
Let us consider the dialogue below:

7.50  (F): *Uhm my brother read to me always.*
     (B): *And uhm I never read the classics or anything like that and now I’m reading them.*
     (B): *I’ve read even Hobbit and I’m reading Treasure Island.*
     (F): *Me neither.*
     (D): *Uhm.*
     (E): *On the subject of classics and so on I think on the whole children are given books to read too young.*
     (E): *and uh so I mean.* (ICE-GB: S1A-013 #222-233)

After speaker B talks about not reading classics in the past, speaker B reveals that he has now read the novels Hobbit and Treasure Island. Speaker E uses the DM so as a cue and I mean to indicate that he assumes other speakers can see the implications of introducing books to children when they are too young.

The function of so of marking implied result is shown below.

7.51 <$C$>: *Ah actually an operation like this has been done in Zimbabwe in nineteen eighty-three eighty-four at at Harare Hospital. But as you know uh every set of conjoined twins is different from the next one. So <$,> you, we can’t say because I’ve done one set before therefore I can do the other one. So uh yeah.* (ZE corpus: Public dialogue 27)

**iv. Marker of a transition relevance place**

In the example below, there is transitions using so after talking about the black pencil to talking about a prepositional phrase.

7.52 *If you want a black pencil, that’s a marker pencil which you have there. Ok. So now we’ve got two trees one of which has got a prepositional phrase underneath the noun phrase.* (ICE-GB: S1B-002 #102).

No data was found that indicates a transition relevance place in the ZE corpus.

**7.6.2.2 Functions of well**

Using the concordance tool in Sketch Engine, all instances were well occurred were identified in the ZE corpus. Instances where well appeared to be used as an adjective or an adverb were excluded from the analysis.
a. Textual level functions

i. Searching for the right phrase

According to Müller (2005: 109) the use of well to search for the right phrase is usually “combined with other means of expressing ‘deliberation’, for example filled and unfilled pauses, truncated words and intonation units, repetition, or other markers”. This can be observed in examples from the ICE-GB below.

7.53 (A): Have you considered the influences of it on your childhood.
   (B): Yeah.
   (B): *Uhm well* first of all *uhm* it undoubtedly made me very single-minded very independent very competitive. (ICE-GB: S1A-075 #034-036)

7.54 (A): What are they about.
   (B): Sorry.
   (A): What are they about *I mean*.
   (B): *Uhm well* uh the one I’m sort of trying to finish at the moment is a play for a company called Quicksilver. (ICE-GB: S1A-096 #023-026)

As evident in example 7.53, speaker (B) starts by using *uhm* as a filled pause and then the DM *well* is used. In extract 7.54 *well* is surrounded by filled pauses and repetition of *uhm* where participant (B) is searching for the right phrase to answer the question.

Extracts from the ZE corpus below show the function of *well* to search for the right phrase.

7.55 <$B>$: Mr Katsinha do you have any proof? Where did you get this information?
   <$A>$: I have the proof. I have the proof. It’s *well* it’s unfortunate that I didn’t *did not* bring it in uh into the studio. But I have the proof. (ZE corpus: Public dialogue 24)

7.56 <$A>$: Pig farmers they say they don’t have a market for their cattle *uhm* for their pigs and uh uh they say buyers traditional buyers such as Colcom refuse to buy from small scale farmers <$#>$ What’s your comment on that?
   <$B>$ <$#>$ We... *well uhm* I’m I’m surprised because we we see we see pork in the supermarket. (ZE corpus: Public dialogue 19)
In extract 7.55 above, *well* is surrounded by repetition of the word *it’s* where participant <$A>$ uses *well* to search for the right phrase. Example 7.56 shows an unfilled pause pre-facing *well* and a filled pause and repetition after *well*.

### ii. Rephrasing/correcting

The use of *well* to correct an utterance or to rephrase it is also highlighted by Schiffrin (1987). In extract 7.57 below from the ICE-GB, speaker (B) uses *well* to correct the first utterance that the speech has to be done at the *Christmas departmental dinner* and not the annual dinner.

7.57 (B): *I might have to do the uh after dinner speech at our annual well not annual, our Christmas departmental dinner.* (ICE-GB: S1A-030 #072)

Let us consider another example:

7.58  (B) *I mean we can never tell uhm,*
(B): *well not we,*
(B): *but they could never tell about how she would grasp concepts.* (ICE-GB: S1B-003 #107-109)

Speaker B uses *well* to correct the sentence from *we can never tell uhm* to *they could never tell about how she would grasp concepts.*

In the ZE context, *well* is also used for rephrasing or correcting. In extract 7.59 below, speaker B first refers to the program as *brand new* and then goes on to correct the utterance to say that the program is in its fourth week. Therefore, *well* is used to correct.

7.59 <$B>$: *Stand by for a brand new program. Well it’s no longer brand new. It’s now four weeks into the program.* (ZE corpus: Public dialogue 30)

Another example that shows correcting or rephrasing is shown below:

7.60 <$GG>$: *I think that the that’s the worst thing that they ever did. Getting beaten by, well, they managed to uhm beat Somalia when they came, when they played here in Zimbabwe.* (ZE corpus: Private semi-scripted dialogue 33)
In example 7.60 above, participant <$GG> uses well to correct the first statement that the players lost to they managed to uhm beat Somalia.

iii. Quotative well

To indicate the starting point of direct speech, well is used. According to Müller (2005) the most common quotative construction is BE + like with other possibilities being go, say, ask and think. For example:

7.61 (A): And the Queen goes well shame Fergie wasn’t here, otherwise we could have hidden the Land Rover. (ICE-GB: S1A-041 #331)

7.62 (A): He says well if you think God doesn’t love you, you only need to look over the border to Edom uh the country . . . (ICE-GB: S2A-036 #020)

In the two examples from the ICE-GB above, well marks the start of direct speech, which is in line with Müller’s (2005) suggestion. This is also manifested in the examples from the ZE corpus shown below:

7.63 <#> Raina collects herself together and says: “Well, don’t forget . . .” and she can’t remember what she wants him not to forget. (ZE corpus: Creative writing 3)

7.64 The Old Man finishes his eating, hands his plate to Betty, washes his hands and after a brief respectable pause says: ‘Well, boy, see you when you come back.’ (ZE corpus: Creative writing 3)

iv. Move to the main story

The use of well to move to the main story is evident in the example below from the ICE-GB:

7.65 (B): Uhm I like to watch sport.
   (B) I enjoy watching it.
   (B): and I enjoy any kind of physical exercise that is you know stretching or uhm working out . . .
   (B): but . . .
   (B): it sort of stops there uhm.
   (B): With working now in movement and dance you can use that.
   (B): you need some strength.
(B): Oh this is terrible sorry.
(A): Right well
(A): Uh ok.
(A): Uhm why do you think physiotherapy is really important? (ICE-GB: S1A-003 #007-017)

After speaker B discusses how he enjoys watching sport in detail, speaker A uses well to move back to the main story, physiotherapy. In example 7.66 below from the ZE corpus, speaker A uses well to move to the main story that is smoking actually having an effect before you actually get pregnant of which speaker B explains.

7.66 <$A>: Alright **well** you mentioned about aahm you mention the issue of ah ah smoking actually having an effect before you actually get pregnant. Right. Highlight that for us.
<$B>: Yes you are very right ah ah Anna Smoking makes it harder for a woman to get pregnant. (ZE corpus: Public dialogue 2)

v. Introducing the next scene

In the ICE-GB, speaker A recalls a conversation he had with the elder statesman of Mrs Thatcher. He then uses well to introduce the next scene, then I spoke to one of the backbench conservative MP as shown below:

7.67 (A): **Well** tonight I talked to that elder statesman of Mrs Thatcher’s first two governments Lord Hailsham. I suggested that the Conservative cabinet and the party in parliament had got themselves into a considerable mess over this whole European issue and it was not just a matter of presentation. **Well** then I spoke to one of the backbench conservative MPs who’s something of a sceptic on the issue of European unity Richard Shepherd. (S2B-017 #078-082)

The use of well for the purpose of introducing the next scene in the ZE corpus is represented below.

7.68 <$GG> <#> And when we finally wore our caps to indicate that we were now graduates, hah that was the most important time in my life. **Well**, then we went to to this uhm pub called uhm it’s Mega Two where we had some drinks and braai. (ZE corpus: Private semi-scripted dialogue 33)
vi. **Conclusive well**

An analysis done by Müller (2005: 120) suggests that “many of the instances of conclusive well carried a notion of ‘I have said enough’ about this scene or topic”. This can be observed in the example from the ICE-GB below.

7.69 (B): *There you go.*
(A): *It’s a very nice shirt by the way well.*
(B): *It’s a very old shirt.*
(A): *It’s very nice. Thank you.* (ICE-GB: S1A-038 #002-005)

In example 7.69 above, speaker A concludes the sentence with *well* and then speaker B responds. If we consider example 7.70 below from the ZE corpus, we can observe that by using *well*, participant <$A>$ gives a cue that he is done talking and participant <$B>$ responds with *ok*.

7.70 <$A>$: *In March nineteen ninety-three he was posted to all Arms battle school as the commander and where he served until thirty-one December nineteen ninety-three and well.*
<$B>$: *Ok. And would you say you learnt from him and the work that he did?* (ZE corpus: Public dialogue 4)

In the above example, *well* can be considered to aid in turn taking among speakers.

b. **Interactional level functions**

i. **Indirect answer**

Insufficient answers, answers that do not directly supply the information required, replies with delayed answers, replies that provide partial answers are categories of indirect answers (Müller, 2005). An example is highlighted below.

7.71 (C): *How many legs would you think we’d be talking about with ply?*
(B): *Well the ends will be OK uhm.*
(B): *If we put in ply, as a shelf I’d put one centre and then one centred in the two halves.*
(B): *So that would be three lengths.* (ICE-GB: S1B-073 #286-288)

The example above shows a delayed answer. Speaker B does not directly tell speaker C the number of legs to be put on the table. Instead, he starts by telling speaker C that the
ends will be ok and then moves on to explain how the legs will be fitted. The answer to the question comes at the end (three).

An example of indirect answer in ZE is shown below.

7.72 <$B>$ Alright thank you very much Doctor Mbuvalyesango $<$#> They were sharing one liver $<$#> So ha... what did you do?
  <$C>$ Well they they had eh a liver that was ah joined but each one had eh a separate biliary system. So eh eh you know the liver makes eh eh bile or gall and each had had a separate system. So what we had to do was to split the liver so that each of the eh patients had a liver with them with a biliary system. (ZE corpus: Public dialogue 25)

After speaker <$B>$ asks “so ha... what did you do?” speaker <$C>$ responds indirectly, starting the sentence with well and first explains that the conjoined twins shared a liver but had a separate binary system. Thereafter, speaker <$C>$ answers the question by saying that they split the liver.

ii. Direct answer

The function of well to indicate direct answers occurs when the required information is given without delays (Müller, 2005). This is illustrated below.

7.73 (A): Uhm can you describe to me if it’s possible uhm a typical day in your home when you were a boy of less than fourteen. What do you remember?
  (B): Mm well I remember I remember sort of there were very hot summers uh for some reason. (ICE-GB: S1A-073 #066-067)

7.74 (A): Are you actually going to bother getting a job?
  (B): Well not for the next two weeks probably. (ICE-GB: S1A-093 #020-021)

In extract 7.73 from the ICE-GB, speaker B provides a direct answer to the question what do you remember? His response starts with well I remember and he goes on to describe the hot summers. Likewise, in response to the question about getting a job, Speaker B uses well to indicate a direct answer, well not for the next two weeks probably in example 7.74.

An analysis of the ZE corpus seems to show the use of well to show a direct answer as highlighted in examples below.
7.75 <$O>$: Ok. Are you not allowed to like food, is there any type of food that you are not allowed to eat?
<$P>$: **Well, not really.** (ZE corpus: Private dialogue 8)

7.76 <$A>$: Right fantastic stuff but of course locally the premiership continues. What’s going on at Luveve?
<$B>$: **Well** at Luveve Chiredzi is ah being entertained by Chicken Inn. (Ze corpus: Public dialogue 42)

Considering example 7.75, participant <$P>$ replies to the question about whether there is any particular food they are not allowed to eat by saying **well, not really.** In this instance, **well** functions as a marker of a direct answer. When asked about **what’s going on at Luveve** in extract 7.76, participant <$B>$ responds directly by saying **well at Luveve Chiredzi is ah being entertained by Chicken Inn.** This example shows **well** as a marker of a direct answer.

**iii. Response to self-raised expectations**

According to Schiffrin (1987: 123), in responses to self-raised expectations “speakers are treating their own prior talk as something to be responded to”. Response to self-raised expectations does not seem to be attested in both the ZE corpus and the ICE-GB.

**iv. Contributing an opinion**

Consider the extract below:

7.77 (A) **What is the difference for you between experiencing it and feeling that you’ve made it up.**
(B): **Because if I made it up because it had a purpose then I should be able to stop it.**
(A): **And will keep you in control.**
(A): Mm.
(A): **Well also that means that’s a bit like carry on taking the tablets you know.**
(A): **Everything’s all right really.** (ICE-GB: S1A-062 #036-041)

After asking speaker B a question and getting a response, speaker A agrees with speaker B, **Mm** and then gives her own opinion, **well also that means that’s a bit like carry on taking the tablets you know.**
The contribution of an opinion in the ZE corpus is shown below:

7.78 <$L><#>Now people are not farming enough because of poor rains. <#>I hope that it rains, well but I don't think it will. (Private semi-scripted dialogue 12)

v. Continuing an opinion/answer

Well is used to continue expressing opinion on a subject matter as done by speaker B in the following extract.

7.79 (A): What sort of activities, physical activities were available?  
(B): I suppose uhm the standard kind of physiotherapy when you asked for it uhm and well sports I guess. (ICE-GB: S1A-003 #002-003)

Another example below from the ICE-GB shows speaker A talking about pictures I saw at the exhibition. The continuation of opinion is preceded by well and then speaker A goes on to say Uh so some of it was not very good quality.

7.80 (A): Well my impression of the pictures I saw at the exhibition was that he wasn’t filtered uh by an actual filter.  
(A): But he was just painting what he wanted to paint.  
(A): And the quality was variable.  
(A): And there was no one standing around saying to him you should junk this.  
(A): And well you know.  
(A): Uhm and  
(A): Uh so some of it was not very good quality. (ICE-GB: S1A-015 #091-097)

Turning to ZE, the following examples illustrate continuing an opinion or answer.

7.81 <$D>: Actually people most of the people are are now buying firewood from nearby farms because the the farms have uh a lot of trees. The area that we live people have actually cut down a lot of trees and well they cut down a lot of trees and build houses there. So, we have very small trees and we cannot find firewood there. (ZE corpus: Private semi-scripted dialogue 4)

7.82 <$B>: Well Calisto Maheya Calisto Maheya to blame here at Ascot Stadium. Well it was a yawning goal in front of him and the the Caps United goal keeper Jorum Much Muchambo was crawling on the ground but for reasons best known to Calisto Maheya well he came up with an effort that went over the bar and well Taurai Mangwiro is to leave a big sigh of relief. (ZE corpus: Public dialogue 41)
In example 7.81, participant <$D>$ is discussing about people cutting down trees in the area where he lives. There is a continuation of an answer after well with the participant further explaining why people cut down trees and the consequences of cutting down trees. The continuation of opinion is illustrated in example 7.82 where well is used three times by speaker <$B>$ to explain what is happening at Ascot stadium.

vi. **Evaluating a previous statement**

In the example below, speaker A first expresses that she doesn’t like EastEnders. She then evaluates the statement and changes it to say she just saw the ending. Speakers A and B go on to discuss that Clyde won even though he looked roughed up.

7.83 (B): *Did you watch EastEnders?*
   (A): *Uh uh I don’t really go a bundle on it.*
   (A): *I just saw well just near the end.*
   (B): *Yeah.*
   (B): *Clyde won.*
   (A) *Yeah that’s . . .*
   (A) *He looked a bit <unclear word> to me.*
   (B): *He looked to me as if he’d been a bit roughed up you know.*
   (A): *Yeah he did rather.*
   (B): *It shouldn’t.*
   (B): *It looked like it should’ve been an RSF there didn’t it but uh.*
   (B): *But obviously he carried on and won.*
   (A): *Yeah.*
   (A): *Yeah well all credit to the lad.* (ICE-GB: S1A-095 #197-210)

The extract below from the ZE corpus shows well being used to evaluate a statement about divas and slay queens.

7.84 *This is why these divas, or slay queens in local parlance, go to elaborate lengths to bath, brush, bleach, comb and dye their pets so that they can ostentatiously out-compete fellow attention-seekers. Well, not where the Bishop comes from.*

*In our world, dogs stay outside, are given one meal per day, are never bathed and are invariably expected to spook ill-willed village pilferers.* (ZE corpus: Newspaper reportage 12)
7.6.2.3 Function of **but**

**a. Contrastive**

As a contrastive DM, **but** shows a sense of dissonance in BrE as will be highlighted from the examples below.

The following examples that show the different functions of **but** are drawn from the ICE-GB. As asserted by Norrick (2001) **but** indicates contrast, which may be conveyed lexically. The examples from the ICE-GB below show the contrastive function of **but**.

7.85 (A): *Well I think that in some ways I am saying things are all right. But I know that they are not alright.* (ICE-GB: S1A-062 #025-026)

In the extract above, **but** contrasts the first statement that speaker A says things are all right and then later on says things are not alright.

If we consider the example below, **but** shows contrast that although a loss is expected in the vote of confidence, VP Singh was allowed to present his arguments first.

7.86 **VP Singh is expected to lose his vote of confidence. But** the speaker allowed him to put his case first. (ICE-GB: W2B-006 #049-050)

Let us consider the example below:

7.87 *They fired rockets and artillery last night but where I was they made no move to advance.* (ICE-GB: W2B-014 #086)

The contrastive nature of **but** is evident because **but** shows that although rockets were fired, there was no move to advance. Likewise, **but** contrasts the first statement about the casualties being few by highlighting that some troops have been wounded and there are reports of a number of deaths.

7.88 *Allied commanders say their casualties have been remarkably light. But some troops have been wounded and there are reports of a number of deaths.* (ICE-GB: W2B-016 #053-054).

Another example from the ICE-GB, which shows the contrastive **but** is shown below.

7.89 (A): *I know you can’t do it overnight,*
(A): **but** I think the way to win her back is by making her feel that she’s not excluded from anything. (ICE-GB: S1A-054 #015-016)

*But* contrasts the utterance in the first statement where speaker A says although she knows it can’t be done overnight, the way to win her back is by making her feel that she’s not excluded from anything.

Turning to ZE, there are several examples below showing the DM **but** used as a contrastive DM. For example, in example 7.90 below, **but** is used to show that even though the plants used to be impressive, after three decades of negligence, the plants are now a shadow of their former selves.

7.90 The original File and Crowborough plants won international engineering awards and were the sort of thing advanced cities in advanced countries were starting to build. **But**, three decades later the first two plants lack maintenance, were never expanded as populations doubled . . . (ZE corpus: Editorials 1)

In example 7.91 below, speaker <$B>$ comments that four goals are needed so that the team advances. **But** in this instance is used to indicate a contrast between expectation and the reality that our defending is atrocious.

7.91 <$B>$: We have to score four goals here to beat them to beat them by four goals to two in this encounter in order to win. **But** our defending is atrocious. We continue to get very silly goals. (ZE corpus: Public dialogue 12)

One of the examples found in the ZE corpus is highlighted below:

7.92 I told you about my problems in my last email. **But** it seems you have decided to ignore me. (ZE corpus: Private business letters 5)

In example 7.93, even though participant <$N>$ was scared in the beginning, **but** is used to highlight that participant <$N>$ later I gained some confidence.

7.93 <$N>$: Yes, at first, I was scared. **But** hah later I gained some confidence. (ZE corpus: Private semi-scripted dialogue 20)

The following example illustrates the use of **but** as a contrastive DM.
7.94 <$F>$: Tomorrow I have to hang around at home. Eh that’s when I have to isolate myself from the rest of the people.

<$E>$: <O>laugh</O>

<$F>$: But the day after tomorrow that’s when I start the harvesting process and hopefully I will be able to finish up the harvesting. (ZE corpus: Private dialogue 3)

But shows that although participant <$F>$ has to stay home tomorrow, he will be harvesting the following day.

b. Sequentially contrastive

As noted above, as a sequentially contrastive DM, but signals a return to the main topic. The following examples provide instances where but was used as a sequentially contrastive DM in the ICE-GB.

When considering Norrick’s (2001) guidelines, the following examples show the use of the DM but to indicate a return to the main topic or point.

7.95 (C): I know the phone number of the chap uhm.

(B): <unclear words>

(B): Yeah.

(B): But what I need is a personal intro to him. (ICE-GB: S1A-027 #019-022)

A shift from talking about scholarships to girls in gowns is highlighted by the use of but in the conversation below.

7.96 (A): I mean I know it sounds awful,

(A): but if they get a scholarship which I don’t think Fran will have, though it’d be nice if she did but she found maths incredibly hard.

(A): Uhm they wear gowns,

(A): but I mean it is rather cute seeing these beautiful girls walking past the Cathedral which has been there for hundreds of years in their gowns. (ICE-GB: S1A-054 #136-139)

Two more examples where but is used as a sequentially contrastive DM is shown below:

7.97 She realised that the star cloth had once covered her table, but had been thrown away by careless people, by vandals and drowned in the mud. But she no longer sobbed because of that. She sobbed because she’d split something on the table and made a shocking stain on the surface. (ICE-GB: W2F-020 #115-117)
7.98 She greeted me with every sign of pleasure and invited me to tell her the latest news from Oxford, but I soon perceived, having accepted the invitation, that her attention was elsewhere. (ICE-GB: W2F-011 #016-017)

In examples 7.97 and 7.98 above, but cancels expectations about what should come next. In this sequential function, “but marks off a digression or other subordinate section of a discourse and signals a return to its main topic or point” (Norrick: 2001: 858).

An examination of the ZE corpus shows the function of but as a sequentially contrastive DM. The examples are explored below.

7.99 So there is still hope, but perhaps this is a window of opportunity that helps us to reconsider and evaluate alternative approaches to using the water resource. But this is also where Command Agriculture or a programme along the lines of Operation Magutu should kick-in, all making Zimbabwe reduce the scope of its reliance on anticipated grain/food imports. (ZE corpus: Editorials 8)

7.100 <$NN>: Uh so obviously yes we stayed in a very nice hotel there, Elephant Hills. But the waterfall, it’s eh water dropping more than hundred metres or around hundred metres forming this nice rainbow especially if you go there during the rainy season. It’s just amazing. (ZE corpus: Private semi-scripted dialogue 40)

The topic under discussion is shifted from alternative approaches to using the water resource to Command Agriculture in example 7.99. In example 7.100, there is a shift in the topic from staying at Elephant Hills to discussing the waterfall.

Another example that shows a shift from greeting to hearing some of the interviews that were carried out by the Star F M chief reporter George Msumba is highlighted below.

7.101 <$B>: Sir welcome to Star F M. <$H>: Thank you I’m glad to be here. <$B>: It’s great to have you. But before we get into uh into our interview we gonna uh hear some uh some of the interviews that were carried out by the Star F M chief reporter George Msumba. (ZE corpus: Public dialogue 20)

But signals a shift in the topic of conversation from thanking the doctors for a successful operation to asking about the number of people present during the operation.
7.102 <$D>$: Yeah first I would like to thank the doctors for oo the good work they have done. It was a miracle to uuu come out with a good result from such an operation. But the question I have for them is eeh how many were they in the operation. (ZE corpus: Public dialogue 28)

In this section, examples of different functions of so, well and but were provided from the ZE corpus and the ICE-GB. The qualitative functional analysis done above helped to determine the frequencies of functions of the three DMs, which is done in the next section.

### 7.6.3 Summary of quantitative and qualitative results for discourse markers

In section 7.6.1, data analysis provides the occurrence- and function frequencies of discourse markers in the ZE corpus and the ICE-GB. A comparative analysis was also done in terms of the log likelihood test to see if there is a significant variation of discourse markers between the corpora, in order to ascertain if the constructions themselves show innovation. This summary thus aims to answer research questions 1.3 and 2.3:

Are the variations between the corpora statistically significant?

The log likelihood test was used to determine whether the observed variations were statistically significant. Regarding the overall frequencies of the three DMs in both corpora, it was observed that the variations observed were statistically significant, with the LL values being higher than 15.13, where $p < 0.0001$. The ZE corpus recorded more instances of so compared to the ICE-GB. Statistically significant variations were noted in the frequencies of well and but, with the ICE-GB recording more instances than the ZE corpus for both DMs.

The spoken and written registers of the two corpora were also compared. It was observed that there were statistically significant variations in the normalised frequency of so in the spoken register, with ZE recording more instances compared to the ICE-GB. No statistically significant variations were observed in the frequency of so in the written register between the two corpora due to the fact that the LL < 3.84, which is a critical value for $p < 0.05$. When considering well, the variations observed in the spoken register
between the two corpora were statistically significant, with *well* occurring more in the ICE-GB compared to the ICE-GB. There were no statistically significant variations in the frequency of *well* in the written register. For *but*, statistically significant variations were observed in both registers, meaning that there are variations in the frequencies.

The **genres** in the ZE corpus and the ICE-GB were also compared. Statistically significant differences were observed in private dialogue, public scripted monologue, creative writing and social letters in the two corpora. There were no statistically significant variations in the frequencies of *so* in public dialogue, popular writing, newspaper reportage, academic writing and editorials because the LL values were below the critical value of 3.84, where $p < 0.05$. Observations were made about statistically significant variations in private dialogue and public dialogue, which showed *well* as being more prevalent in BrE than in ZE. The newspaper reportage genre showed statistically significant differences, with ZE recording higher instances compared to the ICE-GB. There were no statistically significant differences in public scripted monologues, creative writing, business letters and social letters. When considering *but*, statistically significant variations were observed in six out of 10 genres analysed. The genres are private dialogue, public scripted monologues, public dialogues, public dialogues, popular writing, editorials and newspaper reportage. No statistically significant variations were recorded in the following genres: creative writing, academic writing, social letters and business letters.

The LL values for the **function frequencies** were also calculated. Observations were made that there were statistically significant variations in the frequencies of functions of *so*, with the ZE corpus recording higher frequencies than the ICE-GB in the following functions: marking result or consequence, main idea unit marker, summarizing / rewording / giving an example, speech act marker - question or a request. The other function that shows statistically significant variations was the sequential *so*, which was more frequent in the ICE-GB compared to the ZE corpus. There were no statistically significant variations in the frequencies of the boundary marker function, the speech act marker – opinion function, the marker of a transition relevance place function and the
unclassified function. This is because the LL values were below the critical value of 3.84, where \( p < 0.05 \).

For the frequencies of functions of *well*, statistically significant variations were recorded, with the ICE-GB recording higher frequencies of the following functions compared to the ZE corpus: searching for the right phrase, rephrasing/correcting, quotative *well*, move to the main story, indirect answer, direct answer, contributing an opinion, continuing an opinion/answer and evaluating a previous statement. There were no statistically significant variations in the frequencies of the following functions: introducing the next scene, conclusive *well*, response to self-raised expectations and unclassified. This is because the LL values were lower than the critical value of 3.84, where \( p < 0.05 \). If the LL values are considered, only the sequentially contrastive function of *but* showed statistically significant variations between the ZE corpus and the ICE-GB, with the function being attested more in the ICE-GB compared to the ZE corpus. The contrastive function and the unclassified function did not show statistically significant variations.

The following section provides a conclusion for this chapter.

### 7.7 Conclusion

As noted in Section 1.6.3, contextual effect on DMs is one of the least researched topics (Lam, 2009b). This study investigated the frequency and functions of *so*, *well* and *but* in a corpus of ZE in order to establish whether there are variations in the frequency and functions of the DMs. In order to contextualise the discussion, an overview of DMs in StE was done and then DMs *so*, *well* and *but* were discussed in particular.

In this chapter, there were statistically significant differences in the normalised frequencies of *so* in private dialogue, public scripted monologue, creative writing and social letters in the two corpora. Private dialogue reported the highest LL value of 164.51, which is well above the critical value of 15.13 where \( p < 0.0001 \) with *so* occurring 86.1 times per 10 000 words in ZE compared to 48.7 times per 10 000 words in BrE. Statistically significant variations were also observed in public scripted monologue because the ZE corpus has 33.8 instances of *so* and whilst for the ICE-GB the number is
9.1 instances. Creative writing and social letters also show variations in frequencies per 10 000 words (12.3 and 57.7 for ZE and 5.4 and 31.2 for BrE respectively). In both genres, so was attested more in ZE compared to BrE. Another interesting observation is that there were no statistically significant differences in the frequency per 10 000 words of so in public dialogue, popular writing, academic writing, editorials newspaper reportage and business letters.

Noteworthy differences were visible in the DM well, with ZE recording 7.4 instances per 10 000 words whilst BrE has 69.4 instances per 10 000 words in private dialogue with a LL value of 848.58. The results show that well occurred more in the ICE-GB compared to the ZE corpus. In this study, observations about the variation in frequency of DMs will aid in the determination of the stage that ZE can be classified as in the DyM, with specific reference to nativisation and innovation as research question 3 and 4.

One of the key findings in this chapter is that there are differences in the normalised frequencies of DMs in the spoken and written registers, with all three DMs occurring more in the spoken registers compared to written registers. This viewpoint is supported by the fact that the LL value of well in the spoken register is 809.71, which is above the critical value of 15.13 where \( p < 0.0001 \). Well was attested more in the ICE-GB compared to the ZE corpus. In the written register, the normalised frequency for well in the ICE-GB is 2.9, per 10 000 words, which is not different from than that of the ZE corpus (2.1).

In this chapter, an analysis of the DMs but indicates that there were statistically significant differences in the frequency of the DM but in six of the genres analysed namely private dialogue, public scripted monologue, public dialogue, popular writing, editorials and newspaper reportage. The highest LL value is found in public dialogue (176.24) where but occurred 43.9 times per 10 000 words in the ICE-GB compared to 15.1 times in the ZE corpus.

Overall there are statistically significant differences in the frequencies of so, well and but with well and but occurring more frequently in the ICE-GB (32 and 29 times per 10 000 words respectively) than in the ZE corpus (6.5 and 25 times respectively), with LL values of 782.81 and 15.98 respectively where \( p < 0.0001 \). Interestingly, the DM so was used
more in the ZE corpus (44.6 times per 10 000 words) compared to BrE (29 times per 10 000 words).

In this thesis, it was shown that DMs function as coherence relations and that they help to link text units (Schourup, 1999). DMs are believed to play an important role in interpreting texts because they show coherence relations between discourse units (Schiffrin, 1987). This was shown by the different functions of so, well and but in the ZE corpus and the ICE-GB. For example, when considering the functions of so, six functions showed statistically significant differences. Marking result or consequence showed the highest statistically significant value of 209.45 and was attested 26.2 times per 10 000 words in the ZE corpus, which is more than the 13 instances per 10 000 words in the ICE-GB. Three other functions that show significant variations with regard to the LL value include main idea unit marker (46.44), summarising/ rewording/ giving example (49.87) and speech act marker- question or request (16.73) because they are above the critical value for $p < 0.0001$, which is 15.13.

Regarding well, the functions that have the highest LL values include; searching for the right phrase (293.01), rephrasing or correcting (126.25), quotative well (11.43), move to main story (57.85), indirect answer (33.75), contributing an opinion (49.21) and direct answer (241.59). These functions occurred more in the ICE-GB compared to the ZE corpus. Interestingly, the function of well that is response to self-raised expectations was not attested in both the ZE corpus and the ICE-GB. Only three of the functions of well did not show statistically significant variations between the two corpora.

The sequentially contrastive function of but shows statistically significant variations between BrE and the ICE-GB. This is because the function occurred 9.6 times per 10 000 words in BrE, which is higher than 5.2 times in ZE. The LL value is 58.03, which is above the critical value of 15.13 where $p < 0.0001$.

The next chapter concludes this study by answering research questions, discussing limitations for this study and suggesting possibilities for future research.
Chapter 8

Conclusion

8.1 Introduction

The discussion about whether ZE is an established variety or not has received some attention by researchers (e.g. Makoni, 1993; Kadenge, 2009, 2010; Mlambo, 2009; Marungudzi, 2016a). There seems to be no consensus on the subject. Some researchers believe that ZE is an interlanguage that improves as speakers become more proficient in English (e.g. Makoni, 1992, 1993) whilst other scholars are of the opinion that there are several features warranting ZE to be regarded as a variety (e.g. Kadenge, 2009, 2010; Mlambo, 2009; Marungudzi, 2016a).

Drawing from the studies mentioned above, this study sought to add to the discussion by establishing whether there are any variations and innovations in the use of articles, modal verbs, and selected DMs (so, well and but) (cf. Chapter 4 for an explanation of statistical significance in variation and the identification of errors versus innovation).

Since the conceptual framework in this study is the DyM (Schneider, 2003; 2007), accounting for the development of postcolonial varieties of English, ZE was investigated as a L2 in the context of WE, where different varieties of English are reported to exist in countries where English is used as a L2. As noted in Section 2.4.3, the different contexts in which English is used in L2 environments lead to variety formation with English adapting to local contexts (Kachru, 1992).

Considering that there are features reported to show variations in the use of English on the eWAVE, the researcher analysed the ZE corpus for some of the features that are reported to show variation with articles and modal verbs to determine whether the features are attested in ZE. In the case of DMs, comparisons were made between the ZE corpus and the ICE-GB to ascertain whether there are variations in the frequency and
functions of the DMs so, well and but. The ICE-GB was used as a reference corpus and was used for comparison. Some studies have shown variations in the use of DMs by L2 speakers of English (e.g. Müller, 2005; Unuabonah, 2019). Kachru (1992) lists variation in DM use as one of the characteristic features of nativisation in L2 varieties of English.

Research question 1 (1.1 to 1.3) is answered in Section 8.2.1 and research question 2 (2.1 to 2.3) is answered in section 8.2.2. Thereafter, the study focuses on research questions 3 and whether there are innovations in ZE in Section 8.2.3. On the basis of the answers to the three research questions, a determination is made in Section 8.2.4, about the stage at which ZE is in the DyM, focussing on research question 4. Section 8.3 discusses the limitations of this study and possibilities for future research. The chapter concludes in Section 8.4.

### 8.2 Revisiting the research questions

#### 8.2.1 Frequencies: Establishing whether there are variation in ZE

This study firstly aimed at answering the three questions outlined below.

1.1 What are the occurrence frequencies of articles, modal verbs and discourse markers (so, well and but) in the ZE corpus and the ICE-GB? (Cf. point five in Section 1.5.)

1.2 How does the occurrence of articles, modal verbs and discourse markers in ZE compare to that of BrE?

1.3 Are the variations between the corpora statistically significant? (Cf. point one and three in Section 1.5.)

In the separate results sections of Chapter 5 (on articles), Chapter 6 (on modal verbs) and Chapter 7 (on discourse markers) an overall occurrence frequency was provided for the articles the, a/an and the zero article, the modal verbs can, could, may, might, must, shall, should, will and would and the discourse markers so, well and but. These occurrence frequencies were calculated for both the ZE corpus and the ICE-GB. The detailed answer
to question 1.1 can thus be found in the results section of each chapter. These frequencies were also calculated for different registers and genres.

In order to answer question 1.2, the results outlined in research question 1.1 were compared at the hand of the Log likelihood statistical test to see if there is a difference between the occurrence frequencies of each construction and subtype. The detailed p-values are provided in the different subsections of the quantitative results discussion. These p-values thus indicate if there is a statistical significance between the occurrences of these constructions in the two corpora and were subsequently used to answer research question 1.3. Due to the nature of the Log likelihood test there are different critical values associated with the p-values where one asterisk (*) is used where p < 0.05, meaning that log likelihood > 3.84. Two asterisks (**) are used for p < 0.01, meaning that log likelihood > 6.63, whilst three asterisks (***) are used for p < 0.001, meaning that log likelihood > 10.83. Four asterisks (****) are used for p < 0.0001, meaning that log likelihood > 15.13. The following tables provides a summary of the LL and associated p-values for each construction, subtype, register and associated genres as well as a simplified summary to show if each construction, subtype, register and the associated genres show variation between the corpora.
Table 66. A summary of the Log likelihood values and associated p-values for each construction, subtype, register and associated genres to answer research question 1.2 and 1.3.

<table>
<thead>
<tr>
<th>Construction</th>
<th>Articles</th>
<th>Modal verbs</th>
<th>Discourse markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The</td>
<td>A/an</td>
<td>Zero</td>
</tr>
<tr>
<td><strong>Subtype</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>+0.59</td>
<td>-87.63</td>
<td>+14.07</td>
</tr>
<tr>
<td><strong>Spoken register</strong></td>
<td>+9.49</td>
<td>-15.75</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Private dialogue</strong></td>
<td>+238.38</td>
<td>+6.73</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Public scripted monologue</strong></td>
<td>-11.24</td>
<td>-14.31</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Public dialogue</strong></td>
<td>-0.83</td>
<td>-46.70</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Written register</strong></td>
<td>-4.90</td>
<td>-103.34</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Creative writing</strong></td>
<td>+11.46</td>
<td>+0.55</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Popular writing</strong></td>
<td>-15.58</td>
<td>-1.90</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Academic writing: Examination</strong></td>
<td>+0.53</td>
<td>-0.58</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Editorials</strong></td>
<td>-6.10</td>
<td>-8.86</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Newspaper reportage</strong></td>
<td>-10.99</td>
<td>-63.38</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Social letters</strong></td>
<td>-3.05</td>
<td>-21.18</td>
<td>n.a</td>
</tr>
<tr>
<td><strong>Business letters</strong></td>
<td>-56.66</td>
<td>-66.95</td>
<td>n.a</td>
</tr>
</tbody>
</table>

| **Log Likelihood**                     |          |             |                   |           |            |            |
| **Subtype**                            |          |             |                   |           |            |            |
| **Overall**                            | +124.14  | -28.41      | -39.95            | +40.32    | -442.96    |           |
|                                       | ***      | ****        | ****              | ***       | ****       | ****       |
| **Spoken register**                    | +147.13  | -12.93      | -5.47*            | +51.64    | -321.42    |           |
|                                       | ***      | ****        | ***               | ***       | ****       | ****       |
| **Private dialogue**                   | +99.13   | +0.42       | -1.14             | +63.81    | -98.61     |           |
|                                       | ***      | ****        | ****              | ***       | ****       | ****       |
| **Public scripted monologue**          | +2.92    | -4.81       | -0.36             | -0.20     | -21.61     |           |
|                                       | *        | ****        | ****              | ***       | ****       | ****       |
| **Public dialogue**                    | +20.02   | -36.43      | -1.15             | +11.44    | -26.30     |           |
|                                       | ****     | ****        | ****              | ***       | ****       | ****       |
| **Written register**                   | +1.47    | -16.88      | -52.90            | +1.66     | +122.94    |           |
|                                       | ****     | ****        | ****              | ***       | ****       | ****       |

© University of Pretoria
Table 67. Variation and variation direction for each construction, subtype, register and associated genres to answer research question 1.2 and 1.3.

<table>
<thead>
<tr>
<th>Construction</th>
<th>Articles</th>
<th>Subtype</th>
<th>Log Likelihood</th>
<th>Discourse markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>No</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>No</td>
</tr>
<tr>
<td>Spoken register</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>n.a</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td></td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
</tr>
<tr>
<td>Private dialogue</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>n.a</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td></td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
</tr>
<tr>
<td>Public scripted monologue</td>
<td>No</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>No</td>
</tr>
<tr>
<td>Public dialogue</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>n.a</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td></td>
<td>Var. ZE&gt;BrE</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
</tr>
<tr>
<td></td>
<td>Var. BrE&gt;ZE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
</tr>
<tr>
<td>Creative writing</td>
<td>Var. ZE&gt;BrE</td>
<td>No</td>
<td>No</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td>Popular writing</td>
<td>Var. ZE&gt;BrE</td>
<td>No</td>
<td>No</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td>Academic writing:</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
<td>Var. ZE&gt;BrE</td>
</tr>
<tr>
<td>Examination</td>
<td>Var. ZE&gt;BrE</td>
<td>No</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td>Editorials</td>
<td>Var. ZE&gt;BrE</td>
<td>No</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. BrE&gt;ZE</td>
</tr>
<tr>
<td>Social letters</td>
<td>Var. ZE&gt;BrE</td>
<td>No</td>
<td>Var. BrE&gt;ZE</td>
<td>Var. BrE&gt;ZE</td>
</tr>
</tbody>
</table>

© University of Pretoria
For articles, an overall occurrence analysis was done in which it was found that only the indefinite and zero articles occurred in statistically significant variation between the ZE corpus and the ICE-GB. These articles however occur in different directions as the indefinite article occurs more times in the ICE-BG in comparison to the ZE corpus and the zero article occurs more times in the ZE corpus than the ICE-GB. Due to the smaller sample size (cf. subsection c in Section 5.6.1.2), the frequencies of the zero article in the spoken and written registers and subgenres were not calculated.

In the spoken and written registers of the two corpora data analysis revealed that for the and a/an, there were statistically significant differences. In this case definite articles occur more times in the ZE corpus in comparison to the ICE-GB and the indefinite articles occur more times in the ICE-GB than in the ZE corpus. In terms of the spoken register genres, specifically for the private dialog the frequency showed statistically significant variations that were higher in the ZE corpus for the and a/an. For the public scripted monologues genre, the statistically significant variation however was higher in the ICE-GB for the and a/an. The public dialogue genre only showed statistically significant variation for the indefinite article where the frequency was also higher in the ICE-GB.

In the written register, both the definite and indefinite articles showed statistically significant variation where the frequency was higher in the ICE-GB. For the variation in the written register genres occurred where the frequency was higher in the ICE-GB in creative writing, editorials, newspaper reportage and in business letters, while the frequency was higher in the ZE corpus for popular writing and academic writing. A /an showed statistically significant variation only in editorials, newspaper reportage, in business letters and in social letters all with higher frequencies in the ICE-GB.

For the overall occurrence of modal verbs, statistically significant variations were noted in the occurrence frequencies of can and will/’ll, with the ZE corpus recording higher normalised frequencies compared to the ICE-GB. The ICE-GB had higher normalised frequencies of could, may, might, and would compared to the ZE Corpus with LL values of these modal verbs being all higher than the critical value where p < 0.0001, meaning that
the variations were not due to chance and showed **variation**. No statistically significant variations were observed in *must, shall* and *should*.

In the **spoken register**, statistically significant variations were recorded in the frequency of *can* and *will/*ll, with the ZE corpus recording more instances compared to the ICE-GB. In the spoken register, statistically significant variations were recorded in the frequencies of *could, may* and *might*, and *would/*d, with the ICE-GB recording more frequencies compared to the ZE corpus. *Shall* was attested more in the ICE-GB compared to the ZE corpus, with a statistically significant variation where \( p < 0.0001 \). No statistically significant variations were reported in the frequency of *must* and *should*. In the **written register** only *can, must, should*, and *will/*ll did not show statistically significant variations and only *would* had a higher frequency in the ZE corpus.

Regarding the **genres**, the ZE corpus showed statistical significance for *can* in private dialogue and public dialogue in the spoken register and for social letters and business letters in the written register with higher frequency in the ZE corpus in comparison to the popular writing register which showed variation but with higher frequencies occurring in the ICE-GB. The ICE-GB recorded more instances of *could* in public scripted monologue, public dialogue for the spoken register genres while for the written register genres of academic writing, editorials, newspaper reportage, and business letters showed statistically significant variation. Only the academic writing genre showed variation where the frequency was higher in the ZE corpus for the academic writing genre.

For the spoken register genres *may* had a low statistical significance overall but no statistical significance in the genres. The LL value for editorials, newspaper reportage, popular writing, and business letters with higher frequencies in the ICE-GB showed that *may* had statistically significant variations. In private dialogue and public dialogue in the spoken register genres as well as creative writing, editorials, newspaper reportage, and business letters in the written register genres, statistically significant variation occurred for *might* with higher frequencies occurring in the ICE-GB. *Must* occurred more in the ICE-GB compared to the ZE corpus in private dialogue but higher in the ZE corpus for the scripted
dialog in the spoken register genres while for social letters, popular writing, and creative writing there was statistically significant variation for the written register genres where only newspaper reportage had higher frequencies in the ZE corpus.

Due to higher LL values, statistically significant variations were noted in the frequency of shall, with statistically significant variation for private dialogue only in the spoken register genres with higher frequencies in the ICE-GB compared to the statistically significant variation for newspaper reportage and in business letters in the written register genres where there were higher frequencies for the in the ZE corpus. No statistically significant variations were in the frequency of should in private dialogue, public scripted monologue, public dialogue. Whilst statistically significant variations were observable in the frequency of should, in popular writing, academic writing, editorial and business letters, should was attested with higher frequencies in the ICE-GB only in the popular writing and in the ZE corpus for the other genres with variation.

The statistically significant variation of occurrence frequencies of will/’ll were higher in the ZE corpus in private dialogue and public dialogue in the spoken register genres while for the written register genres only academic writing did not show statistically significant variation. Creative writing, popular writing, editorials and newspaper reportage showed higher frequencies in the ICE-GB while social and business letters showed higher frequencies in the ZE corpus. Regarding would/’d, statistically significant variations were noted in private dialogue, public scripted monologue and public dialogue in the spoken register genres with higher frequencies occurring in the ICE-GB while it also only showed no statistical significant variation for the academic writing genre. The ICE-GB showed higher frequencies for creative writing, editorials, newspaper reportage and business letters while the frequencies were higher in the ZE corpus for popular writing and social letters.

For the overall occurrence of discourse markers, the DMs so, well and but were analysed for their frequencies, where there were statistically significant variation for all three DMs with frequencies being higher for so in the ZE corpus in comparison to those of well and but in the ICE-GB. In terms of the spoken register and associated genres all three DMs also
showed a statistically significant variation for private dialogue however only *well* had higher frequencies in the ICE-GB. For public scripted monologue *so* had statistically significant variation with higher frequencies in the ZE corpus in comparison to *but* having higher frequencies in the ICE-GB. For public dialogue both *well* and *but* had statistically significant variation with higher frequencies in the ICE-GB. For the **written register** only *but* showed a statistically significant variation between the two corpora. There were statistically significant differences in the frequency of the DM *but* in popular writing with higher frequencies occurring in the ZE corpus and for editorials and newspaper reportage with higher frequencies occurring in the ICE-GB for the written register genres. Interestingly, the DM *so* was used more in the ZE corpus compared to ICE-GB. These results are comparable to some studies that have shown variations in the use of DMs by L2 speakers of English (e.g. Müller, 2005; Unuabonah, 2019).

The statistically significant variations of the articles, modal verbs and DMs (*so, well and but*) in the spoken and written registers, and in different genres, point to the fact that there are differences in the frequencies of these constructions in the ZE corpus and the ICE-GB. No statistically significant variations in the spoken and written registers and in different genres in these constructions means that the frequencies of these constructions are not different in the two corpora. These reported variations and no variations will aid in the determination of the stage of ZE in the DyM.

**8.2.2 Establishing whether there are variations in function frequencies**

This study secondly aimed at answering the three questions outlined below.

2.1 What are the function frequencies of articles, modal verbs and discourse markers (*so, well and but*) in the ZE corpus and the ICE-GB? (Cf. point five in Section 1.5.)

2.2 How do the functions of articles, modal verbs and discourse markers in ZE compare to those of BrE?

2.3 Are the variations between the corpora statistically significant? (Cf. point one and three in Section 1.5.)
In the separate results sections of chapter 5 (on articles), chapter 6 (on modal verbs) and chapter 7 (on discourse markers) an overall function frequency was provided for the articles *the, a/an* and the zero article, the modal verbs *can, could, may, might, must, shall, should, will* and *would* and the discourse markers *so, well* and *but*. These function frequencies were also calculated for both the ZE corpus and the ICE-GB. The detailed answer to question 2.1 can thus be found in the results section of each chapter.

In order to answer question 2.2, the results outlined in research question 2.1 were compared at the hand of the Log likelihood statistical test to see if there is a difference between the function frequencies of each construction and subtype. The detailed p-values are provided in the different subsections of the quantitative results discussion. These p-values thus indicate if there is a statistical significance between the functions of these constructions in the two corpora and were subsequently used to answer research question 2.3. Table 68 provides a summary of the LL values, which are associated with statistically significant p-values for each construction, subtype and function and are outlined in green while table 69 provides a simplified summary to show if each construction, subtype, and function shows variation between the corpora. This variation is also identified in green and in bold typeface.
Table 68. A summary of the Log likelihood values and associated p-values for each construction, subtype and function to answer research question 2.2 and 2.3.

<table>
<thead>
<tr>
<th>Functions</th>
<th>Articles</th>
<th>Modal verbs</th>
<th>Discourse markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The</td>
<td>a/an</td>
<td>8 other modal verbs</td>
</tr>
<tr>
<td></td>
<td>Quantification</td>
<td>-37.62***</td>
<td>Epistemic</td>
</tr>
<tr>
<td></td>
<td>Quantification</td>
<td>-25.92***</td>
<td>Epistemic</td>
</tr>
<tr>
<td></td>
<td>Ascription (attribution)</td>
<td>-50.91***</td>
<td>Epistemic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Co-text bound non-part.</td>
<td>-50.39***</td>
<td>Deontic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 69. Variation and direction for each construction, subtype and function.

<table>
<thead>
<tr>
<th>Articles</th>
<th>Modal verbs</th>
<th>Discourse markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>a/an</td>
<td>should</td>
</tr>
<tr>
<td>Anaphoric: Direct Var. BrE&gt;ZE</td>
<td>Quantification Var. BrE&gt;ZE</td>
<td>must</td>
</tr>
<tr>
<td>Anaphoric: indirect Var. BrE&gt;ZE</td>
<td>Reference to specific thing Var. BrE&gt;ZE</td>
<td>may</td>
</tr>
<tr>
<td>Immediate Situation Use Var. BrE&gt;ZE</td>
<td>Ascription (attribution) Var. BrE&gt;ZE</td>
<td>might</td>
</tr>
<tr>
<td>Cataphoric reference Var. BrE&gt;ZE</td>
<td>Generic uses Var. ZE&gt;BrE</td>
<td>will</td>
</tr>
<tr>
<td>Sporadic reference Var. BrE&gt;ZE</td>
<td>Idiomatic uses in fixed expressions</td>
<td>would</td>
</tr>
<tr>
<td>Non-Referential Var. BrE&gt;ZE</td>
<td>Indeterminate</td>
<td>should = would</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Subjunctive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dynamic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Should = would</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unclassified</td>
</tr>
</tbody>
</table>
For articles the function frequencies and the subsequent statistical comparison in terms of the LL showed that all articles and all functions of the articles show statistically significant variation between the ZE corpus and the ICE-GB. The LL values were calculated for the articles to determine whether there are statistically significant variations in the frequencies of different functions. For the definite article all functions had statistically significant variations and had higher frequencies in the ICE-GB than in the ZE corpus with the exception of the sporadic reference that had higher frequencies in the ZE corpus. For the indefinite article all functions had statistically significant variations and had higher frequencies in the ICE-GB than in the ZE corpus with the exception of the generic reference that had higher frequencies in the ZE corpus.

When considering the frequencies and functions of modal verbs, a comparative analysis of the ZE corpus data and the ICE-GB data showed no statistically significant differences in the frequencies of must and should per 10 000 words.

For the epistemic function may, might, could, will, would and shall showed statistically significant variation with higher frequencies occurring in the ICE-GB except for will and shall that had higher frequencies in the ZE corpus. For the deontic function, only might occurred with a higher frequency in the ZE corpus and shall with a higher frequency in the ICE-GB. Both had statistically significant variation. Might, could and would had a higher frequency in the ICE-GB while can, will and shall had higher frequencies in the ZE corpus, with all six modal verbs having a statistically significant variation for the dynamic function. Lastly for the indeterminate function only can with a higher frequency in the ZE corpus and will with a higher frequency in the ICE-GB had statistically significant variation.

To determine the frequencies of different meanings of modal verbs, a quantitative functional analysis was done in order to inform on the frequencies of the different meanings. There were no statistically significant differences in the functions of obligation and necessity modals must and should because the LL values were below 3.84 where p < 0.05. Regarding the modals of possibility, permission and ability, some variations were observed in the frequencies per 10 000 words. For example, it was
recorded that the epistemic function of *may* was attested more in the ICE-GB. The dynamic function of *can* was more prevalent in the ZE corpus. BrE speakers use the dynamic function of *could* more frequently compared to ZE speakers. If we consider modals of prediction and volition, the epistemic *will* was attested more in the ZE corpus whilst the epistemic *would* and the deontic *shall* were more frequent in the ICE-GB. Other functions of the modals did not show statistically significant differences. The dynamic and indeterminate functions of *shall* and the deontic and indeterminate functions of *will* did not show statistically significant differences.

In this thesis, it was found that DMs function as coherence relations and that they help to link text units (Schourup, 1999). DMs are believed to play an important role in interpreting texts because they show coherence relations between discourse units (Schiffrin, 1987). Regarding function frequencies, statistically significant variations were evident in the **frequencies of functions of so**, with the ZE corpus recording higher frequencies than the ICE-GB in the following functions, namely, marking result or consequence, main idea unit marker, summarizing / rewording / giving an example and speech act marker - question or a request. Statistically significant variations were noted in the frequency of the sequential *so* function, which was more frequent in the ICE-GB compared to the ZE corpus. Due to LL values being lower than 3.84, where *p* < 0.05, no statistically significant variations in the frequencies of the boundary marker function, the speech act marker – function, the marker of a transition relevance place function, and the unclassified function were recorded.

For the **frequencies of functions of well**, statistically significant variations were recorded, with the ICE-GB recording higher frequencies of the following functions compared to the ZE corpus, namely, searching for the right phrase, rephrasing/correcting, quotative *well*, move to the main story, indirect answer, direct answer, contributing an opinion, continuing an opinion/answer and evaluating a previous statement. There were no statistically significant variations in the frequencies of the following functions, namely, introducing the next scene, conclusive *well*, response to self-raised expectations and unclassified. This is because the LL values were lower than the critical value of 3.84, where *p* < 0.05. Four functions of *well* namely introducing the next scene, unclassified,
conclusive well, and response to self-raised expectations did not show statistically significant variations between the two corpora.

If the LL values are considered, only the sequentially contrastive function of but showed statistically significant variations between the ZE corpus and the ICE-GB, with the function being attested more in the ICE-GB compared to the ZE corpus. The contrastive function and the unclassified function did not show statistically significant variations.

The statistically significant variation of these constructions in functions points to the fact that ZE speakers and BrE speakers do not use the functions of articles, modal verbs and DMs (so, well and but) in a similar way. Where no statistically significant variations were recorded in the frequency of the functions of articles, modal verbs and DMs (so, well and but), this points to the fact that ZE speakers and BrE speakers use these functions the same way. These reported variations and no variations will aid in the determination of the stage of ZE in the DyM.

8.2.3 Determining whether there is innovation in ZE

The following research questions were answered in this section.

3.1 Which features of innovation are present with regards to articles and modal verbs in the ZE corpus?

3.2 Do the features in the corpora show innovation for ZE? (Cf. point one and three in Section 1.5.)

Regarding whether there is innovation in the use of articles in ZE, results indicated that there was innovation in some features whilst in other features there was no innovation. The features studied were grouped in different categories in Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features. It can be concluded that there is innovation in feature 67 as shown by the use of demonstratives for definite articles. There was a considerable number of occurrences of feature 67 (90 times) in the ZE corpus warranting its classification into category B (feature is neither pervasive nor extremely rare). The
fact that the use of demonstratives for definite article was attested in 7 out of 10 genres of the ZE corpus warranted feature 67 to be assigned to B category.

It was found that the use of zero article where StE has definite article (feature 62), the use of zero article where StE has indefinite article (feature 63), and the use of indefinite article where StE favours zero (feature 65) belonged to category C. This shows that the features exist, but are extremely rare.

Regarding articles, there was attested absence of feature 60; the use of definite article where StE has indefinite article, feature 61, the use of indefinite article where StE has definite article, feature 64, use of definite article where StE has favours zero, and feature 66, the use of indefinite article one/wan because their occurrences were fewer, which led to the examples noted in the corpus being regarded as errors. This is in line with Van Rooy's (2013) suggestion that if features are concentrated in a few genres or a few speakers, then they can be regarded as errors. When compared to results from Minow (2010), article omission, (represented by features 62 and 63 in this study), is a less frequent feature in BSAE. The author notes that article insertion (represented by features 64 and 65 in this study) is more common in BSAE. Similarities can also be observed in the results by Minow (2010) regarding article substitution (represented by features 60 and 61 in this study) because results from this study also indicate that the use of definite article where StE has indefinite article and the use of indefinite article where StE has definite article is absent in BSAE. With regard to definite and indefinite article omission, Minow (2010) draws attention to the high suppliance rates as evidence of how relatively marginal article omission is in BSAE. A comparative study by Sand (2004) showed that contact varieties namely Kenyan English, Singapore English, Irish English, Jamaican English, and Indian English, use definite articles in contexts that are different from StE contexts.

Turning to modal verbs, ZE showed less innovation because an analysis of feature 121 (double modals) and feature 123, whereby there is use of present tense forms of modals where StE has past tense forms, showed that the features exist, but are extremely rare. The two features are in category C. There was no evidence of feature 127, which shows
the non-standard use of modals for politeness reasons in the ZE corpus. Therefore, feature 127 belongs to category D. The different categories of the features studied are highlighted below.

*Table 70. Summary of features and their categorisations in ZE*

<table>
<thead>
<tr>
<th>Feature</th>
<th>Category of feature in ZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 60: Use of definite article where StE has indefinite article.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>Feature 61: Use of indefinite article where StE has definite article.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>Feature 62: Use of zero article where StE has definite article.</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>Feature 63: Use of zero article where StE has indefinite article.</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>Feature 64: Use of definite article where StE favours zero.</td>
<td>(D) Attested absence of feature</td>
</tr>
<tr>
<td>Feature 65: Use of indefinite article where StE favours zero [article].</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>Feature 66: Indefinite article one/wan.</td>
<td>(D) Attested absence of the feature.</td>
</tr>
<tr>
<td>Feature 67: Demonstratives for definite article.</td>
<td>(B) Feature is neither pervasive nor extremely rare</td>
</tr>
<tr>
<td>Feature 121: Double modals</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>Feature 123: Present tense forms of modals used where StE has past tense forms.</td>
<td>(C) Feature exists, but is extremely rare</td>
</tr>
<tr>
<td>Feature 127: Non-standard use of modals for politeness reasons.</td>
<td>(D) Attested absence of feature</td>
</tr>
</tbody>
</table>

From the table above, the use of demonstratives for definite article (feature 67) was regarded as belonging to category B in Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features. Features 62, 63, 65, 121 and 123 show some innovation in the way they are used in ZE due to the fact that, even though the features are extremely rare, they exist in different genres, are attested in both spoken and written registers, are used in editorials, newspaper reportage, and the media (norm orientation) and are used by different speakers. It was found that the features fulfil some of the requirements of an innovation (subsection e in Section 4.4.2.6) suggested by Kruger and Van Rooy (2017) and Bamgbose (1998). This corresponds with Botha’s (2012) findings on BSAE showing...
that although some differences exist in the use of specific constructions, overall, these differences are fewer. An examination of the ZE corpus together with results from questionnaires showed that ZE users tend to follow L1 speaker norms when considering features 60, 61, 64, 66, and 127.

8.2.4 The stage of ZE in the dynamic model

This section focuses on answering research question four as outlined below:

What is the stage of adult use of these features in ZE when considering the dynamic model of postcolonial English?

Regarding the stage in the DyM, ZE seems to be at stage 2 when considering some features (as outlined below), which were attested to exist but were extremely rare and at stage 3 when considering feature 67 that showed neither pervasive nor extremely rare (category B). Further explanations are given below.

According to Schneider (2003, 2007), the emergence of innovative features is a hallmark of stage 2 of the DyM. The model posits that at stage 2, differences between indigenous and standard strands of English emerge. In this study, observations were made about some features that were attested with limited frequency. These observations about the variation in frequency of features aided in the determination of the stage that ZE can be classified in the DyM, with specific reference to nativisation and innovation. These features existed, but were extremely rare (level C). The features are; use of zero article where StE has definite article (feature 62), use of zero article where StE has indefinite article (feature 63), use of indefinite article where StE favours zero (feature 65), the use of double modals (feature 121), and the use of present tense forms of modals where StE has past tense forms (feature 123). It can be concluded that the abovementioned features seem to be emerging and have not yet been established in ZE.

ZE can also be situated at stage 3 in the DyM. This is because evidence from the corpus shows that feature 67 shows innovation in ZE because it belongs to category B in Kortmann, Lunkenheimer and Ehret’s (2020) categorisation of features. In addition, there were statistically significant variations in the normalised frequencies in articles,
modal verbs, and DMs (so, well and but). This can be taken as showing that there are variations in the use of these constructions in ZE compared to BrE.

Due to the above observations, there seems to be change at lexico-grammatical level which is a characteristic of stage 3 of the dynamic model. At the same time, some features are not attested (e.g. 60, 61, 64, 66, and 127). ZE seems to follow L1 norms in the use of these features. This finding corresponds with Marungudzi (2016a) who observes that ZE is at both stage 2 and 3 of the DyM. There are no grammar books, dictionaries, and user guides of ZE. In schools, L1 conventions are the norm in teaching and learning (Marungudzi, 2016a). This shows that ZE has not been accepted as a variety in its own right. Scholars are still debating about whether ZE can be regarded as a variety or not (e.g. Makoni, 1993; Kadenge, 2009, 2010; Mlambo, 2009; Marungudzi, 2016a).

8.3 Limitations and possibilities for future research

This study has created a corpus of ZE, which can be used as a linguistic resource in future studies. Real life communication samples can be utilised for speech-to-text and text-to-speech technology. In addition, since corpus studies on ZE is still a less researched area, this study has added to the empirical data on ZE. By exploring articles, modal verbs and selected DMs, the study has provided a better understanding of how ZE speakers use English as a L2. This contributes towards the advancement of knowledge and towards linguistic theory. Despite these strengths, there are several limitations, which present possibilities for further research.

One of the limitations is that although semi-structured interviews were used to determine whether participants can converse in English, the interviews did not provide a clear picture of participants’ linguistic proficiency levels. This is because according to Brière (1971: 385) “proficiency in a language is much more difficult to define and, obviously, much more difficult to measure in a testing situation”. Future research could do a linguistic proficiency test to determine participants’ level of proficiency.

The Observer's paradox is another shortcoming of this study. “The aim of linguistic research in the community must be to find out how people talk when they are not being
systematically observed; yet we can only obtain these data by systematic observation” (Labov, 1972: 209). Labov suggested that a researcher has to find ways of overcoming this limitation. In this study, the researcher recorded some participants. To limit the effects of the observer’s paradox, some participants were requested to record themselves during private conversations. Further research should limit the number of recordings done in the presence of the researcher.

The mixed sampling method used for gathering data (snowball, convenience and purposive) was the most suitable for this study considering time and financial limitations. However, the corpus was not expected to be representative of the way all Shona-English bilinguals use English as a L2. Efforts were made to gather data from different registers and genres in order to have a balanced corpus. This study did not seek to generalise the innovation features of all ZE speakers. The focus was on selected participants and selected features. It would be intriguing if future research could use different sampling methods and include more diverse genres and speakers.

Due to the scope of this study, it was not possible to study all the DMs. Future research could examine other DMs not included in this study.

The absolute and normalised frequency of the zero article in different registers and genres and the function frequencies were not calculated due to the smaller samples of the zero article in 400 nouns. Future research could examine the frequencies of the zero article in different registers and genres and the function frequencies.

Lastly, due to time and financial constraints, the corpus size was limited. It would be interesting to see whether a bigger corpus of ZE could yield different results.

8.4 Conclusion

From the findings in this study, it was observed that there were statistically significant variations in the frequencies and functions of articles, modal verbs in ZE in some instances. In other instances, data analysis showed that ZE speakers seem to follow StE norms. In relation to features 62, 63, 65, 121, and 123, L2 users of English who speak
Shona as their L1 seemed to deviate slightly from L1 English conventions. This is because the features are extremely rare. The use of demonstratives for definite article (feature 67) was attested to be neither pervasive nor extremely rare. Although there may seem to be some innovations in the way English is used in Zimbabwe, attitudes towards the new norms may hinder acceptance (Marungudzi, 2016b; Mareva, Kaburise & Klu, 2016). There was also evidence showing that ZE follows L1 English conventions in the use of articles, modal verbs, and DMs (so, well and but). Features 60, 61, 64, 66, and 127 were not attested in the ZE corpus. These results seem to support the idea that ZE is at both stage 2 and stage 3 of the DyM. Whether ZE moves on to another stage or fossilises, will depend on a number of factors such as attitudes towards ZE and acceptance of the variety. As gleaned from the literature review, there is no consensus on the variety status of ZE. By examining articles, modal verbs and DMs, this study has contributed towards a better understanding of how English is used in an L2 context.
Bibliography


© University of Pretoria


Fraser, B. 1993: Discourse Markers Across Language. Pragmatics and Language Learning
Monograph Series, 4:1-16.


and New York: Mouton de Gruyter.


Hasty, J.D. 2012. We might should oughta take a second look at this: Syntactic re-analysis of double modals in Southern United States English. Lingua, 122: 1716-1738.


Huddleston, K. & Fairhurst, M. 2013. The pragmatic markers anyway, okay, and shame: A


Kadenge, M. 2012. The indigenisation of English in Chenjerai Hove’s novel Bones and


© University of Pretoria


Myllyniemi, S., 2015. Modal verbs can, could, may, might, should and must and semi-modals ought to and have to in spoken Scottish English as compared to spoken English. Unpublished Master's thesis, University of Tampere.


Rayson, P. & Garside, R 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. Hong Kong, 1-6.).


Shinobu, R. 2018. The concept of “Distance” and English modal verbs. *Fora 2*: 64-82.


The Sunday Mail. [https://www.sundaymail.co.zw/](https://www.sundaymail.co.zw/). Last accessed on: 31 December 2019./


Appendices

Appendix A: Consent form for individuals

TO WHOM IT MAY CONCERN

CONSENT FORM

You are invited to participate in a research study whose details follow below.

The research is about an investigation of English as spoken by Shona mother tongue speakers. Your participation involves completing a small biographical questionnaire and participating in informal interviews. The interviews will be conducted in English and recorded for 20 minutes. You may be asked to record your English conversations for 20 minutes. Please read this letter carefully and confirm your willingness to participate in the study by filling in your name and signing in the spaces provided. I undertake to guarantee anonymity and confidentiality.

RESEARCH TITLE

An investigation of L2 English as spoken by Shona mother tongue speakers using a corpus-based approach

RESEARCHER'S DETAILS:  
Faith C. Chapwanya
Department of Afrikaans
University of Pretoria
Pretoria
0002
Email: fchapwanya@gmail.com
AIMS OF THE RESEARCH

● To investigate Zimbabwean English as a second language (L2) spoken by Shona mother tongue speakers. This will be done in order to determine whether there is variation with determiners, discourse markers, and modals of Zimbabwean English as compared to that of Mainstream BrE.

● To examine the influence of the Shona language on Zimbabwean English. In order to do this, speech samples will be collected from Shona-English bilinguals. A corpus will be created using the collected samples. The corpus will be analysed to determine how Zimbabwean English as a L2 is written and spoken. The corpus may be used for further research.

TASKS IN PARTICIPATION

Your participation in the study will involve:

● Completing a small biographical questionnaire.
● Participating in informal interviews. The interviews will be conducted in English in order to determine your mastery of English.
● Recording your English conversations for 20 minutes.

OTHER DETAILS

● The interviews and recordings will take place in your private home or any preferred location that you may choose.
● The interviews and private conversations will be recorded on a smartphone or audiotape for 20 minutes.
● There are no specific financial or material benefits in the participation, but sharing in the experience of the research will be greatly appreciated.
● There are no risks associated with participating in this study.
● The information you give as well as your identity will be treated with utmost confidentiality.
● You have the right to voluntarily withdraw from the study at any point without any repercussions.
● If you choose to withdraw from the study, data collected from you will be destroyed.
● You may contact the researcher at any time should you have any questions.
● If you wish, I will provide the results of the study to you upon completion.

DECLARATION

Having read through the contents of this consent document and understood their implication, I……………………………………..……………………………………….. hereby declare that I will participate in the study.

Signature of researcher: _________________________________

Signature of participant: _________________________________

Date: ___________________________________
Appendix B: Consent form for private business letters

TO WHOM IT MAY CONCERN

CONSENT FORM

You are invited to participate in a research study whose details follow below.

The research is about an investigation of English as used by Shona mother tongue speakers. Your participation involves providing samples of your private business letters. Please read this letter carefully and confirm your willingness to participate in the study by filling in your name and signing in the spaces provided. Complete anonymity in reporting and confidentiality throughout data collection and subsequent reporting in reporting is assured.

RESEARCH TITLE

An investigation of L2 English as spoken by Shona mother tongue speakers using a corpus-based approach

RESEARCHER'S DETAILS: Faith C. Chapwanya
Department of Afrikaans
University of Pretoria
Pretoria, 0002
Email: fchapwanya@gmail.com
Tel: +27 82 5895721
AIMS OF THE RESEARCH

● To investigate how Zimbabwean English as a second language (L2) is written and spoken by Shona mother tongue speakers. This will be done in order to determine whether there is variation with determiners, discourse markers, and modals of Zimbabwean English as compared to that of Mainstream BrE.

● To examine the influence of the Shona language on Zimbabwean English.

In order to do this, samples of samples of business letters will be collected from Shona-English bilinguals. A corpus will be created using the collected samples. The corpus will be analysed to determine how Zimbabwean English as a L2 is written and spoken. The corpus may be used for further research.

TASKS IN PARTICIPATION

Your participation in the study will involve providing samples of your private business letters.

OTHER DETAILS

● There are no specific financial or material benefits in the participation, but sharing in the experience of the research will be greatly appreciated.

● There are no risks associated with participating in this study.

● The information you give as well as your identity will remain anonymous and data gathered will be treated with utmost confidentiality.

● You have the right to withdraw from the study at any point without any repercussions.

● You may contact the researcher at any time should you have any questions.

● If you wish, I will provide the results of the study to you upon completion.

DECLARATION
Having read through the contents of this consent document and understood their implication, I hereby declare that I will participate in the study.

Signature of researcher: _________________________________

Signature of participant: _______________________________

Date: _______________________________
Appendix C: Biographical and language background questionnaire

Questionnaire about biographical information and language background of participants

You are invited to complete this questionnaire as part of your participation in the research about Zimbabwean English. I undertake to guarantee anonymity in reporting with the use of pseudonyms (your name and personal information will not be disclosed) and confidentiality during data collection and reporting. Please place an X in the appropriate box and fill in details in the required spaces. This is not a test so there are no “right” or “wrong” answers.

1. What is your gender?

☐ Male
☐ Female
☐ Other

2. How old are you? _____________ years old.

3. What is your first language? (mother tongue)?

☐ Shona
4. List all the other languages that you speak and rank them according to proficiency.

1. ______________________________

2. ______________________________

3. ______________________________

4. ______________________________

5. What are your educational qualifications? Please tick all the qualifications that you have obtained.

☐ Form 4
☐ Form 6
☐ Diploma
☐ Degree
☐ Postgraduate

6(a) From which country are your educational qualifications? ______________________

6(b) In which language were these completed? ________________________________

7(a) When did you start learning English?

☐ Before starting school
☐ At school
☐ Other Please provide details ______________________

7(b) Where did you learn English as a second language?

☐ At home
7(c) At what age did you start learning English?

7(d) For how many years have you used English?

7(e) Where do you usually use English?

8. Have you ever been taught by someone who spoke English as a first language (a native English-speaking teacher)?

8.1 If yes, for how long?

9. What variety or type of English did you learn?
9.1 If you learned English at school, was the English that you learned at school based on any of the following?

☐ BrE

☐ American English

☐ I don’t know the variety or type that I learned.

☐ Other Please provide details ________________________________

10. How proficient are you in English (How well do you speak English)?

☐ Very good- I am able to converse about general matters of daily life and special topics at a high level.

☐ Good- I am able to converse about familiar daily topics.

☐ Average- I am able to hold simple conversations such as greetings and introducing someone.

☐ Not so good/bad- I am able to give simple greetings using set words and phrases.

11. Approximately how many hours a week do you speak English? _____________

12. English ability: Please rate your current overall proficiency in English:

☐ Upper intermediate level and over: Able to converse about general matters of daily life and topics of one’s specialty and grasp the general meaning of lectures and broadcasts. Able to read high-level material such as newspapers and write about personal ideas.

☐ Intermediate level: Able to converse about familiar daily topics. Able to read material about familiar everyday topics and write simple letters.
Post beginner level: Able to hold a simple conversation such as greeting and introducing someone. Able to read material and write a simple passage in elementary English.

Beginner level: Able to give simple greetings using set words and phrases. Able to read simple sentences, grasp the general meaning of short passages, and to write a simple sentence in English.

13. In what context do you use English? Mark all the appropriate boxes.

☐ Hearing and speaking
☐ Speaking and writing
☐ With co-workers
☐ With my friends and family
☐ With strangers if they speak in English to me first

14. Do you consider yourself as a bilingual, that is, do you consider yourself to be proficient in both Shona and English?

☐ Yes
☐ No
☐ No, I am more proficient in Shona
☐ No, I am more proficient in English
☐ I am a multilingual (I speak more than two languages)

Finally, in the event that I need to contact you again about this research and for record keeping, kindly fill in the following details. Your cooperation is greatly appreciated.

Surname: ____________________________
First name(s): ________________________________

Street address: ________________________________
Telephone number: ________________________________
E-mail address (if available): ________________________________
Occupation: ________________________________

You have come to the end of the questionnaire. Thank you very much for your time.

Appendix D: Semi-structured interview schedule

1. Can you please tell me about yourself?

2. What was your favourite sport at school and why?

2.1 Speaking of sport, what do you think about the Zimbabwe national soccer team’s performance this year?

2.2 If you don’t like sport, what is your favourite hobby?

3. Please tell me about any situation where you felt very proud of yourself.
3.1 What can you say was the scariest situation in your life?

4. What do you usually do during the week?

5. What do you usually do during your free time?

5.1 Follow up question based on the answer.

6. If you use a cell phone what do you think are the most interesting things that can be done on a cell phone?

7. What was your dream career when you were a child?

7.1 What did you need to do in your life if you wanted to follow that career path?

8. Describe a situation where you had to speak to a person of authority?

8.1 How do you think someone else would speak in that kind of situation?

9. Please describe a common problem that you face as a society in the area where you live.

9.1 What steps are you taking as a community to solve this problem?

10. What is the most beautiful place you have ever visited?

10.1 What did you like about the place?

11. Have you ever owned pets in your life?

11.1 If so, what is/was the most exciting things you did with your pet?

11.2 What can you say are the needs of your pet?

12. What is your greatest wish about yourself or your family?

13. If you were a millionaire or if you had a lot of money, how would you spend the money?

13.1 Would you give some money to the poor?

14. Please describe your most treasured possession. In other words, what do you own that you can say is most important to you?

14.1 Why do you say it is important?

14.2 What are the needs of your family?
14.3 Please describe the day you will never forget in detail.

15. Please explain to me in detail what happened during cyclone Idai.

15.1 How was your area affected?

15.2 What happened afterwards in terms of assisting the affected people?

16 If you didn't experience cyclone Idai, please explain what happened during a major disaster or accident.

16.1 What type of assistance did the affected people receive?

17. Please describe the duties of any public figure or any civil servant. For example, a local Member of Parliament, chief, district administrator, teacher, nurse, doctor etc.

18. Do you think there will be normal rains this year?

18.1 Please explain why you say so.

Thank you very much for your time.
Appendix E: Permission letter from North-West University

TO WHOM IT MAY CONCERN

PERMISSION TO USE CORPUS OF ZIMBABWEAN ENGLISH

As part of his PhD at the North-West University, under my supervision, Dr. Thaddeus Marungudzi compiled a corpus of Zimbabwean English in 2011-2013, comprising of samples of letters, published material (newspapers, books, magazines), informal spoken interactions, formal spoken interactions and radio broadcasts.

Dr. Marungudzi sadly passed away on 5 January 2019, and the responsibility to maintain the corpus has now fallen onto me. The North-West University also funded the project, and therefore retains the right to use this corpus for research purposes. All contributors to the corpus signed over permission to use the material for research purposes, but also limited the distribution to bona fide research, and do not allow that the corpus be made available publicly in full. Dr. Marungudzi and I adhered to the ethics requirements of the North-West University that obtained at the time of data collection.

As owner of the corpus, I hereby grant permission for the use of this corpus by Ms Faith Chipwanya and Dr. Joanna Nel as part of the PhD study of Ms Chipwanya under supervision of Dr. Nel. It is standard custom among corpus linguists to make corpora available for research purposes at no fee. There is confidentiality agreement in place, which requires that all names of people and organisations have to be removed when citing examples of the corpus, but otherwise, it is permissible to cite brief extracts from the corpus.

We hope you find this information in order.

Yours sincerely,

Prof AJ van Rensburg

Director: Research Focus Area Understanding and Processing Language in Complex Settings
Appendix F: University of Zimbabwe permission letter

P O Box MP 167
Mount Pleasant
Harare, Zimbabwe
General line: +263-4-303213 Ext 11105
Direct line: +263-4-303284
Fax: +263-4-300943
E-mail: registrar@udein.uz.ac.zw
Website: www.uz.ac.zw

From the Registrar:

DR N A MUTONGORENI

18 March 2019

Ms Faith C Chipwanya
Department of Afrikaans
University of Pretoria
Pretoria
South Africa

Dear Ms Chipwanya

RE: REQUEST TO CARRY OUT A RESEARCH AT THE FACULTY OF ARTS
UNIVERSITY OF ZIMBABWE

I refer to the above matter.

I am pleased to advise you that your request to carry out a research at the Faculty of Arts from 1 March 2019 to 31 December 2019 was approved. By copy of this letter, the Dean, Faculty of Arts is hereby kindly requested to make all necessary arrangements.

Yours sincerely,

DR N A MUTONGORENI
Registrar

Cc: Acting Vice-Chancellor
Dean, Faculty of Arts
Appendix G: Permission letter from The Herald and The Sunday Mail newspapers

ZIMBABWE NEWSPAPERS (1980) LIMITED
(Publishing and Digital Division)

The Chairperson
Faculty of Humanities
University of Pretoria
Pretoria
0002

RE: PERMISSION TO ACCESS THE HERALD ONLINE

This letter serves as permission for your PhD student Faith Chiadza Chidwanya to access and use material from The Herald and The Sunday Mail online during the period of her research which runs from May 3-31 December 2019, whose topic is "An investigation of ZJ English as spoken by Shona mother tongue speakers using a corpus-based approach."

The permission is granted under the following condition which must be adhered to:

1. That all materials should be appropriately cited using the University's academic referencing style, e.g. Harvard Referencing or any style in use at your institute.

Yours sincerely,

[Signature]

Happiness Chikwereka
Group Online Editor - Zimpapers

© University of Pretoria
Appendix H: Results from biographical and language background questionnaires

Table H.1: Participants’ age, gender and occupation (private spoken monologues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$A$&gt;</td>
<td>35</td>
<td>Male</td>
<td>Engineer (Resident)</td>
</tr>
<tr>
<td>&lt;$B$&gt;</td>
<td>23</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>28</td>
<td>Male</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>26</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>30</td>
<td>Male</td>
<td>Self employed</td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>30</td>
<td>Male</td>
<td>Field officer</td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>26</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>22</td>
<td>Female</td>
<td>College student</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>71</td>
<td>Male</td>
<td>Retired teacher, farmer</td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>36</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>25</td>
<td>Male</td>
<td>Graphic designer, music producer, entrepreneur</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>47</td>
<td>Female</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>18</td>
<td>Female</td>
<td>Student</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>46</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>43</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>51</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>30</td>
<td>Female</td>
<td>Banking</td>
</tr>
<tr>
<td>&lt;$R$&gt;</td>
<td>40</td>
<td>Female</td>
<td>Registered nurse</td>
</tr>
<tr>
<td>&lt;$S$&gt;</td>
<td>43</td>
<td>Female</td>
<td>Business owner</td>
</tr>
<tr>
<td>&lt;$T$&gt;</td>
<td>39</td>
<td>Male</td>
<td>Teacher (secondary)</td>
</tr>
<tr>
<td>&lt;$U$&gt;</td>
<td>43</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$V$&gt;</td>
<td>33</td>
<td>Female</td>
<td>Registered nurse</td>
</tr>
<tr>
<td>&lt;$W$&gt;</td>
<td>44</td>
<td>Female</td>
<td>Doctor (gynaecologist)</td>
</tr>
<tr>
<td>&lt;$X$&gt;</td>
<td>56</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$Y$&gt;</td>
<td>66</td>
<td>Male</td>
<td>Retired manager</td>
</tr>
<tr>
<td>&lt;$Z$&gt;</td>
<td>54</td>
<td>Female</td>
<td>Nurse aid</td>
</tr>
<tr>
<td>&lt;$AA$&gt;</td>
<td>32</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$BB$&gt;</td>
<td>39</td>
<td>Male</td>
<td>Prison officer</td>
</tr>
<tr>
<td>&lt;$CC$&gt;</td>
<td>24</td>
<td>Male</td>
<td>Electrician</td>
</tr>
<tr>
<td>&lt;$DD$&gt;</td>
<td>53</td>
<td>Female</td>
<td>Banking officer</td>
</tr>
<tr>
<td>&lt;$EE$&gt;</td>
<td>51</td>
<td>Male</td>
<td>Self employed</td>
</tr>
<tr>
<td>&lt;$FF$&gt;</td>
<td>34</td>
<td>Female</td>
<td>Records assistant</td>
</tr>
<tr>
<td>&lt;$GG$&gt;</td>
<td>28</td>
<td>Male</td>
<td>Medical laboratory scientist</td>
</tr>
<tr>
<td>&lt;$HH$&gt;</td>
<td>53</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$II$&gt;</td>
<td>51</td>
<td>Male</td>
<td>General hand</td>
</tr>
<tr>
<td>&lt;$JJ$&gt;</td>
<td>24</td>
<td>Female</td>
<td>District focal person</td>
</tr>
<tr>
<td>&lt;$KK$&gt;</td>
<td>27</td>
<td>Male</td>
<td>Electrician</td>
</tr>
<tr>
<td>&lt;$LL$&gt;</td>
<td>68</td>
<td>Female</td>
<td>Semi-retired lecturer</td>
</tr>
<tr>
<td>&lt;$MM$&gt;</td>
<td>52</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$NN$&gt;</td>
<td>56</td>
<td>Male</td>
<td>Professor</td>
</tr>
<tr>
<td>&lt;$OO$&gt;</td>
<td>34</td>
<td>Male</td>
<td>Civil engineering technician</td>
</tr>
<tr>
<td>&lt;$PP$&gt;</td>
<td>24</td>
<td>Female</td>
<td>Businesswoman</td>
</tr>
</tbody>
</table>
Table H.2: Summary of participants (private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant name</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$QQ&gt;</td>
<td>35</td>
<td>Female</td>
<td>General hand</td>
</tr>
<tr>
<td>&lt;$RR&gt;</td>
<td>41</td>
<td>Male</td>
<td>Clerk</td>
</tr>
<tr>
<td>&lt;$SS&gt;</td>
<td>37</td>
<td>Female</td>
<td>Crops and livestock extension officer</td>
</tr>
</tbody>
</table>

Table H.3: Summary of participants (private business letters)

<table>
<thead>
<tr>
<th>Participant pseudonyms</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Tawanda</td>
<td>66</td>
<td>Male</td>
<td>Printing business owner/ retired manager</td>
</tr>
<tr>
<td>Emily Mudzingwa</td>
<td>26</td>
<td>Female</td>
<td>Creche owner/ postgraduate student</td>
</tr>
<tr>
<td>Petina Murungwa</td>
<td>64</td>
<td>Female</td>
<td>Events coordinator</td>
</tr>
<tr>
<td>Tendaishe Mufura</td>
<td>29</td>
<td>Female</td>
<td>Shop owner</td>
</tr>
<tr>
<td>Tinoda Guruza</td>
<td>51</td>
<td>Male</td>
<td>Cross border trader</td>
</tr>
</tbody>
</table>
### Table H.4: 18-30 age group (private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$B$&gt;</td>
<td>23</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>28</td>
<td>Male</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>26</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>30</td>
<td>Male</td>
<td>Self employed</td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>30</td>
<td>Male</td>
<td>Field officer</td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>26</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>22</td>
<td>Female</td>
<td>College student</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>26</td>
<td>Female</td>
<td>University student</td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>25</td>
<td>Male</td>
<td>Graphic designer, music producer, entrepreneur</td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>18</td>
<td>Female</td>
<td>Student</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>24</td>
<td>Male</td>
<td>Electrician</td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>28</td>
<td>Male</td>
<td>Medical laboratory scientist</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>24</td>
<td>Female</td>
<td>District focal person</td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>27</td>
<td>Male</td>
<td>Electrician</td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>30</td>
<td>Female</td>
<td>Banking</td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>24</td>
<td>Female</td>
<td>Businesswoman</td>
</tr>
</tbody>
</table>

### Table H.5: 31 to 50 age group (private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$A$&gt;</td>
<td>35</td>
<td>Male</td>
<td>Engineer (Resident)</td>
</tr>
<tr>
<td>&lt;$B$&gt;</td>
<td>36</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>43</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>40</td>
<td>Female</td>
<td>Registered nurse</td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>43</td>
<td>Female</td>
<td>Business owner</td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>39</td>
<td>Male</td>
<td>Teacher (secondary)</td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>43</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>33</td>
<td>Female</td>
<td>Registered nurse</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>32</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>39</td>
<td>Male</td>
<td>Prison officer</td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>34</td>
<td>Female</td>
<td>Records assistant</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>34</td>
<td>Male</td>
<td>Civil engineering technician</td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>35</td>
<td>Female</td>
<td>General hand</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>41</td>
<td>Male</td>
<td>Clerk</td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>37</td>
<td>Female</td>
<td>Crops and livestock extension officer</td>
</tr>
</tbody>
</table>
### Table H.6: Age group 51+ years

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$P$&gt;</td>
<td>51</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>57</td>
<td>Female</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>56</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$II$&gt;</td>
<td>51</td>
<td>Male</td>
<td>General hand</td>
</tr>
<tr>
<td>&lt;$W$&gt;</td>
<td>54</td>
<td>Female</td>
<td>Doctor (gynaecologist)</td>
</tr>
<tr>
<td>&lt;$X$&gt;</td>
<td>56</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$Y$&gt;</td>
<td>66</td>
<td>Male</td>
<td>Retired manager</td>
</tr>
<tr>
<td>&lt;$Z$&gt;</td>
<td>54</td>
<td>Female</td>
<td>Nurse aid</td>
</tr>
<tr>
<td>&lt;$DD$&gt;</td>
<td>53</td>
<td>Female</td>
<td>Banking officer</td>
</tr>
<tr>
<td>&lt;$EE$&gt;</td>
<td>51</td>
<td>Male</td>
<td>Self employed</td>
</tr>
<tr>
<td>&lt;$HH$&gt;</td>
<td>53</td>
<td>Male</td>
<td>Teacher (high school)</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>71</td>
<td>Male</td>
<td>Retired teacher, farmer</td>
</tr>
<tr>
<td>&lt;$LL$&gt;</td>
<td>68</td>
<td>Female</td>
<td>Semi-retired lecturer</td>
</tr>
<tr>
<td>&lt;$MM$&gt;</td>
<td>52</td>
<td>Male</td>
<td>Teacher (primary)</td>
</tr>
<tr>
<td>&lt;$NN$&gt;</td>
<td>56</td>
<td>Male</td>
<td>Professor</td>
</tr>
</tbody>
</table>

### Table H.7: 18 to 30 age group (Private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$C$&gt;</td>
<td>29</td>
<td>Female</td>
<td>Shop owner</td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>22</td>
<td>Female</td>
<td>Aspiring teacher</td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>28</td>
<td>Female</td>
<td>Postgraduate student</td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>20</td>
<td>Female</td>
<td>Polytechnic student</td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>22</td>
<td>Female</td>
<td>Self-employed</td>
</tr>
<tr>
<td>&lt;$R$&gt;</td>
<td>20</td>
<td>Female</td>
<td>Self employed</td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>29</td>
<td>Female</td>
<td>BED student, vendor</td>
</tr>
</tbody>
</table>
Table H.8: 31 to 50 age group (Private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$B$&gt;</td>
<td>41</td>
<td>Male</td>
<td>Mechanic</td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>43</td>
<td>Male</td>
<td>College lecturer</td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>36</td>
<td>Male</td>
<td>Registered nurse</td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>38</td>
<td>Male</td>
<td>Teacher</td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>50</td>
<td>Male</td>
<td>Bus driver</td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>43</td>
<td>Male</td>
<td>Agricultural extension officer</td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>34</td>
<td>Female</td>
<td>Matron</td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>39</td>
<td>Male</td>
<td>Self employed</td>
</tr>
</tbody>
</table>

Table H.9: 51+ age group (Private spoken dialogues)

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (years)</th>
<th>Gender</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$A$&gt;</td>
<td>75</td>
<td>Male</td>
<td>Farmer</td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>51</td>
<td>Male</td>
<td>Pastor</td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>57</td>
<td>Male</td>
<td>Teacher (secondary)</td>
</tr>
<tr>
<td>&lt;$S$&gt;</td>
<td>55</td>
<td>Female</td>
<td>Lecturer</td>
</tr>
<tr>
<td>&lt;$T$&gt;</td>
<td>51</td>
<td>Female</td>
<td>Lecturer</td>
</tr>
<tr>
<td>Participant</td>
<td>Age</td>
<td>Context of English use</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>&lt;$A$&gt;</td>
<td>35</td>
<td>When talking to my superiors.</td>
<td></td>
</tr>
<tr>
<td>&lt;$B$&gt;</td>
<td>23</td>
<td>With kids because they are learning English at school.</td>
<td></td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>28</td>
<td>When I do discussions with friends. With colleagues.</td>
<td></td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>26</td>
<td>I usually use English during lecture times.</td>
<td></td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>30</td>
<td>With clients.</td>
<td></td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>30</td>
<td>English is a common and widely accepted mode of communication.</td>
<td></td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>26</td>
<td>During lectures when discussing with my friends and other students.</td>
<td></td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>22</td>
<td>When doing discussions at school.</td>
<td></td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>75</td>
<td>With grandchildren who don't speak Shona and homework with those at school. Occasionally with friends.</td>
<td></td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>36</td>
<td>I use it whilst teaching pupils at work and when attending official meetings and workshops.</td>
<td></td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>25</td>
<td>I use English whenever and wherever I feel I need to communicate not in Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>47</td>
<td>English is the official language when teaching and learning.</td>
<td></td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>18</td>
<td>I speak English with my friends and teachers.</td>
<td></td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>46</td>
<td>For teaching.</td>
<td></td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>43</td>
<td>When talking to my children, teaching learners at school and talking to the school head.</td>
<td></td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>51</td>
<td>I use English as a language of learning and teaching.</td>
<td></td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>30</td>
<td>At work with my co-workers. At school when I attend lectures.</td>
<td></td>
</tr>
<tr>
<td>&lt;$R$&gt;</td>
<td>40</td>
<td>At school as a learning language, at home, work and with my friends it comes naturally.</td>
<td></td>
</tr>
<tr>
<td>&lt;$S$&gt;</td>
<td>43</td>
<td>With my friends, sometimes we just speak English and at home with my family.</td>
<td></td>
</tr>
<tr>
<td>&lt;$T$&gt;</td>
<td>39</td>
<td>With my students during lesson delivery and with co-workers during meetings and at professional gatherings.</td>
<td></td>
</tr>
<tr>
<td>&lt;$U$&gt;</td>
<td>43</td>
<td>I use English for teaching.</td>
<td></td>
</tr>
<tr>
<td>&lt;$V$&gt;</td>
<td>33</td>
<td>At work speaking with co-workers.</td>
<td></td>
</tr>
<tr>
<td>&lt;$W$&gt;</td>
<td>43</td>
<td>When I see patients who can’t speak Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$X$&gt;</td>
<td>56</td>
<td>Instruction, social issues.</td>
<td></td>
</tr>
<tr>
<td>&lt;$Y$&gt;</td>
<td>66</td>
<td>When I speak to my kids who always want to speak English. When I speak To my friends who don't speak Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$Z$&gt;</td>
<td>54</td>
<td>At workshops and to my superiors.</td>
<td></td>
</tr>
<tr>
<td>Code</td>
<td>Page</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>32</td>
<td>I used it in the school presentations as well as funny communicative functions.</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>39</td>
<td>With friends who speak Ndebele.</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>24</td>
<td>Usually I speak English when doing customer service.</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>53</td>
<td>At school as a medium of instruction.</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>51</td>
<td>Work related matters.</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>34</td>
<td>At home I usually speak English with my children. At school I use it to communicate with my classmates and lecturers. At work I use it when communicating with my superiors and other English speaking clients. I also speak English with my friends here and there.</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>28</td>
<td>Talking to my school and workmates and also in meetings.</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>53</td>
<td>Teaching learners and during staff meetings, seminars and workshops.</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>51</td>
<td>Among home mates, as family always speak, school work.</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>24</td>
<td>Sensitisation meetings.</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>27</td>
<td>Giving orders at work to the subordinates.</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>68</td>
<td>With my grandchildren who cannot speak Shona. At work when I teach. With my Ndebele friends.</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>52</td>
<td>With learners.</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>56</td>
<td>Teaching, socialising and with family.</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>34</td>
<td>At school with teachers and at work with bosses.</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>24</td>
<td>At work: I use English to communicate with my clients.</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>35</td>
<td>When talking to the headmaster at our school.</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>41</td>
<td>With co-workers.</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>37</td>
<td>With my kids and other staff at work.</td>
<td></td>
</tr>
<tr>
<td>Participant</td>
<td>Age</td>
<td>Context of English use</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----</td>
<td>-----------------------</td>
<td></td>
</tr>
<tr>
<td>&lt;$A$&gt;</td>
<td>75</td>
<td>When I go to sell my produce in Bulawayo.</td>
<td></td>
</tr>
<tr>
<td>&lt;$B$&gt;</td>
<td>41</td>
<td>With customers who can't speak Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$C$&gt;</td>
<td>29</td>
<td>When talking with customers at my shop and when I am teaching my kids homework.</td>
<td></td>
</tr>
<tr>
<td>&lt;$D$&gt;</td>
<td>22</td>
<td>When having conversations with my friends.</td>
<td></td>
</tr>
<tr>
<td>&lt;$E$&gt;</td>
<td>51</td>
<td>When I speak with friends.</td>
<td></td>
</tr>
<tr>
<td>&lt;$F$&gt;</td>
<td>43</td>
<td>At home it comes naturally, at work and with friends - they don't speak Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$G$&gt;</td>
<td>36</td>
<td>At work when I attend to patients who can't speak English.</td>
<td></td>
</tr>
<tr>
<td>&lt;$H$&gt;</td>
<td>29</td>
<td>During lectures.</td>
<td></td>
</tr>
<tr>
<td>&lt;$I$&gt;</td>
<td>38</td>
<td>At work when I teach learners.</td>
<td></td>
</tr>
<tr>
<td>&lt;$J$&gt;</td>
<td>50</td>
<td>I speak to my friends in English and with my family who speak Ndebele and not Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$K$&gt;</td>
<td>57</td>
<td>At work.</td>
<td></td>
</tr>
<tr>
<td>&lt;$L$&gt;</td>
<td>43</td>
<td>When teaching and with my family and friends.</td>
<td></td>
</tr>
<tr>
<td>&lt;$M$&gt;</td>
<td>28</td>
<td>During lectures.</td>
<td></td>
</tr>
<tr>
<td>&lt;$N$&gt;</td>
<td>34</td>
<td>I use English language when teaching.</td>
<td></td>
</tr>
<tr>
<td>&lt;$O$&gt;</td>
<td>20</td>
<td>During lectures.</td>
<td></td>
</tr>
<tr>
<td>&lt;$P$&gt;</td>
<td>22</td>
<td>During lectures.</td>
<td></td>
</tr>
<tr>
<td>&lt;$Q$&gt;</td>
<td>39</td>
<td>With family who can't speak Shona.</td>
<td></td>
</tr>
<tr>
<td>&lt;$R$&gt;</td>
<td>20</td>
<td>With friends who cannot speak Shona. At work.</td>
<td></td>
</tr>
<tr>
<td>&lt;$S$&gt;</td>
<td>55</td>
<td>When I teach students.</td>
<td></td>
</tr>
<tr>
<td>&lt;$T$&gt;</td>
<td>51</td>
<td>When I do lectures.</td>
<td></td>
</tr>
</tbody>
</table>
Figure H.1: Summary of private conversations and monologues participants’ gender

![Gender Chart]

Figure H.2: Employment status

![Employment Status Chart]
Figure H.3: Participants’ age groups

Figure H.4: Hours spent speaking English per week
Figure H.5: English ability in private monologues and conversations

Figure H.6: Participants’ bilingual status
Figure H.7: Context in which English is used

Figure H.8: Teaching by English first language speaker
Figure H.9: Other languages spoken besides Shona

Figure H.10: Where English is used
Figure H.11: Age when English was first learned
Appendix I: Transcription conventions

Transcriptions of recordings was loosely based on Jefferson's (2004) transcription conventions.

{PP1, PP2, etc.}: Pilot Participant 1, Pilot Participant 2 (Pseudonyms for participants)

{ND}: Name of participant deleted and replaced with pseudonym, e.g. PP1, PP2, etc.

{R}: Researcher

( . . . ): Pause lasting longer than three seconds

(.): Brief pause

{O}: Overlapping speech

{LG}: Laughter

--: Incomplete sentence

( ): Unfinished word with the unfinished part in brackets

{NS}: Noise in the background

{BC}: Background conversation

< >: Description of action done e.g. clears throat, tongue clicks

{CS}: Codeswitching (Codeswitched words written in italics)

{ }: Hesitation to speak

{EAD}: Email address deleted

{()}: Inaudible

{BR}: Sighs, audible breathing sounds

{CG}: Coughing, grunting
{MW}: Mumbled words

{rp}: Repetition

{c}: Correction