

2014-07-11

# Developing a Model Corpus for Endangered Languages

Hashi, Awil

---

Hashi, A. (2014). Developing a Model Corpus for Endangered Languages (Doctoral thesis, University of Calgary, Calgary, Canada). Retrieved from <https://prism.ucalgary.ca>. doi:10.11575/PRISM/25614  
<http://hdl.handle.net/11023/1632>

*Downloaded from PRISM Repository, University of Calgary*

UNIVERSITY OF CALGARY

Developing a Model Corpus for Endangered Languages

by

Awil Hashi

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES  
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR  
THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAMS IN EDUCATION

CALGARY, ALBERTA

JULY, 2014

© Awil Hashi 2014

## Abstract

World languages are becoming endangered at an unprecedented rate. Linguists estimate that 50-90% of the world's 6700 languages are endangered. Krauss (1992) predicted that as few as 10% of the world's languages may survive by 2100, which means we may be left with only 670 languages. Given the relationship between linguistic diversity, cultural diversity, intellectual diversity and biological diversity, this study is timely because it contributes to maintaining the biodiversity of languages. This research expands upon the previous success of corpus linguistics for certain world languages to develop and document the Somali language.

This educational technology study examines the phenomenon of endangered languages by examining technology and languages in Corpus Linguistics, the study of language in real world texts. Through analysis and a review of the literature and historical precedents, the study demonstrates how the Somali language, spoken by millions, can be labelled an *endangered language*. Using a pragmatic approach and expanding upon the theories and experience of previous corpus constructions, this study developed an iterative design framework in five separate but complementary phases for corpus design in challenging contexts and drafted a sampling frame for text collection. The research designed and built a Somali language corpus as a working prototype for endangered languages that illustrates how it could be used for language development and documentation purposes. Phase one used the Somali language and collected 22 texts from different genres to produce a general corpus of 865,214 words. Aided by quantitative results generated by antConc corpus tools, the study reveals the 20 most frequently used Somali words. The identification of such words has practical implications for teachers, students, curriculum developers, lexicographers, and language policy makers because their high frequency has pedagogical significance. The study illustrates how corpus tools identify the least commonly used words with implications for those who are involved in language documentation and development initiatives. Using types/token ratio analysis, the study indicates that certain genres such as poetry have an inherently richer vocabulary, which has implications for lexicographers and those who are involved in vocabulary and terminology development in languages.

## Dedication

This work is dedicated to my late father *Calidhuux Hashi* and my uncle, *Dr. Xasan Xaashi Fiqi*, who are unfortunately not here with us to rejoice the achievement of this landmark work, which is undoubtedly the culmination of the fruits they sowed by supporting, nurturing and inspiring me not merely to excel but also to contribute to humanity.

## *Hibeyn*

*Waxa aan miraha hawshaan u hibeeyey oo aan kusoo xasuusanyaa Allaha u naxariistee, aabbahay Calidhuux Xaashi iyo adeerkay Dr. Xasan Xaashi Fiqi, kuwaasoo nasiib-darro aan halkaan nala joogin si ay noola qaybsadaan libinta hawshaan oo ay dhab u bogaadin lahaayeen. Tanoo aan shaki ku jirin inay tahay mirihii ay ku beereen taageeradoodii, ababintoodii iyo dhiirrigelintoodii ahayd inaan kaligeys tallaabsan ee aan adduunyada wax uun kusoo kordhiyo.*

## Acknowledgements

This academic work has been one of the most challenging projects I have ever had to undertake. The support, patience, and inspiration from many people have made the initiation of this endeavor as well as its completion attainable. I owe them my utmost appreciation.

- Dr. Michele Jacobsen, my supervisor who despite her many other academic, family and professional commitments, provided me with timely and constructive feedback consistently on each chapter. Her tactful demeanor and commitment to excellence has inspired me to stay focused. I can't thank you enough!
- My supervisory committee, Dr. Susan Crichton, Dr. Ian Winchester and Dr. Darin Flynn, have all given me unwavering support and insight from day one. Thank you all!
- My mother has always been a source of knowledge and inspiration. Her encouragement to excel and to also live my life to the fullest has been inspirational. Thank you for being such a wonderful mother.
- My soulmate, Faiza Xassan Xaashi and my children have all been there for me. Your endless support and patience have taught me so much about empathy, perseverance and love. Special thanks goes to Anas and Anisa Hashi who have worked with me as project co-supervisors during the initial stages of this thesis and as a result learned about the overall structure and the subtitles of the first three chapters. I am so thankful to you all beyond words. My heartfelt thanks also goes to my sister, Ubax Calidhuux Xaashi and my awesome nieces Rumaysa, Ruweyda and Ramla Maxamed Weli who have been so supportive throughout the entire process.
- A grade-school friend and a cousin, Abdirishid Hashi of the Heritage Institute for Policy Studies, has worked with me tirelessly during the data collection in ensuring that I gather

as diverse texts as possible towards constructing a corpus of one million words. I owe you one, Rashid!

- Abdurahman Hashi of Fiqi Publishers has not only donated much-needed texts but has also been a dependable friend and cousin. Ali Mohamed Saeed has always been a staunch supporter during tough times. Thank you, Ali Hashi.
  - The following individuals deserve special gratitude for their encouragement and inspiration: Cabdirisaaq Cabdullaahi Xaashi, Cabdinuur Khaliif, Guuleed Xasan Xaashi, Fuad Hassan Hashi, Dahir K. Hashi, Fiqi Bashir Khalif, Abdulkadir Abdinur Hashi, Ali Bosir, Dr. Afyare Elmi, Siciid Suugaan, Dr. Cabdullaahi Barise and Dr. Abdullaahi Hussein (Muqadin).
  - The following friends and colleagues have generously donated various texts: Ali M. Abdigiir (Cali-ganey), Abdulqadir Khaliif Xaashi, Abwaan Xasan Cabdullaahi Xaashi (Timacadde), and Dr. Qaasim Hersi Faarah.
  - Sylvia Parks, our graduate program administrator has always been supportive, considerate and effective in her efforts. I thank you very much!
  - I am also very grateful to University of Calgary's Graduate Division for Educational Research for funding this research through in-house graduate funding and also through the *Queen Elizabeth II Scholarship*. I am eternally grateful.
  - Muthhir Mohamed and Dr. Hussein Warsame of Haskayne School of Business and all my friends in Calgary have always encouraged me to stay on course. Thank you all!
- There are many friends and colleagues who have helped me one way or another over the course of this journey. I will always cherish your efforts and support in my heart.

## Table of Contents

<b>ABSTRACT .....</b>	<b>I</b>
<b>DEDICATION .....</b>	<b>II</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>III</b>
<b>LIST OF TABLES.....</b>	<b>VIII</b>
<b>LIST OF FIGURES.....</b>	<b>VIII</b>
<b>CHAPTER ONE: OVERVIEW .....</b>	<b>1</b>
1.1 RESEARCH BACKGROUND .....	1
1.2 SIGNIFICANCE OF THE STUDY .....	2
1.3 PURPOSE OF THE STUDY .....	4
1.4 RESEARCH QUESTION .....	4
1.5 METHODOLOGY AND CONCEPTUAL FRAMEWORK .....	4
1.6 THE SCOPE OF THE STUDY.....	5
1.7 ORGANIZATION OF CHAPTERS .....	6
<b>CHAPTER TWO: ENDANGERED LANGUAGES .....</b>	<b>8</b>
2.1 GENERAL OVERVIEW .....	8
2.2 WHY ARE LANGUAGES DYING? .....	17
2.3 WHY DOES IT MATTER? .....	24
2.4 WHAT DOES IT TAKE? .....	26
2.5 HOW CAN LANGUAGES BE REVITALIZED? .....	27
THE REVITALIZATION STORY OF MODERN HEBREW .....	28
REVITALIZATION EFFORTS OF MAORI IN NEW ZEALAND .....	32
2.6 WHO IS INVOLVED? .....	33
MICROSOFT LOCAL LANGUAGE PROGRAM .....	35
GOOGLE CORPORATION—THE ENDANGERED LANGUAGE PROJECT.....	36
A LONG NOW FOUNDATION LIBRARY OF HUMAN LANGUAGE: THE ROSETTA PROJECT.....	37
VOLKSWAGEN FOUNDATION: DOCUMENTATION OF ENDANGERED LANGUAGES.....	38
ROSETTA STONE: ENDANGERED LANGUAGES PROJECT .....	38
UNESCO: ENDANGERED LANGUAGES.....	39
AFRICAN ACADEMY OF LANGUAGES .....	40
SOMALI-SPEAKING PEN CENTRE .....	40
INSTITUTE FOR LANGUAGES DJIBOUTI .....	41

UNIVERSITIES OF CAMBRIDGE AND YALE: WORLD ORAL LITERATURE PROJECT.....	41
UNIVERSITY OF LONDON .....	42
SWARTHMORE COLLEGE: LABORATORY FOR ENDANGERED LANGUAGES .....	42
UNIVERSITY OF MINNESOTA: SECOND LANGUAGE REQUIREMENT .....	42
UNIVERSITY OF ALBERTA .....	43
<b>CHAPTER 3: THE CASE FOR SOMALI AS AN ENDANGERED LANGUAGE .....</b>	<b>46</b>
3.1 OVERVIEW .....	47
3.2 CONTEXT OF THE PROBLEM.....	48
3.3 OVERVIEW OF THE SOMALI LANGUAGE .....	50
3.4 LIFECYCLE OF THE SOMALI LANGUAGE .....	51
3.5 SOMALI LANGUAGE VITALITY ASSESSMENT .....	54
3.6 WHAT IS AT STAKE? .....	62
<b>CHAPTER FOUR: WINDOW ON THE WORLD OF CORPUS LINGUISTICS.....</b>	<b>67</b>
4.1 OVERVIEW .....	67
4.2 THE EVOLUTION OF CORPUS .....	68
4.3 WHAT IS A CORPUS? .....	72
4.4 PRACTICAL USES OF CORPORA.....	74
4.5 CORPUS FOR RESEARCH AND DEVELOPMENT .....	76
4.6 PROSPECTIVE CORPUS USERS .....	78
<b>CHAPTER FIVE: CORPUS DESIGN .....</b>	<b>81</b>
5.1 CORPORA CLASSIFICATIONS .....	81
5.2 DESIGN METHODOLOGY FOR CORPUS CONSTRUCTION .....	87
5.3 OVERVIEW OF PREVIOUS DESIGN APPLICATIONS.....	96
BROWN CORPUS.....	97
BRITISH NATIONAL CORPUS (1980s-1993) .....	97
BANK OF ENGLISH AND COBUILD CORPUS .....	99
OXFORD ENGLISH CORPUS .....	100
CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA).....	101
5.4 DEDUCTIONS FOR CORPUS DESIGN IN CHALLENGING CONTEXTS.....	103
<b>CHAPTER SIX: CORPUS DESIGN IN CHALLENGING CONTEXTS .....</b>	<b>107</b>
6.1 CONCEPTUAL FRAMEWORK.....	107
6.2 METHODOLOGY .....	117



6.3 FIVE STAGES FOR CORPUS DESIGN IN CHALLENGING CONTEXTS.....	119
6.4 CASE STUDY .....	123
6.5 DEFINING THE SCOPE OF THE DATA .....	125
6.6 THE SOMALI LANGUAGE.....	126
6.7 CHOOSING CORPUS CONSTRUCTION SOFTWARE.....	133
6.8 SOMALI LANGUAGE BANK (SLB).....	134
6.9 DATA COLLECTION .....	134
6.10 DATA ANALYSIS .....	139
6.11 COPYRIGHT ISSUES .....	140
<b>CHAPTER SEVEN: RESULTS AND DISCUSSION.....</b>	<b>149</b>
7.1 OVERVIEW .....	149
7.2 POETRY AS A WORDBOOK IN DISGUISE .....	153
7.3 ADDITIONAL RESEARCH INSTRUMENTS.....	161
7.4 THE 20 MOST FREQUENTLY USED SOMALI WORDS .....	163
7.5 LANGUAGE DEVELOPMENT THROUGH KEY WORDS IN CONTEXT (KWIC) TOOL .....	174
7.6 TOWARDS DOCUMENTATION AND DEVELOPMENT OF LESS FREQUENTLY USED WORDS .....	178
<b>CHAPTER 8: SUMMARY, CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS FOR FUTURE WORK .....</b>	<b>185</b>
8.1 SUMMARY OF THE STUDY .....	185
8.2 CONCLUSIONS .....	198
8.3 RESEARCH LIMITATIONS .....	202
8.4 RECOMMENDATIONS FOR FUTURE RESEARCH .....	204
8.5 FINAL THOUGHTS .....	205
<b>REFERENCES.....</b>	<b>209</b>
<b>APPENDICES.....</b>	<b>229</b>
APPENDIX 1: FIVE STAGES FOR CORPUS DESIGN IN CHALLENGING CONTEXTS .....	229
APPENDIX 2: SAMPLING FRAME FOR THE SOMALI LANGUAGE.....	233
APPENDIX 3: PROPOSED SAMPLING FRAME FOR SPOKEN TEXTS.....	234

### List of Tables

Table 1.1 UNESCO’s Levels of Endangerment.....	10
Table 6.1. Text Types of Spoken Data in the BNC .....	135
Table 6.2. SLB Text Classifications Sampling frame.....	135
Table 6.3 Text Category to Specific Texts.....	137
Table 6.4 Text Identification: Additional Metadata About the Text.....	138
Table 7.1. Corpus Statistics.....	152
Table 7.2. Normed Type/Token Ratios.....	154
Table 7.3. Ratio of Single Poem.....	157
Table 7.4. The 20 Most Frequently Used Somali Words in the Corpus.....	163
Table 7.5. Corpus of Contemporary American English and Somali Language Bank...	166

### List of Figures

Figure 3.1. Stages of the Somali Language From Oral to Written Language.....	51
Figure 3.2. Language Vitality Factors Adopted from UNESCO (2003).....	54
Figure 7.1. Composition of the General Corpus.....	150
Figure 7.2. General Corpus Composition.....	151
Figure 7.3. Visual Representation of Type/token Ratio.....	157
Figure 7.4. A Poem Wordlist.....	158
Figure 7.5. Mareeg in Transition.....	159
Figure 7.6. Corpus results of the 20 Most Frequently Used Somali Words.....	168
Figure 7.7. Keenadiid: “Oo” in a Non-Corpus-Based Dictionary.....	169
Figure 7.8. Pugllieli and Mansuur (2012): “Oo” in a Non-Corpus-Based Dictionary.....	171
Figure 7.9. A Sample of KWIC Results for the Somali Word “oo”.....	173

Figure 7.10. Infrequent Word—“ <i>Falkin</i> ” .....	176
Figure 7.11. “ <i>Falkin</i> ” in Context Using the KWIC tool.....	177
Figure 7.12. Keenadiid (1976): “ <i>Falkin</i> ” in a Non-Corpus-Based Dictionary.....	178
Figure 7.13. Puglieli and Mansuur (2012): <i>Falkin</i> in a Non-Corpus-Based Dictionary..	179

## **Chapter One: Overview**

Language diversity and biological diversity are inseparable. In the language of ecology, the strongest ecosystems are those that are the most diverse. That is, diversity is directly related to stability; variety is important for long-term survival. Our success on this planet has been due to an ability to adapt to different kinds of environment over thousands of years. Such ability is born out of diversity. Thus language and cultural diversity maximizes chances of human success and adaptability. (Baker, 2001, p. 281)

### **1.1 Research Background**

Central to this thesis is precisely what Baker (2001) discussed in the opening epigraph, above. This study concerns linguistic diversity, cultural diversity, and intellectual diversity, which are encoded in human languages. It explores ways in which linguistic diversity could be maintained.

Every language in the world embodies human history, culture, identity, and the wisdom of humanity. Therefore, the loss of a language means that knowledge of plants, animals, and ecosystems as well as cultural tradition is at stake (Harmon & Maffi 2002; Nettle & Romaine 2000; UNESCO, 2003). Unfortunately, as Maffi (1998) stated, recent projections indicate that as many as 90% of the world's languages may disappear during the course of the next century. The United Nations Educational and Cultural Organization (UNESCO), (2010) has reported that approximately half of the current 6700 languages have now been classified as endangered. The discussion of endangered languages is at the core of this research and it will be a recurring theme

especially in Chapter Two and again in Chapter Three in which I will address the issue: “Why does it matter?” in more depth.

This research examines how technology can revitalize and document languages to a sustainable status while making them accessible to linguists, material developers, and lexicographers for development purposes. As I will show in later chapters, technology offers a number of ways to document, raise awareness, and develop languages. International efforts aim at reversing the current language crisis. Technology giants such as Google and Microsoft are among the corporations that are involved in the revitalization initiatives. Similarly, as the study documents in Chapter Two, a number of governmental and non-governmental organizations as well as universities have joined forces in the ongoing worldwide campaign that is documenting and revitalizing endangered languages.

Of all the solutions that technology offers, building a corpus for a language is perhaps the most effective means to document a language digitally and to maintain languages. In linguistics and lexicography, a corpus is a body of texts (spoken or written), which is more or less representative of a given language. It is usually stored as an electronic database (McArthur, 1992). Many Indo-European languages have used corpus development to streamline the process of language accessibility and facilitate research into and development of those languages in ways that have not been feasible in the past. Unfortunately, many of the world’s languages do not have corpora or a language database facility nor do they have the language infrastructure on which linguistic research and development initiatives could be based.

## **1.2 Significance of the Study**

Two principal reasons dictate the need for scholarly attention to replicate and extend the revolutionary corpus-based study that has led the advancement of certain world languages. The

first comes from the success of the use of corpora (Dash, 2010; Davies, 2009; Francis & Kučera, 1979; Kennedy, 1998) for different ends ranging from dictionary development, language-learning material development, including grammar textbooks, as well as developing word-processing tools such as spell checkers and grammar and style checkers.

The first and perhaps the best-known national corpus is the British National Corpus (BNC), which was built to represent as wide a range of modern British English as possible (Bernard, 2002). The British National Corpus (BNC) contains 100 million words and is used as a reference corpus (Sinclair, 1991) for various linguistic and non-linguistic studies, cross-linguistic comparisons, and other language development areas including but not limited to the development of spell checkers, lexicography and machine translation devices. More recently, the Corpus of Contemporary American English, which contains 450 million words, has been constructed as a monitor corpus to track how the language is used for development and language-teaching purposes (Davies, 2009). The successful use of corpora for major languages suggests the use of corpus-based language study for less commonly studied languages for development, revitalization, and research purposes.

The second reason for the usefulness of the topic is that most of the world's languages remain understudied and approximately 50-90% are endangered. Linguistic diversity is linked to cultural and intellectual diversity because knowledge and culture (thus different worldviews) are encoded in language. For example, several linguists (Crystal, 2000; Dalby, 2003; Nettle & Romaine, 2000) directly correlated language loss and knowledge loss, which gives the issue even more importance. By digitizing world languages through corpus linguistics, appropriate measures to safeguard languages are being taken while allowing language resources and language learning materials to develop as they may.

### **1.3 Purpose of the Study**

The purpose of this work is two-fold. First, the study develops a model corpus for endangered languages using the Somali language as a case study. Such a model presents a workable model for other endangered languages in an effort to preserve the majority of languages from extinction. Second, the study allows language technologists and researchers to study and devise ways in which to develop languages that are not feasible without technology. The model also assists language researchers to study and analyse languages in cross-linguistic research.

### **1.4 Research Question**

This research examines the following question: In what manner can a model corpus help document and develop endangered languages?

In addition, the following secondary questions will be addressed:

1. How does one learn from the experiences of previous corpus construction and apply those lessons to the context of other languages?
2. What opportunities does a language corpus provide in terms of language development tools such as resource materials, dictionaries, and grammars?
3. What does it take to build a language corpus for languages with limited resources that have ineffective or no supporting national institutions?

### **1.5 Methodology and Conceptual Framework**

The methodology and the conceptual framework that guide the study are presented in Chapter Six in more detail. Davies (2009) and Sinclair (2005), who advanced pragmatic approach to corpus design, inspired the methodology used in this study. The study needed to be meaningful, practical and yet theoretical allowing me to question and re-question so that the chasm between

theory and practice is minimized. I wanted to propose a practical solution to a real-life problem. My vision was to expand upon prior corpus design methodology for certain world languages.

### **1.6 The Scope of the Study**

A number of factors constitute limitations in replicating, analysing, and interpreting the data as well as the outcome of this research. Despite using systematic and consistent methods of data collection, the scope of the study (to build a blueprint for a larger and more comprehensive corpus for the Somali language) has influenced the decision process throughout the study. Ideally, a general, large, and balanced corpus requires a much longer time than the limited time given for doctoral research. Likewise, such a corpus demands large sums of money to collect data from the environments in which the language is spoken. Considering the scope, time, and resources available, the data was collected based on availability. This type of purposeful data collection seems to suffice for the purposes of this study since the goal was to produce a sample, a working prototype, which could eventually lead to a larger, more representative and balanced corpus. My background and passion for languages and technology is a limitation that may have skewed or shaped the design and interpretations in the study. Like any other social science phenomenon, the study of language is complex and therefore the applicability of the results and the design of this case study to any other language will depend on the context of the case and the linguistic variables between the language in question and the language of the case study, Somali. The limitations of the study and how they could be mitigated in the future are revisited in Chapter Eight.

Based on the research goals, instead of collecting samples from extensive and various language texts, the study marked the boundaries of the text types that were accessible. The study does not collect text types from a wide range of dialectal variations, which might have proved



more representative of the language studied. The researcher was cognizant of the fact that a number of factors such as language register, genre, dialects, and situations should be considered when collecting data for a representative language corpus. Nevertheless, the study was delimited to the language types that were readily available to the researcher. The study adopts an iterative design model with five different phases in constructing a larger, more representative corpus for the Somali language. This study aims to produce an exploratory, prototype corpus covering the first stage of the five phases envisioned. A more detailed plan of the proposed five phases and the sampling frame developed for this study will be discussed in Chapter Six.

## **1.7 Organization of Chapters**

**Chapter Two** familiarizes the reader with the multidimensional issue of endangered languages. It gives the definitions of terms and discusses why languages become endangered and how the problem could be managed. Examples of two languages that have been revitalized are given. The chapter ends with a brief account of international institutions that are involved in revitalizing and documenting languages.

**Chapter Three** builds a case for the Somali language being endangered, using UNESCO (2003) language-endangerment rubrics. Along with analysis and a review of the literature and historical precedents, the study illustrates how the Somali language, although spoken by millions, can be regarded as an endangered language.

**Chapter Four** explores corpus linguistics as a field, introducing definitions and tracing its evolution. Discussion involves the nature and the validity of the data it uses along with the practical uses of a corpus.

**Chapter Five** presents corpus design methodology in general. The chapter sets the scene with classifications of existing corpora along with design methodologies employed by corpus

designers for different ends. A review is presented of the thorny issues of representativeness, balance, and diversity in corpus design.

Building on the review of Chapter Five, **Chapter Six** examines corpus design in challenging contexts. This chapter describes (a) the conceptual framework that guides the study, and (b) the design methodology for corpus construction in challenging contexts. An iterative design model in five phases is proposed as an alternative design model so that issues of diversity balance and size become less of a challenge over time. The researcher develops a sampling frame for this study.

**Chapter Seven** presents the results and pertinent discussions. The composition of the corpus is discussed along with the text types represented in the corpus. The relative richness of different text types is analysed using types/token ratio. The chapter includes a list of the 20 most commonly used Somali words based on the corpus built for this thesis. Different ways to develop and document the language are illustrated.

**Chapter Eight** summarizes the study and notes practical implications and recommendations for future work. The study concludes with final thoughts that synthesize the main themes by reconstructing final reflections on the main issues.

## Chapter Two: Endangered Languages

### 2.1 General Overview

It is always prudent to identify a problem in its early stages so it can be dealt with before the situation is out of hand. In the context of languages, “endangerment” status is assigned when a language starts to move towards extinction. At this time of transition, linguists try to identify the language as endangered early enough so the problem can be subdued in a timely manner. Focusing attention on a problem means that publicity is often increased and awareness raised. Central to this study is the conceptual framework that an intellectual effort such as this doctoral study ought to produce a body of knowledge that offers practical application to a real-life problem, preferably tackling the problem before it is too late. The thesis is time sensitive because it deals with a sensitive issue, endangered languages.

What do we mean when we say a language is endangered? UNESCO (2003) stated:

A language is in danger when its speakers cease to use it, use it in an increasingly reduced number of communicative domains, and cease to pass it on from one generation to the next. That is, there are no new speakers, adults or children. (p. 4)

Janse (2003) noted, “Languages in the process of dying are endangered” (p. v).

The two documents agree that the language in question is disappearing; however, it is unclear where we make the decision about when the language is endangered. Metaphorically speaking, when can we call a language healthy, safe or strong? When does a language become sick or endangered? What are the signs? Even more importantly, what are the levels of concern about this state of affairs? Ultimately, when do we consider a language a dead language? As Tsunoda (2006) explained, the concept of language endangerment can be more clearly understood if the language is placed on a scale according to its level of endangerment or its level

of vitality or strength, with safe and strong languages at one end of the scale and extinct languages at the other end.

Wurm (2003) introduced five stages of language endangerment. He argued that, first, a language is *potentially endangered* when young generations begin to favour the dominant language over their native language, which in turn causes speakers to learn and speak the native language less proficiently. Second, a language is *endangered* when young adults are the youngest who speak that language and there are few or no children who speak it. The third stage is when a language is *seriously endangered* because the youngest speakers are in their middle age or beyond. Fourth, it becomes *terminally endangered or moribund* when it reaches a point where all the speakers of that language are elderly. Fifth, it goes beyond this stage and becomes *dead or extinct* if all the speakers are dead.

Tsunoda (2006) proposed similar classifications. The first stage before a language becomes endangered is called *healthy*. The first stage of endangerment is called a *weakening* stage while the stage before it dies is *moribund*. When the language moves beyond being moribund, it becomes *extinct*.

UNESCO (2003) presented a classification system similar to those proposed by Wurm (2003) and Tsunoda (2006). The UNESCO report used six labels to describe the status or the complete life cycle of a language adding a further label to describe the status of being *not endangered*. Of the six labels, four are used to name levels of endangerment or anything other than being safe or extinct. The labels are as shown in Table 1, below.

Based on the intergenerational language transmission factor which is among the nine language vitality assessment factors, UNESCO annotated its six levels of the language lifecycle as follows: safe, unsafe, definitely endangered, severely endangered, critically endangered, and

extinct. The nine factors affecting the vitality of a language will be reviewed in the following section.

Table 1.1 *UNESCO's Levels of Endangerment*

Degree of Endangerment	Grade	Speaker Population
Safe	5	The language is used by all ages, from children up.
Unsafe	4	The language is used by some children in all domains; it is used by all children in limited domains.
Definitively endangered	3	The language is used mostly by the parental generation and up.
Severely endangered	2	The language is used mostly by the grandparental generation and up.
Critically endangered	1	The language is used by very few speakers of the great-grandparental generation.
Extinct	0	No speaker exists.

Other writers use comparable but reduced levels of endangerment to describe the vitality status of a language. For example, Kinkade (1991) used four levels: strong languages, sick languages, dying languages, and dead languages. Likewise, Krauss (1992) proposed four classifications: safe languages, endangered languages, moribund languages and extinct or dead languages.

It seems that the majority of researchers have used similar systems in terms of the terminology and the extent of detail in their classifications. For the purposes of this study, the

classifications developed by UNESCO (2003) will be adopted because they cover a greater range of levels in describing the status of language vitality than do other classification systems.

In the final stage of its life, a language is called extinct when the last person who speaks it passes away, but a language is technically dead or on its deathbed when there is only one person alive who speaks it because s/he would need someone else to communicate with for the language to be in use. In contrast, Latin can hardly be called a dead language although it is not spoken in daily use. Vulgar Latin developed into Romance languages in the 6th to 9th centuries; the formal language continued as the scholarly lingua franca of Catholic countries in medieval Europe and as the liturgical language of the Roman Catholic Church. Krauss (1992) noted that as a language is becoming extinct, it becomes *moribund* which means that the language is in its dying period or the language is on the verge of death or extinction.

Let us now turn our attention to the factors that cause language to become endangered. An expert group on language endangerment and language maintenance assembled by UNESCO (2003) has created a set of guidelines or criteria for evaluating the vitality of languages. The document by this group presents recommendations on revitalization, documentation, and maintenance. The following factors have been adopted to assess the vitality of the Somali language in Chapter Three. The goal of this study is to explore the application of these factors to a real language to shed light on its meanings and usage. For now, the factors are listed with brief commentaries in the next section. Each of the factors has its own scale to evaluate the degree of endangerment of a language. As I will demonstrate later, these factors stem from social, political, and economic factors as well as unpredictable factors such as natural and manmade disasters.

### **1. Intergenerational language transmission.**

According to Fishman (as cited in UNESCO, 2003), the most important factor in assessing the vitality of a language is the level or the extent of the language being passed on to the next generation. This factor refers to whether or not the younger generations are learning and using their mother tongue.

### **2. Absolute number of speakers.**

This factor refers to the number of people who speak the language in question. Although a large number of speakers do not prevent a language from becoming endangered, a smaller speech community is more prone to the external threats to language such as natural and/or manmade disasters, earthquakes or civil wars respectively (UNESCO, 2003).

### **3. Proportion of speakers within the total population.**

Researchers examined the percentage or the number of speakers in relation to the absolute number of speakers who identify with that speech community (UNESCO, 2003). This factor concerns the percentage of the population who speak that language out of the total population of a given speech community.

### **4. Trends in existing language domains.**

This factor surveys different settings and functions which people use for that language (UNESCO, 2003). It is closely related to the first factor, intergenerational language transmission, because the greater the language domains, the greater the degree of language use, which affects how the language is being transmitted to younger generations.

### **5. Response to new domains and media.**

As the life of its speakers evolves, new domains of use surface while some existing domains may be abandoned. UNESCO (2003) cited examples such as schools, new work

environments, new media such as broadcast media outlets and the Internet as new domains for language use. Domains like these tend to strengthen the dominance of the dominant language while weakening the vitality of the threatened language. However, some native languages manage to cope with the new challenges being presented to them by dominant language use, while others fail to respond to the challenge. For example, in education, two areas can be looked at: (a) what is the highest level of use in education? and (b) how many courses are taught using the native language? A language that is the medium of instruction at all levels of the education system should score higher than a language that is taught only as a subject. It is recommended that all new domains be assessed together to understand how the language is coping with the new domains (UNESCO, 2003).

#### **6. Materials for language education and literacy.**

Education in and through the native language is vital for the longevity of any language. For this to happen, educational materials should be available in the native language on a wide range of topics and for an audience ranging from children to adults. The language policy implemented by the prevailing administration and the attitude of the speech community toward their own language influences this factor (UNESCO, 2003). The policy on the use of a particular language in schools and in other official domains enhances or hinders whether materials are produced in that language and made available for use.

#### **7. Governmental and institutional language attitudes and policies including official status and use.**

Governments tend to institutionalize policies on the status and the use of a language in a given territory; however, the laws might be legislated or manifested through the actions of the ruling elite instead. For example, the law might protect all languages equally or some languages



might be given a special official status while others are given a relatively lower status. In certain instances, minority groups may be encouraged to abandon their native language through policies that emphasize the education of children in the dominant language (UNESCO, 2003).

#### **8. Community members' attitude toward their own language.**

Speakers of any language have negative, indifferent or positive attitudes toward their own language. UNESCO (2003) contended that these attitudes might be influenced, among other things, by whether or not it affords economic opportunities. Native speakers might even see their language as a burden to their social and economic mobility. Based on their perceived value of the language, therefore, native speakers might use it proudly, avoid using it or even consciously choose not use it and abandon it altogether. As we will see in a subsequent section, social, economic, and political forces will eventually put a language in danger as language use decreases. When this happens to a speech community, native speakers usually speak their own language alongside the dominant language, which eventually results in a switch to the dominant language. Certain groups have been able to resist the pressures and they often devise their own coping strategies (UNESCO, 2003).

#### **9. Amount and quality of documentation.**

This factor refers to, but may not be limited to, the availability and quality of language resources in the form of printed and/or digital dictionaries, grammar books, textbooks, and audiovisual materials, either monolingual or translated. The availability of such language materials promotes the use of the language because they enable researchers and technologists to design appropriate strategies in reviving, revitalizing, maintaining, or preserving the language. The availability of language resources also helps those who are in the field to prioritize the documentation efforts by looking at the areas of deficiency (UNESCO, 2003).

While the UNESCO (2003) factors have been adopted as the framework for this study, it is important to be aware of other assessment factors and frameworks. For example, Tsunoda (2006) proposed quite similar but slightly reduced assessment factors that include the following:

**1. Number of speakers with particular attention to native speaker communities.** This is similar to UNESCO's (2003) absolute number of speakers. The focus here is the number of native speakers of a given language.

**2. Age of speakers.** This refers to the age groups who speak that language. The previous listing did not mention this criterion but the rationale seems to be that a language with speakers of all ages is stronger than a language with reduced numbers of age groups.

**3. Transmission of the language to children.** Tsunoda's (2006) factors are in line with UNESCO's (2003) intergenerational language transmission factor. Tsunoda (2006) explained, "It is convenient to set it up as a separate criterion, for the survival of a given language crucially depends on whether or not children learn it" (p. 9). This statement underscores the importance of this factor as one of the major if not most important factors in assessing language vitality.

**4. Functions of the language in the community.** This is very similar to UNESCO's (2003) factors number 4 and 5 of the preceding list, which are *trends in existing language domains* and *responds to new domains and media*.

The foregoing assessment criteria developed by (UNESCO, 2003) will be adopted in this study for it seems to be more comprehensive criteria. Similarly, for the sake of being consistent, the criteria will be used for the language endangerment classification also developed by UNESCO (2003).

Before we proceed, I turn our attention briefly to the numbers. In other words, how many languages are being endangered? Unfortunately, the issue is too complex to get accurate and

agreed upon statistics. Even the number of languages spoken in the world is not well established in the literature. For example, Krauss (1998) estimated that the world speaks about 6,000 languages in total while Grimes (2000) estimated that there are about 6,800 languages spoken in the world. The BBC (2012) stated that there are as many as 7,000 languages in the world today. Lewis, Simons, and Fennig, editors of the Ethnologue website (2009, 2014), presented detailed information on all the estimated 6,909 languages spoken in the world. The reason for the discrepancy seems to be that the nature of languages is as dynamic as the lives of the people who use them. New languages are born and others are being discovered while healthy and stable languages become endangered and some leave the world for good.

Linguists and other concerned organizations differ not just on the number of languages but on the number of endangered languages. According to Nettle and Romaine (2000), there are about 6,000 languages spoken, of which about 50% of them will die out in the next century. This means that about 3,000 languages will be spoken by the turn of the next century. Crystal (2000) estimated that approximately every two weeks, one language spoken in the world today dies out. Krauss (1992) echoed this assessment and predicted that of the 3,000 or so languages that are endangered, as few as 600 languages may be sustained in the future. He asserted that about 90% of the world's languages are likely to be lost if the status quo persists. UNESCO (2003) predicted that "in most world regions, about 90% of the languages may be replaced by dominant languages by the end of the 21st century" (p. 4). Crystal (2000) concluded the debate among linguists by stating that pessimistically the number is as high as 90% while optimistically the number of endangered languages stands at about 50%.

Based on the foregoing predictions, either way the number of endangered languages is high and it seems that few languages will have the strength to survive the forces that threaten

languages. Globalization and technology, an issue that will be revisited later in this chapter, are accelerating these factors.

The figures on endangered languages are alarming and require immediate scholarly and collaborative international intervention. Resorting for a moment to analogical reasoning, common sense dictates that when a patient goes to a doctor, s/he has to take the doctor's diagnosis seriously and more importantly follow the doctor's prescribed treatment plan. Patients do this for two reasons. First, depending on the seriousness of the medical concern, the patient knows if the condition is not treated early on, it could become a matter of life and death. Second, the patient knows that the doctor uses a body of scientific knowledge in the deduction of the given diagnosis. In a similar fashion, researchers and educators ought to take the diagnosis, the predictions, and the treatment plan developed by many scientists in linguistics and anthropology to prevent the loss of an unprecedented number of world languages. The loss of languages is a reduction of the biodiversity of the world. Action needs to be taken. These scientists have used scientific inquiry in their fields to arrive at such conclusions. The issue will affect our lives and survival on earth as a human species. Readers can refer back to section 2.3 or Chapter Three for a more comprehensive review of the issue of biodiversity of languages. In what follows, I will investigate the major causes that put languages at risk and eventually cause their death.

## **2.2 Why are Languages Dying?**

In the history of humankind, many languages have disappeared and new ones have taken their place, which replenishes the pool of languages. This natural evolution in languages happens because traditional farmers and animal herders have each created their own territories inhabited generally by speakers of one language family.

Presently, languages are dying at an alarming rate and in large numbers and this tendency is exacerbated by the fact that new ones are not replacing them. In this context, the natural linguistic and cultural biodiversity is being lost (Romaine, 2008). Languages take one of two routes towards extinction, or to put it differently, there two major causes of language death. One is termed *sudden death* and the other is dubbed *gradual death*. The first one often comes in the form of natural or manmade disasters (Crystal, 2000; Krauss, 1992; Nettle & Romaine, 2000; Tsunoda, 2006).

Tsunoda (2006) used different terms to describe how languages die: “Language death may be caused by the death of the population or by language shift” (p. 70). Both factors cause languages to become endangered and may eventually cause death. Sudden death includes natural disasters such as earthquakes and floods. These causes diminish the number of speakers, thereby putting that language in danger. Manmade disasters such as war, which, depending on the scale of the war, decreases the number of speakers by either killing them or driving them from their natural environment (Tsunoda, 2006).

Tsunoda (2006) indicated that gradual change or language shift is much more common than sudden language death. She stated that political, social, and economic factors are the major causes endangering languages, summarized as *socio-politico-economic* factors. The writer further presented a detailed list of these factors in fifteen ways in which they can manifest. As the following fifteen factors indicate in one form or another, the term *social upheaval* will encapsulate all the language endangerment causes. Whether they are social, political, economic, or even technological, the upheaval exhibits some sort of disturbance to the social structure of a community and therefore the term *social upheaval* aptly illustrates the various forms that the

phenomenon of social upheaval can take. The list, below, includes the explanations of the author (Tsunoda, 2006) paraphrased:

1. **Dispossession of the land** means that the speakers are no longer the masters of their own land—examples include foreign invasions, colonization, settlement and/or grazing.

2. **Relocation of the people.** Either forced to relocate or they voluntarily choose to relocate as in the case of immigration.

3. **Decline or loss of the population** caused by natural disasters such as earthquakes, droughts, diseases, war, or emigration.

4. **Breakdown in isolation and proximity to towns.** The boundaries of the natural environment in which a language has lived are broken, exposing it to other languages and cultures. Moreover, it occurs when speakers of a particular speech community are closer to towns in terms of physical proximity or the availability of easier modes of transportation.

5. **Dispersion of the population** occurs when the users of a language are no longer together and are dispersed in a geographically distant area.

6. **Mixing speakers of different languages** in such situations such as boarding schools, settlements, and intermarriages which force speakers to communicate in one of the dominant languages such as English (in Canada, Australia, and the US), and Swahili, people may forgo using their indigenous languages and instead communicate in the lingua franca of countries such as Tanzania and Kenya, which speak Swahili.

7. **Socio-economic oppression, oppressive domination and economic deprivation.** This is a rather powerful factor; people often abandon a language or learn a new language based on the employment opportunities it affords. If an indigenous language is seen to be of no value to

the speaker, s/he will abandon it and shift to the dominant language and will not teach the original language to future generations.

**8. Low status or low prestige of a speech community** might be caused by one or more of the above factors and for some reason the speakers may feel ashamed if they use their own language in front of others.

**9. Language attitude** concerns how people perceive the intrinsic value of their own language vis-à-vis the value they attach to the dominant language.

**10. Assimilation policy and language policy.** The example here is when children are educated in a language other than their mother tongue, which has a major impact on the vitality of their language. French or English immersion programs are good examples of this, as the dominant language is promoted at the expense of the first language of the children. Heller (1994) pointed out that immersion schools in Canada give the English- or French-speaking communities the economic, political, and cultural edge over minority children who are educated in these schools.

**11. Relative lack of indigenous language literature.** Literature is the core of any language and its absence can be a major contributor to its endangerment. This lack refers to when a language loses its oral literature as well as printed resources.

**12. Social development, modernization, and industrialization.** These factors are driven mainly by facilities and tools made available by technology such as TV, the Internet, computers, and infrastructure such as roads, cars, ships, trains, and airplanes. These tools have replaced old technologies and traditions. They have enabled people to become more mobile, both physically and in their communications.

**13. Destruction of the environment/habitat** means that speakers leave their homes and towns when a government decides to take over land ostensibly for development reasons.

**14. Spread of religion.** Religions bring with them languages in which the religion was revealed or written; this might cause a shift to the new language or using the new language in certain domains, replacing the native language in those domains.

**15. Cultural contact and clash.** All of the above factors cause social upheaval and language loss, but the main argument here is when one group conquers another and a cultural and linguistic shift inevitably emerges.

Campbell (1994) presented a rather collapsed list that groups the 15 factors into two major factors: *socio-economic* and *socio-political*. Socio-economic includes causes such as the pursuit of economic opportunities, industrialization and migration, and so on, whereas socio-political factors include war, institutionalized language policy, discrimination and any other form of community repression, including colonization. Tsunoda (2006) asserted that colonization has been the prime cause of languages disappearing through the language and cultural threat posed by colonial powers in the past that falls under the causes listed above.

A new phenomenon but perhaps a more serious threat to language diversity is posed by globalization with its soft-power tools facilitated mainly by technology. Crystal (1997) regarded globalization as the most powerful cause because it facilitates the assimilation of languages into one or a few dominant languages. With regard to language endangerment, technology is arguably the main driving force of globalization because it has secured the status of English as the lingua franca of the world. In illustrating this trend, Crystal (1997) hypothesized, “The biggest potential setback to English as a global language, it has been said with more than a little irony, would have been if Bill Gates had grown up speaking Chinese” (p. 112).



Obviously, such a distinct status does not come from technology alone, but from powerful socio-cultural forces. Complementing the foregoing argument, Dieu (2005) summed up what drives the dominance of the English language:

The United States have consolidated their cultural, economic and technological power: inventions, rock and roll, the first man on the moon, the revolution of the Internet, the country's growing prosperity and commercial aggressiveness have contributed to the further expansion and importance of English in the world today. (p. 2)

Of all the modern tools afforded by technology, Krauss (1992) singled out the television as the “cultural nerve gas” because it takes dominant cultures and languages into the living rooms of speech communities around the world, thereby speeding the process of cultural assimilation. TV and other similar media tools afforded by technology are essentially used to perpetuate what Krauss called “linguistic and cultural genocide.” Therefore, it is evident that technology drives much of the cultural and language assimilation prevalent in many parts of the world. Indeed, all the media outlets including TV and the Internet with their unprecedented influence through music, movies, programs, and shows combine to make the English language the most dominant language in the world today. This dominance provides a seemingly unshakable influence on indigenous languages. Grenoble and Whaley (1998) aptly pointed out, “No language has ever exercised so much international influence as English” (p. 69). As a consequence, this influence is maintained across the world inevitably at the expense of other languages, which become endangered, moribund, and extinct.

Nevertheless, it must be stated that although English is the most dominant language, there are other languages that exert a lesser but noticeable influence and pressure on weaker languages. The political clout and influence of these languages are evident in the international

arena. Of all the 6,700 languages in the world, only six of them enjoy official status at the UN. According to UN official languages (2014) the official languages of the UN are: Arabic, Chinese, English, French, Russian and Spanish. This means that a UN dignitary may choose to speak any of the official languages and his/her talk will concurrently be rendered in other official languages. It also means that the UN releases its publications in these languages and speakers of these languages have better career chances than speakers of other world languages.

These languages have varying influences in different parts of the world, each competing with indigenous languages in their own territory. Djibouti, which was formerly known as French Somaliland, was a French colony until 1977. When Djibouti became an independent country, French remained the official language of the country together with Arabic (Hagget, 2001). Levinson (1998) pointed out that Djibouti had economic and trade relations with France and Middle Eastern countries. Adegbija (1994) argued that the languages left by colonial powers remain unshakable. They are still largely regarded as the official language and medium of education systems in Africa. Berns (2010) illustrated this by pointing out that of the 56 countries in Africa, 22 countries have French as their official language. In Somalia, after the collapse of the central government, English and Arabic have become dominant in schools. As Cassanelli and Abdikadir (2008) stated, “The medium of instruction at the primary level may be Arabic, Somali, or English. Most secondary schools use either Arabic or English” (p. 107). It appears that stronger languages compete with or sometimes supersede indigenous languages in their own territory. As discussed previously, economic ties, religious affiliation, and geographical proximity (among other factors) all contribute to people opting for stronger languages than their mother tongue. In the case of Somalia, people may prefer Arabic (Cassanelli & Abdikadir, 2008) presumably for one or all three factors mentioned over Somali.

### 2.3 Why Does it Matter?

One might ask whether or not it would be easier to live in a world with fewer languages. This issue will be discussed in more detail in Chapter Three but one approach to answering this question is to reverse the question and suggest to the questioner: “Let us not include your language among the chosen few.” It is likely that the answerer would reply defensively, showing among other things exactly what linguists have found about the inextricable relationship between language, personal worth, and identity. Spolsky (1999) stated, “Language is a central feature of human identity. When we hear someone speak, we immediately make guesses about gender, education level, age, profession, and place of origin. Beyond this individual matter, a language is a powerful symbol of national and ethnic identity (p. 181). It is reasonable to equate how language identifies a person through passports and other documents. In most cases, language is more powerful in identifying someone than the passport s/he holds. Anzaldua (1987) further illustrated the point by asserting: “Ethnic identity is twin skin to linguistic identity—I am my language” (p. 59). Therefore, language loss is not only losing a language but the person loosens attachment to her or his culture, identity, and personal values. Skutnabb-Kangas (2000) explored the issue as part of a basic human rights issue: “In a civilized state, there should be no need to debate the right to maintain and develop the mother tongue. It is a self-evident fundamental linguistic human right” (p. 625). It is apparent that languages are linked to personal and cultural identity, which ought to be respected as a basic individual human rights issue.

Language and cultural loss has a domino effect because language and heritage loss is no longer confined to that particular or individual speech community, which affects our diversity on earth. Consequently, as Harmon (2002) argued, the world community loses part of its diversity on earth and this is not only about biological diversity but it also is about losing another

dimension of our natural balance—intellectual diversity (Harmon, 2002). Alcom (1993) noted that a new concept, biodiversity, captures such interplay:

While proof of conservation success is ultimately biological, conservation itself is a social and political process, not a biological process. An assessment of conservation requires therefore an assessment of social and political institutions that contribute to, or threaten, conservation. (p. 424)

As UNESCO (2003) noted, conservation biology is often equated with conservation linguistics. The two are inextricably intertwined because the former is concerned with saving the diversity of living beings on earth while the latter deals with saving the diversity of languages and cultures in the world. The link is important because in order to describe the diversity of life in general, one cannot exclude the diversity of languages and cultures. The combination of biology and linguistic diversity is what Maffi (2005) called *biocultural diversity*. She asserted that for the benefit of conserving endangered species, languages and cultures that carry knowledge about them should also be saved.

Evans (2010) stated that in the Western educational system, there are two main schools of thought. The first tends to make knowledge “universal” under the umbrella of one language with one worldview. The author traces history and points out that Latin, Arabic, French, and English have all had their turn. The second school of thought acknowledges that one language cannot conceivably represent the richness and range of ideas that other speech communities offer. Evans stressed the importance of linguistic diversity and “any reduction of language diversity diminishes the adaptational strength of our species because it lowers the pool of knowledge from which we can draw” (p. 19). Hinton (2001) summarized the gravity of the loss rather more powerfully:

More broadly, the loss of language is part of the loss of whole cultures and knowledge systems, including philosophical systems, oral literary and music traditions, environmental knowledge systems, medical knowledge, and important cultural practices and artistic skills. The world stands to lose an important part of the sum of human knowledge whenever a language stops being used. Just as the human species is putting itself in danger through the destruction of species diversity, so might we be in danger from the destruction of the diversity of knowledge. (p. 5)

In other words, the world is in dire need of the diversity of ideas if society is to tackle the complex global problems we face today, and more so now than perhaps at any other time in human history. Humanity is confronted by a myriad of complex problems such as chronic and deadly diseases, wars, and political instability in many parts of the world, as well as global warming and environmental degradation. The effects of globalization and technology coupled with the widening gap between the rich and poor are complex issues that will take the entire global community to address. These contemporary phenomena would seem enigmatic if they were examined through a microscope with a single worldview. Evidently, such complex issues dictate the need for intellectual, cultural, and linguistic diversity rather than one or more worldviews.

## **2.4 What Does it Take?**

The foregoing statistics reveal that more than one-half of the world's languages are endangered and this paints a picture that the problem is insurmountable. True, the task of language preservation, documentation, and revitalization is a daunting task. Nonetheless, people with strong willpower can undertake difficult and complex tasks such as this. The source of such strong determination comes from inspirational or success stories on the issue. Crystal (2000)

stated that there are cases where languages have been revitalized after they had fallen into the category of endangerment or were identified as languages moving towards extinction. As we will illustrate, current speakers of Modern Hebrew and Maori, spoken by native New Zealanders, have managed to reverse the status of their own language.

Inspirational or success stories might serve as a catalyst or a driving force that gets a language revitalization project initiated. In order to sustain that strong willpower until the task is finished, people need resources including time, money, and relevant information about the task ahead. Crystal (2002) tried to give fairly reasonable numbers so that researchers and practitioners, along with affected speech communities, can gather statistical information to use for project planning purposes. He estimated approximately \$100,000 per year for three years for any given endangered language. He multiplied that amount for the number of endangered languages (at that time) estimated to be about 3,000 languages and suggested the total amount would be \$900 million. This is of course a rough estimate and would depend on the context of the language in question, the resources available, and the documentation and revitalization method employed, among other things, but it serves as a starting point. Once the amount and the time have been established, the next question is the task of resource mobilization. Below are success stories about two languages and how they have managed to turn the tide and have mobilized all the resources needed to reverse a situation that might have seemed hopeless and irreversible to many people.

## **2.5 How Can Languages Be Revitalized?**

The method of intervention depends on the case itself; therefore, a variety of strategies were developed to meet the treatment needs of each of the ailing or endangered languages.

Grenoble and Whaley (2006) observed that the context dictates the treatment:

While in one context a revitalization effort may be centred around formal education, in another it may be focused on creating environments in which the language can be used on a regular basis. Although tremendous variety characterizes the methods of and motives of reinvigorating languages, revitalization, as a general phenomenon, is growing and has become an issue of global proportion. There are now hundreds of endangered languages, and there are few regions of the world where one will not find at least nascent attempts at language revitalization. (p. 1)

Shaul (2014) noted that there are a number of successful language revitalization initiatives, notably Hebrew, Maori, Hawaiian, Greenlandic, Mohawk and others. A brief overview of the success stories of Hebrew and Maori will be given. The revival of Hebrew from a dead language to a language with over five million native speakers serves as an inspirational story while the revitalization of Maori and its rise to official status in New Zealand illustrates that an endangered language can be reclaimed in a country where it is regarded as a minority language.

### **The Revitalization Story of Modern Hebrew**

One of the most impressive success stories comes from the history of the Hebrew language. As Zuckermann and Walsh (2011) explained, Hebrew is probably one of the world's oldest languages, dating back to the 14th century BC. It is a Semitic language belonging to the Afro-Asiatic language family.

Zuckermann and Walsh (2011) asserted, "For approximately 1,750 years Hebrew was *clinically dead*" (p. 114). Although Hebrew was not spoken as a mother tongue, it was nonetheless dead in the strict sense of the word. It had functioned as a language for all religious purposes and it has a written literature (Fellman, 1973; Zuckermann & Walsh, 2011).

Many have credited the revival of Hebrew to Ben Yehuda who came to Palestine in 1881 from Lithuania (Fellman, 1973; Zuckermann & Walsh, 2011). He was exposed to Hebrew at a very young age through his studies of the Torah, the Old Testament. He also studied the History of the Middle East and read books and ancient literature in Hebrew. He later went to France for further studies where he experimented with writing articles in Hebrew. Ben Yehuda travelled to Algeria where he met fellow ethnic Jews and discussed the idea of reviving Hebrew as unifying mother tongue for Israelis. It was only an idea before Ben Yehuda decided to travel to Palestine and test it on the ground. He was multilingual (Fellman, 1997).

Fellman (1997) acknowledged in his book that Ben Yehuda did not write about the process of his revitalization framework but it was systematic indeed. Fellman summarized the steps of Ben Yehuda's revitalization agenda as follows:

*1. Ben Yehuda's family as the first Hebrew-speaking family*

It appears that Ben Yehuda was cognizant of two essential requirements for any community-based initiative. First, it starts at home and if it is initiated and tested in a home, others are likely to follow the practice with confidence. Second, Ben Yehuda also understood the effectiveness of modelling and that if one puts his own theory into practice people are likely to take him more seriously. Ben Yehuda had an agreement with his new wife that their household would adopt Hebrew as the language of all functions. An apparent challenge was that Hebrew had no adequate terms or phrases for everyday communication because it was a language that had not been used for over 1,700 years.

Inevitably, Ben Yehuda and his wife had to coin new words and phrases out of Hebrew root words. Note that Hebrew was largely preserved in religious and literary books, so the enduring legacy of this documentation served as a great source of reference. As they coined new words



they started using them and documenting them in a phrasebook that Ben Yehuda had started. It is appropriate here to cite a Somali poet who rhetorically asked in a poem: “*Wixii la qoraa quruumo haree, muxuu hadal qiima leeyahay?*” In English the line reads as: “Written materials survive for centuries; what value can we attach to oral discourse?” The poet uses conservation as the only criterion for evaluation presumably to underscore the usefulness of documented artifacts especially in a scenario such as the one Ben Yehuda had to confront.

A further challenge arose when Ben Yehuda’s first child was born because the project was still in its infancy and the language was being modernized. As part of their prior agreement, their son was expected to join forces with his parents in enforcing the Hebrew-only policy in the household. In helping their child learn Hebrew only, they had to ensure that the baby was never exposed to words or phrases that might be used by visitors who spoke either Yiddish or Arabic, which were the most common languages at that time.

## *2. A an appeal to both local and Diaspora Jews*

Concurrently with his efforts at home, Ben Yehuda had made his plan known to the wider community in an effort to mobilize his community toward the common goal to revive a lost language dormant for centuries. The aim was to garner support for the revitalization project since the revival of a language largely depends on the support and commitment of its speakers.

## *3. Expanding the number of speakers*

This step was to expand the number of Hebrew speakers. Building on prior stages, Ben Yehuda now sought to find committed community members who were willing to enlarge the base of Hebrew speakers. Through this method, the speakers gradually grew in the local community who were being inspired by the language modelling in Ben Yehuda’s household.

## *4. Hebrew in the schools*

An important step was to penetrate the school system. Ben Yehuda approached a principal in one of the schools with a proposal that he and some of his colleagues help children learn Hebrew in a natural setting where children were immersed in the language.

#### *5. Hebrew in the media*

The objective was to disseminate the developing language through the first newspaper that was published in Modern Hebrew. This seems to have been an effective tool to reach out to a wider audience so that the language moved towards building standards and conventions of use. One of the strategies included the use of rich context so that people understood what a particular word meant. Readers then solidified their understanding of words or phrases, as the words were re-used in the newspaper.

#### *6. The dictionary of contemporary and old Hebrew*

Drawing upon various sources including religious and literary books, the wordbook he created, the newspaper, and other sources, Ben Yehuda published his first dictionary, which included both modern and old Hebrew words.

#### *7. Hebrew Language Council*

Established in 1889, the council was formed with a mandate to oversee the development of the language and the documentation and standardizing of the language. Zuckermann and Walsh (2011) noted that the council was later replaced by the Academy of Hebrew Language in 1953, which is still responsible for language development. This includes researching and developing existing dictionaries and leading the development of the language given its needs and current status.

It took about forty years after Ben Yehuda's arrival for the language to regain its status as a language spoken and written by millions of Israelis in every facet of their lives.

According to the UCLA Language Materials Project (2014), Hebrew now has 5.3 million speakers and it is one of the three official languages in the country together with Arabic and English. Hebrew dominates print and digital media including school textbooks, magazines, and newspapers and is the medium of instruction in schools up to university level.

### **Revitalization Efforts of Maori in New Zealand**

The Maori language is spoken in New Zealand by native New Zealanders. It was interesting to learn that according to Shaul (2014), Maori speakers had a strong literacy tradition up to 1818. In fact, in the early 1830s Shaul reported that Maori speakers had higher literacy than English speakers in New Zealand. At that time Maori was used as the primary language at home, in politics, and even in the media. Wright (1996) noted that as early as 1847 “Anglicization” of education started and in 1967 The Native Schools Act was passed in New Zealand, which legalized English as the only language of instruction in all schools and banned Maori in schools. This had a detrimental effect on the Maori language because many of its speakers had to switch to English and an increasing number of children ceased to speak Maori. In the early 1970s, the language vitality was at its weakest, accelerated by mass immigration to urban areas coupled with the cessation of intergenerational transmission through schools, home, and media channels. The language was considered an endangered language because it had ceased to be used in almost all domains (Shaul, 2014; Wright, 1996).

However, Wright (1996) indicated that the Maori language revitalization movement started with organized public protests and political pressures in 1981, which marked a proposal between the New Zealand government and community activists to establish the “language nests” program for children. The program sought to immerse pre-school children of Maori descent in a setting where Maori elders created activities for the children using only Maori. In 1982, four

pilot centres were opened in which children were observed to have remarkable results. Given the success of the centres, they grew to about 500 in 1987.

In 1986 the Maori language was recognized as one of the official languages in New Zealand along with English and a language commission was established. There are now over 819 languages nests and over 335 other schools in which the medium of instruction is Maori. The Maori language revitalization is viewed as one of the most successful indigenous languages (Hoffman, 2012; Shaul, 2014; Wright, 1996). Considering the position of Maori in the 1970s, the achievement in the last four decades has been commendable. Having been granted official status as well as the increasing pace at which schools have adopted Maori is a remarkable success story.

The success stories of Hebrew as a dead language being reclaimed and Maori, an endangered language, being revitalized, illustrate the fact that while language revitalization might seem hopeless, it is an attainable mission. As McCarty, Romero, and Zepeda (2006) echoed, “There are tangible precedents for such a possibility as exemplified by successful indigenous language revitalization initiatives in New Zealand, Hawai’i, California and elsewhere” (p. 43). Given the limitations of space and time, we have used only two languages to illustrate the point.

## **2.6 Who is Involved?**

In the past two decades, the debate on the severity of the language endangerment phenomenon and ways to develop and mobilize resources to document, maintain, and revitalize languages has gradually drawn international attention. Nettle and Romaine (2000) stated that in the 1990s the issue was limited to academia in which mainly linguists have researched the link between language loss, cultural diversity, and biodiversity. About five years later, Maffi (2005)

observed that the field of bio-cultural diversity, and by extension the field of linguistic diversity, has started to move beyond the boundaries of linguistics to the spheres of other academic areas in an interdisciplinary field. Maffi (2005) argued:

Over the past decade, the field of bio-cultural diversity has arisen as an area of trans-disciplinary research concerned with investigating the links between the world's linguistic, cultural, and biological diversity as manifestations of the diversity of life. The impetus for the emergence of this field came from the observation that all three diversities are under threat by some of the same forces and from the perception that loss of diversity at all levels spells dramatic consequences for humanity and the earth. (p. 599)

Grenoble and Whaley (2006) observed that the issue has drawn the attention of not only academics but also the international community. The *transdisciplinary* nature of the field that Maffi (2005) noted is reflected by the range and the number of international corporations, governments, non-governmental organizations, and educational institutions from many fields and geographical areas who have become increasingly involved in the research, revitalization, and the maintenance of endangered world languages and cultures. The efforts are concerned mainly with confronting the issue practically, while the research on the issue is being accumulated. Maffi (2005) stated, "The field of bio-cultural diversity has developed with both a theoretical and a practical side, the latter focusing on on-the-ground work and policy, as well as with an ethics and human rights component" (p. 599).

It has been encouraging to learn of the numerous organizations that are involved in the documentation, research, and revitalization projects targeted at endangered languages and also at the rate in which they have increased over the years. For example, a number of organizations such as Google have joined the collaborative international effort since this thesis began in 2010.

It would have been useful to give an exhaustive list of all organizations involved but given the limited scope of this study, the involvement of a few organizations and nations are given below based on how their efforts relate to the themes of this research.

### **Microsoft Local Language Program**

As part of its corporate global citizenship efforts, Microsoft helps communities around the world use their technological tools and interfaces by localizing them in their own languages. The program was launched in 2004 to bridge the technological gap among global citizens. According to Microsoft (2004):

The Local Language Program is a global initiative that fosters the development and proliferation of regional language groups, enabling them to preserve and promote their language and culture while benefiting from continuing IT advancements. Through this collaboration with local governments to offer citizens the ability to customize leading, values-based Microsoft software applications with local language capabilities, people around the world will be able to work with PCs—some for the first time—in their native languages. Individuals will be able to build skills, open opportunities and realize overall IT progress. (para, 4)

Microsoft introduced Language Interface Packs, which can be downloaded for free in many languages. Many of these communities have not had any technology facility in their native tongue. Microsoft (2014) stated the Language Interface Packs are available in 108 languages, offering Microsoft products and services in their own language. Of particular interest to this thesis is the inclusion of less commonly studied languages. The list includes languages that are now being developed: Maori (New Zealand), Icelandic (Iceland), Dari and Pashto (Afghanistan). In Africa the list includes Amharic, Tigrinya (Ethiopia), Yoruba, Igbo, Hausa (Nigeria), Swahili

(Kenya, Tanzania, Uganda), and Welsh (Wales). Furthermore, Inuktitut speakers in Canada can also use Microsoft products and supported services in their language.

Through the Local Language Program, Microsoft includes its packs with Windows, Microsoft Office, Microsoft Translator and Microsoft Language Portal among other services. Some of the services provide a platform where IT experts, linguists and communities expand their vocabulary collaboratively through translations of IT terminology. For example, the Microsoft Language Portal collects and disseminates Microsoft terminology while Microsoft Translator Hub allows people to offer IT terminology translations in their language towards a common goal that all languages will eventually be included in the program.

### **Google Corporation—The Endangered Language Project**

Prior to its launch of the Endangered Languages Project in June 2012, Google has been involved in other charitable activities that sought to impact the lives of communities around the world. According to Google (2014), Google's foundation wing donates \$1 billion annually while donating another billion in products along with 60,000 hours in in-kind services grown out of its philanthropic wing. The Endangered Language Project (2014) presents technology tools that facilitate collaborative work in tackling the problem:

Endangered Languages Project puts technology at the service of the organizations and individuals working to confront the language endangerment by documenting, preserving and teaching them. Through this website users cannot only access the most up-to-date and comprehensive information on endangered languages as well as samples being provided by partners, but also play an active role in putting their languages online by submitting information or samples in the form of texts, audio, or video files. In addition,

users will be able to share best practices and case studies through a knowledge-sharing section and through joining relevant Google Groups. (para. 3)

Google's project managers are aware of the effectiveness of collaborative efforts in addressing a problem that needs contributions from all corners of the world. This recognition is evident in its facilitation of collaborative work through offering technology infrastructure that helps users to upload and share files in various formats. Similarly, Google has chosen its project to be under a wider umbrella of the Alliance for Linguistic Diversity. The Alliance brings together members of the media such as Canada's CBC radio, world-class universities, non-profit organizations, libraries, and most importantly many indigenous groups working with endangered languages. The membership is open to organizations and individuals who are interested in getting involved. Notable institutions include the University of Hawai'i at Manoa and Eastern Michigan University, which are engaged in building a digital catalogue of endangered languages. A Long Now Foundation, listed below, is also a member of the Alliance for Linguistic Diversity.

### **A Long Now Foundation Library of Human Language: The Rosetta Project**

A Long Now Foundation in collaboration with Stanford University Libraries and the National Science Digital Library based in the United States has established the Rosetta Project with the aim of building a freely accessible digital library of human languages. Inspired by how the discovery of Egypt's Rosetta stone became a window to the life, culture, and language of the ancient Egyptian civilization (Addison, 2009), A Long Now Foundation plans to replicate the same concept on a larger scale. This is another collaborative effort between technologists and native speakers of languages that are endangered and under-documented. The project has already digitized over 1,500 languages on a disk that fits on a human palm, containing over



13,000 pages. It is clearly a modern-day Rosetta stone for future generations to learn from, in case some or many of these languages die out (The Rosetta Project, 2014).

### **Volkswagen Foundation: Documentation of Endangered Languages**

As its tagline “A Foundation of Knowledge” suggests, Volkswagen is a non-profit organization that focuses on research and knowledge dissemination in particularly challenging areas. According to Volkswagenstiftung (2014), the foundation added to its portfolio the theme of documenting endangered languages, which is not confined to Europe but takes in documentation proposals from every corner of the world. It funds and mentors documentation and research projects that aim to document or add a research contribution to the field. Since 2002, it has funded projects ranging from endangered languages in the Republic of Georgia, North Australia, and from East Timor to Papua New Guinea. The foundation is based in Germany.

### **Rosetta Stone: Endangered Languages Project**

Rosetta Stone is a digital language-learning materials developer. Based in the UK, the company provides technological knowhow for language documentation projects helping speech communities develop versions of content in their native language.

According to Rosetta Stone (2012), “Previous success stories include the Mohawk community of Kahnawake in Quebec, the Seminole tribe of Florida and NANA Corporation’s Inupiat shareholders in the Arctic. Rosetta Stone has been selected by these communities as the technology of choice for language revitalization” (para. 2). For example, Rosetta Stone has collaborated with the Mohawk community in Quebec, Canada in the production of an interactive language-learning software for Mohawk language learners, which is available on a CD or online for use in schools, at home or in community centres. Rosetta teaches languages through

technology and it invites individuals and communities to develop editions of their language. According to its website, while Rosetta Stone offers technology expertise, it refers language revitalization projects to funders such as Canada Heritage and Administration for Native Americans.

### **UNESCO: Endangered Languages**

The United Nations Educational, Scientific and Cultural Organization (UNESCO) is arguably the largest global institution exclusively promoting issues related to world languages, culture, heritage, and education. It is involved in a wide variety of projects that strategically target the issue of language endangerment from different fronts. It provides a comprehensive list of endangered languages in its Atlas of the World's Languages in Danger, available on its website in English, French, and Spanish. The latest publication was the 2010 edition, which classifies languages into five degrees of endangerment. It regularly holds expert and ad hoc meetings on matters that relate to language endangerment and vitality issues. For example, a notable contribution by its ad hoc expert meetings is UNESCO's (2003) methodology for assessing language vitality and endangerment, which this study adopted for its language vitality assessment. Collaborating with communities and other concerned institutions and governments, UNESCO carries revitalization and documentation projects around the world.

UNESCO serves as an advocacy body for endangered languages by holding regular general conventions on the issue as well as drafting declarations about the preservation of cultural and linguistic diversity. In 2001, for example, UNESCO drafted and tabled the UNESCO Universal Declaration on Cultural Diversity, which was adopted and signed by all UNESCO member states. As part of its advocacy and awareness campaign, UNESCO holds special events and conferences. Furthermore, its website has a collection of resources including

websites and online resources, audiovisual resources, brochures, and flyers as well as a selected bibliography on endangered languages. UNESCO supports regional bodies engaged in local language development.

### **African Academy of Languages**

Under the auspices of the African Union, the African Academy of Languages was established in 2006 with the vision of integrating Africa's diverse languages into its development spheres and promoting their use in all domains of life. The academy acknowledges the rapidly declining status of Africa's languages and the negative attitudes that most Africans have towards their language, stating that "not many Africans believe that meaningful education is possible in their languages beyond the early years of primary education" (African Academy of Languages, para. 11). According to African Academy for Languages (2014), the core projects include offering master's and doctorate level degrees in applied linguistics focusing on research on African languages. The degrees are offered in conjunction with three universities located in three African countries—South Africa, Ethiopia, and Cameroon. Other major projects include collecting and publishing stories across Africa for children as well as a Lexicography and Terminology Development Project. The Academy works closely with UNESCO as well as institutions and AU member countries (African Academy for Languages, 2014).

### **Somali-Speaking PEN Centre**

Established in 1997, the centre has its headquarters in Djibouti with satellite offices in the UK, Norway, and Ethiopia; it has three offices in Mogadishu, Hargeisa, and Garowe, Somalia. It aims to coordinate, collect, and disseminate the efforts of Somali writers. The centre serves as an advocacy body for the Somali language and its development, holding international events and

working closely with organizations such as ACALAN, UNESCO and other regional institutions including the Institute for Mother Languages in Djibouti.

The centre is a member of PEN International, which is an international organization with members in over 100 countries, with the aim of fostering literature and freedom of expression around the world (Somali Speaking PEN Centre, 2014).

### **Institute For Languages Djibouti**

The institute is engaged in research and development of the two mother tongues in Djibouti: Somali and Afar. The institute regularly holds conferences, research symposiums, and book launch events. Marking the 40th anniversary of the implementation of Somali orthography, the institute together with Somali PEN unveiled new books in the Somali language including a comprehensive dictionary in Somali (Somali Speaking PEN Centre, 2014).

### **Universities of Cambridge and Yale: World Oral Literature Project**

These universities have joined forces in forming the World Oral Literature Project. According to the project website, it is “an urgent global initiative to document and disseminate endangered oral literatures before they disappear without record” (World Oral Literature Project, 2011, para. 1). The projects aims to collect “verbal art” including poems, folktales, songs, myths, legends, word games and many others that the project classifies as endangered artifacts in many parts of the world. The project funds fieldwork and research in collecting endangered oral literature. It has an outreach component through which workshops, lectures, and a database of world oral literature are shared with the public (World Oral Literature Project, 2013).

### **University of London**

The School of Oriental and African Studies offers graduate programs in which students specialize in one of the related areas in their Endangered Languages Academic Programme. According to the University of London (2014), “Language documentation is a new sub-discipline within linguistics that has emerged as a response to the growing crisis of language endangerment” (para. 1). This school is an indication that the field is beginning to solidify its position in academia as major world universities start to offer graduate degree programs in degrees that are tailored to address the growing trend in supporting endangered languages. Zuckermann and Walsh (2011) argued that as a new domain in linguistics, “Revival Linguistics” need to be established in universities.

### **Swarthmore College: Laboratory for Endangered Languages**

Swarthmore College supports endangered languages in small communities. The college documents, researches, and develops talking dictionaries in those languages. Some of the talking dictionaries include Tuvan (Siberia), Ho (India), Sora (India), Siletz (Native American in Oregon), Chamacoco (Paraguay), and Remo (Peru). These are all endangered languages. Furthermore, in collaboration with the National Geographic and Enduring Voices project, Swarthmore plans to create a map showing endangered language hotspots.

### **University of Minnesota: Second Language Requirement**

Students in the College of Liberal Arts must meet a second language requirement in order to graduate. Students choose from a menu of 26 languages including many less commonly studied languages such as Dakota, American Sign Language, Hmong, Somali, Swahili, Latin, Hebrew, Ojibwe, and Greek as well as other stronger world languages such as Chinese, French, Swedish, and German. It is interesting to note that these courses are all credit bearing, so

students in other colleges can take them as an elective course. The Somali course was introduced in 2009 and has been a popular course with younger Somali–American students as well as non-Somali speakers. Abwaan Said Salah, a Somali and English teacher, poet and playwright and Somali language activist, teaches the course.

### **University of Alberta**

The university houses the Canadian Indigenous Languages and Literacy Development Institute (CILLDI), which offers credit-bearing courses to students in learning an indigenous language or taking a course related to language revitalization and documentation. The institute offers:

An intensive, annual summer school held at the University of Alberta whose goal is to train First Peoples speakers and educators in endangered language documentation, linguistics, language acquisition, second-language teaching methodologies, curriculum development and language-related research and policy making. Our mandate is the preservation of endangered languages by developing research skills and teaching resources in the speakers of these languages themselves. (University of Alberta, 2014, para. 1)

The institute was established in 2000 in a joint initiative with the Faculty of Arts, Faculty of Education, and Faculty of Native Studies. It appears that the interdisciplinary nature of endangered languages has brought about such a partnership. For instance, students who are enrolled in the Indigenous Languages Instructors Certificate would require teaching methodology courses from the Education Faculty, courses in indigenous culture and art from Native Studies, and linguistics from the Faculty of Arts. As this partnership indicates, it has been

a recurring theme in this thesis that efforts in endangered languages require skills and knowledge that include linguistics, education, art, and culture as well as technology.

This chapter has given a general overview of the issue of endangered languages by discussing the endangered languages crisis in light of the factors that perpetuate the status quo. The significance and the magnitude of the issue, which warrants immediate international intervention, were illustrated. The chapter concludes with an annotated, non-exhaustive list of those who are involved in the cause of language revitalization. It is an apt reminder to end the chapter with a statement borrowed from Google's Endangered Project:

Humanity today is facing a massive extinction: languages are disappearing at an unprecedented pace. And when that happens, a unique vision of the world is lost. With every language that dies we lose an enormous cultural heritage; the understanding of how humans relate to the world around us; scientific, medical and botanical knowledge; and most importantly, we lose the expression of communities' humour, love, and life. In short, we lose the testimony of centuries of life. (The Endangered Language Project, 2014, para. 1)

The quote encapsulates the essence of what fuels and sustains the efforts of those who are involved in the cause of language revitalization. The main themes of the quotation will be revisited in the following chapter. In particular, the correlation between medical knowledge and languages will be illustrated through the evaluation and analysis of the Somali language as representing endangered languages. Similarly, creative ways of expressing "humour, love and life" through Somali poetry will be explored vis-à-vis the Story of Canada's iconic novelist and

writer, Margaret Laurence and how her artistic and creative talent was inspired by her contact with the Somali language and literature.



### Chapter 3: The Case for Somali as an Endangered Language

*Afkaygow, tabaaliyo*  
*Maxaad tow ku nooleyd!*  
*Maxaad belo u taagneyd!*  
*Shisheeyuhu tab iyo xeel*  
*Muxuu kugu tuntuunsaday,*  
*Tisqaadkaaga dhaawacay!*  
*Maxaa gabay tilmaannaa*  
*Maahmaaho toolmoon*  
*Maalmuhu tireenoo*  
*Qalbigaygu tebayaa!*  
*Maxaa erey tafiir go'ay!*  
*Maxaan maanso teeriya,*  
*Tacab ba'ay ka joogaa!*  
*Sida ay u taxan tahay*  
*Murtidaadu tabanaa*  
*Teelteelna badanaa! ~Maxamed Xaashi Dhamac "Gaariye"*

Think back to the time when our language suffered  
 One onslaught after the other: invasions,  
 Armies crossing our borders, the songs  
 Our fathers once sang destroyed or derided,  
 Our epics fading in memory, even  
 Our idioms gradually losing their meanings.

Every lost syllable tells in my heartbeat,  
 Every lost line is a scar on my heart.  
 Poems go hand-over-hand to create  
 A chain of wisdom, a chain that goes  
 From strength to strength; when this was shattered,  
 When our chain of poems was broken and scattered,  
 We were left with nothing but fragments, nothing  
 But scraps of wisdom - our inheritance  
 Nothing more than a handful of images.<sup>1</sup>

---

<sup>1</sup> This is an excerpt of the poem A to Z composed by one of the great Somali poets, Gaariye, on the issue of the Somali language being endangered and the importance of languages in general. The literal translation of the poem was made by Martin Owen and Maxamed Xasan 'Alto.' David Harsent has done the final translated version. The original poem and its translated version can be found at the *Poetry Translation Center* at [www.poetrytranslation.org/poems/70/A\\_to\\_Z/translated](http://www.poetrytranslation.org/poems/70/A_to_Z/translated).

### 3.1 Overview

Cahill (1980) noted the strategic location of Somalia as one of the underlying reasons for the involvement of various international actors:

Somalia is much in the news as East and West contend for the future of the Indian Ocean. The strategic significance of Somalia's vast coastline, from which the oil shipping lanes at the entrance to the Red Sea and the southern end of the Suez Canal can be blocked, has long attracted the superpowers' attention. (p. 46)

Somali-speaking people, although separated by geographical borders, largely inhabit the Horn of Africa. As Cahill (1980) noted, the 400,000 square miles in the northeastern Horn of Africa is what is now called the Somali Republic. Somali speakers also live in Kenya, Ethiopia, and in French Somaliland (currently Djibouti). In the early 1960s, which incidentally was when most African nations regained their independence from European colonial powers, Segal (1962) predicted:

Of all the problems created or left unsolved by colonial rule in Africa, none is as potentially dangerous as that of Somali division in the east of the continent. From the Horn of Africa itself to the west stretch the Somali people, most of them today in the Somali Republic, some of them French Somaliland, some in Kenya's Northern District, some in the eastern regions of Ethiopia, and nearly all of them passionately loyal to a united Somali nation. (p. 154)

After 38 years, linguists Nettle and Romaine (2000) described what Segal had predicted and its effect on language. Except for the Somali Republic, which now has its own problems and

cannot create a sustainable environment for a language to grow, in the other parts of East Africa where Somalis live, the Somali language has no official status.

The boundaries of modern nation-states have been arbitrarily drawn, with many of them created by the political and economic interests of Western colonial powers. Many indigenous people today, such as the Welsh, Hawaiians, find themselves living in nations they have no say in creating and are controlled by groups who do not represent their interests. (Nettle & Romaine, 2000, p. 12)

Such is the situation for the Somali people—they live in a nation they did not create and their language has no official status.

### **3.2 Context of the Problem**

There has been no central government in Somalia since 1991. Arguably, Somalia has endured one of the bloodiest civil wars in modern history. The situation remains chaotic today and, as a result, the exact number of Somali speakers who have lost their lives during the last two decades is unknown. Justifiably, different sources give different figures with regard to the human cost of the war. For example, a report by Global Security (2001) stated, “Since 1991, an estimated 350,000 to 1,000,000 Somalis have died because of the conflict” (para. 1). Human Rights Watch (2010) gave a bigger picture of the situation:

Some 3.75 million people—roughly half of Somalia’s remaining population—are in urgent need of humanitarian assistance. More than a million people are displaced from their homes within Somalia and tens of thousands fled the country as refugees in 2009. (para. 2)

In this context, Nettle and Romaine (2000) suggested that, for a language to endure, it needs a population who actively use it and pass it on to their children. The authors described the environment in which such a population or community can thrive by stating:

A community of people can exist only where there is a viable environment for them to live in, and a means of making a living. Where communities cannot thrive, their languages are in danger. When languages lose their speakers they die. (Nettle & Romaine, 2000, p. 5)

The turmoil and chaos in Somalia wrought by civil war creates an untenable environment for a people and a language to thrive.

Related to the foregoing argument, linguists have identified two types of language death: sudden and gradual. The first type is termed *sudden death*, which is characterized by a loss of its speakers by means of war or a natural disaster. *Gradual death* means that a language gradually loses its speakers as the language ceases to be used for all linguistic functions (Nettle & Romaine, 2000). Given the precarious situation in which the Somali language finds itself, it is evident that Somali faces both types of language death. It faces sudden death by means of population loss and because of the instability and the hostility caused by the civil war. Somali speakers are driven out of the environment in which their language has flourished for centuries. Evans (2010) cited the Ubykh language of the northwestern Caucasus to illustrate how military intervention and sustained warfare could drive the population to other places where they most likely end up switching to the language of the host country. As Somalis become more widely dispersed through flight and immigration, the environments to sustain the Somali language become smaller.

### 3.3 Overview of the Somali Language

As Lewis (1961) pointed out, the Somali language is a member of the Cushitic languages related to Oromo and Afar spoken in Ethiopia and Djibouti. These languages are descendants of the Afro-Asiatic family of languages such as Arabic and Hebrew.

As mentioned earlier, Somali speakers live in separate nation-states in the Horn of Africa. This means the vitality of the language depends on its status in all areas where Somali is spoken. The US library of Congress (1992) stated that the 1991 estimate of the Somali population in the Somali Republic was approximately 7.7 million. The Daily Nation (2010), as cited by Hiiraan (2010), Kenya's national newspaper, reports the most recent census in which ethnic Somalis in Kenya are reported as 2.3 million. In Ethiopia, a census conducted in 1997 (Ethiopian Government, 2012) stated that the population in the Somali Regional State is 3,439,860 and that 95.6% speak Somali.

A more comprehensive statistic is given by Lewis, Simons, and Fennig (2014) who reported that Somali is spoken in four countries in Africa. In Somalia, there are 6,460,000 speakers while in the Somali Region of Ethiopia there are 4,610,000 speakers. In Kenya's northeastern province 2,386,000 people speak Somali while in Djibouti there are 297,000 Somali speakers. Adding these numbers, the total number of Somali speakers in all four countries stands at 13,753,000. Despite the total figure, Lewis et al. (2014) stated that the total speakers number 14,679,000. The discrepancy between the two totals is 926,000. However, outside of the Horn of Africa, Sheikh and Healy (2009) estimated the Somali-speaking population in the Diaspora to be one million. These Somali speakers are spread around the world including Canada, the US, the UK, Australia, and the Middle East. The one million Somali speakers estimated to live in the Diaspora might explain the disparity found in the statistics. Nonetheless, since the most of the

figures presented are estimates, it is reasonable to claim that between 14 and 15 million people in the world speak Somali as their mother tongue. So, given these figures, why does this researcher claim that the Somali language is endangered? In subsequent sections of this chapter, the life cycle of the Somali languages as well as information on the Somali Language Vitality Assessment are presented in order to shed light on the status of Somali and its speakers.

### **3.4 Lifecycle of the Somali Language**

The Somali language has been through many stages. For the purpose of this review, however, the two most significant phases will briefly be explored. In choosing the following two stages, the modes of language use were considered and, in the Somali language context, the modes were either oral or in written form. Three main phases will fall within the oral mode of the language while the two other phases will be within the written mode of the Somali language. Andrzejewski (1985) coined three terms, the Golden Era, the Era of Fire and Embers, and the Era of the Lute, which he used to describe different stages of Somali literature. He used another term, the New Era, to describe the period when Somalia had military government, between 1969 and 1991. Another important phase of the Somali language and literature is missing from the list, which is the period after the collapse of the Somali central government. For the purposes of this thesis, this period will be termed *The Cane and Cone Era*. The rationale for choosing this phrase will be described. The diagram below displays the stages of the Somali language (Figure 3.1), which will be discussed further.

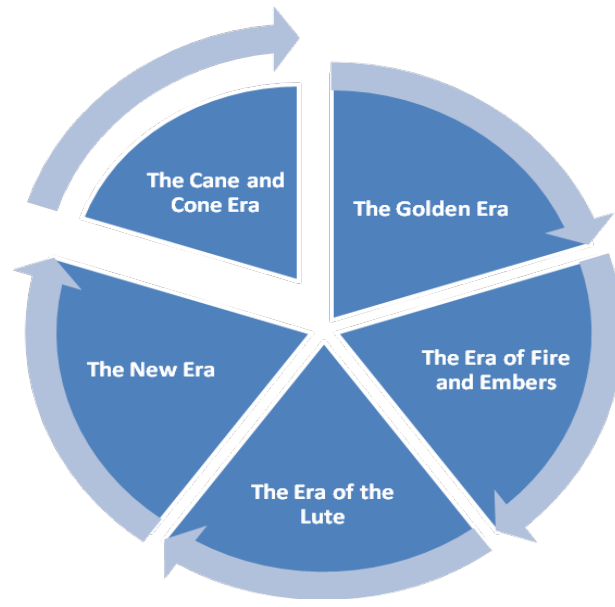


Figure 3.1. Stages of the Somali Language From Oral to Written Language

1. *The Golden Era (the pre-colonial period)*. According to Andrzejewski (1985), this is the period when the Somali language had no contact with the outside world. He noted that literature from this period is largely unrecorded. It is not clear why Andrzejewski called this period the *Golden Era*; it appears that he chose the term for two reasons. First, the literature from this period was largely about simple yet profound aspects of life such as love, friendship, existence, and other philosophical issues. Second, he wanted to capture the fact that both the Somali culture and its language were intact without any significant external influence.
2. *The Era of Fire and Embers (1988-1944)*. Somali literature during this period was dominated by war and conflict, namely the resistance and uprising of the Somalis against the colonial powers. Andrzejewski (1985) asserted that during this period Somali literature flourished and the dominant theme was war. Leaders used oral poetry during

this period to mobilize people. Leaders had a poet who served as what is now known as the “communications director for the president” and perhaps this was the main reason that allowed Somali literature, primarily poetry, to flourish.

3. *The Era of the Lute (1944-1969)*. This is the period when segments of the Somali nomads started to move from the countryside to urban areas. During this time, as reflected in the literature, a new perspective on life had emerged. During the Era of the Lute, poets started to compose poetry about the new addition to the Somali culture. Andrzejewski (1985) illustrated the changes to Somali literature. He recounted that a Somali truck driver came up with a new form of poetry, which the driver-poet called it *belwo*, which in English means “vice.” Andrzejewski named this period the Era of the Lute, as this was a major addition to Somali poetry.

In it he combines the theme of love treated in a somewhat hedonistic manner with the elaborative imagery characteristic in the past of the more serious classical poetry and he introduces new lively tunes to which the poems were sung accompanied by tambourine and flute. (p. 358)

4. *The New Era (1969-1991)*. This is the period when the most effective government (in terms of the contribution it made to the development of the Somali language and literature) was in power. At this time, the writing system of the Somali language was introduced and implemented. The language was standardized and it was made the official language of the Somali Republic; the medium of education was switched from English, Italian, and Arabic to Somali. Textbooks in Somali were made ready for schools to use in the 1980s. The medium of instruction in some of the faculties in the Somali National University was made Somali.



5. *The Cane and Cone Era (1991 onwards)*. This is the period when everything that was accomplished prior to 1991 was practically reversed. The first part of the term metaphorically represents the Somali language, which at this stage of its life has had serious injuries and is unable to stand or walk without a cane. The word “cane,” according to Hashi and Hashi (2004), has three equivalents in Somali, *ul*, *matoobbo*, and *bakoorad*, while the word “cone” has no Somali equivalent. This indicates that the word “cane” has a place in Somali culture as evidenced by the variety of terms given, while the lack of Somali translation for the word “cone” suggests that the concept is foreign to Somali culture. The term “cone” symbolizes the external forces and pressures on the language.

The lifecycles of the Somali language described in this section demonstrate how the Somali language has moved from a primarily oral mode in the pre-colonial period through three phases to the written mode of the Somali language in two phases. The new era signifies the birth of Somali orthography and the remarkable growth that the language has experienced during this period. The cane and cone stage signifies a period of noticeable decline due to a myriad of language-endangerment factors. The following section assesses the vitality of the Somali language by using UNESCO’s (1993) language-vitality factors and reviewing relevant literature.

### **3.5 Somali Language Vitality Assessment**

UNESCO (2003) stated, “A language is *endangered* when it is on a path toward extinction” (p. 4). What is challenging is to determine at which point the language is endangered. In tackling the issue of assessing language vitality and endangerment, UNESCO (2003) enlisted the help of a group of linguists who developed the nine factors to consider when assessing language vitality (see Figure 3.2).

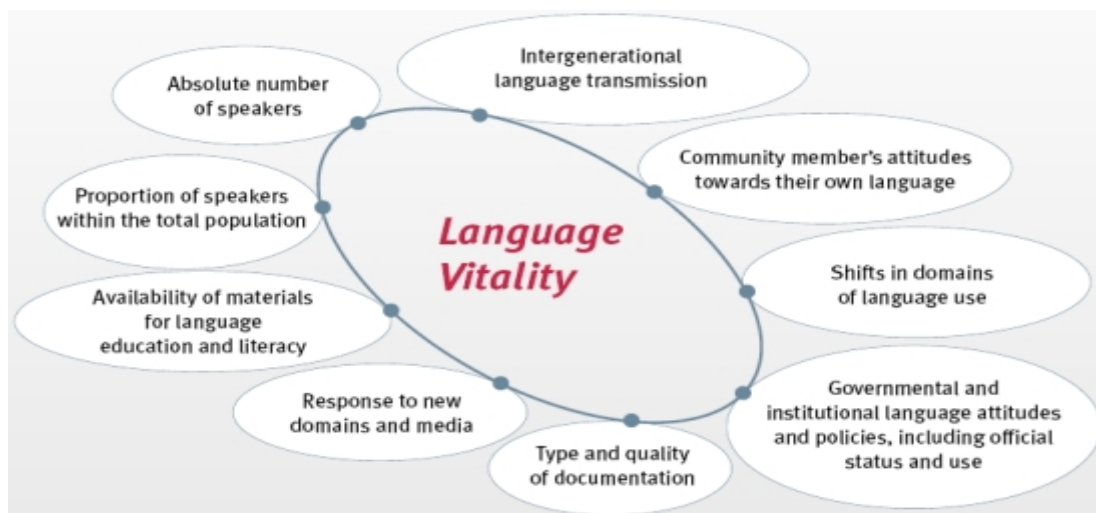


Figure 3.2. Language Vitality Factors Adopted from UNESCO (2003)

Evans (2010) maintained that although the symptoms used by linguists to evaluate threatened languages are salient, it is nonetheless difficult to assess language endangerment accurately. First, given the inexact nature of all predictions, unforeseen factors may alter one's predictions. Second, predictions of language endangerment are often questionable. Furthermore, as Evan (2010) argued, "The second cause of uncertainty in our projections is that there are some regions—particularly in Africa—where the effective collapse of nation states and research infrastructure makes it almost impossible to gather data" (p. 213). With regard to Africa, Nettle and Romaine (2000) wrote, "Kenya is the only country for which reliable estimates exist on this topic" (p. 5). Considering the fact that Somalia has had no central government for decades, it is reasonable to place Somalia at the bottom of the list of African nations in terms of availability of reliable data for language assessment.

Another limitation in the area of language vitality assessment has to do with the methodology used in applying the nine criteria for language vitality assessment. UNESCO (2003) cautioned, "No single factor alone can be used to assess a language's vitality or its need

for documentation” (p. 9). Yet another factor that needs to be considered is the fact that the present-day Somali language has two main dialects: the older oral Somali language and the modern Somali language, which has a written form as well as an oral form that is quite different from the older oral Somali. The former version is the older version from the period before 1972 when the language had no orthography. The latter modern Somali language is the version that developed after the introduction of the writing system. Therefore, there are two varieties of the Somali language and, evidently, they are not equally endangered.

The fact that Somali speakers live in different parts of the world has to be taken into account when assessing vitality, as the vitality of the language in different geographical areas will vary. UNESCO (2003) noted that the nine factors as proposed are to be used as guidelines and therefore users are expected to modify them as the context of the study dictates. Since speech communities know more about the status of their language, UNESCO (2003) encouraged the communities to take the lead in assessing their language: “A speech community may examine these factors first to assess their language situation and to determine whether action is needed, and if so, what to do next” (p. 19). As a Somali speaker, I find this statement adds credibility to the assessment task at hand in this study.

With these limitations in mind, an attempt will be made to use the limited resources available to assess the vitality of the Somali language. This assessment will spark a discussion on how war and instability alters the status of a language in two decades. In what follows, UNESCO’s (2003) nine factors will be applied and a level of endangerment will be assigned to the Somali language. Slight modifications will be made where applicable and when they are made they will be stated, along with the rationale for making such changes.

### **Factor 1: Intergenerational language transmission.**

*Older Oral Somali:* Severely endangered (2); the language is *spoken* only by *grandparents and older generations*; while the parent generation may still *understand* the language, they typically do not speak it to their children.

*Modern Somali:* Unsafe (4); most but not all children or families of a particular community speak the language as their first language, but it may be restricted to certain social domains (such as at home where children interact with their parents and grandparents).

Note that the language transmission depends on the country and context in which speakers live. For example, as I write this chapter, I reflect on the Somali students here at University of Calgary and whether they use the Somali language. There are approximately 40 students of Somali descent at the university that I know of, and these students either prefer not to speak Somali or they are unable to. In the Somali Republic, younger generations speak Somali at home, but they study in either Arabic or English at school. So, the country and context for speaking Somali matters in terms of daily opportunities to use the language.

### **Factor 2: Absolute number of speakers.**

Given the fact that the number of Somali speakers is relatively high, one might argue that the language is not in any way endangered. According to UNESCO (2003), “Languages cannot be assessed simply by adding the numbers” (p. 19). In the case of the Somali language, the effects of war in terms of population loss and ongoing waves of emigration show that the number of Somali speakers cannot be taken for granted. Consider the history of Yiddish and its road to endangerment. A language profile written by the UCLA Department of Germanic Languages indicates that Yiddish speakers were approximately 11 million before the Holocaust. Yiddish was widely used in the media with around 60 newspapers and between 300-400 daily periodicals

published in 30 countries. Presently, the UN considers Yiddish an endangered language with only one million speakers alive (UCLA, 2009, para. 2). Yiddish faced political instability and consequently its speakers were threatened to the extent that most of them fled from where they lived while a great number lost their lives. Nettle and Romaine (2000) suggested using additional factors such as these in assessing the vitality of a language: “A large language could be endangered if the external pressures on it were great, while a very small language could be perfectly safe as long as the community was functional and the environment stable” (p. 41). The Somali language faces war, political instability, and as a consequence, an unprecedented number of its native speakers have either perished or fled from their home country; even worse, the trend may continue if the current predicament persists.

**Factor 3: Proportion of speakers in the total population.**

*Older Oral Somali:* Severely endangered (2); a minority speak the language.

*Modern Somali:* Definitely endangered (3); a majority speak the language.

Regarding the older oral Somali, which is the main language used to compose poetry and other Somali literature, only those who deliberately learn Somali literature and some of the older generations speak that variety; most Somalis would not understand older oral Somali. To many Somalis, the vocabulary and the structure of Somali literature might sound peculiar or foreign to them and so it is not uncommon to hear a Somali speaker who asks for a poem to be explained in a simpler form of Somali or for younger generations to ask for an English translation. Recently, in the process of writing this chapter, while conversing on the phone, a friend referred to lines of a Somali poem to illustrate a point, and for a moment, I thought that he had switched to speaking a foreign language. These anecdotal observations point to the decline of Somali literature being used in social domains. With respect to the other variety of the language, the modern Somali,

given the number of Somalis who live outside the Somali Republic, it is reasonable to estimate that a majority of Somalis speak it with various degrees of facility. The guidelines classify this diversity and range of Somali language use and fluency as definitely endangered. The next grade up, according to factor 3, is *unsafe* (4), which is normally given when some children use the language in all domains and all children use it in limited domains. The Somali language may be classified somewhere between these two.

**Factor 4: Trends in existing language domains.**

*Older Oral Somali:* Limited or formal domains (2); the language is used in limited social domains and for several functions.

*Modern Somali:* Multilingual parity (4); two or more languages may be used in most social domains and for most functions.

The word “language” is used, but in reality it is used to mean a variety of Somali that was predominantly spoken at a particular period. The older oral variety is currently limited to social functions such as weddings, social gatherings, and in Somali literary circles. Given that almost half of Somalis live outside the Somali Republic, it is normal to find Somali speakers who use the languages of their host country with varying degrees of fluency. For example, it is typical to find a Somali speaking Swahili, English, and Somali in Kenya; French, Somali and Arabic in Djibouti; and Amharic, Somali and Oromo in Ethiopia. Therefore, the assessment and the grade given are reasonable. The extent of language use is vital as Grenoble and Whaley (2006) noted, “If a language is used in increasingly fewer domains, it is a sign of a lessening vitality” (p. 9).

**Factor 5: Response to new domains and media.**

*Older Oral Somali:* Minimal (1); the language is used in only a few new domains.

*Modern Somali:* Coping (2); the language is used in some new domains.

The older oral variety is not used as widely as the modern variety in the new technological domains such as the Internet and social media networks. The modern variety is necessarily being used in the emerging technologies, although English is still dominant.

**Factor 6: Materials for language education and literacy.**

*Older Oral Somali:* Grade (2); written materials exist, but they may be useful for only some members of the community; for others, they may have a symbolic significance. Literacy education in the language is not part of the school curriculum.

*Modern Somali:* Grade (2); written materials exist, but they may be useful for only some members of the community; for others, they may have symbolic significance. Literacy education in the language is not part of the school curriculum.

Many materials in the old and modern Somali language have been destroyed and the remaining materials have not been reprinted, as schools largely switched to materials written in other languages. UNESCO (2003) stated, “Education in the language is essential for language vitality” (p. 14). The importance of Somali materials for language education is emphasized by Grenoble and Whaley (2006), “A critical domain for language use is education” (p. 10).

**Factor 7: Governmental and institutional language attitudes and policies, including official status and use.**

*Older Oral Somali:* Practically no official status in many parts of speech communities; in other parts of speech communities dominant languages penetrate public domain, business and education.

*Modern Somali:* (same as above) Practically no official status in many parts of speech communities; in other parts of speech communities dominant languages penetrate public domain, business and education.

**Factor 8: Community members' attitudes toward their own language.**

*Older Oral Somali:* Grade 4; most members support language maintenance.

*Modern Somali:* (same as above) Grade 4; most members support language maintenance.

Although there are Somali speakers who prefer to speak languages other than Somali, it is safe to predict that most of them would support language maintenance. Nevertheless, most Somali parents would prefer to send their children to English/Arabic or French schools for practical and economic reasons. Parents naturally want their children to succeed and they think it is beneficial for their children to be educated in one of the languages with a higher status and one with greater economic opportunities.

**Factor 9: Amount and quality of documentation.**

*Older Oral Somali:* inadequate (1); only a few anthologies, short word lists, and fragmentary texts exist. Audio and video recordings do exist, but they are mainly in the forms of outdated audio-visual formats, or are completely un-annotated.

*Modern Somali:* fair (3); there may be an adequate grammar or sufficient number of grammars, dictionaries, and texts but no everyday media; audio and video recordings may exist in varying quality or degree of annotation.

With respect to the older oral variety, a great amount of Somali literature is alive in the memory of Somali elders. Unfortunately, the warring factions have destroyed a huge number of literary books and cassette tapes in Mogadishu. In a personal communication with one of Somalia's greatest poets, who was a member of the Somali Orthography Committee, he mentioned that around 400 rare literary books were destroyed in the Somali National Museum in Mogadishu in the early 1990s. Although there are books and other forms of Somali literature, these are still a fraction of what needs to be documented. As Grenoble and Whaley (2006) stated,



“In most basic terms, though, the fewer written materials, the less they are taught, and the higher the level of endangerment” (p. 11).

Considering the vitality of the Somali language and all nine factors, the older oral Somali language is critically endangered. This assessment of the vitality of the older oral Somali is because, as the forgoing discussion shows, Somali children are not learning this variety of the language; it is largely in the memory of elders. Modern Somali is moderately endangered or unsafe to say the least. More in-depth research is needed to refine this preliminary study. The Somali language lost close to a million of its speakers in the last two decades according to Global Security (2001). Mass immigration is another factor, as the number of Somalis in the Diaspora now stands at 1 million. Children are not being taught in their mother tongue in schools inside the Somali Republic or out of Somalia. Based on these factors, it is reasonable to label the Somali language an endangered language. It is also safe to predict that if the current political instability and social upheaval in Somalia continue, the language may find itself among the 3,000 world languages that linguists predict will die out in the course of the next century. This is a particular case, as the Somali language has witnessed one of the bloodiest and longest wars in human history, which has transformed its speakers economically, politically, socially, and technologically.

### **3.6 What is at Stake?**

In his book, *Light at the Edge of the World*, Davis (2001) introduced a new term *ethnosphere* to capture the essence and the severity of the cost to humanity if a language dies out. He gives the following definition of *ethnosphere*:

[It is] the sum total of all thoughts, dreams, ideas, beliefs, myths, intuitions and aspirations brought into being by the human imagination since the dawn of

consciousness. It is a sample of all that we've accomplished and all that we can accomplish . . . the *ethnosphere* is humanity's great legacy. (p. 26)

Consider Davis's *ethnosphere* as a living museum of the world where each language deposits its perspectives on the world. In the case of the Somali language, it might offer a scientific contribution in the areas of medicine, marine science, archaeology, and linguistics. Cahill (1980) observed, "The landscape of Somalia ranges from lush river valleys to scrub savannah and ancient towns" (p. 26). Somalia has the second longest coastline in Africa. Note that the area Cahill describes does not include the other parts of Horn of Africa where Somali is spoken, namely Kenya, Ethiopia, and Djibouti. A wealth of knowledge about such a diversified environment and its ecosystems is encoded in the Somali language.

Of all the knowledge that the Somali language could contribute, knowledge about camels is the most intriguing and offers promising breakthroughs in the field of medicine. As some authors (FAO, 1988) wrote:

"Somalia has the largest camel (*Camelus dromedarius*) population in the world with more than six million animals" (as cited in Munz, Moallin, Mahnel, & Reimann, 1989, p. 191). The number is close to the population of the Somali Republic. This fact is reflected in the culture through Somali language and literature; to learn about Somali culture, it is necessary to learn about camels. Xiques (2005) noted that Margaret Laurence was confronted with this challenge. "She wanted to understand, for example, the context in which Somalis made so many references to camels in their poems" (p. 171).

Somali nomads traditionally used camel milk to treat a number of medical conditions. Recently, a team of medical professionals in the Department of Family Medicine at Ben-Gurion University in Israel have conducted research in which they studied eight children with severe

food allergies. The parents of the children, some from the United States and some from Israel, had tried all avenues of modern medicine and wanted to try camel milk. The children were put on a strict camel-milk diet and the result was that all children have fully recovered from their food allergies (Shabo, Barzel, Margoulis, & Yagil, 2005).

In another clinical study, researchers found that camel milk has helped in the treatment of Type One diabetes. Researchers hypothesized that the effects were due to the fact that camel milk contains an insulin-like substance or protein (Agrawal, Dogra, Mohta, Tiwari, Singhal, & Sultania, 2009). Anecdotal stories note that Somalis in the Diaspora with various medical conditions including high blood pressure, high cholesterol levels or even diabetes, who went back to Somalia for treatment with camel milk and camel meat have had positive results. The traditional medicinal use and knowledge of camel products substantiated by research indicate that the knowledge about camels encoded in the Somali language may be of help to researchers because of its vital data collected over many centuries.

Another prominent feature of the Somali people is their creativity with words through poetry, proverbs, and other forms of Somali literature. Somali oral literature made a very special friend back in the 1950s. Margaret Laurence, mentioned above, who later gained iconic status in Canadian literary circles, started her career translating Somali poetry and folktales, which she published in a book entitled, *A Tree of Poverty: Somali Poetry and Prose* (1954). A number of Canadian writers acknowledge that Margaret Laurence's innate talent for writing was somehow cultivated and inspired by Somali poetry. Stovel (2008) asserted, "I argue that translating Somali oral folk literature, poems and stories influenced Laurence's Canadian fiction significantly" (p. 88). Xiques (2005) equated Margaret Laurence's prose to poetry and described her writing as

poetic. “One could select almost any passage from her novels and render it as poetry simply by dividing it into lines” (p. 186).

Laurence herself reflects on her own experience with Somalis and their language. She elaborated on the rationale of Somalis to value poetry over other forms of art. Somalis are a “nation of poets” (Laurence, 1956), echoing what other scholars have previously noted. She mentioned that since Somalis were nomads, poetry is portable and requires no tools other than one’s mind and memory. Laurence observed, “Somalis are poor in possessions but rich in literary culture” (p. 24).

Somali poetry has also inspired and influenced the young Somali-Canadian artist K’naan, whose song, *Wavin’ Flag*, lifted him to international stardom when it was chosen as the theme song for Coca Cola’s 2010 FIFA World Cup ad campaign. He composed rap songs by combining Somali concepts of poetic endowment with Western urban life. K’naan’s method of bringing us all closer through English, art, and blended culture is attractive because we have a lot to learn from each other. Drawing from multiple cultural perspectives is richer than drawing from one. The goal is to expand one’s repertoire through listening to and learning from other cultures. Teal (2010) cited K’naan: “According to K’Naan, in Somalia, everything revolves around [poetry]. Conflict resolution is written in poetry . . . our laws are. Everything about Somali people, the only way we know how to communicate is poetry” (para. 2).

Teal (2010) quoted Said Samatar, a professor at Rutgers University and an expert on Somali literature, who stated in an interview with Teal:

[T]he war has destroyed Somalis’ most precious asset, its poetry. One of the greatest tragedies of the current conflict is that Somalis’ ability to produce poetry has been greatly diminished. Traditionally, there were huge poetic contests throughout the country. If we

can restore those contests, that will go a long way towards resolving the conflict. (para. 13)

As with any other language, the Somali language has a certain perspective on the world and as the examples given here show, it could inspire others to broaden their perspective.

## **Chapter Four: Window on the World of Corpus Linguistics**

### **4.1 Overview**

According to Crystal (1992), corpus linguistics is an interdisciplinary field that integrates linguistics and computer science. Corpus linguistics is both an approach and a practice that uses corpus or computer-readable data in exploring language phenomena. McEnery and Harddie (2011) stated, “It is certainly quite distinct from most other topics you might study in linguistics, as it is not directly about the study of any particular aspect of language” (p. 1). Similarly, Dash (2005) described corpus linguistics as “a multidimensional area. It is an area with a wide spectrum for encompassing all diversities of language use in all domains of linguistic interaction, communication, and comprehension” (p. 1).

This means that, by using corpus data, a researcher can work on research questions on semantics, morphology, sociolinguistics, syntax, lexicography, computer-assisted language learning, and so on. However, researchers in other fields often rely on using manual or small-scale data collection and analysis software so the question remains as to why a corpus, or large sets of computer-readable data, is needed for the collection and analysis of linguistic data. If manual or small-scale data collection is sufficient for other researchers, why use corpora for language research? The answer to this question lies in part in the nature and the characteristics of language. The study of a language is a complex task as humans engage in infinite instances of language production. McEnery and Harddie (2011) explained:

It is the large scale of data used that explains the use of machine-readable text. Unless we use a computer to read, search, and manipulate the data, working with extremely large data sets is not feasible because of the timeframe it would take a human analyst to search through the text. (p. 2)

The need for large data sets and the challenge of working with large data sets is how such a mutually interdependent relationship between computer science and linguistics started. Both disciplines needed a solution to an identified problem. The next section will examine the stages of corpus linguistics development prior to the advent of computers and how the field has evolved.

## **4.2 The Evolution of Corpus**

Bennett (2010) stated that although corpus linguistics has existed for over a century, what is currently known as the modern-day corpus started with the advent of computers about a half a century ago. In the early stages of corpus-linguistics development, researchers tended to conduct their linguistic research using manual techniques.

According to McEnery and Wilson (2001), corpus linguistics started in the late nineteenth century as the basis of linguistic studies that included first-language acquisition, spelling conventions, and second-language teaching. For example, notable works in the area of first-language acquisition included data that was manually collected between 1876-1926 to document language development utterances of children, by their parents. The data was later used and is still being used by researchers to study language acquisition. The late German linguist, J. Kading carried out another landmark study. In 1897, he manually collected about 11 million words to learn more about the spelling conventions in German by looking at frequency lists of the words he collected. Given the fact that computers did not aid Kading's study, the collection of 11 million words was impressive indeed. In the area of foreign language study, Kennedy (1992) argued, "The corpus and second language pedagogy had a strong link in the early half of the twentieth century, with vocabulary lists for foreign learners often being derived from corpora" (as cited in McEnery & Wilson, 2001, p. 4).

Despite the remarkable work done by researchers without computer assistance, manually collecting and analysing huge linguistic datasets has inevitably discouraged researchers from employing corpus linguistics as a methodology, thus slowing the growth of corpus linguistics as a field. With the advent of personal computers, the field of corpus linguistics took a major step forward. Svartvik (1992) noted:

While the manual excerpting of textual data has been the regular means of gathering information for linguistic description, its modern form, which only recently has come to be known by the name of corpus linguistics—the use of large collections of text available in machine-readable form—only dates back to the early 1960s. (p. 9)

Furthermore, like any other form of scientific enquiry, corpus linguistics has been the subject of controversial debate between empiricists and rationalists as to what constitutes data in corpus linguistics. MacEnery and Wilson (2001) reported, “Chomsky suggested that the corpus could never be a useful tool for the linguist, as the linguist must seek to model language competence rather than performance” (p. 6). To appreciate the essence of the debate, the distinction between performance and competence can be seen as one that stems from the ongoing debate between empiricists and rationalists. Drawing from the rationalist school of thought, Chomsky asserted that linguistic performance, which is the underlying knowledge of a language, may or may not always mirror what the speaker actually knows. Linguistic competence, on the other hand, is the data that linguists ought to work with, for it is the only model that represents the speaker’s true knowledge of the language, argued MacEnery (2006). Favouring performance over competence, Chomsky (1965) explained:

Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows his language perfectly and is unaffected by



such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. (p. 3)

Chomsky (1962) made another criticism that the natural language data collection methodology employed by corpus linguists does not correlate with the kind of linguistic data that a native speaker would gather. He elaborated on this issue:

Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. (p. 159)

The point that Chomsky made is that corpus data will not fully reflect the language it purports to study since language is infinite by nature. Despite the criticisms levelled at the developing field of corpus linguistics in its early stages, corpus linguistics as a discipline has flourished in the last 50 years or so. The growth in the field has been partially due to the works of many linguists who refined the role of corpus in linguistic research. It also appears that Chomsky contributed to the development of the field of corpus linguistics by providing sensible and thought-provoking criticism, which might have inspired linguists to revisit and refine their methodologies. Although Bennett (2010) did not respond directly to Chomsky, she discussed what corpus linguistics is *not* meant to do, which responds to Chomsky's reservations about the role and nature of data in corpus linguistics:

1. The corpus is not meant to give prescriptive rules or predict what is possible or not possible in a language. This means that the data is meant to describe or show what is there or not there in the corpus.

2. The corpus is not meant to give answers as to why a certain linguistic phenomenon exists. Its function is merely to show what is. Researchers and language users are supposed to use other linguistic theories and/or their intuition to explain the linguistic instances shown in the corpus.
3. The corpus is not meant to cover all language instances. Rather, a corpus is meant to be representative of a language through a carefully planned sampling strategy that aims to cover a wide range of the language it purports to represent.

Accordingly, the two schools of thought (empiricists and rationalists) in the debate over the respective merits of competence and performance over collecting data for corpus construction have come together. It should be two approaches that complement each other. For example, Wasow (2002, as cited in McEnery & Hardie, 2011) explained:

While data from corpora and other naturalistic sources are different in kind from the results of controlled experiments (including introspective judgment data), they can be extremely useful. It is true that they contain performance errors, but there is no direct access to competence; hence, any source of data for theoretical linguistics may contain performance errors. And given the abundance of usage data at hand, plus the increasingly sophisticated search tools available there is no good excuse for failing to test theoretical work against corpora. (p. 163)

With the advent of personal computers and the refinement of methodology, corpus linguistics has become a popular and useful methodology among linguists and other scholars who work with linguistic data for different ends. The first machine-readable corpus, created by Francis and Kučera (1979) at Brown University, marked the beginning of what is known today as the field of modern corpus linguistics. The Brown University Corpus contained one million

American English words. As Johansson and Stenstrom (1991) noted, the field has grown steadily since that time. For example, there were 10 corpus-based studies from 1961-1965 while from 1976-1980 there were 80 corpus-based studies. The real growth in corpus linguistics research started after 1985 when researchers of languages other than English started building their own corpora.

### **4.3 What is a Corpus?**

The term “corpus” (plural “corpora”) is known to many in the fields of language and technology integration, particularly in the area of corpus linguistics or computational linguistics. The term might be a little confusing to non-specialists because it has another common meaning in everyday English. The Merriam-Webster Dictionary defined corpus as “The body of a human or animal especially when dead” (cf. corpse). In the world of corpus linguistics, however, the term has a specialized meaning. As will be shown in subsequent chapters of this dissertation, the linguistic definition similarly includes the word “body” or a related term. In the context of linguistics though, the word “corpus” is often associated with the word “language” as in the collocation, “a body of language,” which could be the origin of the confusion.

In the research literature, various authors have defined the word “corpus” with regard to one or more of the characteristics of a corpus. For example, Bernard (2000) defined a corpus as a machine-readable database in the form of a large searchable collection of language, which occurs in a natural context in a variety of text types. McEnery and Wilson (1996) defined corpus as “a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration” (p. 24). Sinclair (2004), a leading scholar in the field of corpus linguistics, suggested a more detailed definition: “A collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a

language or language variety as a source of data for linguistic research” (p. 19). Leech (1991) described the size and type of data: “A sufficiently large body of naturally occurring data of the language to be investigated” (p. 9). Teubert and Cermakova (2007) offered a definition that adds yet another characteristic of a corpus: “A collection of naturally occurring language texts in electronic form, often compiled according to specific design criteria and typically containing millions of words” (p. 140). Dash (2005) gave a more apt description of a corpus:

As a language, corpus is (C)apable (O)f (R)epresenting (P)otentially (U)nlimited (S)elections of texts; it may be defined acrostically from the letters used for construction of the term in the following way: “Representative of a language or variety, Processable by man and machine, Unlimited in amount of data, and Systematic in formation and representation” (p. 4).

Based on the preceding definitions, the characteristics of a corpus can be summarized as follows:

1. *Natural Language*: A corpus is a naturally occurring body of language generally in the form of texts. Natural language means that the data should be drawn from language used in natural contexts. The term “naturally occurring language,” which is considered an important criterion, refers to spontaneously spoken or recorded data and written texts produced for a real-life communicative purpose and not for the purpose of artificially producing data for corpus inclusion (Gries, 2009). Therefore, newspaper articles, TV shows, and blog posts all qualify for inclusion.
2. *Follows sampling criteria*: The data is not collected randomly but follows a selection criteria defined before the data collection starts.

3. *Preferably large*: A corpus should ideally be large, preferably in the millions, but may be as small as 10,000 words. The decision as to the size of a corpus is dictated by the purpose of the corpus construction.
4. *Representative*: A corpus should be representative of the language under study, containing a wide variety of texts from diversified domains of language as used by its speakers.
5. *Machine-readable texts*: It should be organized in a manner that is suitable for computers to read and process. By definition, spoken data should be turned into text so that the data fits the machine-readability criterion.
6. *For language research*: A corpus is used for linguistic research and for identifying and addressing language problems.

A more detailed treatment of the six characteristics of a corpus will be given in Chapter Five in which the design methodology of a corpus is discussed. The chapter will draw from the experiences of major previous corpora and discuss how researchers have made design decisions.

#### **4.4 Practical Uses of Corpora**

Just before the turn of the 21st century, Kennedy (1998) aptly observed, “With a corpus stored in a computer, it is easy to find, sort, and count items, either as a basis for linguistic description or for addressing language-related issues and problems” (p. 11). This has evidently been the case for certain world languages because the application of corpus use resulted in language research with implications for the production of language tools that ease the use of those languages. The authors who spearheaded a project in the UK dubbed “Enabling Minority Languages Engineering—EMILLE” argued:

Corpus building in Europe has traditionally focused on languages that are indigenous to European countries: English, French, Spanish, Italian, etc. Users of such languages benefit from an extensive range of computational resources: fonts, word processors, spell-checkers, online dictionaries, thesauri, and automatic translation utilities. (McEnery, Baker, Gaizauskas, & Cunningham, 2000, p. 1)

The main goals of the project were to inspire research and development initiatives by providing the data or the foundation needed to plan language engineering primarily for South Asian languages.

Despite the initiatives such as the foregoing project and others, which have resulted in the creation of a number of corpora for certain world languages, it is unfortunate that many world languages have not had the opportunity to develop their own corpora, including the vast majority of African languages. Maxwell and Hughes estimated that that only 20-30 of the world's 6,900 languages have corpora (as cited in Abney & Bird, 2010, p. 88).

As the world moves increasingly toward the use of digital modes of communication, the non-represented or underrepresented languages need to catch up by creating their own corpora or they will lag behind and face further risks of becoming endangered. In this context, Ostler (1999) predicted, "Languages which do not take a full part in the electronic media are doomed to stagnate, if not atrophy" (p. 3). It is evident that there is a looming danger faced by many world languages that lack the digital resources needed to communicate in the digital world. This study aims to bridge that digital gap among world languages. Through the construction of a model corpus, the study hopes to inspire similar research and development work for many languages that lack the infrastructure to conduct research and support language engineers in developing digital tools to keep these languages current and digitally communicable.

Dash (2005) discussed the rationale or the motivation behind the use of corpora:

The basic philosophy behind corpus linguistics has two wings: (a) we have a cognitive drive to know how people use language in their daily communication activities, and (b) if it is possible to build up intelligent systems that can efficiently interact with human beings. With this motivation, both computer scientists and linguists have come together to develop language corpus that can be used for designing intelligent systems (e.g., machine translation systems, language processing systems, speech understanding systems, computer-aided instruction systems, etc.) for the benefit of human knowledge and application (p. 1).

The foregoing discussion is in alignment with the research purpose of this study. Both research and development work on a language could facilitate communication in that language which helps restore the confidence and morale of its speakers. Similarly, the field of corpus linguistics is viewed as a bridge connecting linguistics to wider scholarly initiatives. The purpose is to develop intelligent systems in order to enhance and add flexibility to the ways we learn, teach and shop, among other things, through the use of sophisticated smart tools.

#### **4.5 Corpus for Research and Development**

Zampolli (1995) outlined major areas that use corpora for research purposes: language learning and teaching, computer-aided instruction, and lexicography, to name a few. As noted earlier, corpus linguistics is used for virtually all aspects of language-related research, so the areas listed above are not exhaustive. Below is a snapshot of the main areas of use:

1. *Lexicography and Materials Development*: Dash (2010) described the use of corpora as a “knowledge source,” meaning that corpus data can be used to develop various learning/teaching materials. The area of lexicography has seen tangible results from the

use of corpora. Summers (1995) noted how research based on the British National Corpus has helped the compilation of some of the most widely used English dictionaries, notably Longman and Oxford. Likewise, Sinclair (1996) noted the role that the Bank of English played in the publication of Cobuild dictionaries. As Rundell (2002) reported, Macmillan dictionaries, which are also very popular, used corpus data from the World English Corpus. Macmillan used corpus research findings to define words and to give them in example sentences drawn from real contexts. Given the effect that corpora have had on English language development, it becomes clear that the use of corpora by other languages could offer similar results. Tapping into the data of language in use, corpora have also been used to write language teaching materials and reference materials.

2. *Natural Language Processing Tools (NLP)*: It is a field active in the research and development of technological tools to facilitate computer-human interaction. Liddy (2001) acknowledged that there is no agreed definition of NLP but she defined it as:

Natural Language Processing is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. (p. 1)

3. Since the aim of NLP is to achieve “human-like language processing,” it is considered a sub-field in the wider domain of artificial intelligence (Liddy, 2001). The Natural Language Processing Group at Microsoft stated their goal to “design and build software that will analyse, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing



another person” (Microsoft Research (2014, para 1). Applications developed using NLP techniques include:

- *Word-processing Tools:* These tools include spell-checkers, grammar and style checkers, text editing, text summarization, etc.
  - *Translation Tools:* Corpus is used for the development of various translation support systems including machine-translation devices, multilingual information access systems, cross-language retrieval systems and other translation support systems (Dash, 2010).
  - *Human-Machine Interaction:* A corpus is used for the production of higher-level human speech processing including speech recognition, automated customer service assistance, and speaker identification.
4. *Linguistic Research:* According to Dash (2010), corpora are used for all sorts of language study including semantic study, language learning, pragmatics, and dialectology, the study of different dialects in a language, sociolinguistics and the generation of terminology databanks, among others.

#### **4.6 Prospective Corpus Users**

Given the various practical uses of corpora, it is anticipated that both specialists and non-specialists would be interested in using corpora for different ends. Non-specialists are considered to be people who speak the language contained in the corpus and who might be interested in exploring certain aspects of their language or might want to use the corpus as learning or a reference tool. As for the specialists, Dash (2010) described the main users of corpora:

1. *Language specialists*: The main users of a corpus are linguists. People such as lexicographers, terminology developers, and other applied linguists use them for different but complementary purposes.
2. *Content specialists*: Archaeologists, historians and sociologists use corpora to trace how knowledge on certain aspects of life has evolved, and the date and place in which something originated.
3. *Media specialists*: Dash (2010) explained that media specialists are those “who are interested in the corpus as a test bed for electronic devices” (p. 20). These are technologists who are interested in the development of various media tools such as intelligent personal assistants, translation platforms, etc. Two examples of these tools would be the Siri application of the Apple and Google translation platform.

In summary, corpus linguistics is a multidisciplinary field that brings together various disciplines. It is well established in the literature, as demonstrated in this chapter, that corpus linguistics is a field that attracts researchers and practitioners from a wide range of disciplines notably from computer science, linguistics, and sociology, among others. Of particular interest to this study is how corpora could be used in teaching and learning as a form of educational technology. Human beings teach through language and learn through language. Therefore, education and language are inseparable. Technology is a tool that can enhance the quality of education. In discussing the role of technology around the issues of language and knowledge dissemination, Eisele and Ziegler-Eisele (2002) stated:

Natural language is the prime vehicle in which information is encoded, by which it is accessed and through which it is disseminated. With the explosion in the quantity of online text and multimedia information in recent years there is a pressing demand for

technologies that facilitate the access to and exploitation of the knowledge contained in these documents. (p. 21)

This is precisely where educational technology research such as this study might bridge a gap in the literature on the pressing issue of documenting, preserving and studying endangered languages through the use of educational technology. Since knowledge is inherently encoded in languages, this is an issue that brings together many intersecting disciplines, notably computer science, corpus linguistics and educational technology.

## Chapter Five: Corpus Design

### 5.1 Corpora Classifications

Depending on the intended use as well as the content of the corpus being designed, corpus linguists have created many types of corpora. Before we delve into the issue of corpus design, brief descriptions of the most common types will be helpful in putting the discussion into perspective. The most common type of corpus is a “general corpus,” which is sometimes called a reference corpus. The purpose of a general corpus is to represent a language by including a variety of texts, both spoken and written. The British National Corpus and the Brown Corpus of Standard American English are notable examples of general corpora that represent British and American English respectively. A corpus can also be classified by looking at the languages it contains, which means that a corpus can be monolingual, bilingual or even multilingual (Kennedy, 1998). This means that the language dimension is used to describe a corpus because it could contain one or more languages: monolingual versus bilingual or multilingual.

Sinclair (2004) examined the size factor when he compared a sample corpus to a monitor corpus. A sample corpus contains fixed or limited samples of data whereas a monitor corpus is developed and updated as the language grows. These are the most widely discussed types of corpora in the literature.

At times, the naming of corpora is inter-related. As the names suggest, *fixed or sample corpus* means a corpus that contains language samples taken at a certain time. This type of corpus is referred to as a *reference corpus*, especially when it is sampled carefully to represent the language under study. It can also be called a *general corpus* (Tognini-Bolleli, 2010).

In contrast, a *monitor corpus* is updated regularly in order to track changes in a language (Sinclair, 2005; Teubert & Cermakova, 2007). The British National Corpus (BNC) is an example

of a sample or general corpus, while the Corpus of Contemporary American English is regarded as a monitor corpus. A sample corpus or static corpus becomes a historical or a diachronic corpus with age; for example, the Brown Corpus, which was created in the 1960s, is now considered to be dated and the BNC will in the near future be considered a *historical or diachronic corpus*. Regarding a monitor corpus, Dash (2005) noted, “[It] is an ever-growing collection of text samples. In fact, it enjoys the scope of regular augmentation to project the ongoing changes occurring within a language over time” (p. 14). A good example of a monitor corpus is the Bank of English and the Corpus of Contemporary American English, which are updated and expanded regularly. A brief description of the Bank of English and other major corpora will be given later in this chapter.

According to Dash (2010), the size and the content of a *special corpus* depend on the purpose, which is often to study a particular variety of a language. It is not representative of the entire language so it should be used only for the purpose for which it was created. A special corpus is different from a general corpus, which is probably the most commonly used corpora, as it purports to represent a language. As its name suggests, it contains general texts from various text types and registers. According to Kennedy (1998) another important distinction in naming a corpus is whether the corpus is annotated or not. This means that a corpus can remain raw with no additional information; alternatively, a corpus can be annotated. The name depends on the intended use of the corpus.

Dash (2010) offered a more comprehensive classification that is conceptually easier to comprehend. It is based on data, design, and application attributes: genre of text, nature of text, type of text, purpose of text, and nature of application. The following corpora types are adapted from Dash’s typology and limited to the relevant types for this study.

***Genre of text.***

- *Written Corpus*: The corpus contains data gathered from various written sources either in print or electronic format. The sources are normally gathered from a variety of written texts whether they are hand-written, printed or published; more recently, written samples are collected electronically. Brown Corpus of American Written English is a good example of this type.
- *Spoken Corpus*: This corpus contains transcribed texts of spoken data. It is sometimes called a *speech corpus* and ideally includes both formal and informal naturally occurring language that ranges from conversations to dialogues, sermons, telephone calls, and public speech. Although it is much more difficult to gather speech samples, it is nonetheless considered superior data, as it is the most commonly used medium of communication showing how language is used naturally. Dash (2005) noted the importance of a speech corpus:

The application value of a speech corpus is widely acknowledged in linguistic studies, since it is accepted that informal and impromptu speech is the most important variety, which is the closest possible representation of the core elements of a language found in its most natural environment. (p. 8)

The genre of a text would help a corpus designer to conceptualize data sources by looking at the main data sources: spoken or written. Furthermore, depending on the purpose of the corpus construction, a corpus contains one or both of the genres, as we will see below. London-Lund Corpus of Spoken English is an example of a speech corpus while Brown Corpus of American English is an example of a written corpus.

### *Type of text*

The language dimension is the defining factor and the main classifications under this criterion are:

- *Monolingual Corpus*: The corpus contains a collection of texts representing how a particular language is spoken and/or written by its speakers. Based on the uses desired by the corpus users, the data may be limited to a period of time or it may not. The British National Corpus is an example of a monolingual corpus. As we will see later in this chapter, the former type is often called a *diachronic or historic corpus* while the latter is termed a *synchronic corpus*.
- *Bilingual Corpus*: It contains texts from two languages, which may be arranged in different ways depending on the intended use of the corpus. As discussed below, the two languages may somehow be aligned, showing words, phrases, sentences and their translations. Conversely, a bilingual corpus may contain data from two languages, which are basically translations of each other, but the two languages are aligned.
- *Aligned Corpus (parallel corpus)*: This type of corpus is a translation corpus and is often used for cross-linguistic research as well as developing bilingual dictionaries and translations tools. Dash (2010) defined this type as “a kind of bilingual corpus where texts in one language and their translations into the other language are aligned sentence-by-sentence, phrase-by-phrase or even word-by-word” (p. 5). Dash gave the Canadian Hansard Corpus as an example of an aligned corpus. This type of corpus is often called a *parallel corpus* or *translation corpus*.

## Purpose of Design

In this category, Dash (2010) listed an annotated versus un-annotated corpus. Essentially this means that additional information can be added to the data or the data can remain raw. Dash (2005) distinguished the two corpora:

An un-annotated corpus represents the plain text while an annotated corpus is encoded with varieties of intra-linguistic and extra-linguistic information. Usually, an un-annotated corpus is found in a simple raw state, because it aims to represent the actual state of the language without addition of any linguistic or non-linguistic information, thus an un-annotated corpus has been and is of considerable use in all kinds of linguistic studies. (p. 12)

Dash (2005) went on to state that, depending on the intended use of the corpus, annotation usually adds extra value to the corpus. *Extra-linguistic information* refers to such information as the genre the text belongs to, the heading, demographic information about the writer and the like. In contrast, *intra-linguistic annotation* refers to information that linguistically describes the text such as part-of-speech tagging, semantic information, morphological information and other relevant linguistic information.

## Nature of Application

The intended application of the corpus is the prime factor in defining the following types of corpora. In this group, a *reference corpus* is compared to an *opportunistic corpus*.

- *Diachronic Corpus*: This type of corpus is also referred as to *historical corpus* and it contains language samples that span certain periods. It is used to trace and compare changes in a given language.



- *Synchronic Corpus*: Often contrasted with *diachronic corpus*, this corpus aims to show how a language is used at a given period without looking at how the language has evolved historically.
- *Reference Corpus*: It aims to provide extensive data about a given language. It is designed to cover a wide variety of registers so that it represents the language as a whole. Ideally, a reference corpus is a *general corpus*, which includes both genres, written and spoken, and contains a variety of genres including language used in both formal and informal settings. As discussed earlier, a general corpus that is designed to represent a language can in fact be called a *reference corpus* and this type of corpus is used to develop standard grammar books and monolingual dictionaries as well as other reference materials and tools. As we will see later in this chapter, the issue of representativeness and whether it is achievable or even necessary is not yet resolved (Biber, 1998; Sinclair, 2005; Tognini-Bolelli, 2010).
- *Comparable Corpus*: Dash (2010) defined this type (e.g., Corpus of European Union) as collection of similar texts in more than one language or variety. It contains texts in different languages where texts are not the same in content, genre, or register. These are used for comparison of different languages. It follows some composition pattern but there is no agreement on the nature of similarity. (p. 6)

This type of corpus is often used for cross-linguistic studies as well as generating bilingual and multilingual resources for languages.

- *Opportunistic Corpus*: “An opportunistic corpus stands for an inexpensive collection of electronic texts that can be obtained, converted, and used free or at a very modest price; but it is often unfinished and incomplete” (Dash, 2010, p. 6).

Dash (2010) stated that some see a *monitor corpus* as an *opportunistic corpus*. However, Sinclair (1991) argued that a monitor corpus is not synonymous with an opportunistic corpus if it takes into account the issues of representativeness and balance in its design. A monitor corpus often attempts to sample a language as used in predefined additions of various texts, which remain constant every year. This is because the corpus, as its name suggests, aims to monitor changes in the language at any given time. As we will see later in the chapter, a monitor corpus is often used for lexicography.

## 5.2 Design Methodology for Corpus Construction

According to Atkins, Clear, and Ostler (1992), “a corpus is a text assembled according to explicit design criteria for a specific purpose, and therefore the rich variety of corpora reflects the diversity of their designers’ objectives” (p. 13).

This means that corpus design decisions should be based on the intended use of the corpus. For example, designing a corpus (especially one that can be used as a reference) should have wide coverage of linguistic data while attempting to balance the different sections of the language data being collected (Kennedy, 1998). Alternatively, Biber (1993) argued that a corpus that intends to gather a specialized area of a language could be represented using samples from that variety alone. Therefore, achieving representativeness is very important for corpora that claim to be representative of a given language.

In this study, the focus is on the latter type of representativeness. First, it is a desirable feature for general corpora designs. Second, the focus of this thesis is general language studies, which aims to shed light on ways to represent languages through corpus construction. Key factors determine representativeness including balance, variety, and size. These will be the recurring themes dominating the following discussion on design issues.

Biber (1995) defined a *balanced corpus* as “one that includes proportions of a range of different text types of a language as they are reflected in the language studied” (p. 4). A balanced corpus is often compared to an opportunistic corpus. In terms of the latter, Leech (2002) stated that an opportunistic corpus is generally not regarded as good practice because the data collection is conducted opportunistically by collecting texts that are easier to gather. Texts are collected based on availability and the issue of including representative registers in a balanced manner is not accounted for.

Teubert and Cermakova (2004) pointed out, however, that an opportunistic corpus “is based on the assumption that each and every corpus is imbalanced” (p. 67). The argument made here is that, given the impossibility of creating a balanced corpus, an opportunistic corpus is inevitable and serves as an alternative design model especially when creating a large corpus. Clearly, there is an ongoing, unresolved debate among linguists on such issues as balance, representativeness and the importance of size when developing corpora.

The second major issue is whether the corpus has diverse registers. According to Biber (1995), diversity is vital since a corpus that contains a large amount of data from a single register cannot capture the variety and the diversity of the way a language is used in different contexts and registers. Biber (1995) argued that unless size and diversity are taken into consideration, the corpus content could not be used to generalize the language it purports to represent.

In the early years of computerized corpus design, diversity and size were considered equally important. According to Francis and Kučera (1979), during the initial stages of designing one of the first general corpora, the Brown Corpus, the range of texts to be included in the corpus and the ratios of each text were considered to be the main design decisions. The Brown Corpus aimed to represent varieties and registers of written American English. At the outset, the decision

was that the Brown Corpus would contain one million words. Once the size of the corpus was decided, a diversified sampling strategy was adopted to collect 500 different texts and from each text 2,000 words were taken randomly. The designers were trying to include both diversity and balance. This meant that diversity was achieved by sampling from 500 different types while actual samples from each text were all equal at 2,000 words each. Aston and Burnard (1998) captured the essence of how the BNC tackled the issue of balance in their design framework by looking at the importance of each text as exemplified by the productive and receptive factors of the text:

Account was taken of both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have wide distribution. Thus, having chosen to sample such things as popular novels, or technical writing, best-seller lists and library circulation were consulted to select particular examples of them. (p. 28)

According to Leech (2002), while diversity is an attempt to get the samples from diverse settings, genres, and topics, the concept of balance can be relatively difficult to grasp. It is almost impossible to measure and evaluate balance. For example, there are those who argue for proportional sampling of texts. Gledhil (2000) explained that texts are considered balanced when they are sampled unequally according to their respective use in the language. “Genres are by their very nature unequal and it is perhaps unreasonable to describe the whole language on the basis of equally represented text-types” (p. 81). This approach advocates for text or genre selection based on its relative use in the language. Biber (1993) stated that if proportional sampling methodology is employed, then daily conversations, for instance, should account for the greatest percentage of all the texts in the corpus.

However, two main weaknesses have been identified with a proportional sampling strategy. First, it tends to produce a corpus with a greater level of homogeneity, which in turn reduces text variety. This occurs because few texts are given sufficient status to dominate the composition of the corpus. The other weakness has to do with the fact that determining or quantifying the relative importance of a text is almost impossible. Atkins et al. (1993) were convinced that since language is infinite, it is reasonable to quantify the amount of language that is produced or heard at any given time. Library catalogues, bestseller lists, and other statistics give us only a rough estimate of texts but it cannot give us the exact readership of a particular text. A record of a book indicating that it was borrowed or purchased does not necessarily guarantee readership.

The question becomes: How does one measure the relative importance of a text? Atkins et al. (1992) discussed this issue by using production vis-à-vis reception factors of a text as a method, which has been used to approximate the relative importance of a text and estimate the proportion of its use in the language:

The corpus builder has to remain aware of the reception and production aspects.

Although texts, which have a wide reception, are by definition easier to come by, if the corpus is to be a true reflection of native speaker usage, then every effort must be made to include as much production material as possible. For a large proportion of the language community, writing is a rare language activity. Judged on either of these scales, private conversation merits inclusion as a significant component of a representative general corpus. Judged in terms of production, personal and business correspondence form a valuable contribution to the corpus. (p. 9)

Atkins et al. (1992) discussed the nature of using production and reception as a method of evaluating the amount and the text to be sampled. Essentially, what determines the relative value of a text is its production rate. The reception factor comes into play when the designer looks at production in terms of reception (how many people read or hear the produced language). This means that the selection of a text is based on language production. Kennedy (1998) defined *reception* as the language that users read and hear, while *production* is the language that users write and speak.

Evidently, using these parameters can only serve as a guide because evaluations based on these two criteria are largely subjective, based on assumptions and estimates. Kennedy (1998) clarified the issue: “No one knows what proportion of the words produced in a language on any given day are spoken or written” (p. 63). Leech (2007) echoed this stance noting, “The difficulties of determining the size of the textual universe and its sub-universes from which a corpus is to be sampled is formidable” (p. 139). While some argue for *proportional sampling*, Biber (1993) favoured *stratified sampling*. This method surveys language registers in existence on a list and then identifies texts under each register. The list is used to sample each text, which ensures the inclusion of various texts. Proportional sampling, in contrast, has the tendency to leave out a variety of texts. To achieve variability, Leech (1997) stressed that having various registers should be considered as important as the size of the corpus or maybe even more important.

Stratified sampling involves creating levels or strata in which all known language registers are organized and samples are gathered from each stratum. However, the most important and first step before sampling is to define the scope of the data or as Biber (1993) described it, “definition of the target population.” In doing so, the scope of the language under

study should be clearly defined by deciding a list of the known text categories followed by a definition of detailed text types that are considered for inclusion or exclusion. This is often called *a sampling frame*, which forms the basis for developing data-selection criteria. The definition of the registers in a language is based on situational or external factors such as the setting, the purpose of the text, the channel or the medium of the text, topics, etc. Another sampling technique is *random sampling*. In this approach, numbers are assigned to the selected texts using a well-defined sampling frame; random numbers are then selected, which means that all texts have an unbiased probability of being selected (Biber, 1993; Sinclair, 2005).

It seems that the impracticality of the design issues in corpus construction (namely representativeness, size, and balance) are widely acknowledged. Therefore, one might ask if these issues are not practically feasible, why bother with corpus design? Sinclair (2005) proposed a compromise, which seems to have settled the debate: “The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components” (para, 11). A word of caution here derives from the experience of the Bank of English, which has 2.5 billion words. It has written prolifically in the area of corpus linguistics and has played a pioneering role in the construction of one of the largest corpora in our times. The caveat here is that while the concepts of representativeness and balance are desirable, they are nonetheless elusive in the sense that they are not practically feasible. Atkins, Clear, and Ostler (1992) acknowledged this problem. They stated, “Because of the sheer size of the population and given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample” (p. 9). The word

population refers to the language being studied; therefore, the point is that language is inherently infinite, so there can never be a true representation of a language in its entirety.

Kennedy (1998) agreed with Atkins, Clear and Ostler's (1992) position in that "it is not easy to be confident that a sample of texts can be thoroughly representative of all possible genres or even of a particular genre or subject field or topic" (p. 62). Kennedy (1998) pointed out a plausible argument in the sense that while it may be possible for other fields to gather a representative sample of a given population, languages by their very nature are infinite so it is impossible to include all the possible spoken or written instances of a language through sampling.

Despite these challenges, the issue of balance and representativeness is at the core of corpus design. Biber (1995) clarified the importance of size and diversity in corpus design. He viewed them as the two main factors affecting the representativeness of a corpus. Biber asserted:

In the social sciences generally, issues of representativeness are dealt with under the rubric of external validity, which refers to the extent to which it is possible to generalize from a sample to a large target population. There are two kinds of error that can threaten external validity: random error and bias error. Random error occurs when the sample is not large enough to accurately estimate the true population; bias error occurs when the selection of a sample is systematically different from the target population. Random error can be minimized by increasing the sample size, and this is the reason that large text corpora are important. By contrast, bias error cannot be reduced by increasing the sample size because it reflects systematic restrictions in selection. (p. 130)

Biber (1995) stressed that bias error is a major flaw in research since a biased sample or a sample taken from one source threatens the external validity of the research, casting doubts as to



whether generalizations can be made about the language being studied. Therefore, for a representative corpus, both size and diversity should be considered.

Sinclair (1991) advocated for a larger corpus size:

In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are the clues to the lexical structure of the language. (p. 171)

It is understandable that Sinclair favoured a larger corpus size since he spearheaded the creation one of the largest monitor corpora, The Bank of English, mainly for lexicographic purposes. Sinclair (1991) provided the rationale for building large corpora. They are updated and expanded regularly, and thus are suitable for lexicography because they keep track of how languages evolve and grow.

Where does one start? Atkins et al. (1992) proposed an iterative design model:

Since it is theoretically suspect to aim at achieving a perfectly balanced corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weakness identified and reported. In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually. (p. 7)

It has been widely acknowledged that there is a need for a balanced and representative corpus, yet notable linguists (Atkins et al., 1992; Biber, 1993; Sinclair, 2005) advised a pragmatic approach to design.

Biber (1993) summed up the design process: “The actual construction of a corpus would then proceed in cycles: the original design based on theoretical and pilot study analysis, followed by collection of texts, followed by further empirical investigations of linguistic variation and revision of the design” (p. 243). Furthermore, given the impossibility of attaining true representativeness in corpus construction, Sinclair (2005) proposed six more elaborate corpus design criteria to be followed if the designer wishes to claim the unattainable but desired goal of representativeness.

1. Draw up situational criteria, which the designer uses to base the sampling framework for the data collection, outlining the main categories for the corpus construction;
2. Then draw up a list of all the possible text types in each category;
3. Prioritize the list in each category based on importance of language use;
4. Decide the approximate sample size of each text based on total sample size of each category considering the number of text types on each list vis-à-vis its importance and availability;
5. During the construction stage, monitor how the actual data matches or fails to match the original design framework; and
6. Document all the design decisions as future reference for development or usage purposes.

These six steps seem to capture most of the key themes discussed in this chapter. It should be noted that the iterative design model has been proposed by a number of researchers. For example, Hunston (2002) pointed out the rationale for adopting an iterative and ongoing design model, which is particularly applicable to languages: “Any corpus that is not regularly updated rapidly becomes unrepresentative” (p. 30). Hunston (2002) considered that languages do in fact grow and change; consequently, corpus design should reflect that reality.

### 5.3 Overview of Previous Design Applications

While the theoretical debate on challenges faced by corpus developers continues, scholars were developing many different corpora by trying to find solutions to some of the unresolved design issues. Using the experiences of four major existing corpora, these issues will be discussed briefly to gather insight into how practitioners in the field dealt with these issues.

The Brown Corpus, the British National Corpus, and the Corpus of Contemporary American English will be surveyed in greater detail while the Bank of English will be briefly sketched. This approach is taken because the first three corpora will yield a panoramic view of corpora types offering a variety of design ideas. The Bank of English is considered one of the most successful corpora ever constructed, but much of the design ideas it offers are also found in the design model of the Corpus of Contemporary of American English since they are both monitor corpora.

Another large monitor corpus is the Oxford English Corpus, which monitors real-time change in the English language as it occurs. As I write this chapter, the news media reports the corpus has just released “selfie” as the winner of the 2013 Word of the Year, based on the fact that the word’s usage increased exponentially (by 17,000%) in 2013 (CNN, 2013). Selfie is a newly coined term and means a photo taken of oneself using a Smartphone or similar device. The term has not yet entered standard dictionaries except for the online version of Oxford Dictionaries (BBC, 2013; CNN, 2013). Hui (2013) wrote in the Chronicle Herald:

Researchers behind the renowned dictionaries pick a prominent word or expression in the English language each year that best reflects the mood of the times. Previous words of the year have included “unfriend” in 2009, “credit crunch” in 2008, “carbon footprint” in 2007 and “Sudoku” in 2005” ((Oxford crowns, November 19, para 3).

Monitor corpora became popular with the availability of a vast number of online texts, while pioneering computer-based corpora like Brown Corpus adopted a more static or sample approach.

### **Brown Corpus**

The Brown Corpus of American English is considered a pioneering work in the area of computer-based corpora. Hardy (2007) noted, “The Brown Corpus, originally made available in 1964, contains one million words of written American English in excerpts of around two thousand words. Although not the first corpus, it was the first computerized corpus” (p. 53). Although it was developed in the early 1960s, given its size and design at that time, it is still considered a standard, balanced corpus especially when designing a sample written corpus intended to represent a written language at a given period. In this sense, many people still consider the Brown Corpus as a relatively good example of a representative corpus, which, to a certain degree, succeeded in representing how written American English was used in the 1960s (Biber, 1993; Kennedy, 1998). Its size, which was and still is one million words, was considered large, given the technological and other limitations of that time.

Two major categories or genres were first identified—imaginative and informative prose. These categories were then analysed into sixteen subgenres from which 500 samples were drawn. A decision was made to include 2,000 words from each sample in the corpus, not the entire text (Francis & Kučera, 1979).

### **British National Corpus (1980s-1993)**

The construction of the corpus started in 1991 and the project was finished in 1994 covering texts from 1980s to 1993. It contains 100 million words of which 90% are from written sources and 10% are from spoken sources. It was a joint-venture project in which major UK-

based publishing houses joined forces with Oxford University, Lancaster University, and the British Library. The project cost was approximately 1.2 million UK pounds. UK government agencies gave full funding to the British Library and the two universities, while the private publishers received partial funding, matching the proportion of the total cost they incurred. The BNC was used by other research and development initiatives, including the compilation of the Longman Dictionary of Contemporary English.

In the 1990s, it was considered to be the largest general corpus. Even though the BNC raised the standards in terms of size, as we will see below, there are now corpora that are in the billions. The BNC was seen as a well-designed reference corpus, first, because it attempted to represent both spoken and written language, although the spoken component was only 10% of the corpus. Second, it followed carefully designed criteria, which tried to represent language variety by including various registers.

The design of the spoken part of the corpus was based on two criteria. First, since there were no guiding resources such as library records, books in print, or bestseller lists on which production and reception language were to be based, the BNC design team devised alternative methods. The first was to sample speakers according to pre-determined demographic information such as age, sex, region, and social class. The target was to gather conversational English. In all, 124 volunteers aged 15 years and above from the four demographic categories were supplied with portable recorders to audiotape all kinds of conversations as they went about their daily activities. The audio files, which were about 700 hours long, were then converted to digital texts for inclusion (Burnard, 1995).

The second method involved the context or the setting in which language was produced. This criterion was used to capture a variety of language types that were not covered by using

demographic information alone. These language types included court proceedings, lectures, TV shows and other speech texts, which are characterized by few producers and many receivers. Four main contextually defined categories were drawn: (a) education and informative talk, (b) business, (c) institutional, and (d) entertainment talk. Each main category was then divided into their speech type: either monologue or dialogue. The texts were sampled from three geographically distinct areas in the UK. Under each contextual category a list of text types was written. The text types were identified intuitively and further texts were added as deemed appropriate during the planning stage. For example, education texts such as lectures, classroom discussions, and student conversations were some of the possible texts identified.

The guiding methodology was to gather balanced samples from each category, keeping the desired percentages in check. The sampling methods were suitable for each category respectively. For example, while educational talks were sampled based on topic, educational level, speaker and age, and region, business meetings were selected based on business type, company size, and the nature of the meeting (Burnard, 1995).

### **Bank of English and Cobuild Corpus**

Cobuild is an acronym that stands for Collins Birmingham University International Language Database. As its name suggests, it is a partnership project, which brought UK-based CollinsHarper Publishers and the University of Birmingham together to build such a large corpus. The corpus is currently considered to be the largest corpus ever created, with 2.5 billion words. It was used to compile the popular series of Cobuild dictionaries (Sinclair, 1991).

## **Oxford English Corpus**

As of spring 2010, the corpus contained over 2 billion words, which makes it the second largest corpora of its kind after the Bank of English. The official website of Oxford Dictionaries claims:

The Oxford English Corpus gives us the most accurate picture of the language today. It represents all types of English, from literary novels and specialist journals to everyday newspapers and magazines and from Hansard to the language of blogs, emails, and Internet message boards. (Oxford Dictionaries, 2013a, para. 5)

The purpose of the corpus is to track language evolution; it is considered a major component of a 35 million pound project at Oxford. As for the contents of the corpus, it gathers its content mainly from the Internet with the exception of some printed materials used to supplement certain genres such as academic journals. Oxford justifies the use of the Internet as follows:

The extensive use of web pages enables us to build a corpus of unprecedented scale and variety. The Oxford English is intended to be as wide-ranging as possible in its representation of the English language. Development was planned to ensure a balanced range of material from different subject areas, regions of the world, and type of writing.

Structuring a corpus this way produces a panoramic view of language use in every area of human life. (Oxford Dictionaries, 2013b, para. 2)

It contains 20 main registers including academic papers, technical manuals, newspapers, novels and short stories, underground and counterculture websites, personal websites, blogs, and message board postings. The rationale is that texts such as academic journals give a picture of how standard language is used while texts such as blogs track how nonstandard language

evolves, including newly coined terms. While the corpus contains 80% British and American English, other varieties are included from Canada, the Caribbean, and Australia to South Africa, India, and Singapore. Texts are gathered using a customized web crawler. Oxford described its job description as follows:

A configuration file is used to direct the crawler to a particular website or an area of a website and to define the behaviour of the crawler within that site: the navigation route it should follow, and the type of pages it should collect along that route. (Oxford Dictionaries, 2013c, para. 1)

Oxford uses customized software for analysis. It seems that this corpus has exploited technology and money to surmount some of the major challenges faced by many corpus builders.

### **Corpus of Contemporary American English (COCA)**

Corpus of Contemporary American English (COCA) is one of the largest corpora, standing at 450 million words. In contrast to BNC and the Brown Corpus, which are sample corpora, COCA is a monitor corpus, expanded and updated annually. Davies (2009) pointed out that although BNC was used extensively for language-based studies in the 1990s, certain limitations have emerged, mainly associated with time. First, the corpus has not been updated in terms of size and there is no plan to expand it, which is not expected because it was not designed to be a monitor corpus. Second, even though the BNC was considered a very large corpus in the 90s, with current technology, the potential to create relatively larger corpora is within our means.

Davies (2009) stated that COCA was envisioned to address such limitations, which are found in sample corpora such as the BNC. Acknowledging and attempting to address its limitations, Davies (2009) noted, “The corpus was designed to be roughly comparable to the BNC in terms of text types” (p. 161). In terms of its content, texts in COCA were selected evenly



among main genres with each genre representing 20% of the corpus composition. The representing genres include: spoken, fiction, popular magazines, newspapers, and academic journals. Each major genre is categorized into subgenres. For example, Davies (2008) stated the spoken genre is composed of “transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), News Hour (PBS), Good Morning America (ABC), Today Show (NBC), 60 Minutes (CBS), Hannity and Colmes (Fox), Jerry Springer, Oprah, etc.) (p. 161).

The newspapers were divided into 10 popular newspapers such as USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle and the like. Furthermore, an attempt is being made to represent different sections of each newspaper sampling from sections such as the local news, opinion page, sport commentaries and the business section (Davies, 2009). The corpus stays balanced because it is replenished yearly with approximately 20 million words, which are divided equally among the five main genres. In terms of diversity, Davies (2009) claimed that COCA is the first large and diverse corpus freely available.

Copyright is considered one of the main limitations when building language corpora. We’ve seen that some of the large corpora designers have settled on using limited samples to avoid infringing the intellectual property of others. COCA (2014) argued that the use of copyrighted materials in COCA is believed to be within the parameters of US Fair Use statute as determined by the following four conditions. According to COCA (2014), the project has consulted intellectual property lawyers who ascertained the legality of the following four criteria.

1. Only limited portions of text are accessible to the public rather than the full text of the material. Users can get access only to words and sentences used in context.
2. Texts are used for educational research and not for business purposes.

3. While 80% of the corpus contains texts that are not considered creative work, such as TV shows, newspaper articles, and academic journals, there are also creative works such as novels and short stories whose proportion of the corpus is still within the fair-use law.
4. There is no profit or loss because the access offered to the public does not in any way compete with other media where the material can be accessed in part or in its entirety.

Even though these arguments are understandable, copyright issues can easily get very complex. For example, what counts as creative work? Similarly, since the exact amount of fair use has not yet been quantified, copyright owners can at any point disagree with what constitutes fair use. It is therefore advisable to be cautious when using copyrighted material.

#### **5.4 Deductions for Corpus Design in Challenging Contexts**

As Atkins et al. (1992) pointed out, the BNC model was a standard model for many in the 1990s although the context in which the corpus was constructed and the challenges faced at that time have changed. The crux of the argument is about the importance of assessing the context in which a particular model was developed, especially if the model is to be replicated in a different environment. Such assessment could lead to awareness of lessons that may or may not be applicable to contexts in which a researcher wishes to replicate or apply the lessons. The main issues of corpus design have been discussed in this chapter along with examples of major existing corpora. As scholars have acknowledged, the challenge of achieving representativeness and balance in corpus design has been widely discussed in the literature (Biber, 1993; Atkins et al.; Sinclair, 1995). As these challenges have been universal, in this study they will be termed *context-independent challenges* because these issues have to be faced regardless of the context of the corpus design.

In contrast, there are *context-dependent challenges*, which are considered much more serious in this study than the universal challenges faced by all corpus designers. The *context-independent challenges* have been debated widely for nearly a century and consequently a theoretical and practical knowledge base, which designers can learn from, has been developed. The context-dependent challenges are found in relatively new frontiers, and thus pose real challenges and limitations to researchers who work with less commonly studied languages, endangered languages, and threatened languages, which are found in resource-constrained environments. Such languages are the focus of this study and they have their particular challenges.

The following three areas, gleaned from the chapter, exemplify the main areas in which great variability is expected in terms of awareness and the motivation level of corpus construction participants, the resources available, and the relative ease by which data can be obtained. Theoretically, awareness and motivation are conceptualized as a prerequisite for the initiation of any corpus construction project, while availability of resources and data enhance the success and sustainability of the project.

1. *Awareness and Motivation*: Any successful project starts with defining the need and importance of such a project, and corpus construction is no exception. Since it is a relatively new field, awareness is even more important. Consider the variety and the sheer number of government institutions, universities, libraries, publishers, academics, students, and volunteers who all joined forces in building the British National Corpus. I argue that with a certain level of awareness of the need and the importance of a British National Corpus, the initiators of the project were able to organize the resources and the human capital they needed to carry out such a challenging task. For example, the

publishers who later used the corpus for dictionary compilation were motivated by the commercial viability of the need for dictionaries. Other participants may have had different motivating reasons, but it is evident that all participants were cognizant of the need and had the motivation and energy to get involved. It will be a challenge to reach a level of awareness of language development and revitalization projects such as corpus building for languages that have less commercial potential.

2. *Resources*: Awareness and motivation can start a project but the project needs resources of various types to move it forward and eventually finish it. During the planning stages, the BNC and Brown corpus had records and statistics about books in print, bestseller lists, and library-borrowing records, all of which aided designers to gauge the range of books, academic journals, and newspapers that were available as well as the popularity of certain genres of texts. Another major contributing factor to the success of corpora construction is the availability of funds. The UK government fully funded the cost incurred by the two universities and the British Library while the private publishers received partial funding. Availability of relevant resources is crucial for the success of corpus construction projects.

3. *Data Availability*: Atkins et al. (1992) stated that collecting data is one of the areas in which cost is inevitable, especially if a conversion from print- to machine-readable is one of the options. It is often cheaper to collect electronic data than convert print to electronic format. It is even more expensive and time-consuming when transcription of audio files is involved. Stable and healthier languages have numerous text types that are used in many contexts. This is not the case with languages that have limited use of domains.

In conclusion, corpus design is a complex undertaking. Endangered and threatened languages have particular challenges in addition to the universal or context-independent issues of representativeness and balance, with which all corpus designers have to deal. In the following chapter, the issue of corpus design in challenging contexts will be discussed using the Somali language as an example.

## Chapter Six: Corpus Design In Challenging Contexts

### 6.1 Conceptual Framework

Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem, which it was intended to solve. (Karl R. Popper, 1972)

Theories provide us with a particular lens through which we view problems, devise appropriate methodologies, and analyse data. Theories can even be used to substantiate well informed but often-subjective recommendations. Up to this point, this study has reviewed the relevant literature discussing different perspectives on endangered languages, corpus linguistics and in particular theories on corpus design. Henceforth, I will be describing the decision-making process and how I approach the research problem of this thesis in a quest to find the most appropriate solutions. Epistemologically, these decisions have been informed and enriched by the myriad of works that I have read (Biber 1993; Davies 2009; Sinclair, 2004) which were complementary in many ways, yet sometimes contradictory. When confronted with scenarios like these, choices are inevitable. The reader needs clarification about the theoretical framework (or metaphorically speaking, the binoculars), which I used to sift through theories and perspectives to arrive at my decisions. Corpus-design methodology varies according to the philosophical stance of the designer and the aim of the corpus, so it is necessary to situate this research in a clear conceptual framework.

It is interesting to note how strikingly Popper (1972) captured the evolutionary nature of science and research paradigms in the title of his book, *Objective Knowledge: An Evolutionary Approach*. This is true when one looks at how theories of research have gone through a process

of gradual growth in terms of numbers of theories available and the degree of maturity they have reached over the last few centuries. According to Lincoln and Guba (1985), and well before the positivist paradigm became the dominating research paradigm, there was a pre-positivist era in which scientists assumed that knowledge could be obtained simply by observing nature. If the researcher intervenes in the laws of nature, s/he will cause an unnatural phenomenon, which will distort the truth. Therefore, scientists were passive observers and their main goal was to describe the reality as it uncovered itself without any intervention. Later, as Creswell (2007) noted, the philosophical stance of the main research paradigms that came after pre-positivism aimed to refine and develop the contributions made by their predecessors. Positivists, for example, hold that there is an objective truth, which can be discovered by the researcher often using quantitative and experimental studies. For them, the researcher gets involved in conducting the research in an objective manner and must avoid subjectivity.

Post-positivists brought about another paradigm shift, again in an effort to carry the torch of knowledge forward. They argue that yes there is an objective truth out there, but people are unable to discover it with certainty. Therefore, the main objective of research is to get closer to the truth. Since absolute truth is unattainable, the outcome of any research may or may not be true. Post-positivists contend that the quantitative method alone is not sufficient for research because different knowledge domains require different methodology. Based on these assumptions, they introduced qualitative methods as yet another research approach, which they believed was more appropriate for social science research as opposed to research in the natural sciences. Additionally, post-positivists acknowledged the need to use mixed-methods research in which both quantitative and qualitative approaches are used concurrently (Creswell, 2007; Lincoln & Guba, 1985).

Another research paradigm that seems to share a greater affinity with post-positivism is the constructivist paradigm, which advocates for inclusion of broader and multiple perspectives. For constructivists, knowledge is relative, based on factors such as context, cultural, and even the historical reality of a given phenomenon. As a result, since knowledge is believed to be relative, the researcher and the research participants need each other to co-produce knowledge, co-constructed by society. Qualitative inquiry is employed as a research methodology by both post-positivists and constructivists to capture the diversity of such multiple perspectives. Both post-positivists and constructivists agree that the researcher bias could be minimized but it is nonetheless unavoidable, as the researcher brings his/her own worldview to the research, and awareness and acknowledgement of this positionality is warranted (Creswell, 2007).

A major paradigm shift surfaced when pragmatists introduced a new perspective to the research world that attempts to bridge the gap between theoretical and methodological stances and their relative practicality in the real world. At the outset, as Creswell (2007) noted, the pragmatist paradigm embraces the worldview that attempts to shift the research focus away from the philosophical debate on what reality is, towards the concept of what is practically feasible. This school of thought contends that real world problems necessitate theories and methodologies that are both relevant and practical. Creswell stated:

In practice, the individual using this worldview will use multiple methods of data collection to best answer the research question, will employ both quantitative and qualitative sources data collection, will focus on the practical implications of the research, and will emphasize the importance of conducting research that best addresses the research problem. (p. 23)



Consequently, research methodologies are chosen based on practicality and relevance to the research questions and problem, and the researcher is free to choose any methodology that works. Creswell (2007) pointed out that pragmatists and researchers who tend to use mixed methods share the concept that there is more than one way to collect and analyse data.

Advocacy or the participatory research paradigm is one of the main alternative research paradigms as opposed to one of the paradigms discussed (the pragmatist approach or constructivist paradigm). A researcher may employ this alternative worldview because other theoretical frameworks do not appropriately fit research questions that are meant to bring about social change. Creswell (2007) stated, “The basic tenet of this worldview is that research should contain an action agenda for reform that may change the lives of participants, the institutions in which they live and work or even the researchers’ lives” (p. 21). Clearly, advocacy research or, as it sometimes called, participatory research, seems to have the theoretical orientation that there are social, economic, historical, political or even technological forces that bring about social ills including, but not limited to, how certain societies abandon their language out of desperation as we have seen in Chapters Two and Three. A participatory research approach, therefore, examines such complex social issues with a clear vision of mobilizing masses and influencing relevant institutions to push for an action agenda to mitigate or even reverse a particular social problem.

This range of educational research paradigms has, to a greater or lesser extent, informed and enriched the way I have approached this research. However, my own worldview equally or perhaps more profoundly shaped my conceptual framework.

As a researcher, I started my research with a set of worldviews and assumptions. As someone who has been exposed to different worldviews and cultures, I consider myself a researcher who sees the world through multiple perspectives. I was educated and have lived in

three culturally distinct continents—Africa, America, and Asia. A comprehensive account of my lived experiences in these environments and how they shaped and enriched my worldview is not feasible here; therefore, I hope the following snapshots of such a journey will suffice to illuminate the point I am trying to make.

I took my undergraduate degree in Canada and went to the United States for my master's degree, later coming back to Canada for my doctoral studies. This has had a profound impact on my outlook and the way I view the world through extensive reading, learning, and relearning and later teaching and working with colleagues from different parts of the world. As a child, I had the opportunity to learn in a culturally and linguistically rich environment where a child learns not only from his/her parents and siblings but also from a multitude of sources in the neighbourhood, both formally and informally. My worldview has been shaped through exposure to such multifaceted resources and growing up in a nation which scholars and early travellers nicknamed “the nation of poets” for its literary endowment (Andrzejewski & Lewis, 1964; Burton, 1894; Laurence, 1954). Hashi (2001) further illustrated this reputation:

Most observers of Somali history and culture noted the value that Somalis attach to oratory skills and facility with language. Traditionally, politicians, clan elders, and judges were chosen for their verbal adroitness and their ability to compose or substantiate their points by citing a poetic verse or proverb. (p. 31)

Ontologically, one learns rather powerful philosophies about the nature of life and reality through these literary teachings. For instance, proverbs are seen in the Somali culture as a prominent institution with its own school of thought that has developed over the centuries together with theoretical ways of life through praxis. Testimony is found in the proverbs themselves. I will take an example of the proverb which Georgi Kapchits, a Russian who is a

Somali literary analyst, has chosen for the title of his recent collection of Somali proverbs:

*“Somaalidu been way sheegtaa, beense ma maahmaado”* which roughly translates as: “Somalis can lie, but they do not lie in proverbs” (Kapchits, 2012, p. 12). Georgi’s collection is one of the few books available on Somali proverbs. Like all Somali literature, the proverbs have been transmitted orally through generations. Still today, the older generations commonly use proverbs to substantiate or clarify their viewpoint. Epistemologically, proverbs epitomized by this example are based on the assumption that we learn through experience by observing the world around us and testing our hypothesis through praxis. In the eyes of many Somalis, proverbs are considered to be like a theory that has repeatedly passed life’s experimental testing. A proverb that fails to conform or respond to the realities of life in terms of its practicality is rejected as an invalid proverb. In Somali discussion circles, people reject a proverb categorically if someone uses a simulated proverb.

I would like to share here my favourite proverb: *“Saddex ayaa rag ugu liita: Nin markuu maqan yahay aan la tebin, mid markuu joogo aan la tirin iyo mid markuu tagayo aan la celin”* which is translated in Hashi and Hashi (2014) as: “Of all people, three are the worst: one whose absence is unnoticed, one whose presence is unacknowledged and one whose departure is tacitly welcomed (by not asking the person to stay).” This trilogy is based on the philosophy that acknowledgment and recognition is based largely on merit and practical contribution to society. Such lofty expectations, including achievement, distinction, and practical contribution inevitably call for resourcefulness, open-mindedness, and the use of a pragmatic approach to issues. Somalis will find a relevant and appropriate proverb for almost any topic they discuss. The nature and nurture debate in the social sciences is even addressed in Somali proverbs, although not written but transmitted orally through generations. On this issue, the Somali proverbial

teachings would align with theorists who argue that nature is more powerful in shaping the individual. As this concise proverb indicates, “*waano abuuris baa ka horreysay*” which translates as: “creation of the person came well before counselling.” The essence of the proverb is that while the role of nurture is evident in the contribution of the person’s personality and behaviour, nature made its contribution well before that and has already taken the lion’s share, so to speak. Imagine how a student from Somalia in applied health or in educational psychology would come to class with a solid theoretical view of the subject.

While the Somali literary genres that have shaped my worldview are many, I will briefly cite poetry as another dominant source of my knowledge base. Maxamed Ibraahim Warsame, also known as Hadraawi, is arguably one of the most gifted living poets in Somalia. I will take an excerpt of his masterpiece, translated as *Life’s Essence* or *Secrets of Life*, which is entitled *Sirta Nolosh*a in the Somali language. According to Jama (2013), Hadraawi “composed more than 200 short and epic poems over the past 40 years, all of irresistible appeal to the Somali masses since they reflect the nation’s struggle, its present predicament and future aspirations” (p. 12). The poem is 17 pages single-spaced and it touches on various facets of life including but not limited to the art of rhetoric and logical reasoning, the nature and value of knowledge vis-à-vis their practical implications to real world issues. For example, Hadraawi ends his philosophy on life’s essence with a powerful conclusion:

Each morning brings its own misfortunes

So don’t waste the day bewailing it

To be plain, it is your duty to solve them

Then plan for tomorrow

**Life requires your clarity**

**Fit rules, sound methods**

And none of this is impossible.

**Treasure knowledge, which is seemly,**

**Combine it with apt action**

And you'll lack no asset—

This is what it is to be civilized

And that alone is life's essence. (Jama, 2013, p. 63)

Here the axiological stance of the poet is that knowledge ought to be valued and its value is enhanced when it solves a practical issue, which means that an attempt should always be made to bridge the gap between theoretical knowledge and practice using appropriate procedures and sound methods.

On the thorny issue of whether there is an absolute truth or not, in the same poem, Hadraawi asserted: "Anyone who claims perfection lacks it; there is always someone worthier." Ontologically, Hadraawi maintained that perfection is impossible but knowledge could get closer to that stage if the person acknowledges that there is always room for improvement and refinement. As we will see later in this chapter, I will provide advice about claims of perfectionism in corpus building by Sinclair who is considered a pioneer in contemporary linguistics. On the issue of logical reasoning, the art and style of rhetoric, Hadraawi argued:

Always weigh your words well

Make things clear to the uncomprehending

Don't forget your similes and figures

Nor get into confusion's cul-de-sacs  
Losing your argument's thread  
Don't swallow its essence  
Avoid hesitation—the clarity  
of your facts should not sit in the shade  
Your argument must be plain  
So take care of its coherence  
Your approach must be reasoned  
So limit your questions  
Whenever making key points  
If three words suffice  
Don't stretch things to thirty  
Leave boastfulness behind:  
Don't speak haughtily  
Or wave your arms dismissively  
—Never utter an unbecoming speech. (Jama, 2013, p. 50)

The poem was composed in 1986, and I fell in love with its imagery, artistic approach and the way it touches on a wide range of topics. I memorized half of the poem and I can still recite most of what I learned at that time. The snapshots of my own learning journey are useful to demonstrate that we all have our own culturally based repertoire of knowledge anchored in early childhood socialization and the knowledge exploration process we go through at different

stages in our adult life. Thus far, two dimensions that have shaped my conceptual framework for this research have been discussed: (a) research paradigms, and (b) my worldview.

The first was the review of relevant research paradigms to understand and conceptually articulate the needs and the dynamics of this research in light of the major existing research paradigms. The second dimension concerned my own worldview and how it allows me to synthesize theories, which affects how research decisions are made. These two dimensions together inform the third dimension, which is the research methodology that I have devised for this research including data collection and the final analysis.

Based on the foregoing discussion, the present study employs corpus-based methodology, which has been used in the past by corpus designers in the field (Biber 1993; Davies, 2009; Sinclair, 2005). As we have seen in the preceding chapters, designers have approached corpus design in different ways and with different purposes. Inspired by a pragmatist paradigm, this study focuses on the problem at hand and approaches it from multiple perspectives, employing research paradigms that provide flexibility and sensitivity to the research problem and its context. Wearing different hats at different stages of this study, therefore, should not surprise the reader. For example, the rhetoric, the tone, and even the way I approach the issue of endangered languages is more in line with advocacy and participatory approach to research. Furthermore, the implications of the research may call for an action agenda to mitigate the problem of language endangerment by the use of technology. A constructivist lens is used to describe the nature of language and its complexity. I subscribe to the notion that language is a complex phenomenon, which relies on the creativity and the diversity of its speakers. Based on this tenet, I agree with scholars who view the role of linguists as descriptivist (Hockett, 1958; Kaplan, 1995; Sinclair, 2005) rather than prescriptivist on how language ought to be used. I see that since language is co-

constructed by its speakers and its livelihood depends on them, documentation alone is not enough. Therefore, I argue that a community-driven revitalization program should be undertaken to complement the work on the corpus. The research resorts to both quantitative and qualitative modes of data collection and analysis, which in this sense bears allegiance to a post-positivist paradigm.

As a pragmatist, I often concur with corpus designers who have advanced a pragmatic approach to research. While I keep the issues of representativeness and balance in mind, I approach corpus design in a more pragmatic fashion using an iterative design model. This methodology finds support from a number of scholars in the field (Biber, 1993; Davies, 2009; Sinclair 2004). To conclude this section of the chapter, the pragmatist framework holds this research together conceptually, allowing flexibility to learn from and draw upon multiple perspectives in order to articulate, describe, and answer the research question. As the epigraph at the beginning of this chapter reveals, a scholar who approaches research with an open mind and a degree of flexibility invariably understands the research topic through the insight gleaned from both opponents and proponents of the perspective with which s/he starts the research.

## **6.2 Methodology**

For a research project to stay focused, the research question must be revisited at different stages of the study. The research activities in this study ask the following question: “In what manner can a model corpus help document and develop endangered languages?” In answering the main research question, the following secondary questions will also be addressed:

1. How does one learn from the experiences of previous corpus construction and apply those lessons to the context of other languages?



2. What opportunities does a language corpus provide in terms of language development tools such as resource materials, dictionaries, and grammars?
3. What does it take to build a language corpus for languages with limited resources and ineffective supporting national institutions?

Keeping the primary and secondary research questions and the theoretical framework of this study in the foreground, and the lessons from the literature review in the background, the study now sets out its methodology agenda. Many scholars have adopted a pragmatic approach to corpus design, at the same time taking every possible step that would eventually lead to a representative, diversified and balance corpus (Biber, 1993; Atkins et al., 1991; Sinclair, 2005). A leading figure in the field of corpus linguistics, Sinclair offers the following caveat especially to novice corpus linguists:

It is important to avoid perfectionism in corpus building. It is an inexact science, and no one knows what an ideal corpus would be like. Compilers make the best corpus they can in the circumstances, and their proper stance is to be detailed and honest about the contents. From their description of the corpus, the research community can judge how far to trust their results and future users of the same corpus can estimate its reliability for their purposes. We should avoid claims of scientific coverage of a population, of arithmetically reliable sampling, of methods that guarantee a representative corpus. The art or science of corpus building is just not at that stage yet. (p. 98)

Complex research topics require the use of multiple lenses and research methodologies to help the researcher conceptualize the problem and devise practical solutions. The case in point is that language is a living, dynamic, and infinite entity. Graddol (1994) stated, “Language like many terms used by linguists is one which is taken from everyday language where it describes a

complex and shifting human experience” (p. 3). The use of the words “complex” and “shifting human experience” indicates that languages are complex and ever-changing phenomena.

Considering these realities is perhaps what has led researchers to embrace practical measures when working with languages. This design advice is true for all languages regardless of their status as either stable or endangered. In addition to these universal challenges, endangered languages have particular challenges, as they are resource-strained in many ways. Considering these universal and contextual challenges, a pragmatic and iterative approach to design is seen to be the most appropriate design strategy for endangered languages. Such a design is viewed to be more practical and context sensitive, allowing more time to collect uncatalogued, scarce and often scattered text categories. Moreover, a monitor corpus is seen as an appropriate corpus, which not only reflects the dynamic and changing phenomenon of a language but also allows a longer time for corpus construction to accommodate building a corpus in challenging contexts. Therefore, the corpus design recommended for endangered languages is a monitor corpus, which optimally should go through five distinct but complementary and iterative stages.

### **6.3 Five Stages for Corpus Design in Challenging Contexts**

1. **Phase One—Prototype Stage:** Given the status of vulnerable and endangered languages and considering that they are resource-strained in terms of availability of language sample and funds to carry out the work, opportunistic data collection suits the first and second phases of the corpus construction. The first phase is to be used as a blueprint for subsequent design phases. Definition of the scope of the corpus is to be stated and then the development of a framework of criteria-based text classifications and a guiding list of text categories follow. Text collection is for exploration purposes, analysing the data quantitatively and qualitatively in order to demonstrate how they could be used for

development and research purposes. Given the limited time and space of this thesis, a working prototype will be furnished to demonstrate how a corpus can help an endangered language in terms of documentation and revitalization initiatives. The aim of this thesis is to finish phase one. The study will help inspire a wider collaborative effort which will take this study forward to implement the following four phases.

2. **Phase Two—Incubation Stage:** At this stage, the designer allows time for the corpus to grow without any design limitations but at the same time strives for the collection of as wide a variety of text samples as possible. This effort includes widening spontaneous speech samples across more settings as well as scanning or retyping rare and/or older texts, and even harvesting the web. The main objective should be the collection of a variety of text types as well as the identification and development of a more comprehensive sampling list of known text categories available in a stratified fashion. The secondary objective is for the designer to build on the guidance and the recommendations of this thesis to refine and develop a diversified stratified sampling list for the third phase by looking at the sampling framework to redefine available text categories and dialect classifications previously employed in the first phase, based on the feedback from phase one.
3. **Phase Three—Monitor Corpus:** This phase turns the opportunistic corpus built thus far into a monitor corpus. At this stage, using the sampling frame developed during the incubation stage, designers should redesign the corpus considering ways to represent the language under study: variability, balance and size. This means the focus should now be on (a) how to represent the language as a whole in terms of variety, (b) balancing samples among texts, and (c) targeting a sufficient size for a large monitor corpus of at least 20

million tokens to start with. The aim is to grow the corpus by adding the same variety of texts in a balanced manner annually. At this stage, based on the needs of the language, specialized corpora can be developed, namely bilingual, literary, or media corpora on which specific and relevant language studies could be based. The bilingual corpus, which starts with English and Somali, is seen as important because it will facilitate cross-linguistic studies and the development of bilingual dictionaries and machine translation tools. Somali literature and its poetry per se, as illustrated in this thesis, are considered to be gemstones buried in the language. It is a comprehensive body of knowledge that Somalis can offer to the knowledge treasures of humanity. For journalists, a corpus is vital because the Somali media industry is arguably the only robust institution that reflects contemporary Somali language. Standardizing and disseminating various learning materials on how Somali is used in the media in relation to how journalists in the past used the language will unify their efforts towards using a Somali that is standardized and more accessible to all Somali-speaking peoples. This point will be revisited in Chapter Eight where the implications of the study will be described. A written corpus is another specialized corpus that could be developed to hold a wide variety of formal written samples collected in an effort to strengthen the standardization of the Somali language. Brown Corpus is an excellent model for this.

**Phase Four—Annotation:** As Sinclair (1991) pointed out, for a corpus to be useful and meaningful, a large collection of words is not enough; therefore, annotation adds to the richness and usability of the corpus. In general, two types of annotation are considered useful for research or development purposes: *linguistic tagging* and *demographic information annotation*. One basic type of tagging is part of speech (POS) tagging, for

which different software taggers are available. Lexical classification such as POS tagging propels the development of spell-checkers and other word-processing tools. Likewise, recording information about the writer of the text or the speaker (e.g., gender, age, geographical information) will inform studies on dialectal variations and standardization efforts. In order to ease the readability of the corpus, the annotated segments of the corpus are better separated from the raw data or the original text. The rationale is that researchers are given the option to access the language in its original version if they so desire.

4. **Phase Five—Representative Monitor Corpus:** The corpus grows and continues to support the needs of a given language. Sustainable growth and redesign, which targets a representative, larger monitor corpus, is the main objective at this stage. While the corpus continues to grow, the immediate needs of the language should be addressed, such as the development of language tools and materials that could be developed. This process stimulates the designers to refine and redesign the corpus to align the needs of the language and the intended uses of the corpus, by building on the recommendations from the previous four phases. This iterative design continues as the corpus grows. At this stage, accessibility to the public is desirable. First, for the corpus to grow and refine its design, it needs the input and feedback from the research community and others who may need to use the corpus for development or research purposes. Second, a wider accessibility could restore the status of the language and the morale of its speakers, which strengthens the documentation and revitalization objectives of this thesis. A simple user-friendly portal with a search function will suffice. Compatibility, usability, and user-experience considerations should be the guiding principles of the portal design. Enlisting

the help of specialists in different areas of expertise is always highly recommended, as the use of technology is multifaceted and requires a collaborative effort in the implementation of projects such as this.

#### **6.4 Case Study**

The title of this study, *Building a Model Corpus for Endangered Languages*, gives the impression that one language is used as a case study since all endangered languages cannot be the subject of this study. It is therefore imperative to clarify how the term “case study” is used in this research. Creswell (2007) viewed a case-study approach as a research methodology. He asserted, “I choose to view it as a methodology, a type of design in qualitative research, or an object of study” (p. 73). The fact that Creswell chose a particular view implies that there are other conceptions about how a case study is used. In support of this position, Yin (2003) agreed, “The case study is but one of several ways of doing social science research” (p. 1). Creswell (2007) noted the use of multiple data collection techniques that are often associated with case studies such as observations, interviews, audiovisual material, and documents.

These perspectives indicate that a case study is often conceived of as a research strategy with its design conventions of data collection as well as data analysis techniques (Creswell, 2007; Merriam, 1998; Yin, 2003). Employing this sense of the term, I used a case study as a research methodology for my MA thesis (Hashi, 2001). For example, I used an in-depth interview and a rich description of what I had learned about the case. I had recurring key themes in presenting a key finding which was a new hypothesis in the area of motivation and second-language learning. For this study, I intend to use “case study” not as the guiding research methodology but rather as a type of study. This stance is supported by Stake (2005) who stated, “Case-study research is not a methodology but a choice of what is to be studied” (as cited in

Creswell, 2007, p. 73). This study serves as an example of how a particular case can inform other similar cases.

The rationale for choosing this conception of the case study is twofold. First, by virtue of the fact that language itself is a complex phenomenon, the study calls for research methodology that captures not only the multifaceted nature of the language being studied but also the practical aspects of the design process. Second, the study employs a corpus-design methodology that it builds on by replicating tested but context-bound design methodologies in challenging and less-commonly studied contexts.

To my knowledge, this is the first study that has been undertaken to present the challenges and the potential benefits of developing a model corpus for endangered languages using the Somali language as a case study. It is my hope that this research will spark discussion in the field or even more importantly inspire language development in the context of other less-commonly studied languages including endangered languages.

Drawing upon how particular corpus-design cases approached corpus construction, I define the boundaries of the topic in terms of (a) the language to be studied, (b) the scope of the data, and (c) the appropriate software for the corpus. I do this by reviewing and analysing previous corpus-development cases and conferring with subject-specific literature. I then create a prototype in an effort to answer the research questions of the study. The challenges and the design decisions are discussed in both graphical and textual representations.

The prototype could later be refined, modified, or discarded. The study recommends subsequent iterative design phases, which mean that the proposed corpus uses a cyclical process of prototyping, analysing, and refining as the work goes through various stages of design, redesign, and development, as described before.

## 6.5 Defining the Scope of the Data

The scope of the data to be collected is constrained by the availability of time and resources. With this in mind, the study uses the Somali language as a case, thereby narrowing the topic to one language. To further define the scope, the definition of the Somali language includes countries where Somali is not considered the official language. These areas include Somalia, Somali Regional State of Ethiopia, Djibouti, Northeastern Province in Kenya and the Diaspora where over a million Somalis currently reside. The Somali language has no official status in any of these countries except Somalia where the language has practically no institutions that can enforce the privileges that come with the official status afforded to Somali in the constitution.

This study proposes the construction of a Somali Language Bank, which aims to be a representative, large, monitor corpus. There are two reasons for the choice of a more inclusive and comprehensive Somali-language corpus. First, generally endangered and threatened languages do not have library records or other reference resources that keep track of books in print, so available resources tend to be scattered. Accordingly, a more inclusive approach is considered, assuming that it yields diversified data, which is the prerequisite of a representative corpus. Second, since the Somali language is used in different countries with different government policies, a collaborative effort among countries and their speakers is seen to enhance the documentation and revitalization objectives that are central to this thesis. In terms of the scope of data, every Somali text available is to be regarded as a candidate for inclusion. This broad inclusion rule will lessen the effects and most likely compensate for the lack of readily available data. At the same time, it helps the inclusion of a wider text variety, which is considered a defining feature of a representative corpus. For a language that has limited resources or few established library institutions, identifying, locating, and collecting text types



that would represent existing language varieties is a contextual challenge for endangered languages and less-commonly studied languages.

## **6.6 The Somali Language**

The Somali language is a member of the Cushitic languages such as Oromo and Afar, which are spoken in the Horn of Africa. These languages are descendants of the Afro-Asiatic family of languages such as Arabic, Hebrew, Aramaic, and Egyptian (Ohio University Center for International Studies, 2010).

It is not feasible to get the exact number of Somali speakers because of the civil war that has engulfed Somalia, where the majority of Somali speakers live, since 1990. Lewis (2009) estimated that close to 14 million people speak Somali worldwide. It is generally estimated that 25 million people around the world speak Somali (UCLA Language Materials Project, 2010). A considerable number of Somali speakers now live in many parts of the world (the Diaspora) and this might explain why the estimates are so far apart. Lewis (2009) listed the main countries where Somali is spoken: Djibouti, Kenya, Ethiopia, Canada, Finland, Sweden, UK, Italy, UAE, Oman, Saudi Arabia, and Yemen. According to the University of Ohio's Center for International Studies (2010), which offers Somali language courses, Somali is considered a regional language in the Horn of Africa. For example, in Djibouti, in the Horn of Africa, Somali is the most widely spoken language in the country but French and Arabic are the official languages and the medium of instruction in schools, including higher institutions (Lewis et al, 2014).

As touched upon in Chapter Three, one of the colonial legacies is the fact that its speakers now live in different nation states in the Horn of Africa. The majority of Somali speakers live in Somalia where Somali is considered the official language and the lingua franca, but it has gradually lost its status as the language of instruction in schools. Djibouti, formerly known as

French Somaliland, is an independent country and although the majority speaks Somali, it has no official status in the country and it is not taught in schools. Somali is considered the language of the majority in the Northeastern Province of Kenya but again it has no official status and children do not learn it in schools. Kiswahili is the official language of the country while English is the working language and the dominant language of instruction in schools (Muthwii & Kioko, 2004). The Somali Regional State of Ethiopia has recently chosen Somali as the working language of their state while Amharic remains the official language of the country and the language of instruction in most schools is Amharic together with English (Hameso, 2001). In addition to these mainland areas, Somali is widely spoken in the Diaspora. The first generation of Somalis in the Diaspora who were fluent in Somali when they came to their host countries are now aging. This scenario paints a rather bleak future for the language because younger generations intentionally or unintentionally abandon their mother tongue. Some still speak the language with different levels of proficiency but that can only take the language further for one more generation. In the context of endangered languages, languages are seen to carry fresher perspectives when they enrich and advance the quest for knowledge development.

Incidentally, while exploring the existing literature on the Somali language, I was struck by a book entitled, *A Tree of Poverty*, by the highly respected Canadian author, Margaret Laurence (1954), mentioned above. In the opening paragraphs of the book, it feels as though Laurence's appeal was a call to spearhead a project like this thesis. She wrote the following lines more than a half a century ago:

One of the most unfortunate aspects of Somali literature is that all poetry and folktales are unwritten . . . the work of recording the stories, which are far more numerous than the poems, will be a gigantic task, and it has not, as far as I know, been started yet. It is to be

hoped that more and more literate Somalis will begin recording the literature of their people. (p. 2).

Sixty years later, Laurence's wishes have not been fully attended to, at least to the extent that technology affords. The scope of this study is to initiate the collection and the digitized samples that contain various forms of the Somali language, both prose and poetry. The main aim is to build a framework that others can build upon in the future. It is my hope that, in the future, a larger digitized bank of the Somali language will be built when the infrastructure of such a model becomes available.

As Laurence noted in 1954, Somali people are well known for their poetic endowment. In 1894, British explorer Sir Richard Burton expressed this defining characteristic in a rather astonishing passage:

The country teems with "poets." . . . Every man has his recognized position in literature as accurately defined as though he had been reviewed in a century of magazines—the fine ear of this people causing them to take the greatest pleasure in harmonious sounds and poetic expressions, whereas a false quantity or prosaic phrase excites their violent indignation. . . . Every chief in the country must have a panegyric to be sung by his clan, and the great patronize light literature by keeping a poet. (p. 14)

Approximately a century later, Andrzejewski and Lewis (1964) echoed Burton's assessment and labelled Somalia the "nation of poets." Likewise, most observers of Somali history and culture noted the value that Somalis attach to oratory skills and facility with language. Traditionally, politicians, clan elders, and judges were chosen for their verbal adroitness and their ability to compose or substantiate their points by citing a poetic verse or proverb. Although Somali was the language of poets, it was completely oral. In 1972, an official

orthography of the Somali language was introduced. Immediately after the inception of Somali orthography, the government launched a massive literacy campaign throughout the country to nomads and urban dwellers alike. By the late 1970s, the Somali school system had adopted the Somali language as the medium of instruction from the elementary to the high school level. In the 1980s, most university faculties adopted the Somali language as their medium of instruction. The mass literacy campaign following the introduction of Somali language orthography increased the national literacy rate from 5% to approximately 55% (Laitin & Samatar, 1987) in the mid-1970s. The absence of caring and effective national institutions coupled with a host of other factors discussed in Chapter Three has reversed much of what has been accomplished since the implementation of the Somali orthography just before the Somali civil war broke out.

In our quest to represent the Somali language as a whole, a consideration of the dialect variations is needed. The most recent report on the subject of Somali language and its dialects has been put together by Norway-based Landinfo (2011), which supplies Norwegian government authorities with information pertaining to human rights as well as country-specific information about foreign countries. The report relies mainly on Somali dialect research conducted by Lamberti (1986a) along with rather dated Somali language studies carried out in the early 1900s.

Those previous studies offer tremendous diachronic insight on the language and can certainly be used as a guide in developing sampling documents during the initial stages of the project. Given the change that languages go through over time, a more recent study in the area of Somali dialectology is to be conducted. This is one of the areas where a corpus can shed light on the relative convergence and/or divergence of the Somali language with regard to dialect variations over the years. Until the outcomes of such a study become available, the earlier studies

can be used as a guide. Lamberti (1986a), as cited in Landinfo (2011), classified Somali dialects into five major dialects while acknowledging that there are more than 68 sub-dialects of Somali:

1. **The Northern Dialects:** These dialects are presumably combined, based on their relative linguistic similarities. This group of dialects includes western (Somali Regional State of Somalia and Djibouti), northwestern and northern areas as well as the eastern and northeastern areas (such as Mudug, Nugaal, northern parts of Bakool and Gedo regions) all the way to lower Juba and its neighbouring northeastern province of Kenya.
2. **The Benadir Dialects:** This includes parts of Galgadud in central Somalia, Hiiraan, Middle and Lower Shabele all the way to most districts in the capital city of Mogadishu and its neighbouring regions such as Afgoye, Qoryoley, Marka, Diinsoor and certain parts of Saakow, Jamaame, and Bu'aale.
3. **The Asharaf Dialects:** This dialect includes Af-Shingani spoken in the Shingani district in Mogadishu, Af-Marka, Af-Gendershe and Af-Jilib, all spoken by tribes who belong to the Ashraf clan.
4. **The Maay Dialects:** This is a group of dialects that share linguistic similarities. They extend from Baardheere in Gedo to Bay and Bakool, all the way to the Lower Juba River area in places such as Jamaame and Afmadow.
5. **The Digil Dialects:** Although this group shares relatively greater similarity, they also exhibit greater differences than any other dialect group. This includes the Tunni dialect in Dhinsor, Brava, and Jilib. Dabarre, Garre, Af-Oroole dialects are spoken in parts of Bay and in lower Shabele River areas such as Qoryoley and certain areas along the Juba River.

Within the Somali speaking communities, Somali language is grouped into two major dialects: *Maay* and *Maxaa Tiri* dialect. Lamberti (1986a) and Landinfo (2011), group the first three dialects together under the *Maxaa Tiri* dialect while the Digil dialect, which is more closely related to *Maay*, comes under the *Maay* dialect. Using this distinction as a basis, the Somali federal government recently added a clause in the constitution that the official language of the Federal Somali Republic is Somali, noting in brackets—*Maay* and *Maxaa Tiri*. Lamberti (1986a, as cited in Landinfo, 2011) noted that these two dialects exhibit certain phonological, lexical, and grammatical differences. The most notable difference is that two pharyngeal sounds which are represented in the Somali orthography as /c/ and /x/ are found in all other dialects but not in *Maay* dialects. A well-respected Somali linguist, Mansur (1988) contends that the *Maay* dialect has in fact retained most of the older Somali vocabulary while *Maxaa Tiri* has lost it and gone through a considerable lexical shift through loan words from Arabic, English, French, or Italian, which explains the phonological and lexical differences that now exist between the two dialects. Therefore, he argued that the *Maay* dialect embodies the origins and the historical characteristics of the Somali language, so it can be used as a basis to trace the historical development of the Somali language, history, and culture.

The issue of standardization is another area that has been challenged since the collapse of the central government in Somalia. Kenadiid (1972) noted in the introductory remarks to his landmark Somali dictionary that the standardization of the Somali language will take a long time until that fateful time when Somali dialects are fully studied and then all find their way into one comprehensive dictionary. Since the implementation of the Somali orthography, the northern dialect has been used as the standard Somali language and has been the language of print materials including school textbooks, the language used by the media, and government

institutions. It has been and still is the language of broadcasting for the BBC and VOA Somali Service. As it stands now, the growing trend to standardization since the implementation of Somali orthography seems to have been disrupted. This is due mainly to the absence of effective national institutions, notably the fact that Somali has ceased to be the medium of instruction in schools. It is gradually shifting away from a unified standard Somali to a situation in which a large number of Somalis who have not been educated in that era, tend to write Somali intuitively based on pronunciation rather than the standard forms. As this trend grows over time, the written Somali language may lose its vitality as a language of communication through written means, throwing the language back to its historical function, a language of communication through oral means only.

A general, representative corpus sheds light on the current status of the language, illustrating how the language is currently written and spoken, giving us data-based perspectives in the areas of Somali dialectology and offering ways to strengthen its standard system. As Biber (1993) noted:

The use of computer-based corpora provides a solid empirical foundation for general-purpose language tools and descriptions, and enables analysis of a scope not otherwise possible. However, a corpus must be representative for generalizations concerning a language; for example, corpus-based dictionaries, grammars, and general part of speech taggers are applications requiring a representative basis. (p. 1)

Therefore, the collection of as wide a variety as possible is of paramount importance and the guiding principle is gathering texts representing not only a variety of registers but also dialectal variations.

## 6.7 Choosing Corpus Construction Software

A corpus uses a computer through specialized software called concordance software. There are a number of software resources currently available on the market. For the purposes of this thesis, three similar software resources with slightly different features have been identified. Each of the three can be used to answer the research questions of this study. Wordsmith Tools and MonoConc Pro are commercially available for reasonable prices while AntConc is available for free if it is used for research and non-commercial purposes. While they all come with the standard features commonly used by corpus linguists, some come with additional features and useful tools that are relevant to this study. The common features include the concordance feature, which gives the context in which a particular word is used in the corpus and is often used to learn how words behave lexically and grammatically. Another common feature is called Keywords analysis, used to determine words that are particularly frequent or not so frequent in a given text. Yet another common tool is Wordlist feature that allows one to generate words in alphabetical order or in terms of their frequency of use in the corpus of a given text. Some commercially available software such as WordSmith Tools comes with more advanced tools such as the Webgetter tool, which is used to gather text from the web using keywords that might be characteristic of a text one desires to collect. For the purposes of his thesis, AntConc was chosen. The reasons for this choice are twofold. First, after reviewing its features and the level of usability, AntConc was found to be appropriate in answering the research question of this thesis. Through this research, I aim to illustrate how a corpus can aid researchers and material developers to preserve and study endangered languages, for which AntConc can suffice. The second reason is to illustrate how corpus construction can be initiated with freely available



software to inspire communities to take charge and rekindle their languages without the constraints of finding sufficient funds.

### **6.8 Somali Language Bank (SLB)**

As discussed earlier, the proposed Somali Language Bank is a larger project, which will eventually produce a large monitor corpus. For phase one, text samples that allow me to address the research questions were collected opportunistically. Data collected at this stage is meant to help designers for the subsequent phases to make better-informed design decisions using the data analysis as feedback. While the design of data collection reflects the characteristics of a representative corpus, the actual data collection for phase one is based on the extent to which it can illustrate the research questions of this thesis. In what follows, the data collection strategy is outlined. However, for the purposes of this thesis, the actual data collection aims to produce a prototype as outlined earlier in phase one.

### **6.9 Data Collection**

The Somali Language Corpus aims to collect 20 million tokens as its first target with the intention to grow the corpus continually. This target size is contingent on practicality and therefore room for adjustment is allowed in the four stages. Drawing upon and consulting with previously used text-collection strategies (Biber, 1993; Davies, 2009; Sinclair, 2005; Frances and Kučera, 1979), and considering the language and the context of this study, the SLB text classifications framework is based on four main criteria: (a) genre, (b) topic, (c) medium, and (d) dialects.

Using genre as the main defining criterion, the corpus aims to balance the genres with equal weighting of 50% spoken genre and 50% written genre. There are two reasons for this. First, scholars are generally in agreement that a balanced corpus should at least cover the two

main genres equally (Atkins et al., 2002; Biber, 1993; Sinclair, 2005). Second, given the oral tradition of the Somali language and the fact that it has a relatively low repertoire of written registers, the spoken genre has more robust and established registers. This is because the Somali language has been used as an oral language for a long time in contrast to the rather younger written tradition, which started on October 21, 1972. This phenomenon is evident when looking at how Somalis divide language or talk in two main registers: poetry and non-poetry. It is indicative of the central position that poetry occupies in the Somali language and indeed in many languages with longstanding oral traditions but with more recent written registers. Poetry is classified as spoken genre while languages with established written traditions classify as written genre. In the process of designing this corpus, I propose a new definition of what constitutes written and what constitutes spoken. For the purposes of this thesis, the written register is defined as texts that originate and are largely available in written form while spoken registers are texts that originate and largely remain in spoken form. Using this definition, poetry is classified as spoken genre because it is composed orally and it is largely available in oral mediums. While a few anthologies and collections of poetry have been published, most Somali poetry remains in oral form. These definitions are deliberately based on the medium in which the texts are available to ease the text-collection process by surveying the medium in which each register is found.

By way of comparison, the following table shows how BNC has designed its spoken component, which has no mention of poetry in its spoken registers. While this table has been useful in designing a text-classification framework for Somali, the organization of the framework has to be redrawn to fit the needs of the Somali language.

Table 6.1. *Text Types of Spoken Data in the BNC (adapted from Burnard, 2000)*

Domain Text types	Text types
Educational and informative	Lectures, talks, educational demonstrations, news commentaries, classroom interactions, etc.
Business	Company talks and interviews, sales demonstrations, business meetings, consultations, etc.
Public/ institutional	Political speeches, sermons, public/government talks, council meetings, religious, court proceedings, meetings, parliament proceedings, legal proceedings, etc.
Leisure	Speeches, sports commentaries, talks to clubs, broadcast chat shows and phone-ins, club meetings, etc.

Conversely, in consideration of the context and the nature of the Somali language, the text-classification framework is drawn as follows:

Table 6.2. *SLB Text Classifications sampling frame*

Criteria	Criteria-based Classifications		Text Categories in English	Text Categories in Somali
Genre	<b>Written (50%)</b>	<b>Dhab (facts-informative)</b>	Non-fiction books, articles, ads, memos, constitution, laws, letter, email, text message, agreements, proposal, dairy, translations	Buug aqooneed, Maqaal, Wareegto, , Iidheh qoran, Dastuur, Xeer, Warqad, Imeyl, Farriin-qoran, Heshiis, Hindise, Xusuus-qor
		<b>Dhalanteed (Fiction-imaginative)</b>	Novels, folktales, play scripts, movie scripts, and all other imaginative works	Buug khayaali ah, sheeko-xariiro qoran, riwaayad, filin, ilaqosol/maad iyo wax kasta oo la curiyo oo dhalanteed ah

	<b>Spoken (50%)</b>	<b>Tix (poetry)</b>	Poems, ballads and other types of Somali literature	Gabay, Geeraar, Jiifto/masafo, Afarrey, Buraanbur, Hees ciyaareed, Heeso-hawleed, Suugaan kale
		<b>Tiraab (talk)- Formal</b>	Lecture, meeting, traditional court deliberations, traditional marriage proposal ceremonies, sermons, debates, interviews, press conference, news, commentary, parliament proceedings	Muxaadaro/cashar, Shir, Gar, Goggolgal, wacdi/khudbo, Dood, waraysi, war-saxaafadeed, war, faallo, doodaha golaha shacabka
		Informal	Chat/conversations, courting, entertainment, sayings, jokes, Riddles, talk-shows, classroom talks	Sheekaysi, Haasaawe, madaddaalo/baashaal, maahmaahyo, ilaqosol, caraatan, doodwadaag, wadal-hadal arday
<b>Topic</b>	<b>Social Sciences</b>			Maamulka, Siyaasadda Taariikh, Juquraafi, Afafka, Sheek-xariiro, Ganacsiga, Madaddaalo, warfaafinta
	<b>Natural Sciences</b>			Xisaab, Fiisigis, Beeraha, Bayooloji, Kimistri, Caafimaadka
<b>Medium</b>	Newspapers, magazines, academic journals, general books, textbooks, audiotapes/CDs, videos, spontaneous speech, lecture, conversations, oral literature, Online texts			
<b>Dialect</b>	<b>Maxaa Tiri Dialects</b>	Northern Dialects		
		Asharaaf Dialects		
		Banader Dialects		
	<b>Maay Tiri Dialects</b>	Maay		
		Digil		

Texts are further classified according to their function, setting or whether they are formal or informal. For example, the non-poetry spoken genre is first divided into formal and informal registers. Then if we take, “conversations” as an example, it could be further classified using two major settings: urban and country. Each of these settings is further classified using settings or locations where the talk takes place. Sinclair (2005) has inspired the use of this binary system and classifies texts on the basis of their language function.

Table 6.2 *Text Category to Specific Texts*

Genre				Setting/function as texts
Spoken	Tiraab (Talk)	Informal: Sheekaysi (conversations)	Urban Setting (Magaalo)	<ol style="list-style-type: none"> <li>1. At coffee shops/restaurants: Fadhikudirir, Caadi, shir</li> <li>2. At home</li> <li>3. At school: tertiary/ schools</li> <li>4. At work</li> <li>5. In the car/bus</li> <li>6. At the market</li> <li>7. In entertainment venues</li> </ol>
			Country Setting (Miyi)	<ol style="list-style-type: none"> <li>1. At home</li> <li>2. At the farm</li> <li>3. At the grazing land</li> <li>4. At the village market</li> <li>5. At the animal market</li> <li>6. In the entertainment venues</li> </ol>

A brief description of each text will be recorded and should be available for corpus users so that they have the metadata about texts in case they require further information about a text. The following table is an example of how this could be developed. A similar one could be developed for the spoken genre.

Table 6.4 *Text Identification: Additional Metadata About the Text*

Text ID		
	Text Name	Ayaan Daran
	Publisher	Madbacadda Qaranka
	Date	1985
	Author's birthplace	Gaalkacyo
	Edition	First Edition

### 6.10 Data Analysis

Data is analysed using both quantitative and qualitative approaches. Employing AntConc, the data is quantified in a number of ways. The developer, Laurence Anthony of Waseda University in Tokyo noted, “AntConc is a freeware, multiplatform tool for carrying out corpus linguistics research and data-driven learning” (2011, p. 1).

For example, the software is asked to display frequency lists of various words and their relation to other words in the corpus as well as carrying out other linguistic analysis.

There are seven main tools of analysis. These tools are used selectively, based on their effectiveness to inform and illuminate the research questions of this research. As Anthony (2011) stated, the main tools in AntConc are useful in analysing any corpus quantitatively and qualitatively, as explained below:

1. *Concordance Tool*: This is used to display how a given word is used in context, better known as Key Word In Context (KWIC). The tool retrieves the word in question in a series of different sentences extracted from texts in which the word is used in the corpus.
2. *Concordance Plot Tool*: With this tool, the researcher is able to view words researched.
3. *File View Tool*: Texts can be viewed individually for further investigation or to learn more about search results given by other AntConc tools.

4. *Clusters (N-Grams)*: This is used to view a summarized result of the Concordance Tool and Concordance Plot Tool. Another useful feature is how it shows words and phrases that are commonly used with a given word. This is often used to identify common expressions and phrases in a language.
5. *Collocates*: This shows the collocates of any given word statistically measuring how related the word is to its collocates by displaying words that occur to the left and to the right of the search word.
6. *Word List*: This tool quantifies all the words in a corpus and lists them alphabetically. It is used to learn which words are the most frequent in a language as illustrated by the texts in the corpus. Although a small corpus, like the one developed for this study, can give a general idea of which words are the most frequent, the issues of size, diversity and balance will play a critical role in giving a true picture of the most commonly used in a language.
7. *Keyword list*: This tool is useful to determine which words are commonly used in each genre or domain.

Furthermore, using these quantitative data, I can then illustrate qualitatively how these numbers are relevant to the research question by showing how they inform the design and the production of language-development tools that facilitate accessibility and ease of use for languages that now lack such facilities.

## 6.11 Copyright Issues

Copyright is a delicate matter posing a real challenge for researchers for they often find themselves in a dilemma. On the one hand, researchers must to respect the intellectual property of individual copyright owners while on the other hand researchers often require the use of such

copyrighted materials in a quest to advance knowledge for the greater benefit of humanity.

Countries have attempted to strike a balance between these conflicting rights. In the United States of America, people who feel that they have used copyrighted material in a reasonably fair manner do so in accordance to how their use meets four guidelines under the doctrine of fair use. Courts use these guidelines to decide whether a given case is deemed to be within the fair-use parameters. Stim (2010) outlined the four factors as follows:

1. **The purpose and the character of use:** In the US, the purpose and character of use is the most important factor in evaluating cases. With regard to the *purpose of use*, judges look at motives such as profit versus non-profit use. For example, education and research are among legitimate purposes that are allowed under fair-use provisions. The rationale is that these works are seen to benefit the greater needs of the public because researchers require them as foundational work. The *character of use* deals with the extent by which the user transforms his/her work so that the new work is distinctly different from the copyrighted material used.
2. **Nature of copyrighted work:** This factor considers whether the work is creative or fictional in nature as opposed to nonfiction or works that are more factual in nature. A consideration concerns whether the work in question is published or unpublished. Creative and unpublished works tend to work against fair use presumably in consideration of the creativity and the mental work involved in producing such work. In contrast, unpublished work is given more protection, permitting the creator to decide when and how s/he wishes to release the work.



3. **Amount and the substantiality of the portion used:** This factor concerns whether a particular use could be fair use, based on the amount and significance of the portion used. First, the amount of the work used is weighed against the work in its entity. This means that the smaller the amount of use is relative to the entire work, the greater the chance the selection is considered fair use. Closely related to this is the centrality of the work, which looks at whether the amount of work used is considered an integral part of the work in question. The importance of the portion of work used, which may favour unfair use is considered together with the amount of work used.
4. **The effect of the use on the potential market for or value of the work:** This factor concerns whether the use has a negative impact on the sales of the original work. Among other things, this rests mainly on whether the user had commercial intent in the first place and the amount and the extent to which the user transforms the work so that it does not serve as a substitute for the original work, hence depriving the rightful copyright owner of profit.

Another example relevant to this thesis is Canada's fair-dealing provision for copyrighted works. Canada has recently modernized its copyright laws in search of better ways to find a balance between the individual rights of copyright owners and the rights of the public to access and build upon an existing knowledge base for research and educational purposes. In 2012, Canada amended its copyright act (clause 29 in particular), which outlines the concept of fair dealing under the title *Bill-C11*, which is also known as the Canadian Modernization Act. In section 29.1, it asserts, "Fair dealing for the purpose of research, private study, education, parody or satire does not infringe copyright" (Government of Canada 2014, section 29).

Canada has developed six factors to use as evaluation criteria for whether a given use of

copyrighted material is fair dealing. These guidelines are largely similar to the US fair-use doctrine, so only the areas where difference exists will be elaborated on.

1. *The purpose of the dealing:* The US fair use has this factor and the second factor collapsed as one factor. However, the main difference is that Canada restricts the purposes while the US fair use places no limitations.
2. *The character of the dealing:* Although this factor is similar to the US fair use, it is interesting to note that Canadian fair dealing might also consider customary practice of how a given work is shared among people in deciding fairness.
3. *The amount of the dealing:* This is usually considered in light of the intent of the dealing, so a reasonable amount of use in relation to the purpose of use is examined.
4. *Alternative to the dealing:* This factor does not exist in the US fair use guidelines. It basically expects the user to exert a reasonable search for an alternative or freely available work.
5. *Nature of the work:* This factor deals mainly with the format and the status of the work such as whether the work is published or not.
6. *Effect of the dealing:* This factor is similar to US fair use.

Although the two concepts of *fair dealing* and *fair use* are very similar, there are a few major differences to note. The first major difference is the fact that Canada's fair-dealing provisions allow users to exercise fair dealing as the user's right to use copyrighted work for a limited number of purposes, while the fair use Act in the US places no restrictions on the purposes of use as long the use is deemed fair.

Even with this major difference between the US and Canadian approaches to copyright, it appears that since Canada has delimited the boundaries by which users are allowed to use

copyrighted materials, it has nonetheless shown greater confidence in establishing a better balance between the rights of the owner with relation to the rights of the user. In illustrating this distinction, D'Agostino (2008) stated:

The real difference between Canada and its UK and US counterparts ultimately lies in the policy preoccupations of their respective courts, with Canada's top court alone concerned with championing user rights above all other rights, or at the very least with not championing owners' rights above all others. (p. 315)

Canada's recent copyright amendment in which education was added as one of the legitimate uses under fair dealing, signals the importance it places on serving the interests of the public. Commenting on Bill-C11, the Government of Canada (2011) stated on its Balanced Copyright Website: "For educators, students and researchers, this Bill opens up greater access to copyright material by recognizing education as a legitimate purpose for fair dealing." Under the allowed areas, Canada seems to focus more on the extent to which a particular use falls within the fair-dealing parameters even though the motive for the use might be for business.

Strba (2012) stated, "The commercial character of a 'use' normally leads to rejection of the defence of fair use in the United States. On the other hand, dealing with a work even for profit may be found as fair by Canadian courts if it serves the public interest" (p. 115). Although it is true that a commercial motive is usually seen as unfair use in the US, as we will see below, Google in its book digitization program has won the lawsuit against it even though its motive was at least partially driven by commercial objectives.

As this thesis examines an international matter, the question of how widespread the concepts of fair use and fair dealing are in the world needs to be explored. According to Strba (2012), both fair use and fair dealing laws are generally in line with the international copyright

treaty better known as *The Berne Convention for the Protection of Literary and Artistic Works*, which stipulates that countries respect and add to their copyright laws fair-use exceptions. World Intellectual Property Organization administers the treaty. It was first signed in Berne, Switzerland in 1886 but all 167 countries that are signatories of the Berne Convention have since updated it. Strba (2012) further stated, “It may be observed that a number of states have either officially endorsed the concept of fair use/dealing by including it in their national legislation or have relied on it for guidance in defining other concepts like fair practice” (p. 112).

Google’s digitization book program exemplifies how the concept of fair use is employed in US courts. This case is particularly important for this thesis because the way texts are stored and used in corpus construction closely resembles Google’s efforts in which it digitized copyrighted books. Google digitized over 20 million books without even attempting to obtain copyright permission. Presumably, Google was convinced that they were not legally infringing copyright in accordance with fair use exemptions to copyright law. The books digitized serve as a corpus or database, which holds all books on its server. Google uses an indexing facility, which allows the public to search key words or phrases and returns with a result of a number of books that contain such keywords. Instead of displaying the book in its entirety or displaying a considerable amount of any of the books, the result will contain only a snippet from the books.

The corpus created for this thesis will use a similar approach in the way copyrighted materials are stored and used for analysis. Using the same rationale, the designers of the largest freely available corpus, the Corpus of Contemporary American English, argue that their use of copyrighted materials does not infringe copyright on the same grounds as Google claims (COCA, 2014).

After careful study of the provisions of fair use and fair dealing as well as drawing upon

cases such as Google's book digitization program and the Corpus of Contemporary American English, I argue that the use of copyrighted materials for this thesis is fair dealing. Purdue University noted that users ought to "evaluate the use for each factor and then make a good-faith determination as to whether the use, in light of all four factors, can be considered fair use. If it is not fair use and none of the other exceptions apply, then permission must be obtained to use the material" (2014, para. 1). In what follows, I put to the test my own use in consideration of the six Canadian fair-dealing factors. I do this for two reasons. First, through this test, I intend to show that copyright is not infringed within the legal boundaries of the Canadian Fair Dealing Doctrine and that no copyright owner is negatively affected by the use of his/her material. Second, since this research builds a Model corpus for endangered languages, subsequent studies on the subject can build on this discussion in making their own decision when tackling the issue of copyright, which has always been a challenge for researchers. Below is my analysis using the six factors used to assess Canadian fair-dealing exceptions to copyright:

1. **The purpose of the dealing:** The purpose of the dealing with regard to this thesis is purely educational research and for non-profit use.
2. **The character of the dealing:** Randomly created sentences and words are used for linguistic analysis and therefore it is impossible for anyone to recreate the original work from the disjointed sentences and words manipulated in this thesis. Canadian fair-dealing doctrine also considers customary law as to how the text in question is normally shared. It is worthy to note that poetry, which contains a fair share of this corpus, is normally considered to be in the public domain.
3. **The amount of the dealing:** Only snippets of words and disjointed sentences that are not in any way central to the original work will be used in this thesis.

4. **Alternative to the dealing:** Corpus construction requires language samples. This means that it is not concerned with the content of any particular work. This makes all texts in the target language essential for a corpus to attain its research and development goals, which in turn gives no reasonable alternative.
5. **Nature of the work:** Corpus uses technology that transforms the original text to a mere manifestation of language use so whether the original text was published or unpublished, creative or factual, that original representation is no longer intact. The work has now taken a new form that bears little resemblance to its original version.
6. **Effect of the dealing:** Given the limited access and exposure to the material, no one can use the disjointed phrases, words or sentences quoted in this thesis as a substitute for the original work; therefore, there is no market effect of the dealing.

In a landmark victory at a federal court in the United States, *Authors Guild, Inc. v. Google Inc.* on November 14, 2013, the court ruled that Google had not infringed copyright in digitizing more than 20 million books without permission. The court was cognizant of the fact that Google's intent to digitize books under its Google Books Program was mainly commercial. Since the case is similar to corpus construction in general and to this thesis in particular, the concluding remarks by Judge Denny Chin summarize the main features and the rationale of the fair-use doctrine:

In my view, Google Books provides significant public benefits. It advances the progress of the arts and sciences, while maintaining respectful consideration for the rights of authors and other creative individuals, and without adversely impacting the rights of copyright holders. It has become an invaluable research tool that permits students, teachers, librarians, and others to more efficiently identify and locate books. It has given

scholars the ability, for the first time, to conduct full-text searches of tens of millions of books. It preserves books, in particular out-of-print and old books that have been forgotten in the bowels of libraries, and it gives them new life. It facilitates access to books for print-disabled and remote or underserved populations. It generates new audiences and creates new sources of income for authors and publishers. Indeed, all society benefits. (Authors Guild v. Google – Summary judgment decision, p. 26, accessed from [library.osu.edu](http://library.osu.edu))

Corpus construction especially for endangered languages does the same thing. It preserves languages, cultures and it promotes research and development of languages by making threatened languages and knowledge encoded in them more accessible to researchers and language technologists in a transformative nature, which does not affect the market for copyright owners. It is another Google-like initiative, which creates a win-win situation for everyone.

## Chapter Seven: Results and Discussion

### 7.1 Overview

This chapter gives the main results and their interpretation in light of the main objectives of this research. Practical implications of the findings are discussed. The main text components of the corpus as well as their relative representations in the corpus are described.

In addressing the research question, both quantitative and qualitative data analysis is used. Computer-generated statistics and lists are used qualitatively to interpret the meaning and significance of the results. It must be stated at the outset that the qualitative approach is inherently subjective, as it relies on experience and personal knowledge on the subject as well as my own bias. Bennett (2010) described how the analysis of a corpus uses both methods:

The corpus approach depends on both quantitative and qualitative analytical techniques.

This characteristic of the corpus approach highlights the importance of our intuition as expert users of a language. We take the quantitative results generated from the corpus and then analyse them qualitatively to find significance. (p. 8)

In analysing data for this thesis, corpus tools are selectively chosen for their ability to illuminate the research question in a way that generates hypotheses, sparks discussion, and steers future researchers in the direction where more study is needed.

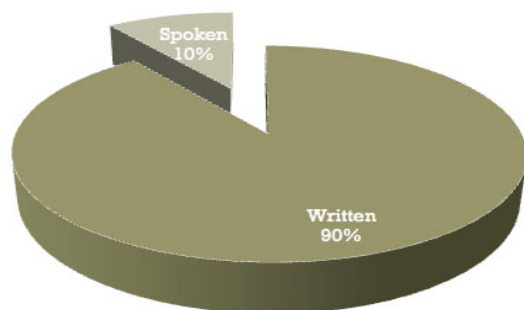
A prototype version of a proposed larger corpus for the Somali language (as discussed in Chapter Six) was constructed for this thesis. A total of 262 documents of various lengths were collected. Each text was included in its entirety. This decision is based on Sinclair's (2005) recommendation for whole text inclusion. For this thesis, size was considered to be of paramount importance since a large corpus gives a clearer picture of how a language is used in different contexts in texts and across texts. Through networking and the use of personal contacts, Somali



publishers and writers donated over half of the texts that are included in the corpus. Larger text types such as novels, short stories, poetry, translations, and religious works are among the texts donated. The remaining texts were harvested from the Internet. These online texts were purposefully gathered so that they contribute and add to the diversity of the donated texts. As discussed in Chapter Six, reviewing the Canadian fair-dealing and the US fair-use provisions, it has been determined that for corpus construction purposes, copyright has not been violated. Given the limited access and exposure to the copyrighted material used, no one can use the disjointed phrases, words or sentences in this thesis as a substitute for the original work, and therefore there is no market effect on the dealing.

Although constraints of text availability, time, and resources were unavoidable, a guiding principle was to keep issues of diversity and balance in mind during data collection. The end product is a corpus that contains 22 different genres. Together these text types form a general corpus of 865,214 tokens. A token is the total number of instances that are found in a text. For example, in the preceding sentence, we have 14 tokens. It contains two sub-corpora: written sub-corpus and spoken sub-corpus. Each sub-corpus, written or spoken, can stand alone from its relevant text. Of the 865,214 tokens, 777,261 tokens are from the written sub-corpus while the remaining 87,953 tokens are from the spoken sub-corpus. As shown in the pie chart below, these figures translate to 90% for the written and 10% for the spoken. Capturing spoken data has always been a challenge for corpus builders. Incidentally, BNC, which is considered one of the largest corpora, contains 90% written and 10% spoken (Atkins, Clear, & Ostler, 1992). The challenge lies in gathering spoken samples and transcribing them for inclusion.

Figure 7.1. Composition of the General Corpus

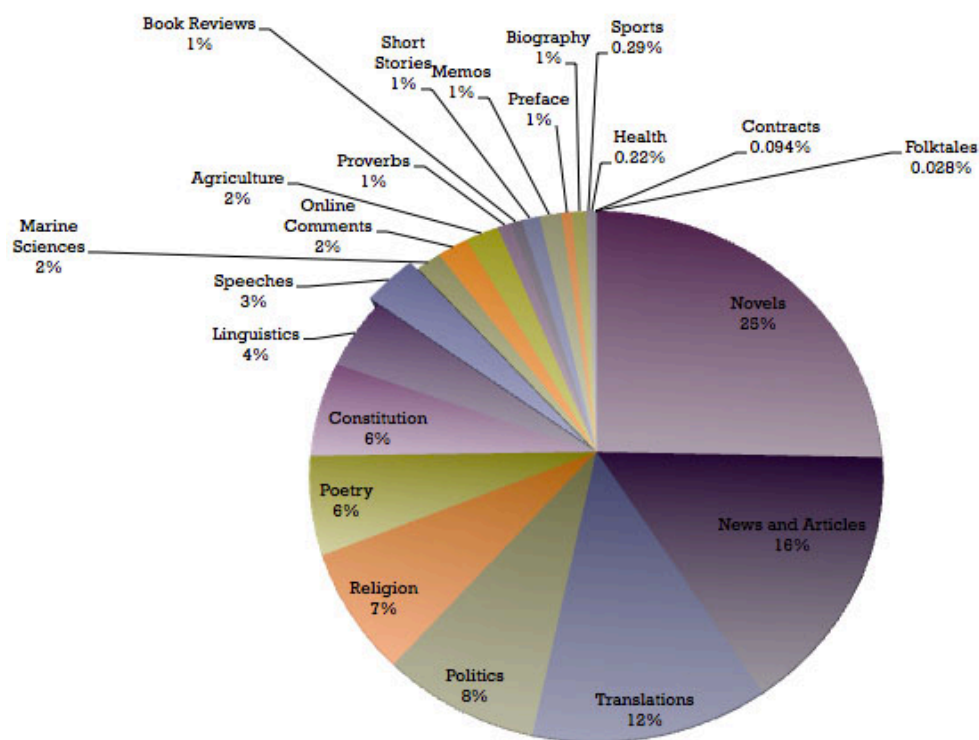


For practical considerations, all texts were chosen based on availability. However, issues of diversity, balance and size were the guiding principles in the search and selection of texts. Nonetheless, at this foundational stage, the latter two criteria (balance and size) have not been adhered to as desired, due to resource and time constraints. Every possible effort has been made to target a sense of diversity among texts, which has reasonably been achieved. This compromise is often seen as inevitable especially at early stages of corpus design (Atkins, 1991; Biber, 1993; Sinclair, 2005).

In all, 22 main text categories were collected. Each text category contains individual texts with identification tags showing the sub-corpus that the text belongs to as well relevant file names such as the topic of the text, the author, the date, the dialect and the source, whenever they are available. Each file was saved with such identification information. This strategy was followed to enrich the analysis of the results since they provide additional and insightful information, which assists the researcher to interpret corpus-generated results more accurately. For example, in examining how a particular word has transformed in terms of meaning or usage over a particular time period, the dates when the document was written or published were

crosschecked to illuminate how the word was used diachronically in different texts across time. The leading text type in terms of size is the novel with 28%, followed by news articles and translations with 16% and 12% respectively. These three text categories make up 53% of the corpus. Politics, religious works, poetry and the constitution make up 27% while all other text categories together form the remaining 10% of the corpus.

Figure 7.2. General Corpus Composition



Despite the imbalance in terms of size across texts, underrepresented texts have nonetheless contributed to the diversity of the corpus giving it a representation of 22 different text categories. It seems that a reasonable diversity has been achieved given the time and resource challenges. Pioneering works had to contend with similar challenges. For example, according to Kučera and

Francis (1979), the Brown Corpus of American English has an overall representation of 15 different text categories.

## 7.2 Poetry as a Wordbook in Disguise

In this section, we first consider the statistics of the corpus. Of the 262 files, 114 are written and 146 are spoken. They indicate the diversity of the texts. This means that different authors (with the exception of poetry) wrote the texts. There are a number of poets whose poems appear three or four times; however, each poem was saved separately since the topics and the dates might have been different.

Table 7.1. *Corpus Statistics*

Corpus Statistics	General Corpus	Written	Spoken
Files	262	114	146
Tokens	865,214	777,261	87,953
Types	87,453	74,601	24,409
Type/Token Ratio	10.10%	9.6%	27.7%

As Table 7.1 reveals, the general corpus contains a word count of 865,214 or as they are referred to in the field, tokens. On the other hand, the general corpus has 87,453 unique words or types. The general corpus has a type/token ratio of a little over 10%. This means that, on average, for every 10 words used in the texts, one new word is introduced. In other words, 10% of the corpus represents unique words or types while 90% are recycled or words that have been used in the texts repeatedly. The written corpus has a slightly lower type/token ratio of 9.6%; however, the spoken corpus has a considerably higher type/token ratio of almost 28%. As Sinclair (2005) pointed out, a higher ratio means a richer vocabulary, which also means more dense text. This indicates that the spoken corpus has almost three times as many types as there

are in the general corpus or in the written corpus. The discrepancy could further be interpreted that the spoken corpus has richer and lexically denser texts.

Studying it further, poetry, which dominates the spoken component, is likely to be a major contributing factor in the diversity of tokens. It is important to restate that (as discussed in Chapter Six) for this thesis, poetry is regarded as a spoken register since it originates in oral mode and is largely available orally. Owen (2001) corroborated this stance:

It is important to bear in mind that Somali poetry is still primarily experienced through listening rather than reading; there has been some publication of collections of important poetry (mostly of poets from the past), but these are not widely distributed at all. (para. 2)

This needs to be clarified because previous corpus constructions have regarded poetry as a written genre. Although this thesis builds on the design techniques of previous corpora, especially those built for the English language, one of the main objectives of this thesis was to adapt such techniques to other languages, in this case Somali, so that the design principles can be expanded or modified for use in challenging and diverse contexts. Therefore, incongruent design principles that defy application in the context of this thesis were expected.

There are a number of possible explanations for the higher type/token ratio produced by the spoken corpus. As noted above, the spoken corpus has a substantially lower overall word count when compared to either the general corpus or the written corpus. For example, the spoken corpus represents only 10% in the general corpus. Such disparity might have contributed to the finding that the spoken corpus has a relatively higher type/token ratio than the written corpus. Scholars in the field noted the effect of disproportionate text sizes, acknowledging that, as the sample size increases, the type/token ratio of the corpus decreases (Biber, 1993; Sinclair, 2005;

Tognini-Bolleli, 2010). Therefore, there is a possibility that size could be the main factor or could, at least, have partly contributed to this finding.

Nevertheless, in order to explore other possible causes, a further investigation was needed. To neutralize the effect of size, a decision was made to compare an abridged genre from the written corpus against the poetry genre. The closest genre in terms of size, “constitution,” was selected from the written sub-corpus for re-evaluation and comparison. The two genres were reduced to equal sizes of 50,420 tokens each, and their type/token ratios were then recalculated. It was interesting to learn that the two normed texts still revealed similar results. Once again, poetry demonstrates a significantly higher type/token ratio of a little over 34% relative to the constitution genre, which stands at about 11%. Contrary to our previous hypothesis (that size might have contributed to the higher type/token ratio of poetry texts), the outcome of the latter experiment suggests that the effect of unequal sizes might have contributed only marginally to the discrepancy found in the type/token ratios.

Table 7.2. *Normed Type/Token Ratios*

Genre	Token	Types	Type/Token Ratio
Constitution	50420	17369	11.1%
Poetry	50420	5628	34.4%

Yet another possible explanation is that poetry, which dominates the spoken corpus, is generally considered to have relatively greater lexical density and richer vocabulary. This difference might inherently exist between poetry and other texts. In analysing the difference between poetry and prose, Guuled (1976) observed:

*Waxay kale oo ay kaga geddisan tahay dhiska ereyadu weedho u aloosaan marka tix maasno la curinaayo taas oo keenta cufis, koobnaan iyo madhaxsi hadal. Si kale waxa ay u tahay miisaanka iyo shubaasha maansada ayaa soo oodaya awooddii ereyada qubanaha la taraarixin jirey. (p. 13)*

Another factor that sets it [poetry] apart is the structure by which words formulate a set of rule-governed lines, which in turn yields lexical density, brevity and efficiency. In other words, its meter and alliteration delimit the means through which words are used freely, as in the case of prose composition. [Translation is mine]

Note that the difference Guuled described exists not only between poetry and one particular genre such as “constitution,” but the dichotomy he illustrates is true for poetry on the one hand and prose on the other. Essentially, the poetic language indicates that since poets have to follow established rules, they are forced to use words resourcefully so that they convey the intended meaning in an effective manner. Conveying the intended meaning effectively might well mean transmitting a line in a poem with multiple meanings. In prose, the writer has the freedom to use words, recycle them or structurally connect them, as s/he deems appropriate.

This analysis strengthens the finding that poetry is characterized by a richer and more diverse vocabulary than prose. It is beyond my understanding how some Somali poets still compose such sophisticated and rule-governed poems spontaneously. Such poems have a name in the Somali language *Gabay Golekafiul ah* which roughly translates as “spontaneous poem.”

In the preceding quote, Guuled (1976) described the formula-governed word structure that characterizes Somali poetry. Unlike prose writers whose main task is to convey meaning through pragmatically appropriate words in their respective register, poets do not have the luxury to choose words as they want; they have to conform to strict rules and conventions. Another scholar who has written about Somali literature, Johnson (1988), echoed Guuled's analysis, substantiating it with an example of how such constraints play out in composing poems:

Apart from the units of quantitative scansion, another constraint, which is truly amazing about classical and other types of Somali poetry, is its alliterative pattern. The *gabay*, *jiifto*, and *geeraar* [these are names of different types of poetry] require that only one sound per poem may alliterate, a sound being defined as a single constant or all the vowels collectively. A 150-line *gabay* [Somali word for poem], for instance, will have 300 words that alliterate with the same sound. (p. 49)

The requirement of alliterating sounds requires the poet to use a new word. This rule sheds some light on the statistically higher type/token ratio and our latter hypothesis that poetry has an inherently rich and lexically dense vocabulary. The implication is that poetry could contribute a variety of words to a corpus. This thesis does not concern the lexical analysis of Somali poetry but the line of argument we are pursuing is how poetry could enrich the vocabulary of a corpus. One of the practical implications of this finding is the development and the documentation of Somali vocabulary and other endangered languages through vocabulary and other richer genres such as poetry. It appears that poetry has rich vocabulary to offer a corpus and therefore its inclusion in corpus construction is essential. Poetry offers words that are infrequent and these are words that might be at risk of dying.



At this point, a more focused experiment will help in triangulating evidence presented thus far in favour of poetry as source of rich vocabulary. For this investigation, a poem composed by Maxamed Maxamuud Fidhin in the 1990s about reconciliation and peace in Somalia is used. The type/token ratio supports previous findings and the analysis given by the authors cited. The poem has a word count of only 434 tokens with 321 types. The relative ratio of types to tokens is that 74% of the tokens are types or unique words. In other words, this means that there are 7.4 types in every 10 tokens. As expected, we observe that there is a discrepancy between this ratio (74%) and the one we previously generated (34.4%) for all the poems in the corpus. As discussed previously, the difference might be due to the substantial difference in size. The former is based on the token count of all the poems in the corpus while the latter is based on the token count of only one poem.

Table 7.3. *Ratio of single Poem*

Genre	Types	Tokens	Type/Token Ratio
Poem composed in 1990s	321	434	74%
By Maxamed M. Fidhin			7.4:10

For this investigation, I wanted to go beyond the numbers to study the poem visually. Table 7.3 illustrates that out of the 321 types, only 49 of them have been recycled throughout the poem. This means after the cut-off at the word 49, the poet starts to introduce new words. This pattern continues until the last new words are used at number 321.

Figure 7.3. Visual Representation of Type/token Ratio

44	2	wada
45	2	wadaad
46	2	walaahoobay
47	2	wanaag
48	2	waxaad
49	2	wedkiyo
50	1	aan
51	1	aasiyo
52	1	abid

The point I have been trying to emphasize is how poetry can serve as a good source of language development and the construction of a corpus. As the evidence indicates, vocabulary can be a useful tool for language development. Browsing the following wordlist from a single poem, a clear resemblance to a wordbook surfaces. All the words on the list are in alphabetical order and each word is entered only once, as in a dictionary.

Figure 7.4. A Poem Wordlist

Word Types: 321			Word Tokens: 434	Search Hits: 0
Rank	Freq	Word	Lemma Word Form(s)	
153	1	kolleey		
154	1	laba		
155	1	lagu		
156	1	lahaa		
157	1	laynaga		
158	1	laynay		
159	1	laynoo		
160	1	laysku		
161	1	laysla		
162	1	leh		
163	1	loogaga		
164	1	loogu		
165	1	madhinay		
166	1	maray		
167	1	mareeg		
168	1	marka		
169	1	markii		
170	1	meel		
171	1	midba		
172	1	mooda		
173	1	nabad		

Studying these words further, the word *mareeg* at number 167 caught my attention. Translated as “rope or strap,” this word is particularly interesting because it has no use for urban dwellers. However, *mareeg* serves a functional purpose in the country. A rope or strap is used to put around the neck of an animal to tie it to a tree overnight. A new function has recently emerged and I was interested to examine through concordance lines the status of the new meaning of the word. In the last two decades the word has evolved to take on a new meaning. Many native speakers now use *mareeg* to refer to “website.”

Figure 7.5. Mareeg in Transition

LINE	CONTEXT
1	Barre sare Maxamed Bulaale oo u waramayey bahda mareegta Farshaxan. Abwaan Xasan Sh.Muumin (IHUN)
2	Maansadii Abwaan Jamaal Cali Xuseen Raadraac: -Mareegaha Baardheere, Farshaxan, Hadhwanaagnews, O
3	Wareysi Qaade: Bahda Mareegta Farshaxan - Garnaqsi Miisaanka Maansada
4	Wareysi Qaade: Bahda Mareegta Farshaxan - Buug : Hal-bixinta Erayada
5	Fu,aad Sheekh Bahda MareegTa Farshaxan www.farshaxan.com ama ww
6	inta maansada. Hal-abuurkeedu wuxuu ku soo baxay mareegaha (website) Soomaaliyeed sannadkii 2008.
7	ida Gaarka ah Loogu soo Diro ee aan Lagu Daabicin Mareego kale. Gooni isku-taaga tuulooyinka iyo j
8	urin isugu jirta, kana mid ah kuwa ugu da'da wayn Mareegaha Soomaalida. Waxayna aad isugu howshaa i
9	ada cayadu moostay iyo meydalka hillaabay Ililada mareegaha ku dhigay miss cadkiiyo ciidda Iyagoo m
10	a intii laga xidhoo waadi lagu tuuray Sidii waxar mareeg galay maxaan kaga wijiiaqiinay Alla maxaan

Of the ten contexts listed, two examples from different poems refer to how the word is still used by livestock herders [lines 9 and 10]. All the other examples refer to the newer meaning, website. In line 6, a writer who seems to be conscious of the fact that the word has not yet been widely accepted, types the equivalent word “website” beside the word *mareegaha*. The addition of the suffix “aha” serves as a determiner indicating a particular website. Words constantly evolve in a living language and their changes ought to be tracked and documented to encourage language growth. This is a major objective of this thesis, so we will revisit this issue later in the chapter.

### 7.3 Additional Research Instruments

This thesis employs two Somali dictionaries as references. They are both considered to be among the best monolingual dictionaries available in the Somali language. In this chapter, reference will be made to these works to illustrate ways in which the dictionaries could be improved through the use of technology. The first one is *Qaamuuska Af Soomaaliga* by Yaasiin C. Keenadiid (1976). Incidentally, this dictionary was published about four years after the Somali orthography was implemented and so is considered a foundational work on which subsequent dictionaries were based. Keenadiid (1976) predated the computer age so the dictionary was laboriously compiled with no support from technology. The second dictionary,

which is also entitled *Qaamuuska Af Soomaaliga*, was published in 2012 by a team of lexicographers and content specialists led by Annarita Pugllieli and Cabdalla Cumar Mansuur of the Centre of Somali Studies at the Università Degli Studi Roma Tre in Italy (Pugllieli & Mansuur, 2012). The second dictionary was part of a joint project called the Somali Research Project between the Italian Ministry of Foreign Affairs and the Somali National University. This project commenced in 1977-78 and the dictionary was among joint projects on Somali language research and development initiatives. The compilation of the dictionary was based on previous works such as the 15,000 words from Keenadiid's (1976) dictionary as well as words gleaned from textbooks and terminology documents published by the then Somali Ministry of Education. The project was unfortunately halted as a result of the collapse of the central government in 1991. Some of the documents were housed at the Somali National University and were lost during the civil war. The project was then reinstated in the late 1990s with the documents that had been saved (Pugllieli & Mansuur, 2012). The introduction contains no reference to the dictionary being corpus-based; therefore, it is appropriate as a reference to illustrate how a corpus as a tool enhances the quality of these and similar dictionaries based on human knowledge and intuition.

While undoubtedly the lexicographers of these two landmark dictionaries have contributed to the development of the Somali language, their work could nonetheless be improved through the use of technology. Editors of the latter dictionary acknowledge the following in their preface:

“Qaamuuskan, dabcan ma aha mid matalaya heerka kamadambeyska ah ee loo samayn karo qaamuus af-soomaali, laakiin waxaan filaynaa in uu noqdo mid si wacan looga faa’iidayso karo mustaqbalka xagga diyaarinta qaamuusyo midkan ka baaxad weyn. Si

kastaba ha ahaatee, qaamuskan halka luqadeed qaabkiisa, marka la eego xagga daraasaadka guud ee afafka Afrikada madow, waxaa la oran karaa in uu ka mid yahay qaamuusyada fara kutiriska ah. (Pugllieli & Mansuur, 2012, p. XI)

[T]his dictionary by no means represents the ideal or the final version of a monolingual dictionary for the Somali language, but we hope that it becomes highly beneficial especially for the compilation and the publication of improved and larger dictionaries in the future. However, this monolingual dictionary is arguably one of the few of its kind in terms of its overall linguistic studies when compared to other dictionaries available in other languages in the sub-Saharan Africa. (Pugllieli & Mansuur, 2012) [Translation is mine.]

As the authors rightly stated, their dictionary has been instrumental in identifying how traditionally compiled dictionaries like theirs could be improved. This study expands on their work by identifying areas where improvements are needed and how a corpus as a technological tool can aid material writers.

#### **7.4 The 20 Most Frequently Used Somali Words**

One of the most useful features afforded by corpus for language research and development is the generation of frequency lists. A frequency list reveals the words that are used in a corpus together with the number of times a word is used in the corpus. It has been of particular interest to uncover the words that would be included in the most commonly used Somali wordlist based on a corpus of over 865,000 words or tokens from a wide variety of Somali texts. During the course of data collection for this research, there has been speculation about the words that would occupy the top 20 positions on the frequency list. More specifically, the researcher and colleagues who were aware of this research as well as family members have

all informally presented a word that they hypothesized as a candidate for the top position in the frequency list. Apparently, such interest and discussion among community members who knew about this study was sparked by the fact that this vital information about the Somali language is, to my knowledge, unavailable. Therefore, among other important points of interest, the most highly anticipated finding of this study was to learn the words that are most commonly used in the Somali language. In the literature, Kilgarriif (1997) substantiated the importance of frequency lists:

A central fact about a word is how common it is. The information is particularly valuable for language learners, as it immediately indicates how important it is to learn. With the advent of large computerized language corpora, it is for the first time possible to meet the demand. (p. 135)

It was rather surprising to learn that the top position in the frequency list is occupied by a word other than the words presented by the community as candidates. The research finds that the Somali word “*oo*” is in fact the most commonly used word in the Somali language, based on the corpus built for this thesis. However, all the words that were put forth were found to be among the top 50 words. I was probably influenced by my exposure to English frequency lists, so I hypothesized “*oyo*”, which is equivalent to ‘and’ in English to top the list. In fact, “*oyo*” occupies the second position on the list. Other words presented as candidates included “*waxaa, maxaa, soo*”, which may be translated as “that, what, towards/to” respectively. These words all serve as grammar or function words in the language. The table below shows the 20 most frequently used words. The first column displays a frequency list generated from the General Corpus or Somali Language Bank (SLB). Henceforth, this corpus will be referred to as SLB. In the second column are the written and spoken sub-corpora respectively. From each of these sub-corpora, the two

genres that have the highest representation in each sub-corpus in terms of tokens are also shown. It is evident that the word “oo” ranks first on five out of the seven lists shown. It was a little surprising that the word was pushed down the list to number seven in the poetry sub-corpus, which may explain why the word was also pushed down to number two in the Spoken Corpus in which the poetry genre belongs. Note that “speeches” which is one of the two genres together with poetry that dominate the Spoken Corpus has the word “oo” in the top position. Therefore, the anomaly seems to rest with the poetry genre. One possible explanation for this might be that poetry is structurally and lexically different from other genres in the corpus. For example, Somali poetry is known for its strict metrical scansion and alliteration as well as its lexical richness (Johnson, 1988).

Table 7.4. *The 20 Most Frequently Used Somali Words in the Corpus*

Rank	General Corpus	Freq	Written	Freq	Spoken	Freq	Poetry	Freq	Speeches	Freq	News & Artic.	Freq	Novels	Freq
1	oo	22557	oo	21021	iyo	2027	iyo	1178	oo	885	oo	3719	oo	617
2	iyo	17827	ka	16189	oo	1536	ka	744	iyo	736	ka	3076	ku	489
3	ka	1762	iyo	15800	ka	1432	waa	729	u	588	ay	2873	ka	453
4	ku	16649	ku	15323	u	1406	u	720	ka	557	ku	2640	u	450
5	u	14927	ay	13864	ku	1326	ku	665	ah	550	iyo	2569	ay	434
6	ay	14491	u	13521	waa	1121	la	615	ku	531	ah	2479	iyo	367
7	ah	12854	ah	12034	la	1083	oo	612	in	529	u	2251	in	336
8	in	12133	in	11423	ah	820	soo	374	ay	442	in	2050	soo	335
9	la	11561	la	10478	aan	756	lagu	344	aan	379	la	1849	uu	269
10	soo	9343	ee	8742	in	710	aan	336	ee	362	ee	1699	la	255
11	ee	9282	soo	8689	soo	654	loo	315	la	333	soo	1292	baa	224
12	uu	8724	uu	8328	ay	627	ma	300	soo	246	waa	1093	aan	217
13	waa	7610	waa	6489	ma	599	e	281	waa	213	waxa	985	waa	191
14	aan	6601	aan	5845	ee	540	baa	276	waxaa	202	uu	984	ah	184
15	waxaa	5783	waxaa	5404	lagu	488	waxaan	262	waxaan	199	aan	945	bay	177
16	lagu	4736	lagu	4248	loo	485	ah	253	uu	168	waxaa	838	ma	127
17	waxay	4441	waxay	4242	waxaan	465	lahaa	244	aad	162	loo	781	ee	126
18	loo	4325	aad	3897	uu	396	laga	233	si	158	lagu	778	ugu	124
19	aad	4200	loo	3840	baa	388	kala	230	badan	137	ayaa	749	buu	117
20	ugu	3893	ugu	3736	waxaa	379	wuxuu	217	wax	135	waxay	732	aad	111
		193699												
		22%												

It has been found that these 20 words make up a remarkable 22% of the words in the SLB corpus. The result was calculated by dividing the total frequency of these 20 words with the total



tokens or words in the general corpus multiplied by a hundred:  $193699/865214*100$ . Only 20 words to constitute 20% is a significant proportion, which might surprise some readers. Given that the top word alone, “*oo*”, is used in the corpus 22,557 times, a further investigation was warranted to discover whether a finding of this nature correlates with previous corpus constructions. For this investigation, the Corpus of Contemporary American English (COCA) was chosen and the frequency of the top 20 words in COCA was listed. As discussed in previous chapters, COCA is considered one of the most diversified, up-to-date and largest corpora available with 450 million tokens (Davies, 2009).

A similar procedure was followed in the calculation of the proportion held by the top 20 words in COCA by dividing the total tokens in the corpus by the total frequency of the top 20 words:  $450,000,000/119399109*100$ . As the following table reveals, the top 20 words in COCA represent 26.5%. Even though the two corpora are considerably different in terms of size, diversity and balance, it was interesting to note that the two corpora are within reasonable range of the relative proportion of the top 20 words in their corpora, 22% and 26.5% respectively. Schmitt (2000) reported that the most frequent 50 words represent 36% of the all the texts in the Bank of English with a corpus of over 200 million words. It appears that SLB-generated finding is closely aligned with findings generated by other corpora. This finding indicates how human languages share remarkable similarities.

The most significant finding is not how the numbers are within a reasonable range of 4% but how the words in the top 20 lists in the two corpora share remarkable similarities. This thesis is more concerned with the identification of ways in which languages could be developed, and this finding validates the previous discovery that function words are in high demand (Biber, 1993; Kilgariff, 1997; Sinclair, 2005). Underscoring this phenomenon, Schmitt (2000) stated, “A

second insight is that the most frequent words in English tend to be grammatical words also known as function words” (p. 73).

A brief discussion of the similarities of the two lists is desirable. First, both lists are occupied exclusively by words that have grammatical functions in their respective languages, English and Somali. It has been interesting to note how the words share a close affinity in terms of their function and meaning despite the dissimilarities of the two languages. In the COCA list the word ‘the,’ which ranks first has a similar function to “*ka*” in the SLB list, which is in the third slot. For instance, in the Somali language, “*nin*” means ‘a man’ but when the man is known, the suffix “*ka*”, which has a function similar to the definite article ‘the’ in English is added. It then becomes “*ninka*”, which translates as ‘the man.’ Another striking similarity is that ‘and,’ which ranks third in the COCA list has the same meaning as “*iyoo*” in the SLB, which ranks second. Even the word “*oo*”, which ranks first in the SLB list has functions in the Somali language similar to the words ‘that,’ ‘and,’ ‘of’ which are all in the top 20 in the COCA list. A more in-depth discussion is devoted to the word “*oo*” later in the chapter.

Table 7.5. *Corpus of Contemporary American English (COCA) and Somali Language Bank (SLB) – COCA source: Davies (2008).*

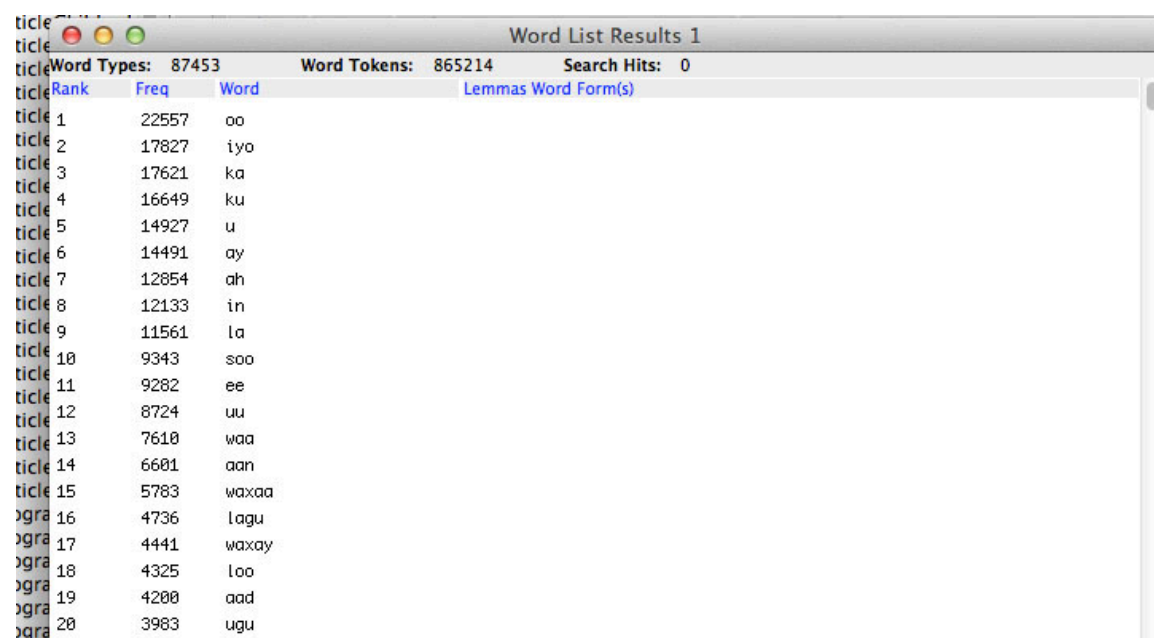
<b>RANK</b>	<b>COCA</b>	<b>FREQ</b>	<b>SLB</b>	<b>FREQ</b>
1	the	22038615	oo	22557
2	be	12545825	iyo	17827
3	and	10741073	ka	1762
4	of	10343885	ku	16649
5	a	10144200	u	14927
6	in	6996437	ay	14491
7	to	6332195	ah	12854
8	have	4303955	in	12133
9	to	3856916	la	11561
10	it	3872477	soo	9343
11	I	3978265	ee	9282
12	that	3430996	uu	8724
13	for	3281454	waa	7610
14	you	3081151	aan	6601
15	he	2909254	waxaa	5783
16	with	2683014	lagu	4736
17	on	2485306	waxay	4441
18	do	2573587	loo	4325
19	say	1915138	aad	4200
20	this	1885366	ugu	3893
		<b>119399109</b>		<b>193699</b>
		<b>26.5%</b>		<b>22.3%</b>

The implications of this discovery are vital because they address the language development issues that this thesis set out to pursue. First, the identification of the most commonly used words is beneficial for the development of language-learning materials such as dictionaries, grammars, and language-learning textbooks. Pedagogically, these are the words that learners need to learn first or early in their language learning. For curriculum writers and lexicographers, for example, these are the words they need to give more comprehensive coverage by providing the learner with all possible functions that each word has in the language as well as

giving extra information such as how common the word is. In illustrating how this discovery could be exploited in terms of language development, the word “*oo*” is given as an example. The hypothesis is that technological tools such as corpus tools are needed to generate a list like the preceding one with relevant statistical information so that corpus tools can be exploited for language development purposes in a way that the human mind is incapable of.

The following is a visual representation of the frequency list query derived from the general corpus. The list first displays general statistics about the word types and their corresponding tokens. It then shows the rank of each word in descending order, the frequency count, and the word itself. Note that lemmas or word forms are not given because this corpus used raw data, which means that the words have not been annotated or lemmatized. As discussed in Chapter Six, for practical design considerations, annotation is reserved for phase four of the SLB development. At this stage, using raw data is sufficient for the purposes of this thesis. The objective was to explore and illustrate language development and documentation possibilities that corpus construction can offer for endangered languages.

Figure 7.6. Corpus results of the 20 Most Frequently Used Somali Words



Rank	Freq	Word	Lemmas	Word Form(s)
1	22557	oo		
2	17827	iyo		
3	17621	ka		
4	16649	ku		
5	14927	u		
6	14491	ay		
7	12854	ah		
8	12133	in		
9	11561	la		
10	9343	soo		
11	9282	ee		
12	8724	uu		
13	7610	waa		
14	6601	aah		
15	5783	waxaa		
16	4736	lagu		
17	4441	waxay		
18	4325	loo		
19	4200	aad		
20	3983	ugu		

For this investigation, the two dictionaries discussed earlier are compared with the SLB corpus.

The discovery of the most frequent words has a number of practical implications. In pursuit of our argument that corpora can serve as development and research tools for endangered languages, we proceeded to look up the most frequently used word “oo” in the most widely used Somali dictionaries. The experiment was to compare intuition with the knowledge-based entry of the word to an entry aided by corpus tools. The wordlist or frequency list feature was used to explore words that deserve a more comprehensive coverage in the dictionaries due to their high demand in the language.

We first looked it up in Keenadiid (1976). As suggested earlier, Keenadiid is considered a pioneer in the fledgling industry of Somali lexicography. Consistent with our hypothesis, there is a disparity between the prominent position the word is given by the frequency results and the

rather brief and underdeveloped coverage it receives from both dictionaries. Keenadiid (1976) gave two definitions, as follows:

Figure 7.7. *Keenadiid: Exploring “oo” in a non-corpus based dictionary*

**Oo** — 1. Erey labo tilmaamood ama labo tixood xiriiriya (sida «nin wanaagsan *oo* deeqsi ah»; «aqal weyn *oo* magaalo ku yaal»); ama inoo sheega si kolkaas lagu sugan yahay (sida «Cali *oo* buka»; «lo’dii *oo* cabtey»); ama ulajeeddooyin kale abuura (isu eeg: «*wiilka kale*» iyo «*wiilka oo kale*»). ~ ee. 2. Xarafka «o» **oo** dheer; (ō).

An English translation of the definitions is given below:

1. A word that connects two adjectives (*such as “man nice **oo** is generous; this is the literal meaning and therefore the closest translation could read like this: “A nice man **who** is generous.” Another example is “A large house **which** is in a city”). “Oo” also gives us information of how things are at any given moment (like “Ali **who** is sick”). “Oo” also has a different purpose (Compare: “the other boy” and “the boy like the other.”)*
  2. “Oo” is a long vowel corresponding to the short vowel “o.”
- [Translations are mine. My interpretations are italicized].*

In an effort to differentiate translations directly from the entry and my own renditions, all italicized words and phrases are my own interpretations of what I thought the author meant. However, dictionary definitions are presented verbatim although I interjected my own

commentary so that the definitions make sense for the non-Somali reader. For example, a word that connects two adjectives is a direct translation of how Keenadiid puts it in the entry. It appears that the connotation gleaned from the definition is that the word “oo” is equivalent to the conjunction “and” in English. However, the example sentences make more sense when used with relative pronouns “who” and “which,” respectively. The use of the relative pronouns is emphasized to show that they may be more appropriate than a conjunction, which is an issue we will explore further in the following paragraphs.

Now, I turn our attention to an entry of the same word, “oo”, in a more recent and expanded publication. Below is a screenshot of the entry in Pugllieli and Mansuur (2012), which is perhaps the largest Somali-Somali dictionary available. Interestingly, the larger dictionary also gives two definitions of “oo” although the definition it gives is somewhat different from Keenadiid’s (1976) definition. A translation of the entry is rendered below:

**OO** Conj 1. (Grammar) Conjunction used to connect two independent clauses.

Example: “Ali is a nice man oo is generous” *Conjunctive words like “and” seem not to give the intended meaning, therefore a more suitable translation reads: “Ali is a nice man who is generous.”* Another example is given, and here is the literal meaning:

“Ali oo was in Mogadishu I met.” *A holistic meaning of the same example is: “I met Ali who was in Mogadishu.” A third example could be translated literally as follows:*

“At a time when you don’t have shade, has someone sought shade near you?” A holistic meaning reads like this: *“At a time when you don’t have a roof over your head, has someone sought refuge with you?”*

2. (Grammar) Conjunction used to connect two actions, which are performed simultaneously. *No examples are given for this definition.*

Evidently, the two dictionaries seem to be in agreement that the grammatical function of the word *oo* is mainly a conjunctive function. However, as noted in the commentary in the translated text, new functions emerge. The grammatical function of a relative pronoun seems to be at work. Relative pronouns in the sentences, such as “who” and “which,” are highlighted in the translations. Review the entry taken directly from Pugllieli and Mansuur (2012) below.

Figure 7.8. Pugllieli and Mansuur (2012): *oo* in a non-corpus based dictionary

**oo** *xi*. 1. (*nax.*) Xiriiriye loo adeegsado in laysku xiro labo weer tabineed oo abyoon. Tus. “Cali waa nin wanaagsan oo deeqsi ah”, “Cali oo Xamar joogay ayaan la kulmay, adoo harsan waayay ma lagu soo harsaday?”  
2. (*nax.*) Xiriiriye isku xira laba fal oo isku mar la qabtay.

Thus far, the experiment has shed some light on how manually compiled dictionaries have covered frequently used words. Despite the high frequency of the word *oo*, we observed that both dictionaries give only two definitions. Our expectation was to find more comprehensive coverage with more meanings. This is corroborated by Vadasy and Nelson (2012), who stated, “The more frequently a word appears, the more meanings it is likely to have” (p. 34). Bogaards and Laufer-Dvorkin (2004) elaborated on why this is the case: “Words with several meanings, polysemes or homonyms may appear higher up on frequency lists than monosemous words by virtue of the combined frequencies of their multiple meanings” (p. 91). This argument has been supported by how frequent words in English are similarly covered in corpus-aided dictionaries.



Collins Online Dictionary (2014) gives the word “that,” in the top 20 words, 19 different definitions. It also gives some of the most frequent words such as “and,” “the,” and “be” 10, 15, and 9 different meanings respectively.

With this hypothesis in mind, we now use another useful tool Keywords In Context (KWIC), which gives queried words in real contexts.

### **7.5 Language Development Through Key Words In Context (KWIC) Tool**

While the frequency tool helps us to explore the significance of words in the language, the concordance tool or Key Words In Context (KWIC) helps us to learn how words behave in a language. Below is a visual representation of the 22,557 instances generated for the word *oo*, which is a long list displaying naturally occurring sentences in a variety of contexts. Figure 7.6 is intended to give the reader an idea of what KWIC results look like. More in-depth research is perhaps needed to study such a long list to derive conclusive results. Such a study is beyond the scope of this thesis. Therefore, our objective is to explore the possibilities with some of the tools available and describe how these findings could be used to improve existing language materials such as the two dictionaries used in this thesis and other relevant language materials. The proposed improvement also serves as a catalyst for discussion on the topic. It is hoped that the debate and the discussion that this thesis generates will lead to further studies that produce more conclusive results. For this thesis, an attempt will be made to deduce from the corpus KWIC results, a tentative dictionary entry for the word *oo*, which can be used to aid grammar textbooks.

Figure 7.9. A Sample of KWIC Results for the Somali Word *oo*

AntConc 3.3.5m (Macintosh OS X) 2012	
Concordance	
Concordance Plot	
File View	
Clusters/N-Grams	
Collocates	
Word List	
Keyword List	
Concordance Hits 22557	
Hit	KWIC
2220	kii 77 waxaa sida uu general Maxamed Cali Samatar oo ahaa ninka dhinaca ciidanka ugu darajada sarree
2221	liyada, 5. Madaxweyne Cabdirashiid Cali Sharmake oo ahaa ninka ugu dhawaa xagga Diinta iyo akhyaarn
2222	axa xusid mudan inaan halkaan ku amaano Wiil waal oo ahaa ninki dejiyey maahmaahda qaabkeeda dambe e
2223	u sheegay in sannadkii 2001dii, dhalinyaro badan oo ahaa qaxooti, aan cidina wehelin, oo dalka Ing
2224	ulkeeduna wuxuu ahaa mid liita. Dakhliga dawladda oo ahaa qiyaastii 5-8% ee wax soosaarka dhaqaalah
2225	dii Barnaamijka Ballaaran ee Xasilinta IGAD, kaas oo ahaa qorshe labadaas waddan ee dariska ah- iyag
2226	axaa beddelay oo xilkii kala wareegay F.W. Dekler oo ahaa qunyarsocod aaminsan in dadka madow si dad
2227	belaaayo asiibto. Nin la oran jiray Yuusuf Badane oo ahaa reer Aadan-Yabaal ayaa ku shirbay: Allihi
2228	gteer nimanka geliya taariikhda. Gabayaa Dallaayad oo ahaa reer Hargeysa ayaa isna seben hore ku maan
2229	) ee MSBarre, waxaa iska horyimid Ayaanle MSBarre oo ahaa RW dahsoon iyo Maslax MSBarre oo ahaa tali
2230	markii aay albaabka soo garaaceen ka furay . Cali oo ahaa saaxiibka koowaad oo Caddaawe aad buu u so
2231	Sagal aan wada rumaysan bay waxay wacday Axmed oo ahaa saaxiibkay aan baray anigu. Salaan ka dib
2232	iis qabaaalka magacuusu ka wada muuqdo iyo Walshe oo ahaa sarkaalkii ugu horeeyay 1884 ee maamulka m
2233	ajesty's Consul-General, Somali Coast Protectorate oo ahaa sarkaalka Cadan ka maamulayay xeebaa So
2234	xaan soo xaqiifiyay inay ahaayeen dad la yaqaanno, oo ahaa shaqaale dib u dhis ku sameynaayay guryo i
2235	dankeedii iska kala tageen ayuu amiirkii culimadu oo ahaa Sheekh Cabdirashiid Macallin soo qaban qaa
2236	iisa muraad ka jiro oo la dhigayo xuska aabbihiis oo ahaa sheekh weyn oo magaalada oo dhan laga qadd
2237	asada laga eego hadba marxalada ay ku jirto, kaas oo ahaa; sugitaanka ammaanka gobolka iyo la dagaal
2238	nle MSBarre oo ahaa RW dahsoon iyo Maslax MSBarre oo ahaa taliyaha XDS. Markii la xamili-waayey kade
2239	are qaarkeed u wadeen in uu Beddelo Gebre Salasse oo ahaa taliyihii guud ee sirdoonka, oo toddobaata
2240	a kooban yahay labadoodii magac ee ay kala wateen oo ahaa; Tanganika iyo Zanizibar oo mid walba sadd
2241	- waa xilligii Shillinka Soomaaligu qiimaha lahaa oo ahaa US\$1 = 7Sh.So.). Haddii xad gudubku soo no
2242	u daari jirtay. Shariif Macow iyo Xaajiyo Xaliimo oo ahaa waalidkii Canab waxba kama ogayn amuuraha
2243	cabsida ay qabaan carruurta aan cidina wehelin oo ahaa waalidkood oo hortooda lagu dilayo, ilmo

The possibilities with KWIC results are numerous. First, the KWIC results provide us with abundant real-life examples in different contexts. The implication is that a language developer with this facility has the means to view the language under study in a way that is not feasible using a manual approach. Second, KWIC results can be used in the language-learning classroom. A teacher with this facility is able to teach students with examples taken from real-life contexts, giving the students a more up-to-date and comprehensive coverage of how the language is used by its speakers. For our purposes, however, we illustrate our point by providing a non-exhaustive list of the possibilities that could be explored with KWIC results. Frequency information could be given by using various indicators such as a star system, check marks, and a colour-coding system; alternatively, a word or a short phrase could also be used. Examples that illustrate

common usages of the word were studied. Through this examination, a pattern, which suggests a particular meaning or function, emerges from the set of examples studied. A tentative meaning is then derived and each meaning is supported by two examples that were taken from the corpus concordance. Most common phrases associated with the word are given. We illustrate this with an improved dictionary entry for the word “*oo*” with a putative entry:

**OO – Grammar functions \*\*\*\*\*Most Frequent Word (Rank1)**

1. **Relative Pronoun:** Refers to and modifies the head noun—who/which/that

Ex: *Cali oo ahaa saaxiibka koowaad oo Caddaawe ayaa u yimid Asli/* Ali **who** was Caddaawe’s closest friend came to Asli.

Ex: *Muqdisho oo ku taal xeebta Badweynta Hindiya/* Mogadishu, **which** is located on the coast of the Indian Ocean.

2. **Conjunction:** And/ in addition to—used to connect two adjectives and to connect a sequence of verbs.

Ex: *Barayaashu waa kaalmayn jireen carruurta dalka Sweden, waa kaxeeyaan oo xannaaneeyaan/* Teachers used to assist children in Sweden **and** they take them out and they take care of them.

Ex: *Yagleel golaha waxbarashada ee sare oo hawlgaliya/* Establish the committee for higher education **and** get them started.

Ex: *Way aamusantahay oo baqaysaa/* She is quiet **and** scared.

3. **Coordinating Conjunction:** So—used to connect two independent clauses

Ex: *Markaas qof Soomaali ah weli lama kulmin, oo xaaladdaydu waa murugsan tahay/* At the time, I hadn’t met any Somalis yet, **so** my situation was complicated.

4. **Preposition: Of:** used to indicate the whole or part of something

Ex: *Waxaas oo dhan ii daa, mushkilad ma lahane/* Leave all **of** that up to me; no problem at all.

5. **As/ while/ in a state of/ at a time when:**

Ex: *Axmed isaga oo 15 jir ayaa London loo diray/* Ahmed was sent to London **when** he was 15 years old.

Ex: *Aniga oo shaqaynaya ayaan iyagana caawinayey/* **While** I was working, I was also helping them.

### Common Phrases

1. **Oo ku saabsan/Oo ku habboon:** Which is about/ Which is suitable

Ex: *Daraasadaha la sameeyey mid ka mid ah oo ku saabsan xoolaha/* One of the studies conducted **which is** about livestock.

2. **Oo la taaban karo:** Which is tangible

Ex: *Waxqabad degdeg ah oo la taaban karo/* Quick action **that is tangible**.

3. **Taas oo qayb ka ah/** That which is part of

Ex: *Taas oo qayb ka ah qorshaha dowladda/* **That which is part of** the government plan.

4. **Sidaas oo ay tahay:** Even though, Nonetheless/ Inkastoo

Ex: *Sidaas oo ay tahay, hadal waa run kama rayste/* **Nonetheless**, truth has to be stated.

Ideally, a corpus-based entry such as this should be complemented by expert input from qualified linguists who have ample time to conduct a thorough study of how words behave in the

rich contexts that KWIC affords. The point here is that neither technology nor human knowledge can achieve what the two are able to produce in complementary fashion.

## 7.6 Towards Documentation and Development of Less Frequently Used Words

While adequate coverage and proper documentation is vital for extremely frequent words, reviewing the status of words that might be at risk of becoming obsolete in the language is yet another avenue we could explore with the corpus KWIC tool. To do so, a complete scan through the words that were infrequent in the corpus was conducted. For this experiment, the word “*falkin*” was chosen. The original function of “*falkin*” is weaving as craftwork. First, it fits the profile of being infrequent in the corpus; it ranks in the 53,239th position. Second, in my lived experience, I have not encountered it in either spoken or written use for a long time.

Figure 7.10. Infrequent Word—*Falkin*

Rank	Freq	Word
53235	1	falki
53236	1	falkiisaas
53237	1	falkiisii
53238	1	falkiisiyo
53239	1	falkin
53240	1	falkiso
53241	1	falkiyey

Once again, we will repeat the experiment that we did with the word *oo*: (a) examine how *falkin* or any of its word forms are used in the corpus, and (b) compare it to the definitions given in the reference dictionaries. We hypothesize that the dictionaries may not reflect current uses of the word in their entries. Our rationale is that since the word is among infrequent words, the existence of current uses can be captured only through the use of technology. We observe that *falkin* and its variants, which are *falki*, *falkiso* and *falkiyey* all have the same frequency of occurring only once in the corpus. The following is a concordance result for the word.

Figure 7.11. *Falkin* in Context Using the KWIC tool

Concordance Results 2:	
Concordance Hits 7	
Hit	KWIC
1	iyaaan xirfado kala duwan ee farshaxanka sida <b>falkinta</b> cawda, <b>tolidda</b> <b>dambiilaha</b> iyo koleyc
2	magacaabo Rag wax Xalin-kara awoodna u leh <b>falkinta</b> cilmiga <b>siyaasada</b> <b>la</b> xariirta sharci
3	Muhiimadda koowaad ayaa noqonaysa sidii loo <b>falki</b> lahaa <b>shax</b> <b>looga</b> shaqaynayo falanqaynta
4	idho-la-aana garabka u saari lahaa ama aan u <b>falkin</b> lahaaba <b>aragti</b> <b>cusub</b> oo aan loo baahn
5	Dhismaha, hannaan wanaagga, farsamaynta, sii <b>falkinta</b> , qurux-ku-dheehidda, iyo mawduucyo
6	ladsaday, waxaanse u qaatay in woxoo dhan la <b>falkiyey</b> ; si <b>dadka</b> <b>diidan</b> colaadda ehliga ah
7	lhawr qisoo walinimadiisa caddaynaysana ay u <b>falkiso</b> !; waayo <b>gar</b> <b>maaha</b> inay qolo kale awo

The concordance result reveals seven hits. All seven sentences exhibit different uses indicating that the word has seven different but related definitions. We presented definitions based on the examples generated by the KWIC tool; now we present a brief review of how the dictionaries have defined it. First, we looked it up in Keenadiid, and a translation of the definitions follow.

Figure 7.12. Keenadiid (1976): Falkin in a non-corpus based dictionary

*falkin*

**Falkin (-ta)** — 1. Caws ama mayrax  
ama dun ama wax la mid ah, gogol  
ka samayn. 2. Hawl isasuro badan  
qabasho, ka xoojin.

The word *falkin* has traditionally been used to refer to tasks that were common among Somali nomads and town folks. It was used to refer to crafts, especially ones that involved weaving. The skill has gradually become uncommon and so has the word *falkin*.

The Keenadiid (1976) dictionary gave two definitions. As shown below, the first one refers to the work of weaving while the second one appears to be an offshoot of the original meaning.

1. To weave mats, baskets, hats or fabrics out of dry leaves or knitting threads.
2. Performing a complex task, manually doing or making something. [Translation is mine]

The larger and more recent dictionary by Pugllieli and Mansuur (2012) gave only the original definition of the word. However, it gives it two main grammar formations. The first definition is the verb form while the second definition is the noun form. It also gives variations in spelling.

Figure 7.13. Pugllieli and Mansuur (2012): Falkin in non-corpus based dictionary

**falki** *f.g2* (-iyay, -isay) Falko samayn. *ld* **falag<sup>2</sup>**.

**falkid** *m.dh* *ld* **falkin**.

**falkin** *m.f.dh eeg* falki. *ld* **falkid, falkis**.

**falkis** *m.l* *ld* **falkin**.

**falko<sup>1</sup>** *m.dh* (-ooyin, *m.l*) 1. Daar, derin iwm oo caw ka samaysan oo aan midab lagu xardhin. 2. *ld* **falag<sup>1</sup>**.

**falko<sup>2</sup>** *f.g3* (-aday, -atay) Falag samaysasho. *ld* **falagso**.

As expected, the two dictionaries have documented the original meaning of the word with the exception of the former dictionary that gives an additional meaning referring to actions involving a complex task.

Below is a corpus-based entry of the word. It was interesting to learn that the word has evolved and taken on new meanings. In order to illustrate how such an evolution could be documented in a corpus-aided dictionary, five definitions have been derived from the seven examples. The definitions are deduced using the contexts in which the words were used. Interestingly, examples are from five different texts. The fact that the word is used in different contexts by different authors indicates that although its frequency is very low, speakers and authors are still using it.

***Falki—falkin, falkinta, falkiyey***

1. ***Tolidda ama farsamaynta darmada, koofiyadaha iyo dambiilaha/ To weave*** mats, baskets, hats or fabrics out of dry leaves or knitting threads.



Ex: *Ardayda waxaan u furaynaa xirfado kala duwan ee farshaxanka sida **falkinta** cawda, tolidda danbiilaha iyo koleyga/* Students will be offered hands-on skills in craftwork such as **weaving** mats, baskets, and handmade bags.

2. *Horumarinta iyo tayeynta arrin wax ama shay/* Strengthening or improving the quality of something.

Ex. *Intii ka danbaysana, waxaa si weyn isu soo tarayey isbeddelka ku dhacayey maansadeeda tiro iyo tayo ahaanba: dhismaha, hannaan wanaagga, farsmaynta, **sii falkinta**, iyo qurux kudheehidda/* Afterwards, her poetic talent has changed gradually in terms of quantity and quality: structure, better configuration, perfection, **improvement** and glorification.

3. *Awoodi kara lafagurka iyo fahanka hawl culus ama cakiran/* Able to **analyse** complex issues.

Ex: *Danta guud waa in xaajada maanta ka socota Soomaaliya loo magacaabo rag wax xallin kara awoodna u leh **falkinta** cilimiga siyaasdda la xiriira sharciyada dastuurka la doonayo in lugu dhaqo ummadda Soomaaliyeed/* The interest of the common good is duly served if men can come up with solutions and **analyse** the science of politics as it relates to the constitution being adopted for the Somali people.

4. *Dejinta qorshe hawleed/* Pre-plan, draw up a framework, strategize work.

Ex: *Muhimadda koowaad ayaa noqonaysa sidii loo **falki** lahaa shax looga shaqaynayo falanqaynta iyo dib u eegidda qaabka federaalka ah/* The first priority is **to draw up** a detailed framework reviewing the adoption of a federal system.

5. *Beenabuur ama in la maleego sheeko aan jirin/* **Fabricate, concoct** or make a story up

Ex: *Hadalkan nin Soomaaliya oo meel iska jooga ma keenin, hadduu keenayna uma badna inuu isagu idaacadda VOA soo raadsaday; waxaanse u qaatay in waxoo dhan la falkiyey*/ This story was not presented by an innocent Somali bystander and even if he had, he would not have publically stated it on the VOA radio, so I concluded that the whole thing was **concocted**.

Although all definitions could all be traced to the older use of the word, it has been remarkable to note how the word has taken up diverse meanings: to improve, to analyse, to strategize, and to concoct or make up a story. It should be stated, however, that the definitions given are not conclusive. An in-depth study using a larger, more representative corpus is ideally needed for conclusive results. However, at this exploration stage, this research succeeds in illustrating how people are using the language creatively and coining new meanings. The main objective here is that words go through stages and they evolve over time. In addition to the original definition of the word, which serves a good purpose in tracing the etymology of the word, other uses of the word need to be tracked and documented. Notice that all examples are taken from the corpus.

This study has explored a number of important issues targeted at ways of documenting and developing endangered and understudied languages. First, the study reveals that vocabulary-texts such as poetry could be wordbooks in disguise. It has been revealed that poetry inherently contains a vast vocabulary. It has been illustrated how infrequent and possibly outdated words can find their way into the world of technology describing relatively new concepts often generated by technology. Second, the research has pioneered for the first time the uncovering of the 20 most frequently used Somali words. Equally important for this study is the status of infrequent words. The next step was to discover ways to examine the status of infrequent words

and how these words evolve. It has been revealed that with the help of appropriate corpus tools such words can be monitored and subsequently documented.

## **Chapter 8: Summary, Conclusions, Implications, and Recommendations for Future Work**

The main objective of this research was to explore ways in which technology can offer endangered languages tools for development and documentation to those who are involved in language revitalization and development plans. **Section 8.1** gives an overview of the study by summarizing the main points in each chapter. It then gives a more detailed account of the main findings. **In 8.2**, conclusions and the implications as well as recommendations for theory and practice will follow synchronously. **In 8.3**, recommendations are given for future work, while in 8.4 the chapter ends with final thoughts in order to synthesize the multifaceted efforts from different stakeholders that are involved in language revitalization.

### **8.1 Summary of the Study**

In **Chapter One**, the study sets the scene by introducing the problem in question and describing the unprecedented language crisis the world faces today. It reports that linguists estimate that 50-90% of the world's languages may disappear in the next century (UNESCO, 2014). The chapter then outlines the significance of the study by revealing what the world stands to lose if the situation persists, noting the timeliness and the importance of this study. It has been argued that languages maintain the cultural and intellectual balance the world needs to sustain its natural diversity for a resource-filled environment. The chapter identifies corpus linguistics as a technological tool for endangered languages. It ends with an outline of the organization of the study.

**Chapter Two** is devoted to the issue of endangered languages. First, the term “endangered language” is defined along with several factors that cause language endangerment. An important distinction is then made between gradual and sudden death, noting that gradual

death is more common than sudden death. The rationale and the motivation that drive the global efforts to revitalize language are discussed. The chapter then considers ways in which languages could be revitalized. Success stories tell of endangered and dead languages that have been revitalized and reclaimed. The chapter concludes with an annotated list of a range of governmental, non-governmental, corporate and educational institutions that are involved in revitalization and language documentation.

Contrary to the popularly held belief that the number of language speakers determines the endangerment status of a language, **Chapter Three** delves into the issue of how a language with over 15 million speakers could be declared an endangered language. Given the precarious situation of the Somali language, the chapter examines how the Somali language faces both sudden death and gradual death. It then puts the problem in context by tracing how it started in the colonial era with the partition of the Somali-speaking populations into different nation states. It then discusses how Somali's prolonged civil war has created a detrimental environment for a language to survive. It touches on the effects of immigration, population death, and the lack of effective institutional support in the countries where the Somali language is spoken. Using UNESCO's (2003) language endangerment rubrics, the study reveals that the Somali language is endangered. Once again, the issue of the significance of endangered languages is revisited, using the Somali language as an illustration and touching on some of the losses to humanity if the Somali language (as representative of other languages in a similar situation) becomes extinct.

**Chapter Four** gives a window on the world of corpus linguistics. The chapter starts with a discussion of corpus linguistics as both an intra-disciplinary and interdisciplinary field. First, it touches on how all other linguistic fields depend on corpus linguistics in conducting linguistic research and analysis. Second, corpus linguistics is an interdisciplinary field, as it blends

linguistics and computer science, offering practical implications in many other fields including educational technology. The chapter then traces the history of corpus linguistics as a discipline and how it has evolved. Theoretical debate between rationalists and empiricists on what constitutes corpus data vis-à-vis its usefulness as a research tool is reviewed. The debate revolves mainly around the relative merits of using language competence versus language performance as corpus data. The characteristics of a corpus and its definitions are presented. The chapter ends its discussion with practical uses of different types of corpora as well as a discussion on some of the researchers and practitioners who use corpora for different ends.

**Chapter Five** discusses in detail the thorny issues of corpus-design methodology. First, an overview of corpus classifications is given. We note that different types of corpora are designed and constructed for different purposes, which dictate the kind of corpus one designs. A corpus is then described or classified according to its characteristics; for example, a general corpus is one that aims to represent a language while a specialized corpus samples only one area of a language. Another example is a bilingual corpus, which samples two languages for translation purposes while a monolingual corpus contains samples of a single language. The issue of corpus-design methodology standards that are often unattainable but nonetheless desirable are discussed. The debate on whether a corpus can be representative of the language it purports to represent is explored. It is well established in the literature that most corpus designers have considered issues of diversity, balance, and size as important design maxims to aim for. Issues of proportional versus stratified sampling are discussed. The chapter further explores how these design objectives have been challenging for corpus designers and how corpus linguists have tackled such design objectives. For example, the pragmatic and iterative design model has

been proposed as an alternative design model so that issues of diversity, balance, and size become less challenging over time.

**Chapter Six** focuses on corpus-design methodology in challenging contexts. The researcher sifts through research theories for guidance while tracking his assumptions and worldviews. The corpus-design methodology for this thesis as well as its requirements and objectives were considered in designing the model vis-à-vis limitations and constraints that had to be contended with. In addition to challenges faced by corpus designers in general, context-dependent design challenges that are particular to endangered languages (with limited support institutions) are considered. The way the term *case study* is used in this thesis is clarified. Five iterative design stages for a monitor, larger and more representative general corpus for the Somali language has been proposed. The scope of the data collection for this thesis has been defined to target collecting samples from as diverse texts as possible for the first prototype stage. The Somali language, its dialects, and the available text types are discussed, and then a tentative sampling frame, which was used as a guide for data collection, has been developed. The chapter identifies appropriate corpus software for data collection and analysis. The issue of copyright as it relates to corpus construction has been considered.

**Chapter 7** presents the main findings of this study and discusses them in terms of their significance and implications. The chapter begins with an overview of the corpus built for this thesis, quantifying the representations of each sub-corpus as well as the general corpus. Corpus texts have been analysed using types/token ratios to explore relative lexical density of texts represented in the corpus. This analysis was undertaken to examine whether certain texts are more suited for vocabulary or grammar development materials. The 20 most frequently used words in the Somali language are presented using an over 865,000-word corpus. The

significance of this finding in addressing the development and documentation objectives of this thesis is illustrated. In particular, it has been illustrated using the Key Words In Context tool (KWIC). These highly frequent words could be developed for the planning of language learning materials such as grammars and dictionaries. The corpus has been used to check the status of least frequently used words along with implications for researchers and language-material writers. The importance of using corpus tools in tracking the status of least frequent words has been noted as well as how they could be developed and disseminated.

**Chapter 8** presents a summary of the findings along with concluding remarks and recommendations. We start with a summary of the five main findings presented in this study. The first finding puts forth a case for the Somali language being endangered in Chapter Two. This was imperative for this thesis because the study uses the Somali language as a case. In Chapter Seven, the other four main findings are presented. The first finding is about how all text types in a language are not created equal in the sense that certain texts have an inherently richer vocabulary than others. This is significant for corpus design especially if the objective of corpus construction is to develop the lexicon of a language. The other three findings illustrate how corpus tools aid language development and documentation initiatives.

The second finding reveals how words that are highly in demand in a language could be identified using the wordlist feature. The third finding demonstrates ways in which highly frequent words in a language could be developed, while the fourth finding illustrates how words that are infrequent could be documented for revitalization and sustainability. As listed below, the findings have been structured around the research questions in order to shed light on different ways in which a corpus could be used to aid those who are involved in the revitalization and



documentation of endangered languages. The findings are ordered in the same way they were presented in the research.

**1. A Case for the Somali Language Being Endangered:** Contrary to the popularly held belief that the number of language speakers determines the endangerment status of a language, the study reveals how a language with over 15 million speakers could be classified as an endangered language. Using the UNESCO (2003) language-endangerment rubrics together with an extensive literature review and historical precedents, the study shows that the Somali language can in fact be categorized as an endangered language. As Nettle and Romaine (2000) pointed out, “A large language could be endangered if the external pressures on it were great, while a very small language could be perfectly safe as long as the community was functional and the environment stable” (p. 14). Although the number of native speakers in a language is one of the assessment factors, it must not be considered the main defining factor. This proposition finds support in a report on endangered languages prepared by an expert group for UNESCO (2003), which stated, “Languages cannot be assessed simply by adding the numbers” (p. 19). Therefore, given the precarious situation in which the Somali language finds itself, the study demonstrates how the Somali language faces both sudden death and gradual death (Nettle & Romaine, 2000). The study puts the problem in context by tracing how it started in the colonial era with the partition of the Somali-speaking populations into different nation states (Segal, 1962).

As discussed in Chapter Three, apart from mainland Somalia that has no effective institutions supporting the language, Djibouti, which was a French colony, has French as its official language and the medium of instruction in education. In northeastern provinces in Kenya where Somali speakers live, Swahili is the official language and English is the language of instruction. The Somali Regional State in Ethiopia has Amharic as the official language and

English and Amharic as the dominant languages of instruction. Somali children in the Diaspora learn through languages other than Somali and many do not use Somali at home. In Somalia, while children use Somali at home, the dominant languages for instruction are English and Arabic. Language policy especially as it relates to education is critical in assessing language vitality (Grenoble & Whaley, 2006; UNESCO, 2003). Furthermore, the study uses historical precedents for strong and healthy languages that became critically endangered as a result of war and instability. The prolonged civil war has created a detrimental and untenable environment for a people and a language to thrive. As Somalis become more dispersed through flight and immigration, the opportunities to sustain the Somali language become smaller.

It is recommended that in order to save the Somali language and reverse the current language shift, two main steps have to be taken before the situation gets worse. Ideally, a comprehensive digital documentation of the old oral and modern Somali language and literature should be undertaken. This is indeed a major task and is of course contingent upon the availability of Somali speakers, research time, and resources. Therefore, as time and resources permit, the degree of endangerment identified in this paper can be used to determine what segments of the language should be documented first. For example, the oral Somali literature (especially poetry) is critically endangered and needs immediate attention. Likewise, the old Somali language with its nomadic culture is critically endangered and may disappear or become moribund in a generation or two.

The documentation effort will serve two main purposes. First, it will facilitate accessibility for those who are engaged in language development or research projects. Such language research and developments may include literary and school textbooks as well as word processing tools such as spelling and grammar checkers. Another major work that could be

derived from the proposed digital database includes a dictionary of old Somali and a more comprehensive anthology of Somali poetry and oral literature. These tools would help speakers of the endangered language to access and use their language with a degree of facility and confidence, and make rich Somali materials available to schools. Second, such documentation would serve as a time capsule in which the language and wisdom of the Somali culture is preserved so that others around the globe can access it and future generations of Somali people will have the opportunity to cherish and learn from the legacy of their ancestors.

Nonetheless, in order to reverse the deteriorating status of the Somali language, documentation efforts alone will not help. Ideally, synchronized efforts of revitalization should be implemented, but revitalization initiatives are more challenging than documentation because they entail various groups and the project depends on how receptive those groups are to change. Politicians can play a major role in changing the language policy in their geographical area, which in the case of the Somali language may be problematic because geographical borders now separate Somali speakers. Revitalization initiatives involve raising awareness among the speakers about the value of their language as well as providing convincing reasons for why they should transmit Somali to their children. The contribution to the *ethnosphere*, to borrow Davis's term, by the Somali people through their language may never be realized, as their language is at the risk of being lost. The issue is time sensitive, as Evans (2010) warned: "Language death potentially gives humankind a generation or two to respond, provided to commit adequate resources to the task" (p. 216).

The study has documented different ways in which technology could revive languages. While the technological tools available are extremely useful, this study has initiated the construction of a monitor, larger and more representative corpus for the Somali language. An

iterative design model and a sampling frame have been developed that could enable the project to proceed with the next four phases. (Please find these documents in the appendices).

To bring the need to preserve the Somali language closer to the Canadian context, in the effort of planning for education with a global perspective in Alberta, students in this province could greatly benefit if they are exposed to multiple perspectives from multiple languages and cultures. Drawing upon multiple perspectives is vital as the world becomes more complex and more interdependent. A recent report from Alberta Education (2010) asserted:

While the industrial model has been successful in educating past generations, will it be enough in a knowledge-based society? The material that we deal with needs to be global and expanded in order to create global learners and global citizens. We need to have a multicultural focus. (p. 13)

As Davis (2008) argued, storytelling is a powerful tool that is a means to spread the word and share the global challenges we face with others. He cited practical examples that people can relate to. He reminded people how concepts that carry human cause have evolved. Consider the fact that *biosphere*, *biodiversity* or even *endangered languages* were terms known only to scientists but are now familiar to most students in junior high. What makes Davis's storytelling strategy attractive is that actions are usually preceded by thoughts and unless people share their thoughts and stories, we may not be able to mobilize the resources and the collective will, needed to save our languages.

**2. Poetry As A Wordbook in Disguise:** Using both quantitative and qualitative analysis, the study reveals that Somali poetry has a remarkably rich vocabulary. Corpus-aided analysis of type/token ratio was used to compute for the written sub-corpus as well as the spoken sub-corpus. The spoken sub-corpus yields a substantially higher type/token ratio of 27.7% than the

written sub-corpus, which gave a much lower ratio of 9.6%. Poetry, which was part of the spoken genre, was hypothesized to have contributed to the discrepancy. However, the texts in both sub-corpora were of different lengths. In order to neutralize the effects that size might have made, a decision was made to compare an abridged genre from the written corpus against the poetry genre. The second experiment uses two equalized texts of 5,040 words from “poetry” in the spoken sub-corpus and 5,040 words from “constitution” in the written sub-corpus. In support of the first finding, the latter experiment revealed a type/token ratio of 34.4% for poetry and 11.1% for constitution. These findings are supported in Guuled (1976) and Johnson (1988) who noted that Somali poetry is characterized by a remarkable and exceptionally rich vocabulary.

Inspired by their comments, a further experiment was conducted to examine the richness of vocabulary in one poem. From the 124 poems in the poetry genre, one poem was randomly selected. The poem has an overall word count of only 434 tokens but relatively very rich types of 321. The type/token ratio confirms the finding with a 74% type/token ratio. This means that the poet has used almost eight new words in every 10 words. It appears that a single poem has a higher type/token ratio than the overall type/token ratio of the entire poetry genre. The effect is probably due to the correlation that is normally found with larger sizes and a lower type/token ratio (Biber, 1993; Sinclair, 2005; Tognini-Bolleli, 2010). Through triangulation of evidence from the literature and conducting a series of experiments, the finding reasonably indicates that poetry has a substantially richer vocabulary than prose.

This finding is vital for corpus designers, material developers, and teachers. First, in the planning stages of corpus construction for the Somali language and other languages that may have a rich oral literature, vocabulary-rich texts such as poetry should be gathered especially if the objective is to develop a corpus for dictionary development or any objective related to

vocabulary. Dictionary compilers can mine the rich vocabulary yielded by poetry texts for designing headwords for the dictionary. Similarly, language teachers may use corpus tools such as KWIC to show words in real contexts when teaching in the classroom. The finding could aid teachers who teach Somali literature through exploring how words are used creatively in the composition of poetry.

**3. The 20 Most Frequently Used Somali Words:** Based on the corpus built for this thesis containing 865,214 words, the 20 most commonly used Somali words were revealed. The wordlist feature in the corpus is a process by which the computer scans and analyses all texts in the corpus and generates the frequency with which each word occurs in the corpus, along with its rank. The discovery shows that these 20 words are unusually frequent. As illustrated by the list generated from the general corpus, the word “*oo*” alone, which is ranked number one, occurs 22,557 times. Motivated by how these 20 words are extraordinarily frequent, a further examination seemed necessary to investigate the exact representation of these 20 words in the entire corpus. As discussed in Chapter Seven, the investigation shows that they actually represent 22%, which means that all other words together make up the remaining 78% in the general corpus. The finding was then crosschecked with how it validates or contradicts previous corpus-generated wordlists. The top 20 words generated by the Corpus of Contemporary American English (COCA) were compared with the top 20 words generated by the Somali Language Bank (SLB). Despite the differences of the two corpora in terms of size, composition, and linguistic differences, the two corpora show that COCA’s 20 top words account for 26.5%; SLB’s top 20 words represent 22.3%. This was seen as a notable correlation with a reasonable difference of 4%. The research further validates the fact that function or grammar words are normally the most frequent words in a language (Biber, 1993; Kilgariff, 1997; Sinclair, 2005). The correspondence

was further demonstrated by the striking similarity of the words in both corpora in terms of their function in their respective language. For example, “the,” which ranks number one in COCA has a function similar to *ka*, which is number three in the SLB. They are both definite articles. In the COCA, “and” occupies number three place, and has a meaning similar to *iyoo* in the SLB. The two words “to” and *ku* also have comparable functions and they occupy number seven and number three places in English and Somali respectively. It is worth noting that the only outlier in the two lists is the word, “say,” which occupies the 19th place in the COCA corpus. With the exception of this anomaly, the two lists are occupied exclusively by words that have a grammatical function in their language.

**4. Language Development Through Key Words In Context (KWIC):** For this finding, any of the highly frequent words previously generated by the wordlist feature could have been examined through visual representation of naturally occurring usages of the word derived from a wide variety of texts in the corpus. The examples reflect how a word behaves in its natural context, as the usages were not artificially constructed for this thesis. In consideration of its position in the ranking, the word *oo* was chosen for this experiment. As previously stated, the word was ranked as number one with 22,557 usages in context. It was not feasible for this thesis to study all examples generated by the KWIC for this word, so a decision was made to scan for recurring themes that could yield meaningful and approximate meanings for development purposes. The first point of enquiry was whether existing and non-corpus aided dictionaries had the entry of this highly frequent word proportionate to its frequency. The assumption was that more frequent words lend themselves to multiple meanings. Bogaards and Laufer-Dvorkin (2004) stated that highly frequent words have invariably multiple uses and functions in the language. The two non-corpus aided dictionaries defined the word *oo* as having a function of a

conjunction. This coverage was not consistent with our hypothesis. It was rather insufficient relative to its high frequency use in the language. In a closer examination of the example sentences given by KWIC, the entry was expanded and refined with useful information for users. As outlined in Chapter seven, the entry now has frequency information marked by stars although other labels might be employed. Second, three more definitions were added including a new function of a relative pronoun, which was supported by usages of the word in many different contexts. Third, a new sub-entry of common phrases that are used with word *oo* was given. The essence of the discovery is how the KWIC tool can serve as a language development tool and rather than the accuracy of the expansion and refinement proposition for the specific words used as experiments.

#### **5. Towards Documentation And Dissemination of Less Frequently Used Words:**

Central to the objectives of this thesis is how specific language skills such as vocabulary could be documented. Again, using the KWIC tool, the word *falkin* was chosen among infrequent words and traditionally used to mean: “weaving as in the crafts like making baskets, hats or fabrics.” It ranks low in 53,219th position with a very low frequency count. Its KWIC window showed only five occurrences of which each displayed a different variant of its word families such as: *Falki*, *falkin*, *falkinta*, *falkiyey*, and *falkiso*. An interesting discovery was how one of the examples in the concordance conformed to its traditional meaning. This base meaning was the only definition given by the two dictionaries, which were cross-referenced with the examples generated by the KWIC tool. The study further revealed that the other four examples exhibited four different uses derived from different texts in the corpus. Studying the contexts in which these different uses were found, the study proposes an improved and expanded entry for the word, intended to illustrate how a corpus can aid material developers to visually study words in



contexts thereby giving them the facility to refine and document such new uses so that they are disseminated through language resources. The study also exposes how certain words take new meanings. The word *mareeg* has had a function in rural Somalia referring to a rope or strap that is used to put around the neck of an animal to securely tie it to a tree overnight. Using the KWIC tool, the word shows that it has taken on a new meaning and it now refers to a “website” as evidenced by the uses generated from the corpus. Such evolution could go unnoticed without technological tools such as those afforded by the corpus.

## 8.2 Conclusions

This interdisciplinary research contributes to the base of knowledge in three different but complementary knowledge domains. Given the nature and the impact of the research problem selected for this study, the research has navigated through theories and practice in different areas of study for guidance towards achieving the scholarly mandate to add to and expand the knowledge base of the theories and practice that inspired it. As discussed in Chapter Six, this reciprocal and incremental nature of knowledge growth is what underpins the epistemological beliefs of the researcher.

First, the study adopted and tested the theoretical endangerment framework for endangered languages to build a new context validating and sometimes challenging the approaches, rubrics and the experiences that were previously applied to certain world languages. While the study validates the body of knowledge available to assess endangered languages, the study presents new evidence that a language with millions of speakers could in fact be classified as an endangered language given the interplay of historical, social, and political status of its speakers. Although such new evidence finds support in the literature, it seems counterintuitive to many, so the study challenges the interpretation of what is commonly conceived as “being

endangered.” The outcome has implications for language policy makers as well as community activists and those who are involved in language development and documentation initiatives.

Given the alarming rate at which more than half of the world’s languages are endangered, this study offers insight on how factors such as the number of speakers and relative use of any language cannot be taken for granted. The implication here is that policy makers are recommended to adopt more proactive measures rather than reacting to alarm bells and disasters when it may be too late to reverse the tide. Inevitably, there are both natural and manmade factors that lead languages to endangerment status. These factors include earthquakes, drought, immigration, technology and wars, to name just a few. However, the lesson here is that a vitality assessment needs to be conducted proactively for languages that face a similar fate as the Somali language. As discussed in Chapter Three, like the Somali language, these languages will likely exhibit signs and symptoms such as the weakening status of the language in schools and at home. Another sign is sustained anarchy and social upheaval as well as the enduring and seemingly irreparable colonial legacy. Ironically, some of the same manmade forces that move languages toward endangerment offer the tools to revitalize, document, and develop at-risk languages.

It is reasonable to argue that the language crisis the world faces today could have been lessened had the languages been monitored and their situations been dealt with proactively. For instance, the Somali language has had no supporting institutions for over 25 years and the symptoms of language endangerment have been evident since the onset of the Somali civil war. If this situation persists for another 25 or 50 years, it might reach a crisis point as in the case of many critically endangered world languages. Zuckermann and Walsh (2011) used a motivating and action-oriented phrase directed at those who work with endangered languages, “stop, reclaim, and revive” (p. 1). Likewise, McCarthy, Romero, and Zepeda (2006) captured the

immediate and strategic course of action needed, calling for: “Strategic action and change” (p. 43). Proactive measures, and timely and strategic action are the key phrases to remember here.

In the following section, we turn our attention to how technology can serve as a double-edged sword. On one hand, it facilitates younger generations embracing dominant languages via the Internet and in schools at the expense of not using their native tongue. On the other hand, as this study reveals, technology offers opportunities to document and develop those same languages in ways that are not otherwise feasible.

As for the technological solution presented here, the research builds and expands upon prior corpus design methodology. The study was informed by the insights and experiences of corpus construction for the English language. The study confirms that previous design methodologies are generally applicable to the new context of this study. However, the study uncovered that certain design approaches applied to the English language corpus construction were not readily applicable to Somali. The challenges stem from a lack of resources and statistics that surfaced in the application of corpus design in the context of the Somali language. Prior corpora design relied on such resources in planning the design and collection of texts in major languages. In the context of this thesis, such resources were either scanty or unavailable. One way of redressing this limitation was to develop an iterative design model that was instrumental in the design of this research and might serve as a guide in the design of a larger, more representative corpus for the Somali language and languages in similar contexts. The iterative design model outlines five phases that languages with limited resources could follow. Likewise, for developing a sampling frame for data collection, the literature recommends the use of resources such as library catalogues, books in print lists, and statistics on bestseller lists from institutions such as book distributors. Since these resources were not available, the study

developed a context-appropriate sampling frame, which was useful in conceptualizing a framework for text collection; it could also be helpful in designing a sampling frame for languages with limited resources.

Third, the research adds to the body of knowledge in corpus linguistics providing feedback to the design theories and practice. The research revealed the most frequently used Somali words and examines the finding against similar findings using the Corpus of Contemporary American English (COCA). The study revealed striking similarities between the two corpora. This observation added to both theory and practice in corpus linguistics. The finding and the subsequent analysis will be useful for practitioners in corpus linguistics.

The finding could be extended to machine translation efforts. Recently, Google launched instant bilingual English-Somali and Somali-English translations. The facility offers instant translations of various texts such as email received or sent through Gmail. However, the translations are awkward and often misleading. It appears that Google wants to experiment with the facility presumably planning to improve its quality in the future. With this assumption in mind, the most frequently used words in any language, as confirmed by this study, account for a proportionally higher percentage of the words in normal use. Based on this reality, Oxford has developed a list of the most frequently used 3,000 words. Using Oxford Corpus, these words are developed in a detailed manner covering all the contexts in which they may occur. In a similar fashion, the 20 most commonly used words list generated by SLB could be expanded to up to 3,000 or 5,000 most commonly used words. Using the KWIC tool, if the main definition of each word is covered and translated into English, the quality of the translations could be substantially improved. The reason is that these are the words commonly recycled in texts. Note that the most commonly used 20 words in the SLB account for 22%, and if the definitions of these words are

covered in a detailed manner with their equivalents in English, this could substantially improve Google's translation facility.

Fourth, the research contributes to the body of knowledge in education, as it provides innovative ways to develop materials for students and technologically aided resources for teachers to use in the classroom in contexts that have not previously been reached. In doing so, the study bridges a gap in the literature, expanding the base of knowledge to new frontiers. Although it was not one of the objectives of this research, the remarkable richness of Somali poetry uncovered in this research might be of interest to students and researchers in comparative literature. They might want to compare the richness of the Somali poetry to those of other languages. Expanding the base of knowledge in corpus linguistics, the study applies and tests previous corpus theory and practice in a context where it has not previously been tested. It validates many of its previous findings while it adds new knowledge to corpus design approach through its application.

### **8.3 Research Limitations**

Even though the study has succeeded in answering the questions that guided this research, there are a number of limitations that need to be considered when interpreting or generalizing the findings of this research.

First, with regard to its first finding in which the Somali language was revealed as an endangered language, the study has relied on existing literature and the knowledge of the researcher in applying UNESCO's (2003) rubrics for language vitality assessment. Complementing this study with primary research could have strengthened the finding of this research. For example, fieldwork that examines each of the nine language vitality factors in light of primary research could have yielded richer data that reflects on what actually happens in the

environments where the language is spoken. However, there are challenges to collect primary data with regards to all nine factors. For example, given the absence of effective institutions in Somalia, statistics on the absolute number of speakers, proportion of speakers with the total population, availability of materials for language materials and literacy are not available. These context-dependent challenges, coupled with limitations of time and resources, are what led this research to rely mainly on secondary resources. Primary data about two out of the nine factors can, nonetheless, be collected. For example, a questionnaire, survey, or any other data collection method that examines community member's attitudes towards their own language could have added to the exploratory finding in this research. Another way to redress this weakness in the future is to collect primary data on the intergenerational language transmission factor. It would be interesting to explore how Somali is being transmitted to younger generations at home and in other settings, especially with Somali families in the Diaspora.

The second limitation of this study has to do with its size, diversity, and balance of the corpus that yielded the results of this thesis. Atkins, Clear, and Ostler (1992) stated corpus construction and what goes into the corpus is dictated by the objectives of the corpus designer. Therefore, the SLB corpus constructed for this thesis was built with the research questions of this thesis in mind. This means that the data collected for this corpus was large enough for the purposes of this thesis because the results yielded were appropriate to address the research questions of the thesis. Therefore, while the results were suitable to answer the research questions, a larger corpus could yield different results. This means that linguistic generalizations should be made with an awareness of such limitations.

The third limitation has to do with the degree of diversity in the corpus. Although every effort has been made to collect as diverse texts as possible, the SLB corpus contains fairly

diverse texts representing only 22 different text categories. This is a limitation because a corpus with much more diverse text categories may have yielded different results. Again, this methodology limitation needs to be considered when interpreting or generalizing the results of this study to results from a corpus that has more diverse texts.

The fourth limitation concerns the imbalance that exists among texts. This is another limitation especially if one wants to compare the results of this study to another corpus that has a dissimilar representation of texts.

It must be stated, however, that existing corpora do not follow one methodology in addressing the issue of balance in corpus design (Biber, 1993; Sinclair, 2005). In addressing such design challenges, this study has proposed an iterative design methodology for building a larger more representative corpus for the Somali language and other languages in a similar position. As discussed in Chapter Six, such design methodology addresses both general corpus design challenges and considers the context-dependent challenges that endangered languages face. (Please refer to the proposed design phases in the appendix.)

#### **8.4 Recommendations for Future Research**

Despite the fact that this study has succeeded in illustrating how a model corpus could assist endangered languages in developing and documenting their language, there are still areas that need further study.

As noted in the iterative design model for this research, the study highly recommends that the subsequent four phases of this study be implemented. The implementation of the remaining phases will result in a large, more representative general corpus. This will both strengthen and draw upon the feedback and the experience of this study. It will address the limitations of size, diversity, and balance that were not feasible for this thesis. In particular, the inclusion of larger,

more diversified spoken samples will enrich its content, thereby producing results that are more representative of the Somali language. Unlike English and other similar languages, Somali has a more developed spoken genre than written genre. Understandably, Somali has very rich oral genres as exemplified by poetry revealed by this study. It has been a written language only since 1972. These rich oral genres will contribute to both theory and practice in various areas of computational linguistics. By using this larger, more representative corpus, language development tools could be developed, namely a spellchecker, which is currently not available for the Somali language. Furthermore, more comprehensive and authoritative dictionaries, grammars, and language-learning materials could be developed.

As noted earlier, further research is needed about the attitudes of Somali speakers towards their own language. In McCarthy, Romero, and Zepeda's (2006) study, *Reclaiming the Gift*, they explored the attitudes and experience of Native American youth toward their native tongue. Native American youth exhibited many attitudes that have been observed in the Somali communities around the world. They view the Somali language as an "emblem of shame" and in many cases some Somali parents encourage their children to speak and use languages other than Somali, namely English, Arabic, or French. An in-depth study exploring the issues of community attitudes and their language preference would be a substantial addition to the field of endangered languages for those who are involved in revitalization initiatives.

## 8.5 Final Thoughts

While technology can serve as a tool for language documentation and development (either the action or inaction thereof), the decision to use such a tool rests with the citizens of the world at the individual, community, national, and the international level. Although this study was concerned with exploring corpora as a tool in language revitalization and documentation, this



research also documented individuals and institutions at different levels that have contributed to the revitalization and development of certain world languages using different tools and methods. As evidenced by the myriad of institutions, individuals, governments, and corporations that are involved in language revitalization projects, the issue requires leadership and collaborative and synchronized effort. Therefore, while this study has presented a new body of knowledge that can both inform and inspire communities to take charge and revive their mother tongue, the complexity of the issue calls for the collaborative engagement of various stakeholders.

For example, at the individual level, the exemplary role of Ben Yehuda (refer to Chapter Two) in the revitalization of Modern Hebrew serves as a source of inspiration and guidance for individual speakers of languages that are endangered. A parent of children whose mother tongue was slipping away, his method of enforcing a policy of “mother tongue only” in the home is a powerful example to follow. It seems that he understood that it starts with the children. The second step he took was to approach schools so that students could learn their mother tongue in the school. Mobilizing the community was another effective strategy. Davis (2001) proposed the same method as a way to reach out to communities for support in what he calls the “storytelling” method. It illustrates how an individual can take a leadership role in mobilizing others towards a revitalization and reclamation campaign.

I acknowledge the concern of some parents in giving their children the best opportunities they could possibly give by encouraging them to learn certain world languages such as English, Arabic, French, or Chinese because they are seen to be economically more viable; however, these languages should not be learned at the expense of the children’s mother tongue. As McCarthy, Romero, and Zepeda (2006) argued, this is costly for the parents because the mother tongue is the link between parents and their children. As the native language is lost, the

connection and the relationships in the family disintegrate. Therefore, it would go against the whole notion of promoting diversity in the world if children were not encouraged to become bilingual or multilingual speakers in order to expand their horizons through learning from other cultures and maximizing their employment opportunities. The danger lies in when they learn foreign languages at the expense of their native language.

The same concept of learning a language as an additional language applies to speakers of stronger languages. In fact, multilingualism has become a feature that has been sought after in education. Alberta Education (2010) called for a curriculum with a multicultural focus. The University of Alberta's summer program based on research and training in Canadian indigenous languages is an initiative that could be expanded to Alberta's schools. As discussed earlier in the chapter, the University of Minnesota's model requires undergraduates to take a compulsory course in foreign languages including understudied and exotic languages such as Dakota, Somali, and Hmong. It is yet another initiative that illustrates how languages could be re-ignited through learning while exposing the students to the diversity of the worldviews that languages offers.

At the local and government level, communities can learn from Maori speakers in New Zealand in their efforts to mobilize their community in advancing Maori language immersion for children. At the same time, they exert pressure on policy makers to recognize Maori as one of the official languages in the country. Linguists, technologists, and other academics that speak an endangered language as their native tongue, could aid communities with tools and technical knowledge. This study is an example of technical knowhow that informs community activists, policy makers, teachers, and language-material developers.

At the international level, non-governmental organizations such as UNESCO and regional and national institutions, ACALAN, Somali PEN, and the Institute for Languages in Djibouti are working toward the common goal of the revitalization and development of indigenous languages. International companies such as Google, Microsoft, and Rosetta Stone are contributing to the cause. In academia, major universities have started graduate programs in language revitalization and documentation. Language revitalization, documentation, and reclamation is a complex undertaking, so it needs a coordinated effort from the readers of this thesis who include individuals, community activists, governments, educational technologists, teachers, and international corporations.

In the opening epigraph of this thesis, Baker (2001) considered linguistic and cultural diversity to be necessary resources for the long-term survival of humanity. The international collaborative efforts aimed at redressing the current language crises are encouraging although they do not seem proportional to the task. If, on the other hand, the biodiversity on earth is not maintained, the problem does not go away but rather waits for future generations. A well-recited poem by the legendary Somali poet, Salaan Carrabey is appropriate as the conclusion of this thesis: “*Dhashaaday sugtaa, xaajadaad dhowrataa abide.*” The line could be translated as: “Your inaction toward matters of importance only leaves a legacy of burden to your children.”

## References

- Abney, S., & Bird, S. (2010). *The human language project: Building a universal corpus of the world's languages*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden.
- Addison, K. N. (2009). *We hold these truths to be self-evident: An interdisciplinary analysis of the roots of racism and slavery in America*. Lanham, MD: University Press of America.
- Adegbija, E. (1994). Language attitudes in Sub-Saharan Africa: A sociolinguistic Overview. Clevedon, UK: Multilingual Matters.
- African Academy of Languages. (2014). *About Acalan*. Retrieved from [http://www.acalan.org/eng/about\\_acalan/about\\_acalan.php](http://www.acalan.org/eng/about_acalan/about_acalan.php)
- Agrawal, R. P., Dogra, R., Mohta, N., Tiwari, R., Singhal, S., & Sultania, S. (2009). Beneficial effect of camel milk in diabetic nephropathy. *Acta Biomed*. Retrieved from [http://www.actabiomedica.it/data/2009/2\\_2009/agrawal.pdf](http://www.actabiomedica.it/data/2009/2_2009/agrawal.pdf)
- Alberta Education. (2010). Inspiring education: A dialogue with parents. Retrieved from <http://engage.education.alberta.ca/uploads/1006/inspiringeducationst92115.pdf>
- Alcom, J. (1993). Indigenous peoples and conservation. *Conservation Biology*, 7, 424-426.
- Andrzejewski, B. W. (1985). Somali literature. In B. W. Andrzejewski, S. Pilaszewicz, & W. Tyloch (Eds.), *Literatures in African languages: Theoretical issues and sample surveys* (pp. 337-400). Cambridge, UK: Cambridge University Press.
- Andrzejewski, B. W., & Lewis, I. M. (1964). *Somali poetry: An introduction*. Oxford, UK: Clarendon Press.
- Anthony, L. (2011). AntConc (Version 3.3.5m) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from <http://www.antlab.sci.waseda.ac.jp/>

Anzaldua, G. (1987). *How to tame a wild tongue. Borderlands/La Frontera: The new Mestiza*.

San Francisco, CA: Aunt Lute Books, 53-64

Aston, G., & Burnard, L. (1988). *The BNC handbook*. Edinburgh, UK: Edinburgh University Press.

Atkins, B. T. S., Clear, J., & Ostler, N. (1991). Corpus design criteria. Retrieved from

<http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf>

Atkins, B. T. S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic*

*Computing. Journal of the Association for Literary and Linguistic Computing* 7/1, 1-

16. doi:[10.1093/lc/7.1.1](https://doi.org/10.1093/lc/7.1.1)

Authors Guild v. Google (2013). Judge Chin grants summary judgment in authors Guild v. Google case. Retrieved from

<http://library.osu.edu/blogs/copyright/files/2013/11/Authors-Guild-v-Google-Summary-Judgment-Decision.pdf>

Baker, C. (2001). Review of Tove Skutnabb-Kangas, Linguistic genocide in education - or worldwide diversity and human rights? *Journal of Sociolinguists*, 5(2) 279-283.

BBC. (2012). *Digital tools to save languages*. Retrieved from [www.bbc.co.uk/news/science-environment-17081573](http://www.bbc.co.uk/news/science-environment-17081573)

Bennett, G. (2010). *Using corpora in the language-learning classroom: Corpus linguistics for teachers*. Michigan, MI: University of Michigan Press.

Berns, P. M. (2010). *Concise encyclopedia of applied linguistics*. Amsterdam, Neth.: Elsevier.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

- Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variations. *Literary and Linguistic Computing*, 5, 257-269.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Biber, D. (1994). Representativeness in corpus design. In A. Zampolli, N. Calzolari, & M. Palmer (Eds.), *Current issues in computational linguistics: In honour of Don Walker* (pp. 377-408). Dordrech, Pisa: Kluwer and Giardini.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.
- BBC. (2013). "Selfie" named by Oxford Dictionaries as word of 2013. Retrieved from <http://www.bbc.com/news/uk-24992393>
- Bogaards, P., & Laufer-Dvorkin, B. (2004). Vocabulary in a second language: Selection, acquisition, and testing. Amsterdam, Neth.: John Benjamins.
- British National Corpus. (2013). *What is the BNC?* Retrieved from <http://www.natcorp.ox.ac.uk/corpus/index.xml>
- Burnard, L. (1995). *User's reference guide to the British National Corpus*. Oxford, UK: Oxford University Computing Services. Retrieved from <http://homepages.abdn.ac.uk/advaith/pages/teaching/NLP/practicals/bnc-doc.pdf>
- Burnard, L. (2000). User's reference guide to the British National Corpus. World edition. Oxford: Oxford University Computing Services.
- Burnard, L. (2002). Where did we go wrong? A retrospective look at the British National Corpus

- In B. Kettleman and G. Markus (Eds.), *Teaching and learning by doing corpus analysis* (pp. 51-71). Amsterdam, Neth.: Rodopi.
- Burton, R. (1854). *First footsteps in East Africa*. London, UK: Tylston and Edwards.
- Cahill, K. M. (1980). *Somalia: A perspective*. Albany, NY: State University of New York.
- Campbell, L. (1994). Language death. In R.E. Asher, (Eds.), *The encyclopedia of language and linguistics* (vol. 4) 1960-68. Oxford: Pergamon Press.
- Cassanelli, L., & Abdikadir, F. S. (2008). *Somalia: Education in transition*. Retrieved from:  
<http://digitalcommons.macalester.edu/bildhaan/vol7/iss1/7>
- Chomsky, N. (1962). Explanatory models in linguistics. In E. Nagel, P., Suppers. A., Tarski (Eds.), *Logic, methodology and philosophy of science: Proceedings of the 1960 International Congress* (pp. 528-550). Stanford, CA: Stanford University Press.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- CNN (2013). *Selfie named word of the year for 2013*. Retrieved from  
<http://edition.cnn.com/2013/11/19/living/selfie-word-of-the-year/>
- COCA. (2014). *Copyright: Fair use*. Retrieved from  
<http://corpus.byu.edu/coca/help/copyright.asp>
- Collins Online Dictionary. (2014). That. In *collinsdictionary.com*.  
 Retrieved from <http://www.collinsdictionary.com/dictionary/english/that?showCookiePolicy=true>
- Corpus. (2014). In. *Merriam-Webster.com*. Retrieved from <http://www.merriam-webster.com/dictionary/corpus>
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

- Crystal, D. (1992). *An encyclopaedic dictionary of language and languages*. Oxford, UK: Cambridge University Press.
- Crystal, D. (1997). *The Cambridge Encyclopedia of Language*. Cambridge, UK: Cambridge University Press.
- Crystal, D. (2000). *Language death*. Cambridge, UK: Cambridge University Press.
- D'Agostino, G. (2008). Healing fair dealing? A comparative copyright analysis of Canada's fair dealing to UK fair dealing and US fair use. *McGill Law Journal*, 53, 309-463. Retrieved from <http://lawjournal.mcgill.ca/userfiles/other/7046615-dAgostino.pdf>
- Daily Nation. (2010). *Somali-Kenyans: Sixth largest of Kenya's 44 ethnic groups*. Retrieved from Hiiraanonline at [http://www.hiiraan.com/news2/2010/sept/somali\\_kenyans\\_sixth\\_largest\\_of\\_kenya\\_s\\_44\\_ethnic\\_groups.aspx](http://www.hiiraan.com/news2/2010/sept/somali_kenyans_sixth_largest_of_kenya_s_44_ethnic_groups.aspx)
- Dalby, A. (2002). *Language in danger*. New York, NY: Columbia University Press.
- Dash, N. S. (2005). *Corpus linguistics and language technology with reference to Indian languages*. New Delhi, India: Mitali.
- Dash, N. S. (2010). *Corpus linguistics: A general introduction*. Proceedings of the workshop on Corpus Normalization, LDCIL, CIIL, Mysore, India, on 25th August 2010. Retrieved from [www.idcil.org/download/Corpus%20Linguistics.pdf](http://www.idcil.org/download/Corpus%20Linguistics.pdf)
- Dash, N. S. (2008). *Corpus linguistics. An introduction*. New Delhi, India: Pearson Education-Longman.
- Davies, M. (2008). *The corpus of contemporary American English: 450 million words, 1990-present*. Retrieved from <http://corpus.byu.edu/coca/>



- Davies, M. (2009). The 385+ million-word corpus of contemporary American English (1990-2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159-190.
- Davis, W. (2001). *Light at the edge of the world: A journey through the realm of vanishing cultures*. Washington, DC: National Geographic.
- Davis, W. (2009). *The wayfinders: Why ancient wisdom matters in the modern world*. Toronto, ON: Anansi Press.
- Dieu, B. (2005). *Some facts and figures about the English language*. Retrieved from [http://the\\_english\\_dept.tripod.com/esc.html](http://the_english_dept.tripod.com/esc.html)
- Edwards, V. (2004). *Multilingualism in the English-speaking world*. Oxford, UK: Blackwell.
- Eisele, A., & Ziegler-Eisele, D. (2002). *Towards a road map on human language technology: Natural language processing*. Retrieved from <http://www.elsnet.org/dox/rm-eisele-v2.pdf>
- Emergency Nutrition Network. (2010). *Somali region Ethiopia: Summary of situation report*. Retrieved from <http://fex.enonline.net/11/somali.aspx>
- Endangered language project. (2014). *About the endangered languages project*. Retrieved from <http://www.endangeredlanguages.com/about/>
- Ethiopian Government. (2012). *Somali National Regional State*. Retrieved from <http://www.ethiopia.gov.et/statesomali>
- Evans, N. (2010). *Dying words: Endangered languages and what they have to tell us*. Oxford, UK: Wiley-Blackwell.
- Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Clevedon, UK: Multilingual Matters.
- Fishman, J. A. (Ed.). (2001). *Can threatened languages be saved? Reversing language shift*,

- revisited: A 21st century perspective*. Clevedon, UK: Multilingual Matters.
- Fellman, J. (1973). *The revival of a classical tongue: Eleizer Ben-Yehuda and the Modern Hebrew language*. The Hague: Mouton.
- Francis, W. N., & Kučera, H. (1979). *Brown corpus manual*. Department of Linguistics, Brown University, Providence, Rhode Island. Retrieved from [www.icafe.uib.no/brown/bcm.html](http://www.icafe.uib.no/brown/bcm.html)
- Gledhil, C. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag Tübingen.
- Global Security (2001). *Somalia civil war*. Retrieved from <http://www.globalsecurity.org/military/world/war/somalia.htm>
- Google (2014). *A better world faster*. Retrieved from <http://www.google.org>
- Gordon, R. G., Jr. (Ed.) (2005). *Ethnologue: Languages of the world*. (15th ed.) Dallas, TX: SIL International. Retrieved from <http://www.ethnologue.com>
- Government of Canada (2014). *The Act to amend Copyright Act*. Retrieved from <http://www.parl.gc.ca/HousePublications/Publication.aspx?Language=E&Mode=1&DocId=5697419&File=4>
- Graddol, D., Cheshire, J., & Swann, J. (1994). *Describing language*. Buckingham, UK: Open University Press.
- Granger, S., & Petch-Tyson, S. (2003). *Extending the scope of corpus-based research: New applications, new challenges*. Amsterdam: Rodopi.
- Grenoble, L. A., & Whaley, L. J. (1998). *Endangered: Language loss and community response*. Cambridge, UK: Cambridge University Press.
- Grenoble, L. A., & Whaley, L. J. (2006). *Saving languages: An introduction to language revitalization*. New York, NY: Cambridge University Press.

- Gries, S. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction*. New York, NY: Routledge, Taylor & Francis.
- Grimes, B. F. (Ed). (2000). *Ethnologue: Languages of the world* (14<sup>th</sup> ed.). Dallas, TX: SIL International. Retrieved from <http://www.ethnologue.com>
- Grimes, B. F. (Ed). (2004). *Ethnologue: Languages of the world* (Web edition). Dallas, TX: SIL International. Retrieved from: <http://www.ethnologue.com>
- Guuleed, C. D. (1976). *Gorfaynta Maansada: Analysis of poetry*. Unpublished manuscript. Somali Academy of Culture and Ministry of Higher Education. Marka: Somalia.
- Haggett, P. (2001). *Encyclopedia of world geography* (2nd ed.). New York, NY: Marshal Cavendish.
- Hameso, S. Y. (2001). *Development, state and society: Theories and practice in contemporary Africa*. San Jose, CA: Authors Choice Press.
- Hardy, D. E. (2007). *The body in Flannery O'Connor's fiction: Computational technique and linguistic voice*. Columbia, SC: University of South Carolina Press.
- Harmon, D., & Maffi, L. (2002). Are linguistic and biological diversity linked? *Conservation Biology in Practice*, 3, 26-27.
- Harmon, D. (2002). Losing species, losing languages: Connections between biological and linguistic diversity. *Southwest Journal of Linguistics*, 15, 89-108
- Hashi, A.A. (2001). *Transformative motivation in second language learning: The case of Keinaan, a Rapnomad*. Master's thesis. Minnesota State University: Mankato, Minnesota
- Hashi, A., & Hashi, A. A. (2004). *Essential English-Somali dictionary* (2nd ed.). Jigjiga, Ethiopia: Fiqi.

- Hashi, A., & Hashi, A. A. (2014). *Essential English-Somali dictionary* (two-in-one edition). Jigjiga, Ethiopia: Fiqi.
- Heller, M. (1994). *Crosswords: Language, education and ethnicity in French Ontario*. New York, NY: Mouton de Gruyter.
- Hinton, L. (2001). Language revitalization: An overview. In L. Hinton & K. Hale (Eds.), *The green book of language revitalization in practice* (pp. 3-18). New York, NY: Academic Press.
- Hinton, L., & Hale, K. (Eds.). (2001). *The green book of language revitalization in practice*. New York, NY: Academic Press.
- Hinton, L., et al. (2002). *How to keep your language alive: A commonsense approach to one-on-one language learning*. Berkeley, CA: Heyday Books.
- Hockett, C. F. (1958). A course in modern linguistics. *Language Learning*, 8, 73-75.
- Hoffman, E. D. (2012) *American Indians and popular culture*. Westport, Connecticut: Praeger.
- Hui, S. (2013). *Oxford crowns "selfie" word of the year*. Retrieved from <http://thechronicleherald.ca/artslife/1168340-oxford-crowns-selfie-word-of-the-year>
- Human Rights Watch. (2010). *Somalia: Events of 2009*. Retrieved from <http://www.hrw.org/world-report-2010/somalia>
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK: Cambridge University Press.
- Jama, M. J. (2013). Maxamed Ibraahin Warsame Hadraawi: The poet and the man (vol. 1). Pisa, Italy: Ponte Invisible.
- Janse, M. (Ed.). (2003). *Language death and language maintenance*. Amsterdam, Neth.: John Benjamins.

- Johansson, S., & Stenstrom, A. B. (Eds.). (1991). *English computer corpora: Selected papers and bibliography*. Boston, MA: Mouton de Gruyter.
- Johnson, J. W. (1979). Somali prosodic systems. *Horn of Africa*, 2(3), 46-54. Retrieved from <https://scholarworks.iu.edu/dspace/bitstream/handle/2022/3257/Somali%20Prosodic%20Systems.pdf?sequence=1>
- Kapchits, G. (2012). *Soomaali been ma maahmaahdo: Somalis do not lie in proverbs*. Isa, Italy: Ponte Invisible.
- Kaplan, R. (1995). The formal architecture of lexical-functional grammar. In M. Dalrymple, R. Kaplan, J. Maxwell, & A. Zaenen (Eds.). *Formal issues in lexical-functional grammar* (pp. 7-27). Stanford, CA: Center for the Study of Language and Information.
- Keenadiid, Y. C. (1976). *Qaamuuska Af Soomaaliga. Wasaaradda Hiddaha iyo Tacliinta Sare*: Mogadishu.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. New York, NY: Longman.
- Kilgariff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135-155.
- Kinkade, M. D. (1991). The decline of native languages in Canada. In R. H. Robins & E. M. Uhlenbeck (Eds.), *Endangered languages* (pp. 157-176). New York, NY: St. Martin's Press.
- Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1), 1-42.
- Krauss, M. (1998). The condition of Native North American Languages: The need for realistic assessment and action. *International Journal of Sociology of Language*, 132, 9-12.
- Laitin, D. D., & Samatar, S. S. (1987). *Somalia: Nation in search of a state*. Boulder, CO: Westview Press.

- Landinfo. (2011). *Somalia: language situation and dialects*. Retrieved from [http://www.landinfo.no/asset/1800/1/1800\\_1.pdf](http://www.landinfo.no/asset/1800/1/1800_1.pdf)
- Lambarti, M. (1986a). *Die Somali-Dialekte*. Hamburg: Helmut Buske.
- Laurence, M. (1954). *A tree for poverty: Somali poetry*. Hamilton, ON: McMaster.
- Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 8-29). London, UK: Longman.
- Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1-18). London, UK: Longman.
- Leech, G. (2002). *The importance of reference corpus*. Retrieved from <http://www.corpus4u.org/forum/upload/forum/2005060301260076.pdf>
- Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 133-149). Amsterdam, Neth.: Rodopi.
- Levinson, D. (1998). *Ethnic groups worldwide. A Ready Reference handbook*. Phoenix, AZ: Oryx Press.
- Lewis, I. M. (1961). *A pastoral democracy*. London, ON: Oxford University Press.
- Lewis, M. P. (Ed). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International. Retrieved from <http://www.ethnologue.com>
- Lewis, M. P., Simons, G. F., & Fennig, C. D (Eds.). (2014). *Ethnologue. Languages of the world* (17th ed.). Dallas, TX: SIL International. Retrieved from <http://www.ethnologue.com/language/SOM>

- Library of Congress. (1992). *Somalia: A country study*. Retrieved from [https://archive.org/stream/somaliacountryst00metz\\_0/somaliacountryst00metz\\_0\\_djvu.txt](https://archive.org/stream/somaliacountryst00metz_0/somaliacountryst00metz_0_djvu.txt)
- Liddy, E. D. (2001). *Natural language processing*. Encyclopedia of Library and Information Science (2nd ed.) New York, NY: Marcel Decker.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- Maffi, L. (1998). Language: A resource for nature. *Nature and Resources, UNESCO journal on the Environmental and Natural Resources Research*, 34(4), 12-21.
- Maffi, L. (2001a). Introduction: On the interdependence of biological and cultural diversity. In L. Maffi (Ed.), *On biocultural diversity: Linking language, knowledge, and the environment* (pp. 1-50). Washington, London: Smithsonian Institution Press.
- Maffi, L. (Ed.). (2001b). *On biocultural diversity: Linking language, knowledge, and the environment*. Washington, London: Smithsonian Institution Press.
- Maffi, L. (2005). Linguistic, cultural, and biological diversity. *Annual Review of Anthropology*, 29, 599-617. Retrieved from [www.terralingua.org/wp-content/uploads/downloads/.../ARA\\_review.pdf](http://www.terralingua.org/wp-content/uploads/downloads/.../ARA_review.pdf)
- Mansuur, A. O. (1988). *Lexical aspect of Somali and East Cushitic languages*. Somali National University: Mogadishu. Retrieved from [http://dspace-roma3.caspur.it/bitstream/2307/1015/5/03\\_A.%20O.%20MANSUR%20-%20A%20lexical%20aspect%20of%20somali%20and%20east-cushitic%20languages.pdf](http://dspace-roma3.caspur.it/bitstream/2307/1015/5/03_A.%20O.%20MANSUR%20-%20A%20lexical%20aspect%20of%20somali%20and%20east-cushitic%20languages.pdf)
- McArthur, T. (Ed.). (1992). *The Oxford Companion to the English Language*. Oxford, UK: Oxford University Press.

- McCarty, T. L., Romero, M. E., & Zepeda, O. (2006). Reclaiming the gift: Indigenous youth counter-narratives on native language loss and revitalization. *The American Indian Quarterly* 30(1), 28-48. University of Nebraska Press.
- McEnery, T., Baker, P., Gaizauskas, R., & Cunningham, H. (2000). *EMILLE: Towards a corpus of South Asian languages*. British Computing Society Machine Translation Specialist Group, London, pp .111-119.
- McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh, UK: Edinburgh University Press.
- McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. London, UK: Routledge.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.
- McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh, UK: Edinburgh University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
- Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass.
- Metz, M. C. (Ed). (1992). *Somalia: A country study*. Washington, DC: GPO for the Library of Congress.



- Microsoft. (2014). *Microsoft enables millions more to experience personal computing through local language program*. Retrieved <http://www.microsoft.com/en-us/news/press/2004/mar04/03-16llppr.aspx>
- Microsoft Research. (2014). *Natural language processing*. Retrieved from <http://research.microsoft.com/en-us/groups/nlp/>
- Mufwene, S. (2001). *The ecology of language evolution*. Cambridge, UK: Cambridge University Press.
- Munz, E., Moallin, A. S. M., Mahnel, H., & Reimann, M. (1990). Camel Papillomatosis in Somalia. *Journal of Veterinary Medicine, Series B*, 37, 191-196. doi:10.1111/j.1439-0450.1990.tb01046.x
- Muthwii, M. J., & Kioko, A. N. (Eds). (2004). *New language bearings in Africa*. Clevedon, UK: Multilingual Matters.
- Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford, UK: Oxford University Press. Ces langues, ces voix qui s'effacent. Paris: Autrement, 2003.
- Ohio University Center for International Studies. (2010). *Somali language*. Retrieved from <http://www.african.ohio.edu/African%20Languages/somali.html>
- Orwin, M. (2001). Introduction to Somali poetry. In D. Weissbort (Ed.), *Modern poetry in translation* (pp. 12-15). No. 17 New Series. Exeter: Short Run Press. Retrieved from <http://www.poetrymagazines.org.uk/magazine/record.asp?id=12334>
- Ostler, N. (1999, April-June). Does size matter? Language technology and the smaller languages. *Elra Newsletter*, 4(2), 3-5. Paris: European Language Resource Association. Retrieved from <http://www.elra.info/nl/newsletters/V4N2.pdf>

- Osler, A., & Starkey, H. (Eds.). (2005). *Citizenship and language learning: International perspectives*. Stoke-on-Trent, UK: Trentham Books.
- Oxford Dictionaries. (2013a). *About the Oxford English Corpus*. Retrieved from <http://www.oxforddictionaries.com/words/about-the-oxford-english-corpus>
- Oxford Dictionaries. (2013b). *The OEC: Composition and structure*. Retrieved from <http://www.oxforddictionaries.com/words/the-oec-composition-and-structure>
- Oxford Dictionaries. (2013c). *The OEC: Technical information*. Retrieved from <http://www.oxforddictionaries.com/words/the-oec-technical-information>
- Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford, UK: Clarendon Press.
- Puglielli, A., & Mansuur, C. C. (2012). *Qaamuuska Af-Soomaaliga*. Rome, Italy: Roma TrePress. Retrieved from <http://dspace-roma3.caspur.it/bitstream/2307/720/1/QAAMUUSKA%20AF-SOOMAALIGA%5B1%5D.pdf>
- Purdue University. (2014). *Copyright exceptions: Fair use*. Retrieved from [https://www.lib.purdue.edu/uco/CopyrightBasics/fair\\_use.html](https://www.lib.purdue.edu/uco/CopyrightBasics/fair_use.html)
- Reagan, T. (2005). *Non-western educational traditions: Indigenous approaches to educational thought and practice* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Romaine, S. (2008). Linguistic diversity, sustainability, and the future of the past. In K. King, N. Schilling-Estes, L. Fogle, J. J. Lou, & B. Soukup (Eds.), *Sustaining linguistic diversity* (pp. 7-12). Washington, DC: Georgetown University Press.
- Rosetta Stone. (2012). *Endangered languages*. Retrieved from <http://www.rosettastone.co.uk/endangered>

- Rosetta Project. (2014). *About the Rosetta project*. Retrieved from <http://rosettaproject.org/about/>
- Rundell, M. (Ed.). (2002). *The Macmillan English dictionary for advanced learners*. Oxford, UK: Macmillan.
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.
- Segal, R. (1962). *African profile*. Baltimore, MD: Penguin Books.
- Shabo, S., Barzel, R., Margoulis, M., & Yagil, R. (2005). Camel milk for food allergies in children. *Israeli Medical Association Journal*, (7) 769-8. Retrieved from <http://www.ima.org.il/imag/ar05dec-12.pdf>
- Shaul, D. L. (2014). *Linguistic ideologies of Native American language revitalization*. New York, NY: Springer.
- Sheikh, H., & Healy, S. (2009). *Somali's missing million: Somali Diaspora and its role in development*. Retrieved from [http://scarab.bates.edu/somalis\\_in\\_maine/54/](http://scarab.bates.edu/somalis_in_maine/54/)
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.
- Sinclair, J. (2004). *How to use corpora in language teaching*. Amsterdam John Benjamins.
- Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxford, UK: Oxbow Books. Retrieved from <http://ahds.ac.uk/linguistic-corpora/>
- Skutnabb-Kangas, T. (2000). *Linguistic genocide in education - or worldwide diversity and human rights?* Mahwah, NJ: Lawrence Erlbaum Associates.
- Somali Speaking Pen Centre. (2014). *Bogga hore – home*. Retrieved from <http://www.sspen.org>

- Spolsky, B. (1999). Second language learning. In X. Fishman (Ed.), *Handbook of language and ethnic identity* (pp. 181-192). Oxford, UK: Oxford University Press.
- Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
- Startvik, J. (Ed.). (1992). *Directions in corpus linguistics*. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin, New York: Mouton de Gruyter.
- Stim, R. (2010). *The craft artist's legal guide to law and business practice*. London, UK: Association of Illustrators.
- Stovel, N. F. (2008). *Divining Margaret Laurence: A study of her complete writings*. Montreal, QC: McGill-Queen's University Press.
- Štrba, S. I. (2012). *International copyright law and access to education in developing countries: Exploring multilateral legal and quasi-legal solutions*. Leiden: M. Nijhoff.
- Stubbs, M. (1996). *Text and corpus analysis*. Oxford, UK: Blackwell.
- Swarthmore College (2014). *Laboratory for endangered languages*. Retrieved from <http://www.swarthmore.edu/SocSci/Linguistics/EndangeredLanguages/>
- Summers, D. (Ed.). (1995). *Longman dictionary of contemporary English* (3rd ed.). Harlow, Essex: Longman.
- Teal, G. (2010). *Somalia is a nation of poets*. Need to Know on PBS. Retrieved from <http://www.pbs.org/wnet/need-to-know/culture/somalia-is-a-%E2%80%99nation-of-poets%E2%80%99770/>
- Teubert, W., & Cermáková, A. (2004). Directions in corpus linguistics. In M. A. K. Halliday, W. Teubert, C. Yallop, & A. Cremakova. (Eds.), *Lexicography and corpus linguistics: An introduction*. London, UK: Open Linguistics Series.

- Teubert, W., & Cermakova, A. (2007). *Corpus linguistics: A short introduction*. London, New York, NY: Continuum.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
- Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 14-27). London, UK: Routledge.
- Tsunoda, T. (2006). *Language endangerment and language revitalization*. Berlin, Ger: Mouton de Gruyer.
- UCLA. (2009). *Yiddish language*. Retrieved from <http://www.germanic.ucla.edu/yiddish>
- UN Official languages. (2014). *UN at a glance*. Retrieved from [un.org https://www.un.org/en/aboutun/languages.shtml](https://www.un.org/en/aboutun/languages.shtml)
- UNESCO. (2003). *Language vitality and endangerment*. Retrieved from <http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf>
- UNESCO. (2009). UNESCO Atlas of the World's languages in danger. Retrieved from <http://www.unesco.org/culture/ich/index.php?pg=0013>
- UNESCO. (2010). *Languages matter: Cultural diversity*. Retrieved from <http://www.unesco.org>
- UNESCO. (2014). Endangered languages. Retrieved from <http://www.unesco.org/new/en/culture/themes/endangered-languages/>
- University of Alberta. (2014). *Canadian Indigenous Languages and Literacy Development Institute*. Retrieved from <http://www.aboriginal.ualberta.ca/en/OurCommunity/CanadianIndigenousLanguagesandLiteracyDevelopmentInstitute.aspx>

- University of London. (2014). *MA language documentation and description*. Retrieved from <http://www.soas.ac.uk/linguistics/programmes/malangdocdesc/>
- University of Minnesota (2014). *Second language requirement*. Retrieved from [http://class.umn.edu/degree\\_requirements/second\\_language.html](http://class.umn.edu/degree_requirements/second_language.html)
- US Library of Congress. (1992). *Country study: Somalia*. Retrieved from <http://countrystudies.us/somalia/36.htm>
- Vadasy, P. F., & Nelson, J. R. (2012). *Vocabulary instruction for struggling students*. New York, NY: Guilford.
- Volkswagenstiftung. (2014). *Documentation of endangered languages*. Retrieved from <http://www.volkswagenstiftung.de/en/funding/completed-initiatives/documentation-of-endangered-languages/>
- World Oral Literature Project. (2013). *About the project*. Retrieved from <http://www.oralliterature.org/about/project.html>
- Wright, S. (1996). (Ed.) *Language and the state: Revitalization and revival in Israel and Eire*. Clevedon: Multilingual Matters.
- Wurm, A. S. (2003). The language situation and language endangerment in the Greater Pacific Area. In Janse, T. (Eds.), *Language death and language maintenance: Theoretical, practical and descriptive approaches* (pp. 15-47). Amsterdam, Neth.: John Benjamins.
- Xiques, D. (2005). *Margaret Laurence: The making of a writer*. Toronto, ON: Dundurn Press.
- Yin, R. K. (2003). *Case study research: Design and methods*. Thousand Oaks, CA: Sage.
- Zampolli, A. (1995). Corpus design criteria. In N. Calzolari, M. Baker, & J. G. Kruyt (Eds.), *Towards a network of European reference corpora: Report of the NERC consortium feasibility study*. Pisa, Italy: Giardini.

Zuckermann, G., & Walsh, M. (2011). Stop, revive, survive: Lessons from Hebrew revival applicable to the reclamation, maintenance and empowerment of aboriginal languages and cultures. *Australian Journal of Linguistics*, 31(1), 111-127. Retrieved from [http://www.zuckermann.org/pdf/Revival\\_Linguistics.pdf](http://www.zuckermann.org/pdf/Revival_Linguistics.pdf)

## Appendices

### Appendix 1: Five Stages for Corpus Design in Challenging Contexts

1. **Phase One—Prototype Stage:** Given the status of vulnerable and endangered languages and considering that they are resource-strained in terms of availability of language sample funds to carry out the work, opportunistic data collection suits the first and second phases of the corpus construction. The first phase is to be used as a blueprint for subsequent design phases. Definition of the scope of the corpus is to be stated and then the development of a framework of criteria-based text classifications and a guiding list of text categories follow. Text collection is for exploration purposes, analysing the data quantitatively and qualitatively in order to demonstrate how they could be used for development and research purposes. Given the limited time and space of this thesis, a working prototype is furnished to demonstrate how a corpus can help an endangered language in terms of documentation and revitalization initiatives. The aim of this thesis is to finish phase one. The study will help inspire a wider collaborative effort which will take this study forward to implement the following four phases.
2. **Phase Two—Incubation Stage:** At this stage, the designer allows time for the corpus to grow without any design limitations but at the same time strives for the collection of as wide a variety of text samples as possible. This effort includes widening spontaneous speech samples across more settings as well as scanning or retyping rare and/or older texts, and even harvesting the web. The main objective should be the collection of a variety of text types as well as the identification and development of a more comprehensive sampling list of known text categories available in a stratified fashion.



The secondary objective is for the designer to build on the guidance and the recommendations of this thesis to refine and develop a diversified stratified sampling list for the third phase by looking at the sampling framework to redefine available text categories and dialect classifications previously employed in the first phase, based on the feedback from phase one.

3. **Phase Three—Monitor Corpus:** This phase turns the opportunistic corpus built thus far into a monitor corpus. At this stage, using the sampling frame developed during the incubation stage, designers should redesign the corpus considering ways to represent the language under study: variability, balance and size. This means the focus should now be on (a) how to represent the language as a whole in terms of variety, (b) balancing samples among texts, and (c) targeting a sufficient size for a large monitor corpus of at least 20 million tokens to start with. The aim is to grow the corpus by adding the same variety of texts in a balanced manner annually. At this stage, based on the needs of the language, specialized corpora can be developed, namely bilingual, literary, or media corpora on which specific and relevant language studies could be based. The bilingual corpus, which starts with English and Somali, is seen as important because it will facilitate cross-linguistic studies and the development of bilingual dictionaries and machine translation tools. Somali literature and its poetry per se, as illustrated in this thesis, are considered to be gemstones buried in the language. It is a comprehensive body of knowledge that Somalis can offer to the knowledge treasures of humanity. For journalists, a corpus is vital because the Somali media industry is arguably the only robust institution that reflects contemporary Somali language. Standardizing and disseminating various learning materials on how Somali is used in the media in relation to how journalists in the past

used the language will unify their efforts towards using a Somali that is standardized and more accessible to all Somali-speaking peoples. This point will be revisited in Chapter Eight where the implications of the study will be described. A written corpus is another specialized corpus that could be developed to hold a wide variety of formal written samples collected in an effort to strengthen the standardization of the Somali language. Brown Corpus is an excellent model for this.

**Phase Four—Annotation:** As Sinclair (1991) pointed out, for a corpus to be useful and meaningful, a large collection of words is not enough; therefore, annotation adds to the richness and usability of the corpus. In general, two types of annotation are considered useful for research or development purposes: *linguistic tagging* and *demographic information annotation*. One basic type of tagging is part of speech (POS) tagging, for which different software taggers are available. Lexical classification such as POS tagging propels the development of spell-checkers and other word-processing tools. Likewise, recording information about the writer of the text or the speaker (e.g., gender, age, geographical information) will inform studies on dialectal variations and standardization efforts. In order to ease the readability of the corpus, the annotated segments of the corpus are better separated from the raw data or the original text. The rationale is that researchers are given the option to access the language in its original version if they so desire.

4. **Phase Five—Representative Monitor Corpus:** The corpus grows and continues to support the needs of a given language. Sustainable growth and redesign, which targets a representative, larger monitor corpus, is the main objective at this stage. While the corpus continues to grow, the immediate needs of the language should be addressed, such as the

development of language tools and materials that could be developed. This process stimulates the designers to refine and redesign the corpus to align the needs of the language and the intended uses of the corpus, by building on the recommendations from the previous four phases. This iterative design continues as the corpus grows. At this stage, accessibility to the public is desirable. First, for the corpus to grow and refine its design, it needs the input and feedback from the research community and others who may need to use the corpus for development or research purposes. Second, a wider accessibility could restore the status of the language and the morale of its speakers, which strengthens the documentation and revitalization objectives of this thesis. A simple user-friendly portal with a search function will suffice. Compatibility, usability, and user-experience considerations should be the guiding principles of the portal design. Enlisting the help of specialists in different areas of expertise is always highly recommended, as the use of technology is multifaceted and requires a collaborative effort in the implementation of projects such as this.

## Appendix 2: Sampling Frame for the Somali Language

Criteria	Criteria-based Classifications		Text Categories in English	Text Categories in Somali
Genre	Written (50%)	Dhab (facts) informative	Non-fiction books, articles, ads, memos, constitution, laws, letter, email, text message, agreements, proposal, dairy, translations	Buug aqooneed, Maqaal, Wareegto, , Lidheh qoran, Dastuur, Xeer, Warqad, Ilmeyl, Farriin-qoran, Heshiis, Hindise, Xusuus-qor
		Dhalanteed (fiction) imaginative	Novels, folktales, play scripts, movie scripts, and all other imaginative works	Buug khayaali ah, sheeko-xariiro qoran, riwaayad, filin, ilaqosol/maad iyo wax kasta oo la curiyo oo dhalanteed ah
	Spoken (50%)	Tix (poetry)	Poems, ballads and other forms of Somali literature	Gabay, Geeraar, Jiifto/masafo, Afarrey, Buraanbur, Hees ciyaareed, Heeso-hawleed, Suugaan kale
		Tiraab (talk) formal	Lecture, meeting, traditional court deliberations, traditional marriage proposal ceremonies, sermons, debates, interviews, press conference, news, commentary, parliament proceedings	Muxaadarocashar, Shir, Gar, Goggolgal, wacdi/khudbo, Dood, waraysi, war-saxaafadeed, war, faallo, doodaha golaha shacabka
		Informal	Chat/conversations, courting, entertainment, sayings, jokes, Riddles, talk-shows, classroom talks	Sheekaysi, Haasaawe, madaddaalo/baashaal, maahmaahyo, ilaqosol, caraatan, doodwadaag, wadal-hadal arday
Topic	Social Sciences			Maamulka, Siyaasadda Taariikh, Juquraafi, Afafka, Sheek-xariiro, Ganacsiga, Madaddaalo, warfaafinta
	Natural Sciences			Xisaab, Fiisigis, Beeraha, Bayooloji, Kimistri, Caafimaadka
Medium	Newspapers, magazines, academic journals, general books, textbooks, audiotapes/CDs, videos, spontaneous speech,			

	lecture, conversations, oral literature, Online texts		
<b>Dialect</b>	<b>Maxaa Tiri Dialects</b>	Northern Dialects	
		Asharaaf Dialects	
		Banader Dialects	
	<b>Maay Tiri Dialects</b>	Maay	
		Digil	

### Appendix 3: Proposed Sampling Frame for Spoken Texts

Genre				Setting/function as texts
<b>Spoken</b>	Tiraab (Talk)	Informal: Sheekaysi (conversations)	Urban Setting (Magaalo)	8. At coffee shops/restaurants: Fadhikudirir, Caadi, Shir 9. At home 10. At school: tertiary/schools 11. At work 12. In the car/bus 13. At the market 14. In entertainment venues
			Country Setting (Miyi)	7. At home 8. At the farm 9. At the grazing land 10. At the village market 11. At the animal market 12. In the entertainment venues

