A PHONETICALLY RICH AND BALANCED LEXICAL CORPUS USING ZIPFIAN DISTRIBUTION FOR AN UNDER-RESOURCED LANGUAGE

AMINATH FARSHANA

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

A PHONETICALLY RICH AND BALANCED LEXICAL CORPUS USING ZIPFIAN DISTRIBUTION FOR AN UNDER-RESOURCED LANGUAGE

AMINATH FARSHANA

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SOFTWARE ENGINEERING

FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Aminath Farshana

Matric No: WGC150023

Name of Degree: Master of Software Engineering

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

A phonetically rich and balanced lexical corpus using Zipfian distribution for an under-resourced language

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 20th October 2017

Subscribed and solemnly declared before,

Witness's Signature Date: 20th October 2017

Name:

Designation:

ACKNOWLEDGEMENTS

In the name of Allah, the most gracious and the most merciful. Peace and blessing be upon Prophet Muhammed ﷺ. Alhamdulillah, all praises to Allah for the strengths and His blessings in completing this thesis. First of all, I would like to pay my gratitude to my supervisor, Dr. Mumtaz Begum Binti Peer Mustafa for her patience and tremendous guidance for this research. This work would not have been possible without her persistent support and encouragement. I am indepted to her for her constant advice and immense ideas for this.

I would also like to thank Maldives Customs Service for providing me the support and rendering all the help I needed throughout my years of study and this research. Also, I would like to express my gratitude to all my friends for their enthusiasm and faithfulness to me till this thesis is completed.

My parents, Mr. Ahmed Aslam and Mrs. Ramla Ibrahim and siblings, Mr. Mohamed Nishan and Ms. Aishath Nishana deserve a special mention for their inseparable motivation and prayers. Special thanks to my grandmother Mrs. Hawwa Abdulla, who passed away while conducting this research. I am thankful to all my family members who was important to the successful realization of this thesis.

Finally, I must express my very profound gratitude to my loving husband, Mr. Ali Dawood. Words will fail to express my appreciation to my husband whose dedication, love and confidence, has taken the load off my shoulders. His unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without him. For myself, this thesis has been an initiation in acquiring knowledge. I pray to the Most Gracious, and Most Merciful for bless.

ABSTRACT

In recent times, speech technology and its related applications are becoming a popular topic among researchers. There are many applications of speech technology developed for businesses, military, transport, aerospace, PDAs, and so on. The importance of speech technology-based applications has prompted researchers to improve the techniques of these applications for many languages around the world. However, only limited number of languages benefited from speech technology applications such as the Automatic Speech Recognition (ASR) system and the Text-to-Speech (TTS) system. One of the main reasons for this technological gap between the languages is the lack of basic resources such as the lexical and speech corpus, which are essential as the foundation for developing this technology. Though researchers have managed to assemble these basic resources for some languages, the methods used for accumulating them are not as efficient as of the established languages. Some of these methods also depend on the types of resources needed for developing lexical and speech corpora. This research emphasizes on developing a lexical corpus for an under-resourced language that lacks the basic resources. This research also focuses on improving the quality of the corpus in terms of phonetic coverage and corpus size for the related under-resourced language. Developing a lexical corpus includes collecting an initial large corpus, and selecting suitable sentences therein. The selected set of sentences must cover all possible phonetic units of the language and ensuring uniform distribution of those units. This research proposed a novel method the development of a lexical corpus for Dhivehi, a language that lacks in key resources for developing speech technology-based applications. This research proposed the use of Zipfian distribution for selecting sentences from the initial large corpus. From 109,208 sentences collected from web sources, 360 sentences were selected to ensure a phonetically rich and balanced lexical corpus. The performance of the developed corpus is evaluated in terms of phonetic coverage and size of the corpus. Phonetic coverage is measured by finding the sum of the sequence of phonemes in the corpus. The size of the corpus is evaluated using the cosine similarity, which measures the frequency distribution of the phonemes occurring in the developed final corpus and comparing them with the large initial corpus. The closer the similarity between final and large corpus, the better is the phonetic coverage. High similarity between the two corpora indicates that the developed corpus using the proposed method can perform as efficient as the initial large corpus. Statistical phonetic unit distribution similarity of selected sentences was 0.988 as compared to phonemes distribution of the large corpus. Since the similarity of the two distributions is close, it means that the optimized corpus can perform as efficient as the larger corpus. The performance of the proposed method was also evaluated by comparing the results with an existing benchmark method (greedy algorithm). The results show that the sentences selected using proposed method cover all the phonetic units and is 14 times smaller than the corpus developed using the benchmark method.

ABSTRAK

Sejak kebelakangan ini, teknologi ucapan dan aplikasi berkaitannya menjadi topik yang popular di kalangan penyelidik. Terdapat banyak aplikasi teknologi ucapan yang dibangunkan untuk tujuan perniagaan, tentera, pengangkutan, aeroangkasa, PDA, dan sebagainya. Kepentingan aplikasi berasaskan teknologi pertuturan telah mendorong para penyelidik untuk memperbaiki teknik aplikasi ini untuk banyak bahasa di seluruh dunia. Walau bagaimanapun, hanya sebilangan bahasa yang mendapat manfaat daripada aplikasi teknologi pertuturan seperti sistem Pengenalan Ucapan Automatik (ASR) dan Sistem Sintesis Pertuturan (TTS). Salah satu sebab utama jurang teknologi antara Bahasa adalah kekurangan sumber asas seperti korpus leksikal dan ucapan rakaman, yang penting sebagai asas untuk membangunkan teknologi ini. Walaupun para penyelidik berjaya mengumpulkan sumber-sumber asas ini untuk beberapa bahasa, kaedah yang digunakan untuk mengumpulnya tidak begitu cekap seperti bahasa yang telah ditetapkan. Beberapa kaedah ini juga bergantung pada jenis sumber yang diperlukan untuk membangun korpora leksikal dan ucapan. Penyelidikan ini memberi penekanan untuk membangunkan korpus leksikal bagi bahasa yang tidak mempunyai sumber asas. Kajian ini juga menumpukan kepada peningkatan kualiti korpus dari segi liputan fonetik dan korpus untuk bahasa yang berkaitan dengan bahasa yang berkaitan.

Membangunkan korpus leksikal termasuk mengumpul korpus besar awal, dan memilih ayat-ayat yang sesuai di dalamnya. Set ayat yang dipilih mesti meliputi semua unit fonetik bahasa yang mungkin dan memastikan pengedaran seragam unit tersebut. Penyelidikan ini mencadangkan satu kaedah baru pembangunan korpus leksikal untuk Dhivehi, sebuah bahasa yang tidak mempunyai sumber utama untuk membangunkan aplikasi berasaskan teknologi pertuturan. Kajian ini mencadangkan penggunaan “Zipfian distribution” untuk memilih ayat-ayat dari corpus besar awal. Dari 109,208 ayat yang dikutip dari sumber web, 360 ayat telah dipilih untuk memastikan korpus lexical yang kaya dan seimbang secara fonetik. Prestasi korpus yang dibangunkan dinilai dari segi liputan fonetik dan saiz korpus. Liputan fonetik diukur dengan mencari jumlah jujukan fonem dalam korpus. Saiz korpus dinilai menggunakan persamaan kosinus, yang mengukur pengedaran frekuensi fonem yang berlaku dalam korpus akhir yang dibangunkan dan membandingkannya dengan korpus awal yang besar. Lebih dekat kesamaan antara korpus akhir dan besar, lebih baik adalah liputan fonetik. Persamaan yang tinggi antara kedua-dua corpora menunjukkan bahawa korpus yang dibangunkan dengan menggunakan kaedah yang dicadangkan dapat berfungsi dengan cekap sebagai corpus besar awal. Kesan pengedaran unit fonetik statistik ayat-ayat terpilih adalah 0.988 berbanding dengan pengedaran fonem korpus besar. Oleh kerana kesamaan kedua-dua pengedaran itu hampir, ini bermakna bahawa korpus yang dioptimumkan dapat berfungsi sebagai korpus yang lebih besar. Prestasi kaedah yang dicadangkan juga dinilai dengan membandingkan hasil dengan kaedah benchmark yang sedia ada (algoritma greedy). Keputusan menunjukkan bahawa ayat-ayat yang dipilih menggunakan kaedah yang dicadangkan meliputi semua unit fonetik dan 14 kali lebih kecil daripada korpus yang dibangunkan menggunakan kaedah benchmark.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ASR	:	Automatic Speech Recognition
TTS	:	Text to Speech
IPA	:	International Phonetic Association
WER	:	Word Error Rate

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

Speech is considered as one of the most powerful and richest tool for communication. It is therefore undoubtedly considered as the ideal strategy to interact with computers. Speech technology research and development has seen a massive progress in recent decades and are gaining attention from users all over the world. Many significant research groups have expanded their scientific works towards improving the quality and naturalness of the speech applications, and presented more and more intelligent high-quality techniques to generate human-like speech. This is especially the case for dialogue machines, smart watches, smart houses, voice verifications systems, machine translators, and virtual personal assistants. This form of technology and its related application are available for major languages such as English and Japanese. However, many other languages (known as the under- resourced languages) that do not enjoy the benefits of speech technologies, which is mainly due to lack of key language-based resources required for the development of speech technologies.

Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) systems are the two prominent focus of speech technology. TTS system is a computer-based system that converts any written text into natural sounding audio in a variety of languages and voices. TTS system is also commonly known as the speech synthesis system, as it generates human-like synthetic speech. On the other hand, ASR system converts an audio into readable text.

One of the important components for the development of these systems is the speech and lexical corpus. Development of a good quality speech and lexical corpus requires expert linguistic knowledge that is currently available for major languages such as English, Chinese, and French. However, the under-resourced languages such as Dhivehi, Mirandese, Scottish Gaelic, and Haitian Creole, there are none or very few expertise is available.

TTS and ASR system are of great benefit to both the organizations and individuals. People with physical disabilities such as Stephen Hawkings, who has Amyotrophic Lateral Sclerosis (ALS), uses TTS system to communicate his brilliant ideas by simply typing the text and the TTS system will read-out loud the text. On top of that, people with learning disabilities and pronunciation issues also make use of TTS system to overcome their inadequacies. Today, TTS and ASR system are embedded in smart devices used by the individuals to assist virtually all everyday activities such as reminders, wake up calls, translators, and so on.

Various business websites use the TTS and ASR system to reach their potential customers. Educational institutions use the TTS system to improve word recognition skills, and increase the pronunciation capabilities of the students.

1.2 Research Background

This subsection describes the fundamentals of the phonetically rich and balanced lexical corpus and the process of developing such corpus.

1.2.1 Phonetically Rich and Balanced Lexical Corpus

Acoustic model is a critical component for both TTS and ASR systems. Acoustic models represent the relationship between the written text and sound. In precise, they convert the phonemes in the texts into sound. Technically, the statistical representations of each of the distinct sound (speech units) that make up the word of a language are stored as the acoustic models. These models require a database of speech (also known as speech corpus) to represent the phonemes. Phonemes are the distinct units of speech of any language They distinguishes the words from each other. For example, English language has 26 alphabets with 44 phonemes each representing distinct sound. The performance of the speech processing systems depends on the quality of the acoustic models. At the same time, the acoustic models depends on the quality of the speech corpus that have a good coverage of a language, to generate human-like speech (Rabiner & Schafer, 2007).

Speech corpus is a collection of recorded speech of a set of phrases or text. However, a text or phrase by itself is not always phonetically rich and balanced. The term phonetically rich refers to text corpus that contain high variety of speech units of a target language, while phonetically balanced refers to text corpus that use specific speech unit at the same frequency.

(Mendonca et al., 2014) found that phonetically rich and balanced corpus is required to estimate the acoustic models. Many researchers have explained the definition and importance of phonetically rich and balanced lexical corpus. Abushariah et al. (2012) stated that a robust ASR system requires a set of recordings that are rich and balanced. It was also stated that the number of occurrence of all the phonemes defines the richness characteristics of a text, while the frequency distribution of the phonemes defines the balanced characteristics of the corpus. (Yuwan & Lestari, 2016), explained that creating phonetically rich and balanced corpus not only makes the system more robust and intelligent, but also saves time and storage capacity. They have developed a phonetically rich and balanced corpus for Quran ASR system by providing minimum amount of

Quranic verses, while at the same time covers all the phonetic aspects of the Quran. The proposed solution saves both the storage space and time for developing the speech corpus for the Quran.

(Malviya, Mishra, & Tiwary, 2016) state that a set of sentences can be considered phonetically rich based on two statistical properties of phonemes, which are the characteristic distribution of phonemes, and the resemblance of the phonemes. They have experimented this concept by developing a phonetically rich and balanced lexical corpus for Hindi, and evaluated its effectiveness by measuring the cosine similarity of the speech corpus. (Abera, Nadeu, & Mariam, 2016) have developed a phonetically rich and balanced lexical corpus for Tigrigna, where phonetic richness is defined as the sentences with high variety of phonemes, while balanced characteristics is defined by the frequencies of the occurrence of the phonemes. On the similar note, (Mendonca et al., 2014) as well as many other researchers state that sentences that closely resemble the phonetic characteristic of the target language and are distributed uniformly, is known as phonetically rich and balanced corpus (Arora, Arora, Verma, & Agrawal, 2004; Raza, Hussain, Sarfraz, Ullah, & Sarfraz, 2009; J. S. Zhang & Nakamura, 2008)..

To train the acoustic models of speech processing applications and to improve the performance of speech-based applications, a phonetically rich and balanced corpus speech database is of high importance. A good quality speech database must represent variety of speech units and this is achieved from using sentences that have high variety of speech units that are distributed uniformly.

Lexical corpus should be phonetically rich and balanced to make the speech-based system more robust. A text corpus is phonetically rich if it contains the maximum speech units of the target language and the uniform distribution of the speech units means that the text is phonetically balanced. The phonetic richness of the texts is measured by counting and comparing the number of phonemes of the corpus with the phonemes of the official dictionary of the target language (Abushariah, Ainon, Zainuddin, Elshafei, & Khalifa, 2012; Nguyen, 2015; Raza, Hussain, Sarfraz, Ullah, & Sarfraz, 2009; W. Zhang, Liu, Deng, & Pang, 2010).

The phonetic balance of the corpus is measured by evaluating the frequency distribution of each speech unit (Gutkin, Ha, Jansche, Pipatsrisawat, & Sproat, 2016; Malviya, Mishra, & Tiwary, 2016; Vorapatratorn, Suchato, & Punyabukkana, 2012).

1.2.2 Process of lexical corpus development

Lexical corpus or text corpus is a large body of structured texts in machine-readable format. It is a collection of individual texts of same or different genre, which are developed for various purposes in the NLP. However, for the development of ASR and TTS system, lexical corpus that is phonetically rich and balanced is required. The process of developing phonetically rich and balanced lexical corpus involves the selection of sentences from a large text data based on various stochastic methods. High-quality speech corpus is then developed by recording the extracted sentences, as shown in Figure 1.1.

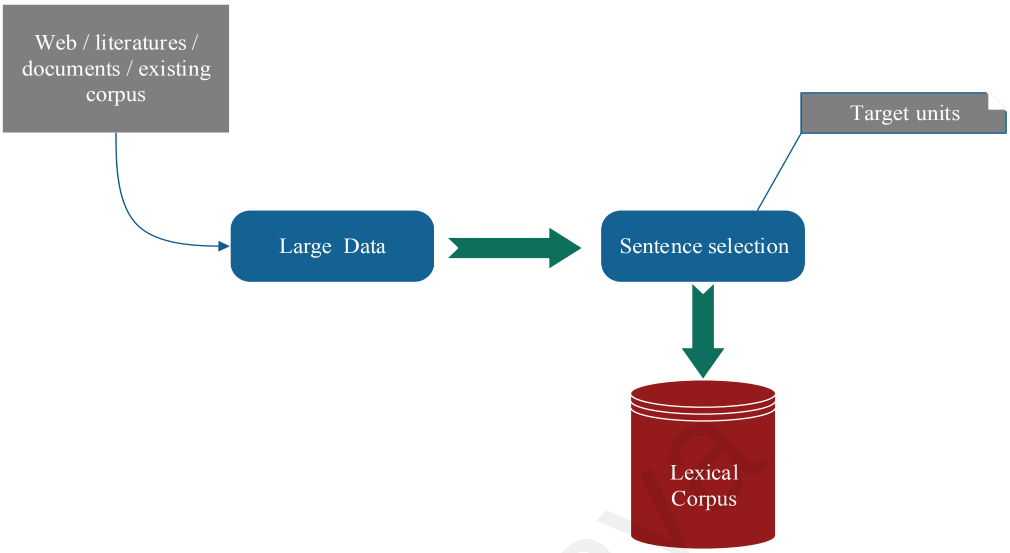


Figure 1.1 Process of lexical corpus development

- **Large Text Data:**

  Before extracting the phonetically rich sentences, a large text data must be analyzed and observed. This large text data is accumulated from various sources such as online sources and physical sources like the dictionary or newspaper of the target language. Some of the languages such as English and Chinese have its commercially available large text data such as TIMID for English, and LCMC for Chinese. Researchers have made use of these text data to extract sentences that are phonetically rich and balanced. However, most of the under-resourced languages do not have any existing text data. As such, the researches need to accumulate the texts from various resources such as web, literatures, crowdsourcing, and other documents. Unlike the commercially available corpus, the text data collected from resources such as web need to be refined to incorporate the linguistic features of the target language. Sentences are onlyextracted once the text data is gathered, processed, and analyzed.

- **Sentence Extraction:**

  The most important stage of developing phonetically rich and balanced lexical corpus is the sentence extraction. Sentences are generally extracted based on a certain criterion (domain, resource availability and needs of the application area) with the aim to maximize the coverage of the speech units of a target language while optimizing the large text data. Most of the researchers has employed words (Tan & Sh-Hussain, 2009), syllables (Abera, Nadeu, & Mariam, 2016; Wang, 1998), and phonemes (Abushariah et al., 2012; Bansal, Sharan, & S.S, 2015; Yuwan & Lestari, 2016) as the basic speech units. However, more recent studies has used contextual units such as diphones (W. Zhang et al., 2010), triphones (Malviya et al., 2016; Mendonca et al., 2014) and n-grams (Habib & Adeeba, 2014) as the basic speech units.

  Researchers have proposed many heuristic methods to a build phonetically rich and balanced corpus (Abushariah, Ainon, Zainuddin, Elshafei, & Khalifa, 2012; Arora et al., 2004; Aubanel, Lecumberri, & Cooke, 2014; Malviya et al., 2016; Uraga & Gamboa, 2004; Wang, 1998; Yuwan & Lestari, 2016), where the focus of these works are on the well-established languages. Very few studies (Abera et al., 2016; Mendonca et al., 2014; Nguyen, 2015) that focuses on the development of lexical corpus for under-resourced languages.

  There is a rapid growth and improvement in speech technology over the recent years, which focus more on the well-developed languages. There are almost 98% of the languages around the globe that lack the basic resources such as speech corpus required as the foundation for developing speech-based applications (Kilgarriff et al., 2008; Scannell, 2007).

Developing a lexical corpus for languages with limited or zero resources can be slow, expensive, and difficult. Under resourced language require more human e Hence, there is a dearth of research in the development of lexical corpus for the majority of the under-resourced languages. On top of that, the existing researches on lexical corpus development for under-resourced languages have not achieved the adequate phonetic coverage when compared with major well-resourced languages.

1.3 Research Motivation

In the past few years, many researchers undertaken significant amount of works to improve the performance of the speech processing systems. One way to achieve this is by developing a phonetically rich and balanced corpus, to make the acoustic model of the speech processing systems to be more robust. Researchers have proposed many methods for developing phonetically rich and balanced corpus to address the issues in developing the corpora for many of the mainstream languages. However, there is lack of research in developing lexical corpus for under-resourced languages, which is the reason that motivates the author to conduct this research.

Currently there are very few researches conducted to overcome the issue of developing a phonetically rich and balanced lexical corpus for under-resourced languages (Abera et al., 2016; Gutkin, Ha, Jansche, Pipatsrisawat, & Sproat, 2016; Mendonca et al., 2014; Nguyen, 2015; Scannell, 2007). Various methods were proposed in these researches to extract sentences that are phonetically rich and balanced. Some have used the large text data, as it is in order to cover more phonemes of the target languages. However, using a large corpus has many issues, and the performance of these corpora is not up to the standard as that of well-resourced languages. A good quality speech corpus will motivate the researchers and developers to improve and to build more speech-based applications for the under-resourced languages, which can reduce the technological gap between the mainstream and under-resourced languages.

While there are several methods proposed in the literature for developing lexical corpus, most of them are language dependent and may not be applicable for other under-resourced language. As such, this is another reason that motivates the author to develop a lexical corpus for my native language, Dhivehi, which is also an under-resourced language, and at the same time proposed a method that can be easily applicable for other languages as well.

1.4 Research Problem

Traditionally, speech corpus is derived from a large text corpus (Habib & Adeeba, 2014; Matoušek, Tihelka, & Romportl, 2008; Wang, 1998). Many researchers have highlighted the issues of developing speech corpus from such a large text corpus, such as long calculation time (Nose et al., 2015), difficulty in identifying relevant linguistic contents (Chevelu, Barbot, Boëffard, & Delhay, 2007), and high calculation cost (Kasparaitis & Anbinderis, 2014). Developing lexical corpus that are phonetically rich and balanced with optimized text to cover the phonetic is now a popular approach (Abushariah et al., 2012; Kasparaitis & Anbinderis, 2014; Malviya et al., 2016; Raza et al., 2009; Uraga & Gamboa, 2004; Vorapatratorn et al., 2012).

Researchers have proposed various methods for selecting the least number of possible sentences using some of the well-known algorithms such as the greedy algorithm(Anumanchipalli et al., 2005; Bansal et al., 2015; Habib & Adeeba, 2014; Matoušek et al., 2008; Nguyen, 2015; Vorapatratorn et al., 2012). Other methods includes sentence selection method (Abera et al., 2016; Arora, Arora, Verma, & Agrawal, 2004; W. Zhang et al., 2010), word frequency method (Raza et al., 2009; Tan & Sh-Hussain, 2009), and the two stage algorithm (Wang, 1998; Yuwan & Lestari, 2016). These methods were found to be effective for most of the languages such as Arabic (Abushariah et al., 2012), Urdu (Habib & Adeeba, 2014; Raza et al., 2009), Punjabi (Bansal et al., 2015), Marati, Tamil, Telugu (Anumanchipalli et al., 2005), and Vietnamese (Nguyen, 2015).

However, this approach is not applicable to many under-resourced languages due to the increased need of human experts to refine and make sure that the sentences contain only words of the respective language, and insufficient resources for using these methods. In addition, the results on few under-resourced languages are not as satisfactory in comparison with well-established languages. The maximum phonetic coverage (phonetically rich and balanced) can only be obtained with the use of large text corpus. A possible cause of this issue is the use of same criteria to extract sentences for both the mainstream and under-resourced languages, as well as the dependency on the basic resources for developing the corpus. The nature of the resources such as source data varies among languages. Thus, the method used must also cope with the phonological variabilities among languages like stressed and unstressed vowels. Hence, a suitable method to extract sentences covering maximum phonetic units with minimal size regardless the nature of the resources used is needed.

1.5 Research Objectives

The aim of this research is to propose a method to develop a phonetically rich and balanced corpus for an under-resourced language with minimal size and covering maximum phonemes of the target language. This research presents a novel method for building a phonetically rich and balanced corpus for an under-resources language.

The following objectives were set to achieve the main aim of this research:

1. To analyze the performance of the existing methods and identify suitable method(s) for lexical corpus development of an under under resourced language, for achieving phonetic richness and balanced (maximum phoneme coverage) with small data size.
2. To develop a lexical corpus for an under-resourced language using the identified method(s)
3. To evaluate the performance of the proposed method(s) with the benchmark method(s)

1.6 Research Scope

This research focusses on the development of a phonetically rich and balanced lexical corpus for an under-resourced language that can be used on various speech-processing systems. This research concentrates on developing a lexical corpus for Dhivehi, an under-resourced language, which also lacks the availability of commercially available resources. Hence, the obvious choice of the kind of data used to conduct this research is the web source such as Facebook, Twitter, and many other Dhivehi websites, due to its abundance, and free availability.

1.7 Research Methodology

To achieve the main aim of this research the following research methodology is adopted.

- **Identifying Problem and Solutions**

The first stage of this research is to conduct the review on the existing literature that focuses on the lexical corpus development. The purpose of the review is to identify the limitations of the existing methods, and reasons for the lack of in the adoption of these methods for under-resources languages. In addition, the reviews focuses on the data and other resources required for the development of speech-based applications. The findings of the review will form the basis for identifying the suitable solutions for developing a lexical corpus for the under-resourced language.

- **Data Collection**

In this stage, the accumulations of resources required for the development of the lexical corpus are performed. The steps involved in data collection are as follows:

- Selecting a suitable source for scrapping web data
- Identify tools and techniques for scrapping web contents
- Refining the accumulated data
- Listing out the desired phonetic units

- **Solution, Design and Implementation**

Based on the findings from the literature review, this research proposes a method to improve the process towards the development of lexical corpus with acceptable quality. The proposed method is used to design and build a lexical corpus for Dhivehi language, a language that is classified as under-resourced languages.

- **Evaluation**

  An evaluation on the performance of the proposed method in term of phonetic richness and balanced is performed by comparing the results of the proposed method with the benchmark method in developing phonetically rich and balanced lexical corpus. The evaluation will includes factors like size and phonetic coverage of the lexical corpus

1.8 Expected research outcomes

By the end of this research, the following outcomes are expected:

- A method to develop a phonetically rich and balanced lexical corpus with limited data, and
- A phonetically rich and balanced lexical corpus for Dhivehi language.

The proposed method allows the accumulation of essential resources for under-resourced language, which then helps in the research and development of speech-based applications for those languages. In addition, this research reduces the technological gap between Dhivehi and other major languages, in term of resource availability, enabling the development of speech-based applications for this language.

1.9 Significance of the research

Today, users are trending on speech-based application from all over the world. Unfortunately, the performance of speech applications for under-resourced languages is not as promising as the well-resourced mainstream languages due to the lack in essential resources. It is crucial to implement a method that improves the accumulation of resources and the performance of the speech processing system of under-resourced languages. A lexical corpus for Dhivehi language developed using suitable method that improves the phonetic richness and balance of the developed lexical corpus. This phonetic rich and balance corpus can overcome the issue of resource scarcity for Dhivehi language and inspire developers to develop various speech- based applications for Dhivehi language.

1.10 Thesis Organization

The rest of the dissertation is organized as follows:

Chapter 2 reviews the relevant literatures on the lexical corpus developments. Essentially, the existing methods for the development of the corpus for different languages are presented, while the resources used to build the lexical corpus, is examined. This section also includes the comparison in the performance of the existing methods.

Chapter 3 is dedicated on the proposed method and its related justification.

Chapter 4 provides the details of the steps of the proposed method used for the development of lexical corpus. This chapter also explains the challenges faced during the development.

Chapter 5 presents the evaluation of the proposed method. It discusses the performance of the proposed method, as well as the comparison against the benchmark method.

Chapter 6 summarizes the major findings of this research, the contribution, the limitations, as well as possible future research direction.

CHAPTER 2: LITERATURE REVIEW

Lexical corpus is an important component for both the development and evaluation of speech processing systems. Other than ASR and TTS system, lexical corpus is also used in natural language processing systems such as speaker verification/recognition and spoken language systems. A phonetically rich and balanced speech corpus is essential for the performance of these applications. For instance, the acoustic models of the ASR and TTS system require a rich speech corpus with adequate representation of all speech units (Rabiner & Schafer, 2007), while language phonologists require such corpus for performing the analysis of speech production and variabilities (Pierrehumbert, Beckman, & Ladd, 2012). In speech therapy, phonetically-rich speech corpus are often used to assess the patient’s speech production (Yang et al., 2014).

To build an efficient speech corpus for these applications, a high-quality lexical corpus is of great importance. The development of a phonetically rich and balanced lexical corpus includes extracting a set of sentences from a large data, allowing the identification of a subset of sentences from a source data that covers all the speech units required. This chapter explains the existing methods on sentence selection, the evaluation of the existing methods, and the comparison on the results of the existing methods.

2.2 Methods for sentence selection

In recent times, researchers have placed significant efforts towards the development of high-quality lexical corpus for many of the languages. Sentences are selected for the speech corpus using different methods and different speech units. Selecting sentences is an iterative approach where the sentences are scored based on a certain criterion. The aim is to maximize the coverage of the target units by selecting a minimum number of sentences from the source corpus. Different methods have been proposed in the literature to fulfill this aim, and each method has its own criteria for the sentence selection. This section discusses the existing methods and criterions used for the sentence selection.

Table 2.1 provides the summary of the existing methods for the development of phonetically rich and balanced lexical corpus

Table 2.1: Existing methods for lexical corpus development

Paper	Methods	Criterion	Language	Domain
Bansal et al., 2015	Greedy algorithm	High ranks to sentences with highly frequent units	Punjabi	TTS
Habib & Adeeba, 2014	Greedy algorithm	High ranks to sentences with highly frequent units	Urdu	TTS
Nguyen, 2015	Greedy algorithm	High ranks to sentences with highly frequent uncovered units	Vietnamese	TTS
Anumanchipalli et al., 2005	Greedy algorithm	High ranks to sentences with highly frequent diphone units	Tamil, Marati, Telugu	TTS
Matoušek et al., 2008	Greedy algorithm (Modified)	Setting a condition that sentences selected must contain n-times the units	Czech	TTS
Vorapatratorn et al., 2012	Greedy algorithm (Modified)	Both phonetic coverage and distribution	Thai	TTS, ASR
Arora et al., 2004	Sentence selection algorithm	High ranks to sentences with unique units	Hindi	TTS, ASR
Abera et al., 2016	Sentence selection algorithm	High ranks to sentences with unique units	Tigrigna	ASR
Zhang et al., 2010	Sentence selection algorithm	High ranks to sentences with rare units	English	TTS
Tan & Sh-Hussain, 2009	Word frequency	High frequent words are selected	Malay	TTS

Table 2.1: Continued

Paper	Methods	Criterion	Language	Domain
Raza et al., 2009	Word frequency + phoneme frequency	High frequent words are selected and then subsequent phonemes are weighted	Urdu	TTS, ASR
Wang, 1998	Two stage algorithms	Statistical analysis of speech units and then score the sentences with their frequency	Mandarin	ASR
Yuwan & Lestari, 2016	Two stage algorithms (Modified)	Statistical analysis of speech units and then score the sentences with their frequency	Arabic	ASR
Malviya et al., 2016	Probabilistic metric + greedy algorithm	Pre-select sentence with heuristic metric for greedy algorithm	Hindi	TTS, ASR
Mendonca et al., 2014	Probabilistic metric + greedy algorithm	Pre-select sentence with heuristic metric for greedy algorithm	Brazilian Portuguese	TTS, ASR
Abushariah et al., 2012	Characteristics and guidelines of the language	Words are manually selected based on the guidelines	Arabic	ASR
Villaseñor-Pineda et al., 2004	Lexicon based phrase selection	Phrases containing words from lexicon dictionary is selected	Mexican Spanish	ASR
Shinohara, 2014	Submodular optimization approach in greedy algorithm	Defined "utility" of the sentence as weighted sum of log- frequency of desired units	Japanese	TTS, ASR
Chevelu et al., 2007	Lagrangian based algorithm for multi-represented SCP (LamSCP) algorithm	Has 3- phases: 1. Sub-gradient phase 2. Heuristic phase 3. Column fixing phase	French	TTS, ASR

The main idea of all these methods is ranking the sentences available in the large lexical corpus based on certain criterion set based on the target units and source data. In some of these methods, criterions are set to overcome the problems of the existing selection methods. While several researches claimed that the use of contextual units improves the robustness of the system, others claims that on non-contextual units also proved to get satisfactory results. From table 2.1, it is clear that most of the methods are used on well-resourced languages.

2.2.1 Greedy algorithm

Greedy algorithm is the one of the classic method used by the researchers to select sentences. In this method, each sentence is given a score based on the high weights to the most frequent speech units. (Bansal et al., 2015), (Habib & Adeeba, 2014) and (Anumanchipalli et al., 2005) have experimented this method, and claimed that this method produced a reduced corpus with better distinct units. (Matoušek et al., 2008) have also used this method but applies an additional condition that the sentences selected must contain the phonetic unit at least n-times, where n ranges from 12 to 50. The study claimed that this method maximizes the overall distribution of diphones in the selected sentences. (Vorapatratorn et al., 2012) have developed an internet-based continuous sentence selection method using the customs phonetic distribution that adapts the greedy algorithm to select sentences with certain perfect target ratio for each phonetic pattern. (Nguyen, 2015) has developed lexical corpus for Vietnamese, where the greedy method was repetitively applied to select a custom-made speech units called di-tonophones until 100% of them is covered, by giving high weights to most frequently used uncovered units.

2.2.2 Sentence selection algorithm

Researchers have also conducted experiments by giving high scores for sentences with unique speech units, where method was used to build lexical corpus for Hindi language using variety of data from GyanNidhi Corpus that contain 46,421 words from 5,147 extracted sentences, which was encouraging (Arora et al., 2004). On similar note, (Abera et al., 2016) to extract sentences but use syllables as the speech units, for the development of the very first lexical corpus for Tigrigna. (W. Zhang et al., 2010) also used the same sentence selection algorithm but gives high ranks to sentences with rare units instead of unique units, for developing a lexical corpus with reduced data for English language. It was found that applying high weights to rare phonemes can improves the performance of the lexical database with complete coverage is feasible This experiment used web data, which was auto-downloaded by web crawler as the source corpus.

2.2.3 Word Frequency method

Whilst most of the methods used contextual speech units as target units, (Tan & Sh-Hussain, 2009), and (Raza et al., 2009) have targeted on words as units to selecting sentences. (Tan & Sh-Hussain, 2009) have developed 381 sentences with 16,826 phonemes covered by choosing 70% of the high frequency words from a large database of 10 million words, covering more phonemes than the Malay phoneme dictionary. (Raza et al., 2009) also used word frequency to select sentences, but by giving priority to the frequency of occurrence of the word in the corpus, and also by assigning the subsequent weight to the number of triphones, which is suitable for languages which lack of strong phonetic inventory. They have developed a speech corpus for Urdu using 725 sentences with 5,681 words.

2.2.4 Two stage algorithm

Analyzing the statistical distribution of the phonetic units and assigning weights to them is one of the most promising methods for sentence selection. (Wang, 1998), introduced a two-stage algorithm to automatically extract phonetically rich sentences for Mandarin, where the Mandarin speech units are statistically analyzed, and sentences are selected by giving scores to the units based on the frequency of their occurrence in the source corpus. In the first stage, phonetically rich sentence with all recognized units are selected, and in the second stage, phonetically rich sentences with a statistical distribution similar to that of the source data is extracted. On the same note, Yuwan & Lestari (2016) developed a phonetically rich and balanced speech corpus for Quran ASR system, where verses are selected continuously until all the phonemes are covered from the QScript.

2.2.5 Method based on phonetic guidelines

(Abushariah et al., 2012) defined “richness” as the word that contains all the Arabic phonemes, which are similar in frequency, and that contains all the phonemes that preserves the phonetic distribution of the target language. They have extracted 663 phonetically rich words based on certain phonological characteristics and guidelines of the Arabic language. Although the experiment produces satisfactory results, two linguistic experts prepared the 663 manually, which us tedious, troublesome, and not well suited for most of the languages particularly the under-resourced languages.

2.2.6 Probabilistic metrics

Probabilistic metric is another common method used for sentence extraction, in which the Greedy algorithm is used to select sentences along with suitable metrics to pre-select sentences. (Malviya et al., 2016) have used the probabilistic Metrix and Greedy algorithm to extract sentences with high triphone coverage for Hindi language. The Greedy algorithm was used for the sentence selection, while the probabilistic metric was devised to pre-select and rank sentences, in which high ranks are given to sentences that contain the least frequent units. In similar note (Mendonca et al., 2014), used the heuristic metrics to pre-select sentences for greedy algorithm to speed up the execution time for selecting Brazilian Portuguese sentences from web sources (wiki dumps).

2.2.7 Submodular optimization approach

(Shinohara, 2014) claimed that the heuristic methods are not the optimal solutions and proposed a new method to develop phonetically balanced sentences. (Shinohara, 2014) introduces the concept of "utility", which refers to the sum and balance of a sentence. On top of that, the performance of the Greedy algorithm was maximized with the use of submodular objective function. A submodular function is a set function that reduces the size of the initial input set of elements (phonemes) as the size of the final set (sentences) increases. Although this method offers a better solution for optimization than the other heuristic methods, it highly dependent on the source corpus with high linguistic features to analyzes the amount and phonetic distribution.

2.2.8 Lexicon based phrase selection method

In this method, sentences are selected when a specific words or lexicons are given as the input. Sentences will be selected based on these words or lexicons. (Villaseñor-Pineda, Montes-y-Gómez, Vaufreydaz, & Serignat, 2004) have developed a phonetically rich lexical corpus from a large data collected from web, in which the selection is based on the words found in two Spanish dictionaries, newspapers, and magazines. Most of the studies on selecting sentences includes phonemes, but this study considered lexicons or set of words to select sentences.

2.2.9 Lagrangian relaxation method

(Chevelu et al., 2007) used the lagrangian relaxation principles to select a subset of sentences from a large French lexical corpus, which was found to be more efficient than the standard Greedy algorithm. This method has three phases, which are the Sub-gradient phase (focusing on optimization), the heuristic phase (focusing on the coverage), and column fixing phase ('promising' sentences are selected).

2.3 Performance of the Existing Sentence Selection Methods

The problem of searching the best sentence set to maximize the coverage of the lexical is one of the key issues in resource accumulation for speech-based application research. Researchers attempted to address this issue by proposing various methods. This section analyses the performance of the existing methods in the literature and examines the links between the performance of these methods with the size and resources of the lexical corpus. Table 2.2 depicits the performance of the existing sentence selection methods towards the development of phonetically rich and balanced lexical corpus for speech processing systems.

Table 2.2 Summary of existing sentence selection methods towards the development of phonetically rich and balanced lexical corpus

Method	**Article**	**Language**	**Language Type**	**Source data**	**Source data (Size)**	**Final Corpus (Size)**	**Target unit**	**Outcomes**
Greedy algorithm (High weights to more frequent units)	(Bansal et al., 2015)	Punjabi	URL	Web	300,000 words	1500 words 300 sentences	Punjabi phonemes	Speech corpus
	(Habib & Adeeba, 2014)	Urdu	URL	Text books	35m words	70,000 words	Urdu phonemes & the corpus itself	99.1999% (ngrams & phonemes)

	(Matoušek et al., 2008)	Czech	URL	Random texts	-	17,378 utterances		Speech corpus
	(Anumanchipalli et al., 2005)	Marati, Tamil, Telugu	URL	CIIL Corpus	Marati: 155541; Tamil: 303537; Telugu: 444292 (sentences)	52 sentences randonmly selected for each speaker (total 200 speakers)	Diphones from CIIL corpus	WER % (Word Error Rate) Marati: 23.2; Tamil: 20.2; Telugu: 28
	(Nguyen, 2015)	Vietnamese	URL	Web	323,934 sentences	Set A: 630 sentences	MEA-SYLDIC	Set A: 95.1% coverage

						Set B: 983 sentences Set C: 334 sentences		of phonemes Set B: 100% of phonemes Set C: 100% coverage of phonemes
	(Vorapatratorn et al., 2012)	Thai	URL	Mixed	3007 sentences	1000 sentences	LVCSR corpus phonemes	99.13% coverage of phonemes
Sentence selection algorithm (High	(Arora et al., 2004)	Hindi	URL	GyanNidhi Corpus	More than millions of	5147 sentences; 46421	Syllables (from the cor	Distribution of syllables in both

weights to unique/rare units)					words	words	pus)	source corpus and selected corpus is both 100 %
	(Abera et al., 2016)	Tigrigna	URL	Web	115,000 sentences	10,212 sentences		The syllable frequency difference between source data and final corpus less than 0.01

	(W. Zhang et al., 2010)	English	HRL	Web	4621 sentences	1000 sentences	CMUDict (39 phonemes)	93.52% diphones
Two stage algorithms	(Wang, 1998)	Mandarin	HRL	Newspapers	22,660,835 sentences	124,845 sentences		Cosine similarity increases as the sentences increases. When tested on 100 sentences cosine similarity is

								0.9979
	(Yuwan & Lestari, 2016)	Arabic	URL	Literature	6236 verses	180 verses	Phonemes from source data	0.9998 cosine similarity with source data
Probabilistic metric	(Malviya et al., 2016)	Hindi	URL	EMILLI Corpus	1,784,784 sentences 42,682,598 triphones	1856 sentences	Phonemes of 70 % freq words	0.87 cosine similarity with source data
	(Mendonca et al., 2014)	Brazillian Portugues	URL	Web	1,229,422 sentences	250 sentences		854 distinct triphones (40.9 %)

Word frequency	(Tan & Sh-Hussain, 2009)	Malay	URL	Web	10,027,126 words	1451 words & 381 sentences	Malay dictionary phonemes	Phoneme coverage higher than malay diphone inventory
	(Raza et al., 2009)	Urdu	URL	Existng word corpus	50,000 unique words	725 sentences & 5681 words		cosine similarity of triphone decreases as the increase in words. 10,133 uniq

								ue triphones in final corpus
Lexicon based selection	(Villaseñor-Pineda et al., 2004)	Mexican Spanish	URL	Web	244,251,605 words and 15,081,123 lines	6082 phrases & 220,776 words		corrrelation coefficient 0.99 with Spanish dictionary phonemes
Submodular optimization approach	(Shinohara, 2014)	Japanese	HRL	Newspaper; Novels	248,530 sentences	18,939 sentences	Phonemes from initial corpus	when compared with random sentence selec

								tion algorithm, proposed algorithm has better uniform distrfibution and more rare triphones hence better performance
Lagrangian based algorithm for multi-	(Chevelu et al., 2007)	French	HRL	Le Monde Corpus	172, 168 sentences	260 sentences	Phonemes from EM	greedy algorithm 334 sente

represented SCP (Lam SCP) algorithm							ILLI corpus	nces which is 10% better solution than greedy algorithm
Characteristics and guidelines of the language	(Abushariah et al., 2012)	Arabic	URL	Web	415 sentences	663 words & 367 sentences	UFPAdic 3.0	100% selected until all phonemes are collected (Handmade sentences)

From Table 2.2, it is noticed that most of the sentence extraction methods are tested on the well-resourced languages like Spanish, English, Japanese, and French. Well resourced languages are languages that have sufficient amount of tools and data for analysis. Initial data required for further research on these data are easily available and doesn't require further refinement. Whereas under-resourced language lacks these resources and has to use words different types of words like compound words, borrowed words and must undergo different processes to refine them. This takes lot of time and cost. But some studies are done on these languages like in Tigrigna (Abera et al., 2016), the final corpus size developed is not as optimized as the well-resourced languages. One of the most optimized final corpuses for an under-resourced language is achieved by in Mendonca et al. (2014). However, the phonetic coverage for the corpus developed in this research is only at 40.9%, which is not an efficient result. However, in Vorapatratorn et al., (2012), the phonetic coverage for the under-resourced language is 99.13%, but the source data used for sentence extraction is a commercially available lexical corpus. In summary, the performance of the existing methods for developing the lexical corpus of under-resourced language depends on the availaibility of external resources such as commercially available lexical database. However, the performance of the existing sentence selection methods is poor for self-generated source database such as web sources, an issue that need to be resolved.

2.3.1 Methods and the Sizes of the Developed Corpus

Developing speech corpus from a large data is usually difficult, intractable and time consuming. The issues of developing lexical corpus from such a large text corpus has been addressed by many researchers using suitable methods to reduce the size of the lexical corpus while preserving the phonetic rich and balance. Table

2.3 shows the reduction in size of the source corpus and the reduced corpus by the existing works.

From Table 2.3, it was found that the majority of the research can reduce the size of the corpus by more than 95%, with some of them achieved an impressive 99% reduction on the source datasets. In (Yuwan & Lestari, 2016), the corpus was reduced to 180 verses from the original 6,236 verses using the two-stage algorithm. However, the two-stage algorithm performance was based solely on the specific domain (Quran ASR), thus the phonetic nature of only specific words was taken into consideration. In (Nguyen, 2015), the percentage of reduction is 99.90% using repetitive greedy algorithm, focusing only on tono-phonemes as the basic unit, developed by the phonetic experts, an expertise that not available for all under-resourced languages.

Table 2.1. Summary of Size and phonetic coverage using the existing methods

Method	**Initial Size**	**Reduced Size**	**Percentage of reduction (%)**
Two stage algorithm (Yuwan & Lestari, 2016)	6236 verses	180 verses	97.11
Probabilistic metrics (Mendonca et al., 2014)	1,229,422 sentences	250 sentences	99.98
Lagrangian based algorithm for multi-represented SCP (LamSCP) algorithm (Chevelu et al., 2007)	172,168 sentences	260 sentences	99.85
Greedy algorithm (Nguyen, 2015)	323,934 sentences	334 sentences	99.90

Guidelines (Abushariah et al., 2012)	415 sentences	367 sentences	11.57
Word Frequency (Tan & Sh-Hussain, 2009)	10,027,126 words	381 sentences	--
Word Frequency (Raza et al., 2009)	50,000 unique words	725 sentences	--
Greedy algorithm (Vorapatratorn et al., 2012)	3007 sentences	1000 sentences	33.49
Sentence Selection (Zhang et al., 2010)	4621 sentences	1000 sentences	78.36
Probabilistic metric (Malviya et al., 2016)	1,784,784 sentences	1856 sentences	99.90
Sentence selection (Arora et al., 2004)	More than millions of words	5147 sentences	--
Lexicon based selection (Villaseñor-Pineda et al., 2004)	244,251,605 words and 15,081,123 lines	6082 sentences	--
Sentence selection (Abera et al., 2016)	115,000 sentences	10,212 sentences	91.12
Submodular optimization approach (Shinohara, 2014)	248,530 sentences	18939 sentences	92.38
Two stage algorithm (Wang, 1998)	22,660,835 sentences	124,845 sentences	99.44
(Bansal et al., 2015)	300,000 words	1500 words 300 sentneces	Speech corpus
(Habib & Adeeba, 2014)	35m words	70,000 words	99.1999% (ngrams & phonemes)
(Matoušek et al., 2008)	-	17,378 utterances	Speech corpus

(Anumanchipalli et al., 2005)	Marati: 155541; Tamil: 303537; Telugu: 444292 (sentences)	52 sentences randonmly selected for each speaker (total 200 speakers)	WER % (Word Error Rate) Marati: 23.2; Tamil: 20.2; Telugu: 28

(Tan & Sh-Hussain, 2009)), have extracted 381 carrier sentences from a very large lexical data of 10,027,126 words using the word frequency method. Although the percentage of reduction cannot be calculated, this method can be considered as very proficient as it can optimize data from a very large database to a minimum size. On the other hand, (Raza et al., 2009) make use of the same method and can optimize the 50,000 unique word into 725 sentences, which can be considered not very efficient. The possible cause for the difference in the optimization performance is due to the differences in the nature of resources used in both research. Some of the works that uses Greedy method reported percentage of reduction of less than 80% (Vorapatratorn et al., 2012; W. Zhang et al., 2010), for languages that can be considered as under-resourced languages. In these studies, though the percentage of reduction of initial corpus is poor, the phonological nature of the final corpus is quite satisfactory.

2.3.2 Methods and Resources for Developing Lexical corpus

With the advent of web, many researchers prefer the web as a good source for developing lexical corpus, to overcome the issue of data scarcity for most of the languages. This section summarizes some of the resources used for lexical corpus development, and identifies the impact of these resources on the nature of the corpus in term of richness and balanced. Table 2.4 provides the summary of the methods and resources for developing a lexical corpus.

Table 2.2 Summary of methods and resources for lexical corpus development

Source	**Resource**	**Language**	**Method**	**Result**
Web	MEA-SYLDIC (Nguyen, 2015)	Vietnamese	Greedy algorithm	334 sentences (100%) 630 sentences (95.1%)
	Source corpus units (Abera et al., 2016)	Tigrigna	Sentence selection	0.01 frequency difference from original corpus
	CMUDict (W. Zhang et al., 2010)	English	Sentence selection	93.52%
	Source corpus units (Tan & Sh-Hussain, 2009)	Malay	Word frequency	16826 unique phonemes
	Spanish dictionary phonemes (Villaseñor-Pineda et al., 2004)	Spanish	Word frequency	0.99 coefficient correlation
	UFPdic 3.0 (Mendonca et al., 2014)	Brazilian Portuguese	Probabilistic metric	854 unique triphones

Web and Pre-built corpus	LVCSR corpus (Vorapatratorn et al., 2012)	Thai	Greedy algorithm	99.13%
Pre-built corpus	Source corpus units (Arora et al., 2004)	Hindi	Sentence selection	100%
	Urdu phonemes (Raza et al., 2009)	Urdu	Word frequency	10,133 unique phonemes
	Source corpus units (Chevelu et al., 2007)	French	Lagrangian based algorithm for multi-represented SCP (LamSCP) algorithm	10% better solution than greedy algorithm
Traditional corpus	Arabic phonemes (Abushariah et al., 2012)	Arabic	Characteristics and guidelines	12.39 WER
	Source corpus units (Malviya et al., 2016)	Hindi	Probabilistic metrics	0.87 coefficient correlation
	Urdu phonemes & source corpus units (Habib & Adeeba, 2014)	Urdu	Greedy algorithm	99.1999%
	Source corpus units	Mandarin	Two stage algorithm	0.9979 coefficient

	(Wang, 1998)			correlation
	Source corpus units (Yuwan & Lestari, 2016)	Arabic	Two stage algorithm	0.9998 coefficient correlation
	Source corpus units (Shinohara, 2014)	Japanese	Submodular optimization approach	Better phoneme distribution

From Table 2.4, two main resources needed are the source data and the types of speech unit. The most common sources of data are the web, pre-build corpus, and other traditional sources like textbooks and newspapers. Most of the sentence selection methods select the sentences from the source data based on the desired units (Abera et al., 2016; Arora et al., 2004; Chevelu et al., 2007; Habib & Adeeba, 2014; Malviya et al., 2016; Tan & Sh-Hussain, 2009; Wang, 1998; Yuwan & Lestari, 2016). Table 2.4 also shows that the methods that uses web corpus as the source data normally relies on the speech units from pronunciation dictionary build by the experts (Mendonca et al., 2014; Nguyen, 2015; Villaseñor-Pineda et al., 2004; W. Zhang et al., 2010). This is because, the linguistic feature of the web corpus is inappropriate, and targeting the units only from web corpus may result in the loss of relevant speech unit that may adversely affect the acoustic models of the speech processing systems, resulting in unintelligent output (speech or text).

2.4 Evaluating the Performance of the Existing Methods

From the literature, researchers applied different approaches to evaluate the performance of the lexical corpus developed using different sentence selection methods. For the greedy algorithm, where sentences are selected iteratively until all the desired units are covered 100%, the performance is evaluated by the coefficient correlation of the desired units in the final corpus and the source units (Abera et al., 2016; Malviya et al., 2016). Alternatively, performance can be measured by calculating the coverage percentage of the phonemes, either by checking the coverage for the whole corpus, or by randomly selecting sentences from the final corpus. Researchers also measure the performance of the sentence selection methods by calculating the unit frequencies of the developed and target corpus, or by identifying number of unique units in the final corpus. Table 4.5 depicts the common evaluation approach for measuring the performance of the sentence selection methods.

Table 4.3: Common Evaluation Approach for Measuring the Performance of the Sentence Selection Methods

Evaluation Technique	**Method**	**Result**	**Remarks**
By Coverage	Greedy Algorithm (Habib & Adeeba, 2014)	99.1999%	Phonemes covered in the final corpus is either compared with initial text or the phonemes of the target language
	Greedy algorithm (Nguyen, 2015)	95.1% (630 sentences) 100% (930 sentences) 100% (334 sentences)	
	Sentence selection algorithm (W. Zhang et al., 2010)	93.52%	
	Probabilistic metrics (Mendonca et al., 2014)	40.9%	
By unit difference	Sentence selection (Abera et al., 2016)	Less than 0.01 syllabic difference	The differences between the speech

	Sentence selection (Arora et al., 2004)	Less than 0.01 syllabic difference	units between the final corpus and other available corpus of the target language is checked
By coefficient correlation	Two stage algorithm (Wang, 1998)	0.9979 (100 sentences)	The similarity between the final corpus and other target corpus is evaluated
	Two stage algorithm (Yuwan & Lestari, 2016)	0.9998	
	Probabilistic metric (Malviya et al., 2016)	0.87	
	Lexicon based method (Villaseñor-Pineda et al., 2004)	0.99	
By unique unit frequencies	Word Frequency (Tan & Sh-Hussain, 2009)	16826 unique phonemes	The more unique speech units available is considered as more efficient
	Word Frequency (Raza et al., 2009)	10,133 unique triphones	
	Probabilistic metrics (Mendonca et al., 2014)	854 unique triphones	

2.5 Zipfian Distribution

Zipfian distribution is a representation of Zipf's law. Zipf's law is an empirical law that describes statistical irregularities, proposed by an American linguistic George Kingsley Zipf. The law states that the frequency of occurrence of words or other items is inversely proportional to its statistical rank in frequency table as shown in Formula 1 below.

$$\boldsymbol{w_n} \sim \frac{\mathbf{1}}{\boldsymbol{n^a}} \quad \textbf{(1)}$$

where w_n is the frequency of occurrence of the nth ranked item.

In other words, there is a constant k such that $w_n.n^a = k$, where higher ranks are given to items with lowest frequency distribution (Weikum, 2002)..

Zipfian distribution has been used in many linguistic phenomenon including the creating lexical corpus, corpus representativeness, and analyzing word frequencies (Moreno-Sánchez et al, 2016; Weikum, 2002). Table 2.6 shows the existing research that used on the Zipfian distribution from the field of computational linguistics.

Table 2.4: Existing Research that uses Zipfian distribution

Article	**Purpose**	**Outcomes / Findings**
(Saloot, Idris, Aw, & Thorleuchter, 2016)	Rank word frequency from a large of tweets (twitter content)	Malay Chat-style corpus with 14,484,384 word instances
(Riyal, Rajput, Khanduri, & Rawat, 2016)	Examine the character frequencies for consonants, vowels for Gawhwali corpus	Frequency of characters using Zipfian distribution makes distinct element grows with stable exponent
(Mohaghegh & Sarrafzadeh, 2016)	Highlighted the features of word frequency distribution using Zipfian distribution	English-Maori parallel corpus to be used in NLP & machine translation tasks
(Piantadosi, 2014)		Statistics of word frequency for multi languages
(Ha, Hanna, Ming, & Smith, 2009)	Combine n-grams of letters,	Large corpus of multiple languages

	phonemes or binary bits for large corpora.	can hold all ranks of words when combines with n-grams using Zipfian distribution

From Table 2.6, one can notice that the Zipfian distribution has been used for various purposes, including building general corpus, analyzing the relation between text, and language models and phonetics units. Zipfian distribution is mostly used in natural language processing area and in this research an attempt is made to use it on speech technology. Zipfian distribution has been used in building general corpus but not in building corpus required for speech applications. This might be because of the lack of importance given to lexical corpus development in speech applications. Hence in this research Zipfan distribution will be used to build lexical corpus and evaluate its effectiveness. As such, the notion of using the Zipfian distribution as a sentence selection method for buiding a phonetically rich and balanced lexical corpus should be explored. This is because, representing phonetic units that are inversely proportional to its frequency will make the lexical corpus to be more balanced, and the iterative process of selecting the words will make the corpus rich.

2.6 Evaluation of Richness and Balanced

Some of the methods used to evaluate the richness and balanced of the corpus in terms are explained below.

2.6.1 Phonetic Coverage

The performance of the sentences selected method is evaluated by measuring the phonetic coverage of the developed lexical corpus. Phonetic coverage refers to the inclusion of maximum number of the phonemes of the target language (richness), and the frequency of statistical distribution of the phonemes (balanced).

- **Phonetically Rich**

A corpus is considered as phonetically rich if it contains the maximum number of phonemes of the target language. The phonetic richness is evaluated by comparing the number of speech units in the developed corpus with the phoneme inventory of the target language, which can be obtained from the source corpus or phoneme set of that language. Table 2.7 provides the existing methods for valuating the phonetic richness of the developed lexical corpus.

Table 2.5: Methods for evaluating phonetic richness

Authors	Method / Formula	Formula
(Habib & Adeeba, 2014; Mendonca et al., 2014; Nguyen, 2015; W. Zhang et al., 2010)	Tools used to count the distribution of the phonemes and compare with a phoneme inventory	Tools to count the number of phonemes
(Malviya et al., 2016; Villaseñor-Pineda et al., 2004; Wang, 1998; Yuwan & Lestari, 2016)	The correlation coeffciency between phonemes of final corpus and phoneme inventory is evaluated	$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\lVert\vec{a}\rVert\lVert\vec{b}\rVert}$ where "a" and "b" are the set of initial and final corpus sets respectively.
(Abera et al., 2016; Arora et al., 2004; Mendonca et al., 2014; Raza et al., 2009; Tan & Sh-Hussain, 2009)	Unique unit frequencies	$s_C = \frac{u_C}{u_t}$ where s_C is the frequency of unique units; u_C and u_t are the no. of unique units in initial and final corpus respectively

From Table 2.7, it can be noticed that, for all these methods, the distribution of the speech units is evaluated and compared with a phoneme set, either for similarity, or for difference between the distributions of the phonemes. Some researchers have also evaluated the richness by comparing existence of unique speech units in the final corpus, where the presence of more unique units enables the corpus to richer, thus improving the performance of speech-based applications.

- **Phonetically Balanced**

The phonetic balance of a corpus is evaluated by assessing the frequency of each speech units distributed in the developed corpus. A balanced corpus is the one with uniformly distributed speech units, which is suitable for building effective acoustic models for both the TTS and ASR system. Research on phonetic balance make use of various heuristic methods (Abera et al., 2016; Mandal, Das, Mitra, & Basu, 2011; W. Zhang, Liu, Deng, & Pang, 2010), extended entropy using phonetic and prosodic contexts (Nose, Arao, Kobayashi, Sugiura, & Shiga, 2017), and submodular optimization approach (Shinohara, 2014). The phonetic balance is calculated as the relative number of occurrences of speech units with the highest values (Abera et al., 2016).

2.6.2 Size

The size of the lexical corpus is evaluated by selecting specific set of sentences and then calculating the phoneme coverage in each set (W. Zhang et al., 2010), or by comparing the size of similar phonetic coverage with the initial text (Vorapatratorn, Suchato, & Punyabukkana, 2012).

2.7 Summary of the Literature Review

One of the objectives of this research is to analyze the existing methods for lexical corpus development. The review of the literatures reveals that there are many various methods proposed by the researchers for developing lexical corpus, where most of these methods are applied to well-resourced languages only. Many aspects such as the size, phonetic coverage, and resources used in the literature are reviewed. In terms of corpus size, the two stage algorithm used in (Yuwan & Lestari, 2016) as mentioned in Table 2.2, produces the most optimum coverage in relation to the size of the corpus (Table 2.2). In term of phonetic coverage (Habib & Adeeba, 2014), reported a coverage of 99.1999% using the greedy algorithm. (Yuwan & Lestari, 2016), and (Villaseñor-Pineda et al., 2004) also reported satisfactory results of 0.9998 and 0.99 coefficient correlation respectively. On the contrary, some researcher reported coverage as low as 40.9% (Mendonca et al., 2014). The varying results can be due to the difference in the nature of the resources used for lexical development. This research will propose a method that maximizes the phoneme coverage with limited text despite the nature of the resources. In this research, sentences will be selected by giving ranks using the Zipfian distribution, in which high ranks will be given to words the with lowest frequency. By doing so, all the phonemes of the target language will be covered including the rare phonemes, and reduces the need of depending on linguistic experts to correct or refine the sentences in the lexical corpus.

CHAPTER 3: RESEARCH METHODOLOGY

This chapter focuses on the research methodology adopted in this research to achieve the research objectives. The main objective being proposing a method to overcome the mentioned research problems. Figure 3.1 depicts the methodology adopted, which comprises several key steps such as the literature review, data collection, development of lexical corpus, and evaluation of the performance of the corpus developed.

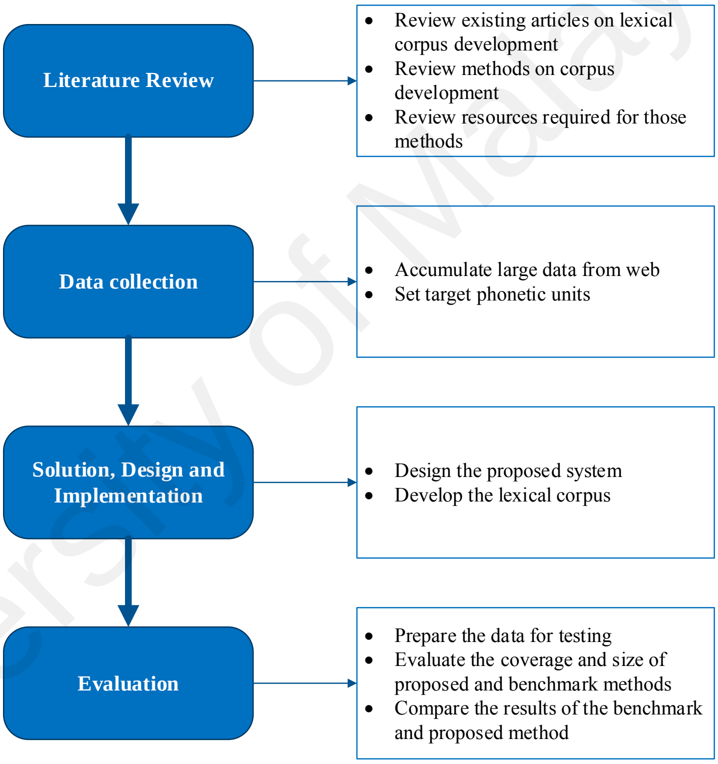

Literature Review
Review existing articles on lexical corpus development
Review methods on corpus development
Review resources required for those methods
Data collection
Accumulate large data from web
Set target phonetic units
Solution, Design and Implementation
Design the proposed system
Develop the lexical corpus
Evaluation
Prepare the data for testing
Evaluate the coverage and size of proposed and benchmark methods
Compare the results of the benchmark and proposed method

Figure 3.1 Research Methodology

- **Literature Review**

The aim of the literature review of related research is to obtain useful information such as the existing sentence selection methods including the merits and demerits, availability of key resources, and evaluation techniques to measure the performance of the existing methods are gathered.

- **Data collection**

  It is a known fact that for most of the under-resourced languages, existing available resources are rare. As such, it is very common that the research and development of speech-based applications will accumulate their own resources, especially the lexical and speech corpus. To develop a lexical corpus, the first step is the accumulation of a very large text data from suitable sources. Web based source is the most ideal source for building a large text data for most of the under-resourced languages. Hence, in this research, web sources such as Facebook, Twitter, and news websites were used for collecting the large text data, from which phonetically rich and balanced lexical corpus is going to be developed.

- **Solution, Design and Implementation**

  By referring to the findings from the literature, this research proposes a sentence selection method using the Zipfian distribution. The proposed method is designed and implemented using the Python script. Details of the development of the proposed method are explained in Chapter 4.

- **Evaluation**

  The performance of the proposed method is evaluated for the phonetic richness and balance, as well as the size of the lexical corpus developed by the proposed method. The performance of the proposed method is then compared with the benchmark method commonly used for developing lexical corpus. The results of the evaluation are explained in Chapter 5.

3.1 Problem Identifications and Solution

3.1.1 Problem Identifications

From the review of the existing literature, it was found that the existing methods for sentence selection were mostly employed on well-established languages (Chevelu, Barbot, Boëffard, & Delhay, 2007; Shinohara, 2014; Villaseñor-Pineda, Montes-y-Gómez, Vaufreydaz, & Serignat, 2004; Wang, 1998; W. Zhang et al., 2010), as compared with the under-resourced languages.

One of the key issues discussed in the literature in relation to the development of lexical corpus is that the coverage of phonemes for a particular language from web-based sources is low (Section 2.3.2), although web is the most ideal source for large text data of an under-resourced languages. As mentioned in Section 2.3.1, another issue identified from the literature is the optimization property of the sentence selection methods, in order to provide the maximum coverage of the developed lexical corpus.

3.1.2 Proposed Solutions

One of the key issues addressed in this research is that the phoneme coverage for lexical corpus extracted from web is not adequate. Hence selecting sentences from the large web data source that contains all the available phoneme of the target language may solve this issue. For optimizing the lexical corpus, priority is given to sentences that have the rarest occurring phoneme, which is reflected in Zipfian distribution.

The proposed method for developing lexical corpus using the Zipfian distribution can improvise the phonetic coverage of the large data collected from web, as the

Zipfian distribution represents each phoneme as inversely proportional to its rank in the frequency table. On top of that, the proposed method is also aimed at improved optimization of the large database.

The development of the phonetically rich and balanced using the proposed method consists of the following stages:

1. Collect data from the web
2. Refine data (remove duplicates, remove alien words, and so on)
3. Select sentences using the Zipfian distribution (Zipf's law)

The proposed method is experimented on Dhivehi, the official language of Maldives, which is also an under-resourced language. Since there is no prior research on speech technology domain for Dhivehi, many challenges are faced such as the lack of language tools, language experts, and speech unit inventory.

Zipfian distribution ranks the words by allocating suitable weights to the desired phonetic units. The main difference between the proposed method and the existing sentence selection methods is that the latter analyze the large corpus first before the sentences are ranked for selection, where in some methods, human experts perform the ranking manually. However, for Dhivehi, the lack of human experts can be resolved with the use of Zipfian distribution for performing the sentence ranking.

In this research, Zipfian distribution is used twice, one to rank the words by giving weights to the phonetic units, and second, to rank the sentences by giving weights to the phonetic units. In this way, the developed lexical corpus will contain a high variety of speech units (richness), and at the same time maintained equal frequency of the phonemes (balanced). On top of that, two pass of the Zipfian distribution will minimize the size of the corpus, and att the same time retain the phonetic nature. In the first pass, words with all the phonemes are covered. Even at this stage, the words will be phonetically rich and balanced as they will contain all the phonemes frequently distributed. And applying zipfian distribution in the second pass to select senteces with words selected in the first pass will make the final corpus more optimized mainlining its richness and balancedness.

3.1.3 Data Accumulation

Accumulating a large lexical data is much easier today due to the availability of the web, as it offers abundant, free, and easily available resources. As such, in this research, web sources such as news websites, Facebook, and Twitter are used as the source for sentence collection. Though the contents of these sources are harder to refine, they are of more conversational type and can accumulate words that are used everyday. Also these are the most easily available sources for many under resourced languages. A python script was prepared to accumulate the data from web to extract the sentences from the web, which are then cleaned, and merged to develop a phonetically rich and balanced lexical corpus.

3.1.4 Dhivehi Language as an under-resourced language

A language is considered as under-resourced language if its lacks some (if not all) key resources such as presence on the web, linguistic expertise, electronic resources for human language technologies, pronunciation dictionaries, vocabulary lists, and so on. Under-resourced language is also described by researchers as *low-density languages, resource-poor languages, low-data languages, poorly resourced languages, and less-resourced languages* (Besacier, Barnard, Karpov, & Schultz, 2013). It is important to understand that the term under-resourced languages does not necessarily refer to minority languages, but can also include well-known languages. In fact, some minority languages such as Catalan language is well-resourced, and is available for Google translate and Google search. On the other hand, there are official languages such as Dhivehi classified as under-resourced language.

Dhivehi, is the official language of Maldives, but is far behind technological development (Gnanadesikan, 2017), as it not represented in many of the well-known global applications such as Cortana, Siri, Microsoft translate, Google translate, and Google voice search. One of the reasons for the lack of progress for Dhivehi is due to unavailability of the key resources such as corpora, and pronunciation dictionary. Chapter 4 provides the detailed description on Dhivehi language.

3.3.2 Web source

Data collection for developing lexical corpus is a difficult task, as it requires a lot of human effort and time. However, with the availability of the Web, the accumulation of sentences for any language is now much easier. Web data offer large and freely access to sentences for many languages including Dhivehi. On top of that, the electronic format of the web makes it much easier for extracting, processing and, selecting the sentences with little human involvement (Kilgarriff, Reddy, Pomikálek, & Pvs, 2008). The internet usage in Maldives has increased drastically and availability of Dhivehi websites has been growing, allowing this research to access and extract data from various Dhivehi websites as well as the contents from Twitter and Facebook.

1.3.3 Stages in Web Data Collection

Figure 3.3 depicts the process of collecting initial large data from the web, which was adapted from (Mendels, Cooper, & Hirschberg, 2016) .

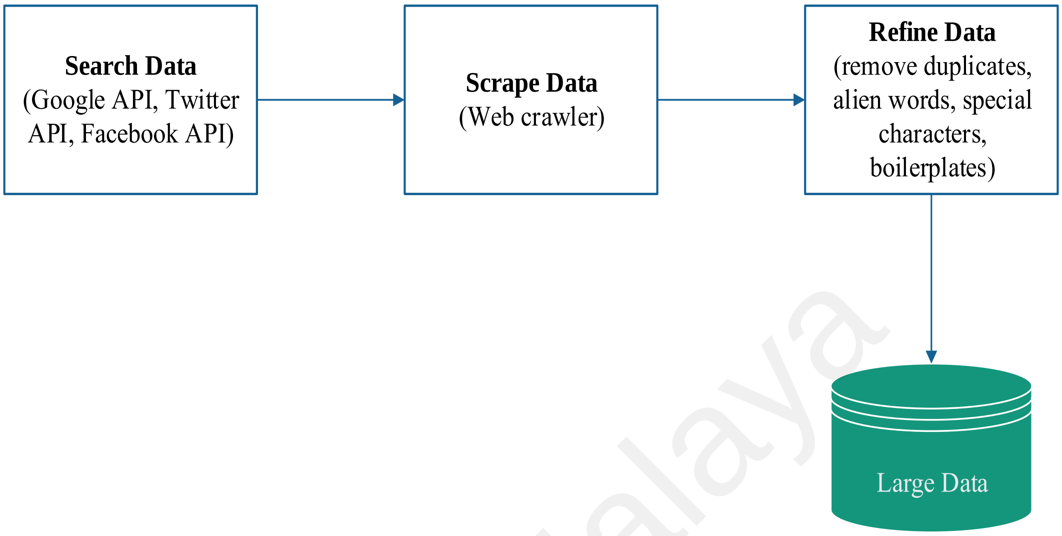

Search Data
(Google API, Twitter API, Facebook API)
Scrape Data
(Web crawler)
Refine Data
(remove duplicates, alien words, special characters, boilerplates)
Large Data

Figure 3.2 Web Data collection method adapted from (Mendels et al., 2016)

- **Search Data**

The first process in web-source data collection is to identify the relevant content using appropriate keywords search related to the target language to obtain the relevant content. The keywords used must be appropriate to avoid the occurrence of words from other languages. In this research, APIs (Application Programming Interface) from mentioned sources are used to get data. These API are set of protocols and tools that allows a programmatic access to data and platform on the web. These APIs are particularly used for authorization to access the data from these sources. These APIs are officially provided by the above mentioned sources (Figure 3.2), hence let us legally access these data.

- **Scrape Data**

This is an important process during the data collection, where this process enables the fetching and extracting data from the source, and the data extracted from various resources are saved in a single file for further refinement

- **Refine Data**

  Due to the enormous size, availability and modernity, of the web, the data extracted contain a large amount of unstructured data. As such, the extracted data must be refined to retrieve the linguistic characteristics of the language. Some of the refinement process applied in this research includes removing duplicates, alien words, special characters, and boilerplates.

3.2 Development of Lexical Corpus for Dhivehi

After collecting all the necessary data for developing lexical corpus, appropriate sentences must be selected to meet the objective of the lexical corpus. As stated earlier, this research will employ the Zipfian distribution to select sentences from the large web data that was refined for developing phonetically rich and balanced lexical corpus for Dhivehi. The development of lexical corpus consists of four stages as depicted in Figure 3.3.

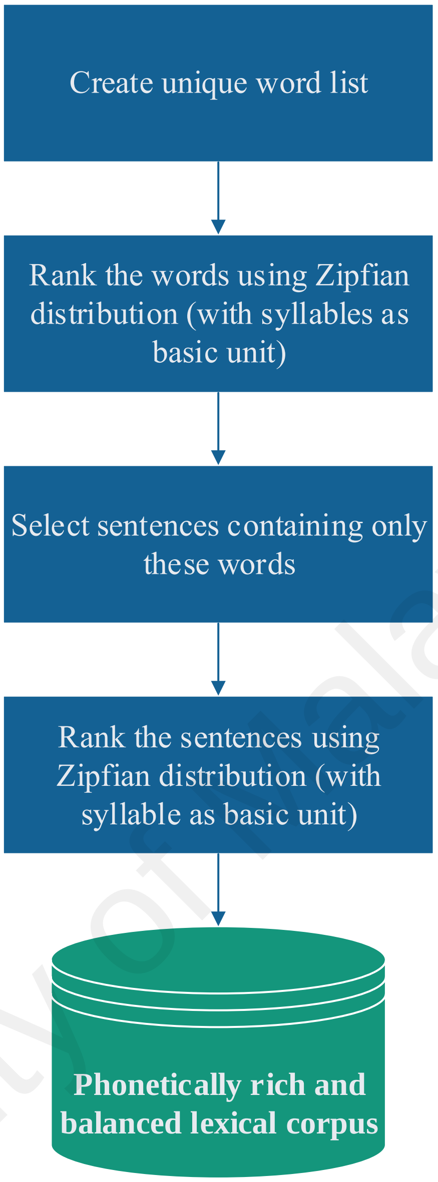

Create unique word list
Rank the words using Zipfian distribution (with syllables as basic unit)
Select sentences containing only these words
Rank the sentences using Zipfian distribution (with syllable as basic unit)
Phonetically rich and balanced lexical corpus

Figure 3.3: Proposed method to develop lexical corpus using Zipfian distribution

3.2.1 Creating Unique Word List

In this stage, a list of unique words from the data accumulated is created with the sole purpose of optimizing the data accumulated.

3.2.2 Ranking words

When the unique word set is created, the Zipfian distribution is applied to rank and select the words that cover most of the phonemes of the target language.

Phonemes, di-phones, tri-phones are some of the units of speech with sylaables being the ost basic unit. In this research, syllables will be used as the basic units as they are the smallest unit of utterances and even inexperienced speaker can use the words with syllables without making mistakes. Syllables will be used for performing the Zipfian distribution, where the words containing the least syllables are given a high rank in the frequency table. For example, if the syllable ‘a’ occurs frequently, then it will be given the rank in the bottommost position of the frequency table. And the syllable ‘z’ which is assumed to be occur the least will be given the highest rank in the frequency table. In this way, the sentences selected contain all the syllables representing the language and is balanced, as the frequency distribution of these syllables will be uniform.

3.2.3 Selecting Sentences

To ensure that the developed corpus is optimized, only sentences containing the words created from the previous stage is selected to ensure that the database is phonetically rich and balanced words with more optimized characteristic.

3.2.4 Sentences Ranking

This is the final process in developing the lexical corpus, where the Zipfian distribution is applied for the second time. During this process, sentences are assigned weights based on the frequency of the occurrence of syllables in the database, where high weights are given to sentences with the rarest syllable and placed at the top of the frequency table.

In order make the final corpus more optimized and balanced Zipfian distribution is applied twice. This way, the initial corpus will be filtered twice giving more efficient output. Since the words and sentences are selected iteratively until all the required syllables are covered, the developed corpus should cover all the syllables of Dhivehi, making the database richer.

3.5 Evaluation techniques

In this research, there are two objectives for evaluation, which are to measure the phonetic coverage of the developed lexical corpus, and to evaluate the size of the corpus. Researchers have proposed several methods for evaluating the performance of the lexical corpus, out of which the method proposed in (W. Zhang et al., 2010), is used in this research as it is the most common and efficient method for evaluating phonetic coverage. In order to measure the coefficient correlation between the source data and the developed corpus a method that is easy, recent and accurate is chosen (Malviya et al., 2016).

3.4.1 Evaluating the Phonetic Coverage

Phonetic coverage of the developed lexical corpus is evaluated using the method proposed in Zhang et al., (2010) to measure the phonetic coverage of the developed corpus, which refers to both the richness and balanced nature of the corpus using the following formula

$$Phoneme\ Coverage = \frac{\sum_{1=0}^{n} P}{s_p}$$

Where S_p : is the sum of the phonemes & P is the phonemes occurring n times

3.4.2 Evaluating the Size of the Corpus

For evaluating the size of the corpus, the cosine similarity of the final corpus with the initial corpus as proposed in (Malviya et al., 2016) is used. Should the coverage of the phonemes is similar to the initial source corpus, then the developed database reflects the original large corpus. The formula of the cosine similarity is follows:

Equation 2: Cosine Similarity

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

where a = initial large data, b = final corpus, $\vec{a}$ = vector of initial large data, $\vec{b}$ = vector of final corpus

3.4.3 Bechmark comparison

One of the objectives of this research is to compare the performance of the proposed method with a benchmark method. Among the most widely used algorithm for sentence selection is the classic greedy algorithm (Anumanchipalli et al., 2005; Bansal et al., 2015; Habib & Adeeba, 2014; Matoušek et al., 2008; Nguyen, 2015; Vorapatratorn et al., 2012).

3.5 Summary

The main objective of this research is to develop a lexical corpus for Dhivehi, an under-resourced language using the proposed method to ensure good phonetic coverage using small lexical database. In order to develop the lexical corpus, several issues need to be considered such as data requirements, and concepts for designing the method. The methodology adopted in this research helps in achieving the objective of this research.

CHAPTER 4: LEXICAL CORPUS DEVELOPMENT

This chapter discusses in detail the process involved in developing a lexical corpus for Dhivehi, an under-resourced language using the proposed method described in chapter 3. It also explains about the nature of Dhivehi as well as the step-by-step processes of the lexical corpus development.

4.1 Dhivehi language

Dhivehi[1] is the official language of the Republic of Maldives that belongs to the family of southernmost Indo-Aryan language as well as southernmost Indo-European language. It is based on Sanskrit foundations and is closely related to the Sinhalese language spoken in Sri Lanka, and the same time has borrowed words from Urdu, Persian, and Arabic such as ސިއްހަތު: sihhatu (health), ކައްފާރަ: kaffara (penance), ސިއްރު: sirru (secret). Dhivehi is inscribed in a unique script called Thaana, which is written from right to left (Fritz, 2002). Unlike other official languages of this world, Dhivehi has received very less technological attention.

Malé, Huvadhu, Mulaku, Addu, Haddhunmathee, and Maliku are some of the major dialects of Maldivian, with Malé being the standard that is widely used in offices, schools, universities, media, newspapers, formal speeches, courtrooms, and all kinds of formal communication. The focus of this research is Malé dialect. So far, there is no speech corpus developed for Dhivehi that can be used for developing speech-based application.

[1] Initially spelled as *Divehi*. It is officially spelled as Dhivehi, after the semi-official transliteration called Malé Latin was developed in 1976.

4.2 Phonetic nature of Dhivehi

Thaana (Dhivehi writing system) is mostly alpha-syllabary in nature, which means that the consonant-vowel sequences are written as a single unit. It somewhat resembles other South Asian scripts known as abugida or abjad of Arabic script. It has a very simple syllabic structure, which consist of some consonant-vowel combinations such as CV, CVV, CVC, and CVVC sequences. (Gnanadesikan, 2017) as described in table 4.1.

Just like the Arabic language, vowels are always written in such a way that indicates the diacritical marks. Most of the Thaana word always carry a diacritic except for ނ (noon) which specifying prenasalizing, has a null consonant އ (alifu), and doesn't have a sound value but act as a sound carrier sign for vowels. On top of that, the letter ށ (shaviyani) act as a carrier vowel frequently used with the ◌ް (sukun), and no word in Dhivehi starts with the letter ށ (shaviyani).

Table 4.1 Classification of Dhivehi letters

Unicode	Thaana	Transcription	Orthography	International Phonetic Alphabet (IPA)[2]
0780	ހ	h	haa	[h]
0781	ށ	-	shaviyani	[ʂ]
0782	ނ	n	noonu	[n̪]
0783	ރ	r	raa	[ɾ]
0784	ބ	b	baa	[b]
0785	ޅ	lh	lhaviyani	[ɭ]
0786	ކ	k	kaafu	[k]

[2] International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin alphabet

0787	އ	-	alifu	-
0788	ވ	v	vaavu	[ʋ]
0789	މ	m	meemu	[m]
078A	ފ	f	faafu	[f]
078B	ދ	dh	dhaalu	[d̪]
078C	ތ	th	thaa	[t̪]
078D	ލ	l	laamu	[l]
078E	ގ	g	gaafu	[g]
078F	ޏ	gn	gnaviyani	[ɲ]
0790	ސ	s	seenu	[s̪]
0791	ޑ	d	daviyani	[ɖ]
0792	ޒ	z	zaviyani	[z̪]
0793	ޓ	t	taviyani	[ʈ]
0794	ޔ	y	yaa	[j]
0795	ޕ	p	paviyani	[p]
0796	ޖ	jh	javiyani	[dʒ]
0797	ޗ	ch	chaviyani	[tʃ]
07A6	އަ	a	abafili	[ə]
07A7	އާ	aa	aabaafili	[əː]
07A8	އި	i	ibifili	[i]
07A9	އީ	ee	eebeefili	[iː]
07AA	އު	u	ubufili	[u]
07AB	އޫ	oo	ooboofili	[uː]
07AC	އެ	e	ebefili	[e]
07AD	އޭ	ey	eybeyfili	[eː]
07AE	އޮ	o	obofili	[ɔ]
07AF	އޯ	oa	oaboafili	[ɔː]
07B0	އް	-	sukun	-

As mentioned earlier, Dhivehi has borrowed words from Arabic language particularly the Arabic letters. The letters of the loanwords are phonetically similar to that of Arabic consonants by means of diacritics. However, these loanwords are mostly used in very formal scenarios such as courtrooms and Friday sermons. Table 4.2 shows the extensions of Arabic letter for Dhivehi.

Table 4.2 Borrowed syllables from Arabic

Unicode	Thaana	Transcription	Orthography
0798	ޘ	ث	ttaa
0799	ޙ	ح	hhaa
079A	ޚ	خ	khaa
079B	ޛ	ذ	thaalu
079C	ޜ	ز	zaa
079D	ޝ	ش	sheenu
079E	ޞ	ص	saadhu
079F	ޟ	ض	daadhu
07A0	ޠ	ط	to
07A1	ޡ	ظ	zo
07A2	ޢ	ع	ainu
07A3	ޣ	غ	ghainu
07A3	ޤ	ق	qaafu
07A5	ޥ	و	waavu

In TTS or ASR system development, the commonly applied speech units are phonemes (Abushariah et al., 2012), diphones (Zhang et al., 2010), triphones (Mendonca et al., 2014), and syllables (Abera et al., 2016) . Based on the analysis of the phonotactics of Dhivehi, syllable is considered as the best phonetic units for selecting the sentences. The use of syllable is justified by the fact that it is the smallest unit of utterance, which an inexperienced native speaker can pronounce in isolation. However, the selection of syllables is strongly favored by the unique syllabic characteristics of Dhivehi. Some of the examples for using syllables are ހާލު (ha-aa-l-u), އަހަރެން (a-h-a-r-e-n) and ގޮފި (g-o-f-i)

4.3 Development Tools and Environments

In this research, the following tools and environments were used:

- Integrated Development Environment (IDE): Visual Studio 2017
- Programming language: Python
- Operating System: Windows 10 Pro
- Processor: Intel® Core™ i7-2637M CPU @ 1.70GHz

4.4 Initial text database

As stated earlier, there are no existing available text sources for Dhivehi, which means that the lexical corpus has to developed from scratch. As such the obvious choice is to scrape the data from Internet using the method proposed in (Mendels, Cooper, & Hirschberg, 2016) as depicted in Figure 4.1. The advantages of using this method is that this method can refine uncleaned and unstructured data from web sources. Also this is the most recent method used to scrape data from web that can be cleaned and structured systematically.

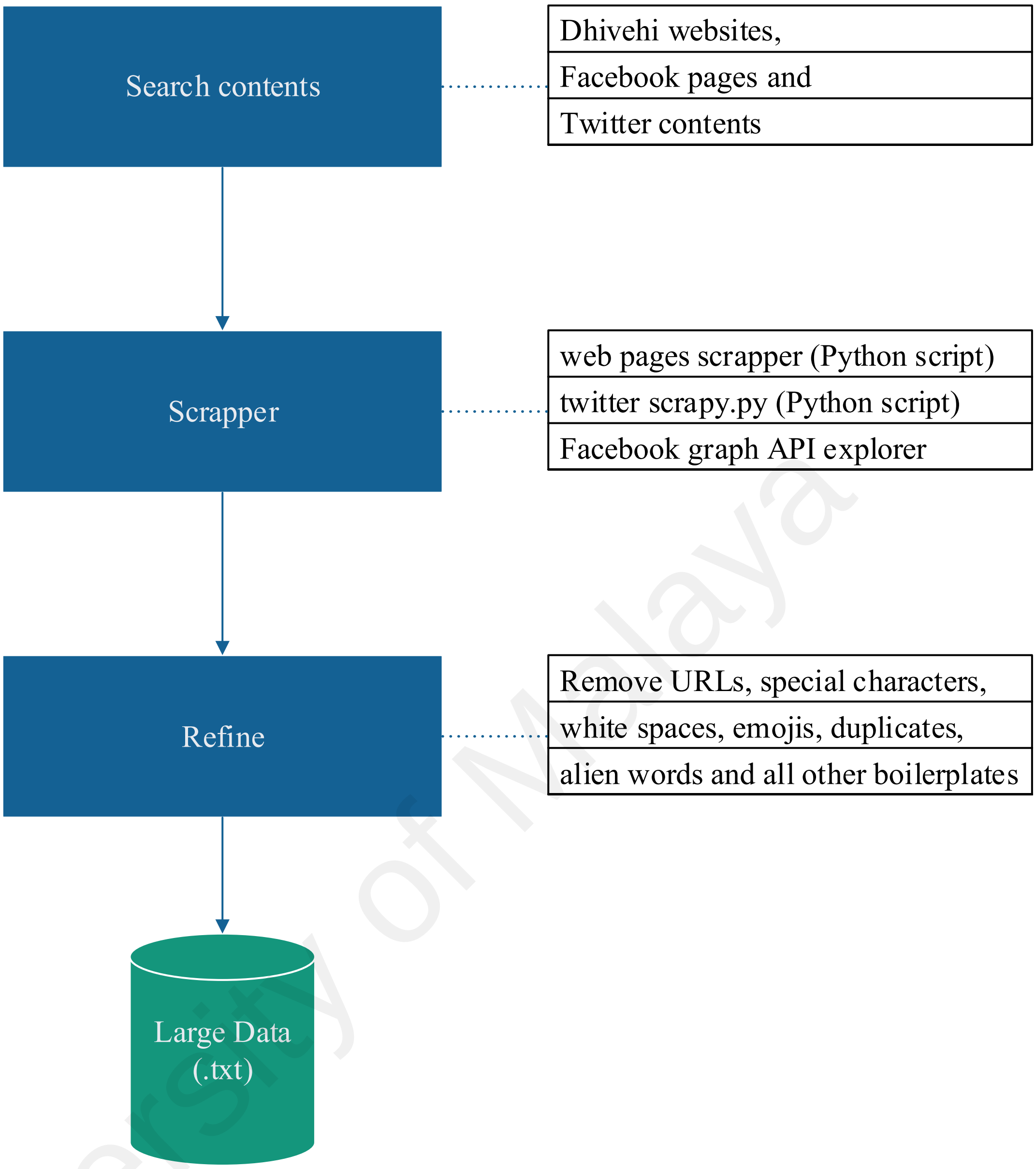

Search contents
Dhivehi websites,
Facebook pages and
Twitter contents
Scrapper
web pages scrapper (Python script)
twitter scrapy.py (Python script)
Facebook graph API explorer
Refine
Remove URLs, special characters,
white spaces, emojis, duplicates,
alien words and all other boilerplates
Large Data
(.txt)

Figure 4.1: Data Scrapping from the Internet (Mendels et al., 2016)

4.4.1 Search and Scrape Content of Web

Web is a huge mine of language data of unprecedented richness and ease of access, which is readily available in machine-readable form. The phonetic inventory consists of the combinations of 24 consonants and 10 vowels. For in-depth analysis, consideration is given to the first position of the syllables, at which it appears in a word.

The web sources from which the data is scrapped are explained below.

- **Websites**: Dhivehi is a language with limited web content as the availability of well-known of Dhivehi websites are rare. In order to maximize the content, a Dhivehi script keyword based query search was applied using the Google's advanced search engine API, allowing the search result to be only in Dhivehi language. A large amount of data from various number of websites are acquired. In this research, latin scripts are used as the corpus for experimentation. hence the contents that are in Dhivehi script, are converted to lating using a C# based transliteration tool.
- **Twitter**: One of the fast-growing microblogs among Maldivians is Twitter, where a large number of audiences from individuals to government organization use Twitter to convey short message. One of the reasons to choose as Twitter as a source for Dhivehi is because the contents are conversational type. With the aid of Twitter API, large number of tweets are collected using a keyword search script created using Python. The keywords include major trending hashtags in the specified geolocation so as to maximize the result.
- **Facebook**: The Graph API explorer of Facebook was used to collect contents from various trending pages with Maldives geolocation, and the results are stored in JSON format.

4.4.2 Refining the Raw Data

After collecting the data from all these sources, those data are merged together into a single large source database. However, the data in this database is raw as there are irrelevant contents or noises that could impair the quality of the corpus.

Hence the raw database is refined by removing urls, special characters, white spaces, emojis, duplicates, alien words, and all other boilerplates using a Python program, as well as the Google’s open refine tool for the conversion of JSON files, and managing data. The refined large database is converted into the .txt format.

4.4.3 Large Source Database

From the scrapping and refining, over 109,208 sentences were acquired, where the sentences ranges from a minimum of two words and no maximum limit. The number of sentences are counted using data analysis tools of MS Excel and Google’s open refine tool. These 109,208 sentences are made up of 1,023,098 words, which means that, on average, each sentence is made up of nine words. Table 4.3 provides the statistics of the large source database, while figure 4.2 depicts the distribution of Dhivehi syllables in terms of occurrence in the large database. It is clearly visible that certain vowels and consonants are used in abundance. The syllable ޏ (gn) has the least frequently used syllables, while ހ (h) is the mostly used consonant. Also, އޯ (oa) and އަ (a) are the least and highly used vowels respectively. This large source database is then further processed to develop a phonetically rich and balanced lexical corpus for Dhivehi.

Table 4.3 Statistics of large data

Total Sentences	**Total words**	**Total unique words**	**Total syllables**
109,208	1,023,098	159,358	2,090,589

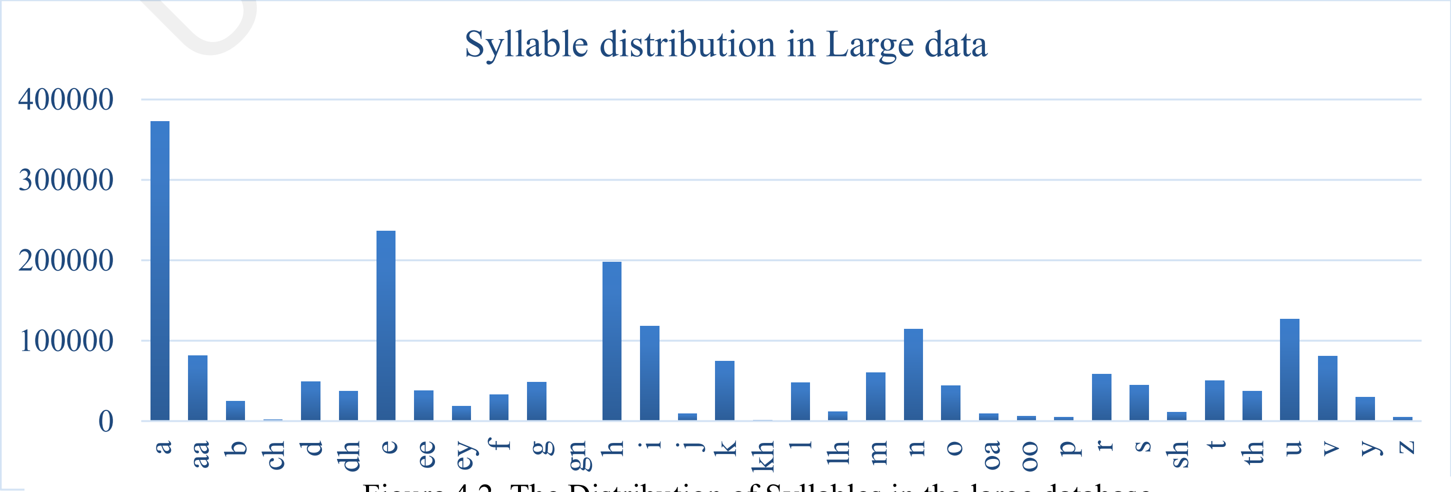
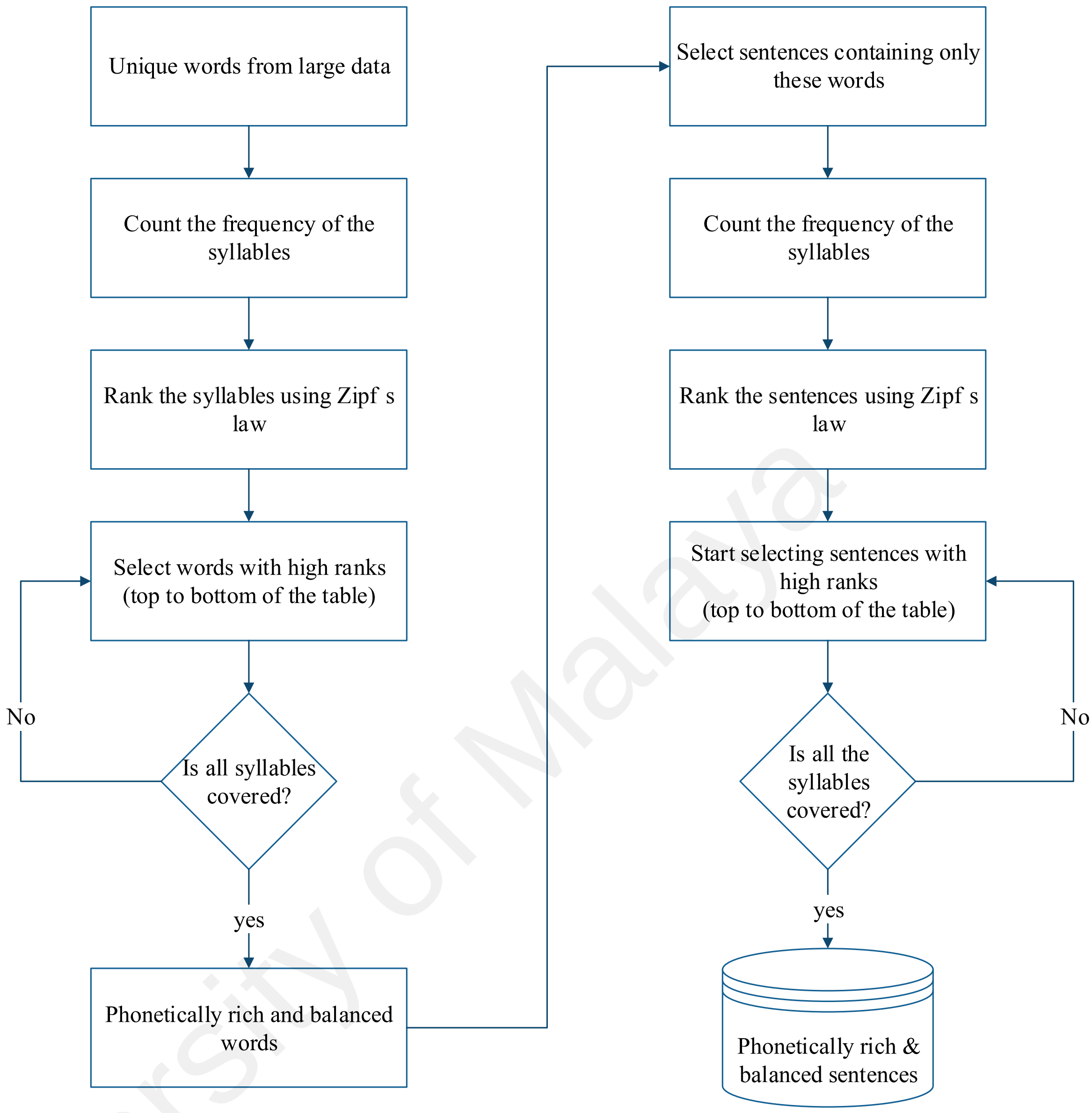


Figure 4.2. The Distribution of Syllables in the large database

4.5 Proposed Method for Lexical Corpus Development

The large database developed earlier serves as the foundation for building the phonetically rich and balanced lexical corpus for Dhivehi. In this section, the process of selecting a subset of sentences from the large data is explained. Figure 4.3 depicts the entire process of sentence selection using the Zipfian distribution for developing a phonetically rich and balanced lexical corpus (small corpus).



Unique words from large data
Count the frequency of the syllables
Rank the syllables using Zipf s law
Select words with high ranks (top to bottom of the table)
No
Is all syllables covered?
yes
Phonetically rich and balanced words
Select sentences containing only these words
Count the frequency of the syllables
Rank the sentences using Zipf s law
Start selecting sentences with high ranks (top to bottom of the table)
No
Is all the syllables covered?
yes
Phonetically rich & balanced sentences

Figure 4.3 Method for Automatic Sentence Selection from the Large Database

The proposed method was executed using python script as shown in Appendix G. The complete scenario is described more clearly below:

i. Given, the large data, L comprising a set of words W = {w_1, w_2, w_3,…….,w_n }. Each word will be made of sequence of syllables in a particular order. Let U be the set of syllabic units, represented as U = {u_1, u_2, u_3, …….., u_n}.

ii. A sentence is considered to be phonetically rich, if it possesses a high variety of syllables. In order to find the high variety of syllables, the frequency of occurrence of these syllables are calculated using the following formula:

$$\mathrm{freq(ui)} = \sum \mathrm{u_i} \in \mathrm{w_i} \qquad \textbf{(2)}$$

where the frequency of syllable $u_{i\,=}$ sum of every occurrence of u_i which belongs to the w_i

iii. The next step is to rank the words based on the frequency of the syllables using the Zipfian distribution. Thus, the least common syllable or the syllable with lowest frequency is given the highest rank in the frequency statistical table.

$$P(u_i) \sim \frac{1}{n^a} \qquad \textbf{(3)}$$

Where the higher the frequency of syllabic unit u_i, the lowest its position n^a in the rank table.

iv. Words are selected from the frequency table from top to bottom and this step is repeated until all the syllables are covered.

v. The steps from (ii) to (iv) are repeated executed again.

vi. A set of sentences containing varied syllables distributed equally is selected at the end of this method, which conforms the characteristics of phonetically rich and balanced lexical corpus.

4.6 The Developed Lexical Corpus using the Proposed Method

A lexical corpus for Dhivehi that consist of 360 sentences containing 868 unique words. By comparing the size of the large database, the percentage of the sentences selected is only 0.33%, a reduction of more than 99.67%. Since the proposed method perform iteratively process to select sentences until all the required syllables are covered, it is highly likely that the developed corpus covers 100% of the targeted syllables, meaning that the 360 sentences contain at least one of the desired syllables of Dhivehi. The developed corpus was also manually reviewed by the author to assure the structure of the sentences, and to make sure that it does not contain any offensive contents. Since Dhivehi is an under-resourced language and getting consultation from human expert is not an easier task, the developed corpus was manually reviewed by the author who is no expert but whose mother tongue is Dhivehi (Gnanadesikan, 2017). Table 4.4 shows the statistics of the phonetically rich and balanced lexical corpus. Figure 4.5 depicts the distribution of the syllabic units of the phonetically rich and balanced lexical corpus, while Appendix A shows the sentences selected by the proposed sentence selection method from the large sentence database.

Table 6 Statistics of the words and sentences constructed using the proposed method

Total words	Total sentences	Total syllables
868	360	27,616

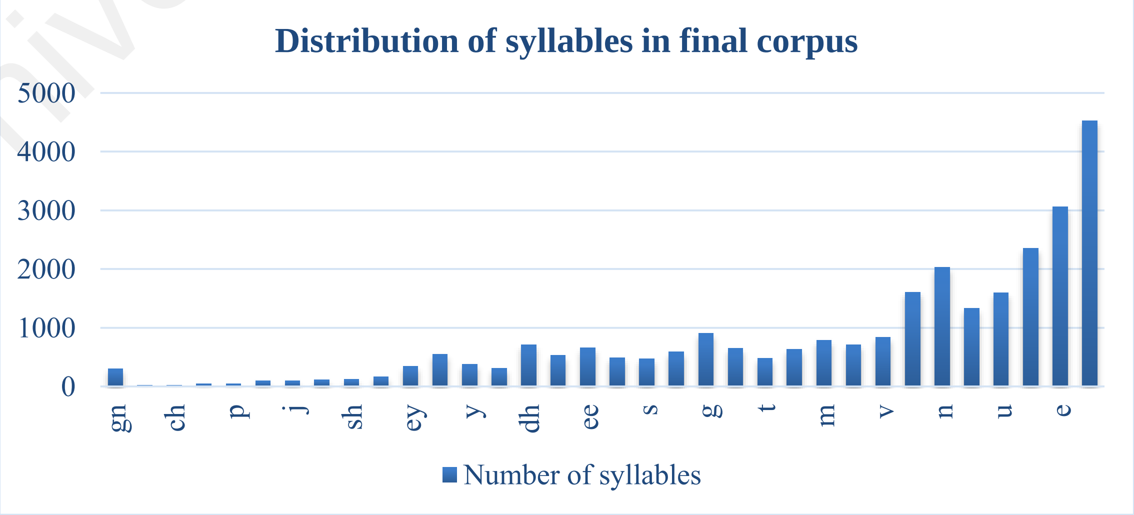


Figure 4.5. The Distribution of Syllables in the Developed Small Corpus

4.7 Summary

This chapter explains the process of developing the lexical corpus for Dhivehi. The initial data is collected from various web sources using the method proposed by (Mendels et al., 2016), modified accordingly for Dhivehi language. A subset of the sentences from the large web data are extracted using the proposed method based on the principles of Zipfian distribution. The phonetically rich and balanced lexical corpus for Dhivehi is made up of 360 sentences from the original database of 109,208 sentences, a reduction of about 99.67%.

CHAPTER 5: EVALUATION, RESULTS, AND DISCUSSION

This chapter focused on the evaluation of the phonetic richness and balanced of the developed lexical corpus of Dhivehi using the proposed method. A major challenge in the evaluation process is the non-availability of existing Dhivehi lexical corpus for comparison purpose. As such, comparison is made using the same large database developed in this research. This is achieved by measuring the phonetic coverage and evaluating the size of the developed corpus, which includes the evaluation on coefficient correlations of the basic units in the corpus.

The developed corpus is also compared with phonetic coverage and optimizing capacity of the greedy algorithm, which is the benchmark method.

5.1 Evaluation Method

5.1.1 Evaluation of the Phonetic Coverage

Zhang et al. (2010) proposed the use of phoneme coverage method for measuring the phonetic coverage of a different set of sentences using the following formula:

$$Phoneme\ Coverage = \frac{\sum_{1=0}^{n} P}{s_p}$$

Where S_p : is the sum of the phonemes & P is the phonemes occurring n times

5.1.2 Evaluation of the Corpus Size using Cosine Similarity

Several researchers have used the coefficient correlation for measuring the size of the corpus. (Malviya et al., 2016; Mendonca et al., 2014; Raza et al., 2009; Tan &

Sh-Hussain, 2009; Villaseñor-Pineda et al., 2004; Wang, 1998; Yuwan & Lestari, 2016). The cosine similarity between the statistical distribution of the small corpus and the source large corpus is compared to determine the closeness of the phoneme distribution between the small corpus and the original large database. The closer the similarity of phoneme distribution, the better is the phonetic richness and balanced of the small corpus.

In this research, the cosine similarity is compared between the initial large database and the final small corpus using the formula below:

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \qquad \textbf{(4)}$$

where a = initial large data, b = final corpus

5.1.3 Comparison with the Benchmark Method

One of the objectives of this research is to evaluate the performance of the proposed method and comparing the results with a benchmark method. This research used the greedy algorithm as the benchmark method, as this method is the most widely used method for sentence selection in many of the existing research. (Anumanchipalli et al., 2005; Bansal et al., 2015; Habib & Adeeba, 2014; Matoušek et al., 2008; Nguyen, 2015; Vorapatratorn et al., 2012). In this research, the greedy algorithm is used for extracting the sentence from the large web-sourced database by favoring the high frequent speech units.

5.2 Results

The results obtained from the proposed method is explained in this section. Data analytical tools of Microsoft Excel and Google's open refine tools are used to identify the syllable distribution in the sentences obtained as the final corpus.

5.2.1 Results of the Phonetic Coverage

Table 5.1 Results of the Phonetic Coverage of the Small Lexical Corpus

Total Sentences	Total Syllables	Phoneme Coverage
360	27,616	100%

Table 5.1 shows the coverage of phonemes in the developed small corpus, where the result shows 100% phoneme coverage due to the nature of the proposed method that perform the sentence selection iteratively until all the desired phonemes are covered.

5.2.2 Results of the Cosine similarity

Table 5.2 shows the results of cosine similarity between the initial large database and extracted final corpus. From the results in Table 5.2, it is observed that the distribution of syllables in the small corpus is very similar to that of the large database. This is due to the nature of the proposed method that select a sentence with low frequent phoneme as the Zipfian distribution gives high priority for these sentences. Evidently, the result also signifies the fact that the quality of the corpus does not necessarily depend on the size of the corpus, where only 360 sentences (representing 0.33% of the large corpus) is needed to represent the entire desired speech units.

Table 5.2: The Results of Cosine Similarity

Large Corpus	Final Small Corpus		$S = \cos \boldsymbol{\theta}$
Total syllables	Total syllables	% in corpus	
2,090,589	27,616	0.00013%	0.988167642

5.2.3 Results of the Benchmark Comparison

Table 5.3 shows the composition of the small corpus selected using the benchmark Greedy algorithm applied on the large database.

Table 7: The Composition of Sentences Selected using the Benchmark Greedy Algorithm

Sentences selected		**Syllables in selected sentences**	
Total	**% in relation to the large database**	**Total**	**% in relation to the large database**
5343	4.89%	554,536	26.52%

One of the least frequent syllabic in the large corpus is ڠ (gn), where it occurs only 401 times. The classic greedy method gives least priority to this syllable, and naturally, its position will be at the bottom of the frequency table. As such, the Greedy algorithm select sentences containing more frequent syllable, and doesn’t stop iterating until all the syllables at the bottom of the table is selected, increasing the number of sentences needed to achieve the phonetic richness and balance. Though both the proposed and benchmark methods perform iteration selection, the different types of prioritization of the syllable resulted in the different in the size of the small database.

Table 5.4 shows the different in the relative size of the small corpus developed using both the proposed and benchmark method. While both method cover all the desired syllables, the proposed method achieve it with only 360 sentences (representing 0.33% of the initial database) as compared to the benchmark method with 5,343 sentences (representing 4.89% of the initial database). The proposed method is 15 times (4.89%/0.33%) more effective than the benchmark in achieving phonetically rich and balanced lexical corpus.

Table 5.4: The Relative size of the small corpus prepared using the greedy method and proposed method

Total sentences in large corpus	Greedy Method		Proposed Method	
	Total sentences	***% in corpus***	***Total sentences***	***% in corpus***
109,208	5,343	4.89 %	360	0.33 %

5.3 Discussions

In this research, the syllables frequencies similarity between the initial large corpus and final small corpus prepared using the proposed method. The cosine similarity shows that the differences between both the corpora is very small. From Table 5.2, it was found that the cosine similarity between the two corpora is 0.988, and at the same time, the size of the small corpus is merely 0.33% of the initial large database. On the other hand, the benchmark Greedy method requires a much bigger corpus to achieve the good phoneme coverage of the initial database, where the developed database is 5,343 sentences or 4.89% of the initial database. This means that the quality of the corpus does not necessarily depend on the size of the corpus.

The better performance of the proposed method over the benchmark method is due to way the syllable is ranked. The proposed method ranks the least frequent syllable first, whereas the benchmark method ranks the more common syllable first. As such, the proposed method achieves the desired syllable in the corpus with much smaller sentence size.

5.4 Summary

In this Chapter, the performance of the proposed method is evaluated. The phonetic coverage is measured using the evaluation method proposed in Zhang et al., (2010) and the results show that the proposed method covers all the desired phonemes with relatively very small corpus size The size of the corpus is evaluated by measuring the cosine similarity between the final and the initial large text database. On top of that, the proposed method is 15 times more efficient as compared to the benchmark method, indicating the effectiveness of the Zipfian distribution in selecting the sentences that can maximize the phonetic coverage of the lexical corpus.

CHAPTER 6: CONCLUSION

This chapter summarizes the major findings of this research towards the development of a lexical corpus for Dhivehi, an under-resourced language. This chapter revisits the objectives of this research as well as the step taken to achieve those objectives. Furthermore, this chapter also discusses some of the research limitations as well as the future works that can extended.

6.1 Research objectives revisited

The aim of this research is fulfilled by achieving the following four main objectives:

6.1.1 Research objective 1

The first objective of this research is to identify and analyze the existing methods towards the development of phonetically rich and balanced lexical corpus. The existing literatures were reviewed to obtain the information such as resources, size, speech units, languages and limitations of the existing methods. The key issues identified from the literatures review includes phonetic the coverage of the corpus developed from web source for the under-resourced language, and the fact that most of the existing sentence selection method is not applied to many of the under-resourced languages.

6.1.2 Research objective 2

The second objective is to identify a suitable method to develop a lexical corpus for an under-resourced language that is relatively small. To achieve this objective, the existing sentence selection methods are analyzed and the criterion set to select sentences are examined. In most of the existing methods, sentences are scored first, and selected iteratively based on the conditions set. In this case, large data is required to cover as many phonemes. However, Zipfian distribution may be able to overcome this problem as it gives priority to the rarest phoneme occurrence in a sentence. As such, even the rarest phoneme will be covered in the developed corpus and selecting the sentences iteratively ensured the balanced distribution of phonemes in the developed lexical corpus, while maintaining a small size.

6.1.3 Research objective 3

The third objective of this research is to develop a lexical corpus for Dhivehi, an under-resourced language using the proposed method. Because of the syllabic nature of the Dhivehi language, syllable is used as the basic unit for sentence selection. A lexical corpus of only 360 sentences, containing 868 words was developed, which achieved the phonetically rich and balanced lexical corpus. As the proposed method perform iteratively selection sentences until all the required phonemes are covered, it can be assured that the developed lexical corpus has 100% phonetic coverage, hence achieving the phonetic richness. On top of that, the proposed method is devised in such way that the required phonemes occurs at least once in the developed lexical corpus, hence they are distributed uniformly making the developed lexical corpus to be phoneme balanced.

6.1.4 Research objective 4

The fourth and final objective of this research is to evaluate the performance of the proposed method against the benchmark method. The research compared the results obtained by the proposed method with the classic greedy method. The findings show that the sentences selected by the proposed method are more optimized, and maintains the richness and balanced nature. The results show that benchmark method was able to achieve richness and balanced by using about 4.89% of the sentences from the initial database, while the proposed method only use 0.33% sentences from the initial database, though both the proposed and benchmark method uses iterative selection. On top of that, the high cosine similarity between the initial large corpus and final small corpus suggest that the small corpus reflect the phoneme structure of the large database.

6.2 Conclusion

Designing a set of phonetically rich and balanced corpus to be used for developing speech-based applications, and in particular for training and testing the ASR and TTS systems, is a critical issue for resourced poor languages such as Dhivehi. In this research, the existing methods for developing lexical corpus were reviewed and a novel method was proposed to develop the lexical corpus for under-resourced languages. The entire process of building lexical corpus for Dhivehi language is presented in this research.

Because of limited text data available for Dhivehi, web-based sources for Dhivehi such as websites, Facebook, and Twitter were used by adopting the method proposed in (Mendels et al., 2016) with minor modification for collecting the web data. After refining the raw data extracted from the web, a large database consisting 109,208 sentences was developed. From the large initial database, a subset of sentences was extracted by ranking sentences using the Zipfian distribution. The small lexical corpus comprises of 360 sentences was successfully developed with cosine similarity of 0.988. The proposed method was also compared with a benchmark method, where the size of the lexical data of the proposed method was 15 times smaller than the benchmark method, indicating the effectiveness of the optimization capacity of the proposed method.

6.3 Contribution

The main contribution of this research includes:

- **Developers**

The accumulated resources for Dhivehi language can help developers to develop speech-based applications in the future. On top of that, the step-by-step procedure of the proposed method in this research can be used as guideline to develop a lexical corpus for many other under-resourced languages.

- **Research community**

Because of the lack of the lexical corpus especially for under-resourced languages, speech-based applications are not available for many of these languages. This proposed method for lexical development will encourage researchers to take more active role in speech-based research, and reduce the technological gap between the languages.

6.4 Limitation

As Dhivehi is a language with limited tools and resources, this research made use of the syllables as the basis unit i.e., mono-syllables. To make the lexical corpus more robust, contextual units such as di-phones, tri-phones, and tono-phones can be used instead of syllables only. However, due to the time and man-power constrain, contextual units such as di-phones, tri-phones, and tono-phones was not applied in this research. On top of that, despite the ability of the proposed method to reduce human involvement, it is not 100% human free. In this research, the researcher needs to manually review the small corpus for any inappropriate contents.

6.5 Future work

The acoustic models of speech processing system require a set of speech corpus that have a good coverage of the language. The sentences constructed in this research can be the basis for developing the speech corpus for Dhivehi language, from which speech-based applications can be developed. On top of that, the proposed method can be suitably adapted for developing the lexical corpus for other under-resourced languages. Another possible future work is to use more complex contextual phonetic units instead of mono-syllables, which can improve the robustness of the lexical corpus.

REFERENCES

Abera, H., Nadeu, C., & Mariam, S. (2016). Extraction of syllabically rich and balanced sentences for Tigrigna language 2 . Phonetic nature of the Tigrigna language. *Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2094–2097.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012). Phonetically rich and balanced text and speech corpora for Arabic language. *Language Resources and Evaluation*, *46*(4), 601–634. https://doi.org/10.1007/s10579-011-9166-8

Anumanchipalli, G., Chitturi, R., Joshi, S., Kumar, R., Singh, S. P., Sitaram, R. N. V, & Kishore, S. P. (2005). Development of Indian Language Speech Databases for Large Vocabulary Speech Recognition Systems. *Proceedings of SPECOM*.

Arora, K., Arora, S., Verma, K., & Agrawal, S. S. (2004). Automatic Extraction of Phonetically Rich Sentences from Large Text Corpus of Indian Languages. *Interspeech*.

Aubanel, V., Lecumberri, M. L. G., & Cooke, M. (2014). The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *International Journal of Audiology*, *53*(9), 633–638. https://doi.org/10.3109/14992027.2014.907507

Bansal, shweta, Sharan, S., & S.S, A. (2015). CORPUS DESIGN AND DEVELOPMENT OF AN ANNOTATED SPEECH DATABASE FOR PUNJABI College of Engineering , Gurgaon Uf [k � i] � W t] [I]. *KIIT College of Engineering, Gurgaon*, 32–37.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2013). Automatic Speech Recognition for Under-Resourced Languages : A Survey d. *SPEECH COMMUNICATION*. https://doi.org/10.1016/j.specom.2013.07.008

Chevelu, J., Barbot, N., Boëffard, O., & Delhay, A. (2007). Lagrangian relaxation for optimal corpus design. *Proceedings of the 6th ISCA Tutorial and Research*, 211–216.

Gnanadesikan, A. E. (2017). *Dhivehi: The Language of the Maldives*. Walter de Gruyter GmbH \& Co KG.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., & Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. *Proc. LREC*, 2005–2010.

Ha, L. Q., Hanna, P., Ming, J., & Smith, F. J. (2009). Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, *32*(1–4), 101–113. https://doi.org/10.1007/s10462-009-9135-4

Habib, W., & Adeeba, F. (2014). Design of Speech Corpus for Open Domain Urdu Text to Speech System Using Greedy Algorithm.

Kasparaitis, P., & Anbinderis, T. (2014). Building Text Corpus for Unit Selection Synthesis. *INFORMATICA*, *25*(4), 551–562.

Kilgarriff, A., Reddy, S., Pomikálek, J., & Pvs, A. (2008). A Corpus Factory for many languages. *Corpus*, 904–910.

Malviya, S., Mishra, R., & Tiwary, U. S. (2016). Structural Analysis of Hindi Phonetics and A Method for Extraction of Phonetically Rich Sentences from a Very Large Hindi Text Corpus, (October), 26–28.

Mandal, S., Das, B., Mitra, P., & Basu, A. (2011). Developing Bengali speech corpus for phone recognizer using optimum text selection technique. *Proceedings - 2011 International Conference on Asian Language Processing, IALP 2011*, 268–271. https://doi.org/10.1109/IALP.2011.16

Matoušek, J., Tihelka, D., & Romportl, J. (2008). Building of a speech corpus optimised for unit selection TTS synthesis. *Language Resources and Evaluation Conference*, 1296–1299.

Mendels, G., Cooper, E., & Hirschberg, J. (2016). Babler - Data Collection from the Web to Support Speech Recognition and Keyword Search. *ACL 2016*, 72–81.

Mendonca, G., Candeias, S., Perdigao, F., Shulby, C., Toniazzo, R., Klautau, A., & Aluisio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *2014 International Telecommunications Symposium, ITS 2014 - Proceedings*. https://doi.org/10.1109/ITS.2014.6947957

Mohaghegh, M., & Sarrafzadeh, A. (2016). Parallel Text Identification Using Lexical and Corpus Features for the English-Maori Language Pair. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on* (pp. 910–915).

Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of Zipf's law in English texts. *PloS One*, *11*(1), e0147073.

Nguyen, T. T. T. (2015). HMM-based Vietnamese Text-To-Speech : Prosodic Phrasing Modeling , Corpus Design System Design , and Evaluation.

Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., & Shiga, Y. (2017). Sentence Selection Based on Extended Entropy Using Phonetic and Prosodic Contexts for Statistical

Parametric Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(5), 1107–1116. https://doi.org/10.1109/TASLP.2017.2688585

Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., Shiga, Y., & Ito, A. (2015). Entropy-based sentence selection for speech synthesis using phonetic and prosodic contexts. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015–Janua*, 3491–3495.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pierrehumbert, J. B., Beckman, M. E., & Ladd, D. R. (2012). Conceptual Foundations of Phonology as a Laboratory Science (reprint).

Rabiner, L. R., & Schafer, R. W. (2007). *Introduction to Digital Speech Processing*. *Foundations and Trends® in Signal Processing* (Vol. 1). https://doi.org/10.1561/2000000001

Raza, A. A., Hussain, S., Sarfraz, H., Ullah, I., & Sarfraz, Z. (2009). Design and development of phonetically rich Urdu speech corpus. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments, ICSDA 2009*, (November 2014), 38–43. https://doi.org/10.1109/ICSDA.2009.5278380

Riyal, M. K., Rajput, N. K., Khanduri, V. P., & Rawat, L. (2016). Rank-frequency analysis of characters in Garhwali text: emergence of Zipf's law. *CURRENT SCIENCE*, *110*(3), 429--434.

Saloot, M. A., Idris, N., Aw, A. T., & Thorleuchter, D. (2016). Twitter corpus creation:

The case of a Malay Chat-style-text corpus (MCC). *Digital Scholarship in the Humanities*, *31*(2), 227–243. https://doi.org/10.1093/llc/fqu066

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers Du Cental*, *5*(1), 5–15.

Shinohara, Y. (2014). A SUBMODULAR OPTIMIZATION APPROACH TO SENTENCE SET SELECTION Yusuke Shinohara Corporate Research and Development Center , Toshiba Corporation. *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 4140–4143.

Tan, T. S., & Sh-Hussain. (2009). Corpus design for Malay corpus-based speech synthesis system. *American Journal of Applied Sciences*, *6*(4), 696–702. https://doi.org/10.3844/ajas.2009.696.702

Uraga, E., & Gamboa, C. G. (2004). VOXMEX speech database: Design of a phonetically balanced corpus. *LREC 2004. Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1471–1474.

Villaseñor-Pineda, L., Montes-y-Gómez, M., Vaufreydaz, D., & Serignat, J.-F. (2004). Experiments on the Construction of a Phonetically Balanced Corpus from the Web. *Experiments on the Construction of a Phonetically Balanced Corpus from the Web*, 3–6. https://doi.org/http://dx.doi.org/10.1007/978-3-540-24630-5_50

Vorapatratorn, S., Suchato, A., & Punyabukkana, P. (2012). Automatic online text selection for constructing text corpus with custom phonetic distribution. *JCSSE 2012 - 9th International Joint Conference on Computer Science and Software Engineering*, 6–11. https://doi.org/10.1109/JCSSE.2012.6261916

Wang, H. (1998). Statistical Analysis of Mandarin Acoustic Units and Automatic

Extraction of Phonetically Rich Sentences Based Upon a Very Large Chinese Text Corpus. *Computational Linguistics and Chinese Language Processing*, *3*(2), 93–114.

Weikum, G. (2002). Foundations of statistical natural language processing. *ACM SIGMOD Record*, *31*(3), 37. https://doi.org/10.1145/601858.601867

Yang, S., Zheng, F., Luo, X., Cai, S., Wu, Y., Liu, K., … Krishnan, S. (2014). Effective dysphonia detection using feature dimension reduction and kernel density estimation for patients with parkinson's disease. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0088825

Yuwan, R., & Lestari, D. P. (2016). Computational Linguistics, *593*, 65–75. https://doi.org/10.1007/978-981-10-0515-2

Zhang, J. S., & Nakamura, S. (2008). An improved greedy search algorithm for the development of a phonetically rich speech corpus. *IEICE Transactions on Information and Systems*, *E91–D*(3), 615–630. https://doi.org/10.1093/ietisy/e91-d.3.615

Zhang, W., Liu, Y., Deng, Y., & Pang, M. (2010). Automatic construction for a TTS corpus with limited text. *2010 International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2010*, *1*, 707–710. https://doi.org/10.1109/ICMTMA.2010.796