

**PRONUNCIATION MODELLING OF PENANG HOKKIEN DIALECT FOR
TEXT-TO-SPEECH SYSTEM**

By
LIM KANG JIE

A REPORT
SUBMITTED TO
Universiti Tunku Abdul Rahman
in partial fulfillment of the requirements
for the degree of
BACHELOR OF INFORMATION SYSTEMS (HONOURS) INFORMATION
SYSTEMS ENGINEERING
Faculty of Information and Communication Technology
(Kampar Campus)

MAY 2022

REPORT STATUS DECLARATION FORM

Title: **Pronunciation Modelling of Penang Hokkien Dialect**
For Text-To-Speech System

Academic Session: **Y4T1**

I **LIM KANG JIE**
(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.



(Author's signature)

Verified by,



(Supervisor's signature)

Address:

27, Lorong Machang Bubok 17
Taman Machang Bubok
14000 Bukit Mertajam
Pulau Pinang

Dr. Jasmina Khaw Yen Min
Supervisor's name

Date: 09 September 2022

Date: 09 September 2022

Universiti Tunku Abdul Rahman			
Form Title : Sample of Submission Sheet for FYP/Dissertation/Thesis			
Form Number: FM-IAD-004	Rev No.: 0	Effective Date: 21 JUNE 2011	Page No.: 1 of 1

FACULTY/INSTITUTE* OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 09/09/2022

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that LIM KANG JIE (ID No: 18ACB02259) has completed this final year project/ dissertation/ thesis* entitled “Pronunciation Modelling of Penang Hokkien Dialect” under the supervision of Dr. Jasmina Khaw Yen Min (Supervisor) from the Department of Computer Science, Faculty/Institute* of Information and Communication Technology.

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.


Yours truly,



(LIM KANG JIE)

DECLARATION OF ORIGINALITY

I declare that this report entitled “**PRONUNCIATION MODELLING OF PENANG HOKKIEN DIALECT FOR TEXT-TO-SPEECH SYSTEM**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :  _____

Name : LIM KANG JIE _____

Date : 9 SEPTEMBER 2022 _____

ACKNOWLEDGEMENTS

Throughout the writing of this thesis, I have received a lot of supports and assistances.

I would like to express my sincere trillions of thanks and appreciations to my supervisor, Dr Jasmina Khaw Yen Min whose expertise was invaluable in guiding me throughout the research questions and methodology. Your insightful feedbacks pushed me to sharpen my thinking and brought my tasks to more advanced level.

I would like to acknowledge my academic advisor, Ts Sun Teik Heng who always encourage me to work hard in my degree. Your encouragement gives me motivation which alongside me during the hard time in my university life.

I would also like to thank musicians and composers that mostly motivate me alongside odyssey of my life to achieve my dreams. Thousands of appreciations to Satoshi Fujihara from Official Hige Dandism, Kenshi Yonezu, Shota Shimizu, Namewee, milet, Yu-Peng Chen, Johann Sebastian Bach, Antonio Vivaldi, Ludwig van Beethoven, Camille Saint-Saëns, EggPlantEgg, and YOASOBI.

Lastly, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me and give me lots of motivation throughout my entire life. An infinite thanks to you my “Oyaji” and my “Ofukuro”.

ABSTRACT

This is academic research about pronunciation modelling of Penang Hokkien for Text-to-Speech System which is under field of study, Speech Synthesis. It is widely known that there are majority of unwritten languages are gradually forgotten by younger generations due to domination of written languages in education and the most significant factor is lack of documentation of the languages. Hence, these hindrances prevent or increase the effort of revitalization on those unwritten languages by implementing current technologies. Penang Hokkien Language, a likely unwritten language spoke in Northern of Southern Peninsular Malaysia is selected as case study of this research where its linguistic resources are partially documented. In order to develop an TTS System for Penang Hokkien, this research project is the first steps to familiarize with this high complexity language. Since this project is part of the effort in revitalizing Penang Hokkien Language, Traditional Chinese Character is opted as standard of writing system and Penang Hokkien Spelling System which created by Hokkien Association of Penang is selected as standard of pronunciation orthography. Listing of phonemes with categorizing them into initials and finals are taken as Penang Hokkien is a tonal language. Moreover, nine tones are marked with the use of diacritics based on Penang Hokkien Spelling System according to the tone marking rules. Tone sandhi rules are also created in orthography standardization phase. The contributions of this project are (1) finding the possible combinations of initials and finals and tones, (2) collect possible graphemes, (3) map graphemes with morphemes, (4) design database to store the processed data and (5) standardizing the tones and tone sandhi rules.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION OF ORIGINALITY	iv
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER 1 Introduction	1
1.1 Introduction	1
1.2 Problem Statement	3
1.3 Motivation	3
1.4 Research Objectives	4
1.5 Research Background	4
1.5.1 Speech Synthesis Techniques	4
1.5.2 Min Languages Origin	5
1.5.3 Penang Hokkien Language	6
1.5.4 Penang Hokkien Facing Extinction	7
1.6 Project Scope	7
1.7 Contributions	8
1.8 Report Organization	8
CHAPTER 2 Literature Reviews	9
2.1 Penang Hokkien Language Writing and Pronunciation System	9
2.1.1 Traditional Chinese Characters with 9 Tones	9
2.1.2 Taiji System with 4 Tones	10
2.1.3 Comparison of Both Systems	10
2.2 Non-Similar Language Corpus Construction Review	11
2.2.1 Cantonese Corpora Construction	11
2.2.2 Czech Corpora Construction	12
2.3 Similar Language Corpus Construction Review	14
2.3.1 Singapore Hokkien Corpus Construction	14
2.3.2 Minnan Child Speech Corpus Construction	15
CHAPTER 3 System Model	17
	vii

3.1	Research Flow Chart	17
3.2	Penang Hokkien Database Design	18
3.2.1	ER Diagram	18
3.2.2	Data Dictionary	18
3.3	Penang Hokkien Database Construction	20
3.3.1	Database Construction Flow Chart	20
3.4	Software and Dictionaries Used	21
CHAPTER 4 EXPERIMENT		22
4.1	Phonemes of Penang Hokkien	22
4.1.1	Initials	22
4.1.2	Finals	22
4.2	Tones of Penang Hokkien	23
4.2.1	Nine Tones of Penang Hokkien	23
4.2.2	Tones Letter and Tones Contour	24
4.2.3	Tones Graph	25
4.2.4	Tones Marking Rules	25
4.2.5	Tone Sandhi	26
4.2.6	Tone Table Creation as Guidelines for Future Research	29
4.3	Database Construction for Penang Hokkien Morphemes and Phonemes	30
4.3.1	Morphemes_Phonemes Table Creation	30
4.3.2	Graphemes Table Creation	32
4.3.3	Mapping Table Creation	33
4.3.4	Implementation of Tables into the Database	37
4.3.5	Database Implementation Issues	41
4.4	Articles and Sentences Collection	44
4.4.1	Taiwanese Hokkien Articles	44
4.4.2	Penang Hokkien Sentences	46
4.5	Audio Guidelines Collection	48
4.6	Concluding Remarks	49
CHAPTER 5 SYSTEM EVALUATION AND DISCUSSION		50
5.1	Data Refining on Pre-Implementation Database Tables	50
5.1.1	Tables of Pre-refining Process	50
5.1.2	Data Mapping and Refining Process	50
5.1.3	Results	51

5.2	Simple Data Analysis Towards Pre-Implementation Database Tables	53
5.2.1	Morphemes_Phonemes Table	53
5.2.2	Mapping Table	54
5.3	Solutions on Errors During Database Implementation	55
5.3.1	morphemes_phonemes.xlsx	55
5.3.2	graphemes.xlsx	55
5.3.3	maps_graphs_morphs.xlsx	55
5.3.4	ER Diagram Redesign with Data Dictionaries	55
5.3.5	Reimplementation of Tables in Database	57
5.4	Project Challenges	61
5.5	Concluding Remarks	62
CHAPTER 6	Conclusion and Recommendation	63
6.1	Conclusion	63
6.2	Recommendations	64
REFERENCE		65
APPENDIX A	Penang Hokkien Spelling System	A-1
APPENDIX B	EVUALATION LOGS	B-1
APPENDIX C	FYP 2 Poster	C-4
APPENDIX D	FINAL YEAR PROJECT WEEKLY REPORT	D-5
PLAGIARISM CHECK RESULT		69
FYP 2 CHECKLIST		72

LIST OF FIGURES

Figure 2.1 Traditional Chinese Character and Romanized Form	9
Figure 2.2 Taiji Writing and Romanized System	10
Figure 3.1 Flow Chart of Research	17
Figure 3.2 Penang Hokkien ER Diagram	18
Figure 3.3 Flow Chart of Database Construction	20
Figure 4.1 Tones Graph with examples of Graphemes and Romanizations	25
Figure 4.2 60 to 80 Age Group Penang Hokkien Speaker Tone Sandhi	26
Figure 4.3 40 to 60 Age Group Penang Hokkien Speaker Tone Sandhi	26
Figure 4.4 20 to 40 Age Group Penang Hokkien Speaker Tone Sandhi	27
Figure 4.5 Standardized Penang Hokkien Tone Sandhi with Tone Contours	28
Figure 4.6 Phonemes and Tones Indexing of Online Taiwanese Dictionary	31
Figure 4.7 Amoy Hokkien Dictionary Page 223 Cropped Image	32
Figure 4.8 Database Tables Importing Process Diagram	37
Figure 4.9 Step1: SQL Script Creation	38
Figure 4.10 Step 2: Import Selected File	39
Figure 4.11 Step 3: Select Import Destination	40
Figure 4.12 Step 4: Select Encoding Type and Attributes Field Type	41
Figure 4.13 Error Snapshot 1: morphemes_phonemes.csv	42
Figure 4.14 Error Snapshot 2: graphemes.csv	42
Figure 4.15 Encoding Types in MySQL	43
Figure 4.16 Article Folder Properties Window	44
Figure 4.17 Articles Unreadable Filename Lists	45
Figure 4.18 Article Lists Snapshot	45
Figure 4.19 Sentences Collected from Speak Hokkien Campaign Facebook Page	46
Figure 4.20 Sentences Collected from Write Penang Hokkien Facebook Page	47
Figure 4.21 Guideline Audios Folder Properties	48
Figure 4.22 Phonemes Guideline Audios Filename Lists	48
Figure 5.1 Total Number of Phonemes Before Mapping	50
Figure 5.2 Total Number of Graphemes Before Mapping	50
Figure 5.3 Total Number of Extra Phonemes	51
Figure 5.4 Total Numbers of Unadded Graphemes and Phonemes	52
Figure 5.5 Total Numbers of Graphemes after Data Refining	52

x

Figure 5.6 Total Numbers of Phonemes after Data Refining	52
Figure 5.7 Total Mapped Records	53
Figure 5.8 Initials and Final Types Analysis	53
Figure 5.9 Percentage Distribution of Final_Types	54
Figure 5.10 Percentage Distribution of Map_Type	54
Figure 5.11 Redesigned ER Diagram for Penang Hokkien	55
Figure 5.12 Successful Import Logs	57
Figure 5.13 Primary Key and Not Null Set Up for graphemes	58
Figure 5.14 Primary Key and Not Null Set Up for maps_graphs_morphs	59
Figure 5.15 Primary Key and Not Null Set Up for morphemes_phonemes	59
Figure 5.16 Foreign Keys Set Up for maps_graphs_morphs	60
Figure 5.17 Foreign Keys Set Up Results for maps_graphs_morphs	60
Figure 5.18 ERD Visualization with Reverse Engineering	61

LIST OF TABLES

Table 3.1 Data Dictionary for morphemes_phonemes	19
Table 3.2 Data Dictionary for graphemes	19
Table 3.3 Data Dictionary for maps_graphs_morphs	19
Table 4.1 Penang Hokkien Initials Phonemes	22
Table 4.2 Penang Hokkien Finals Phonemes	22
Table 4.3 Tones and Diacritic Marks of Penang Hokkien	23
Table 4.4 Tone Naming	23
Table 4.5 Tones Letter and Tones Contour Table	24
Table 4.6 Tone Marking Priority	25
Table 4.7 Sample Phoneme_Morpheme Table from Excel	30
Table 4.8 Sample Graphemes Table from Excel	33
Table 4.9 Sample Mapping Table from Excel	34
Table 4.10 Map Types Abbreviation and Penang Hokkien Translation	36
Table 5.1 Redesigned Data Dictionary for morphemes_phonemes	56
Table 5.2 Redesigned Data Dictionary for graphemes	56
Table 5.3 Redesigned Data Dictionary for maps_graphs_morphs	57

LIST OF ABBREVIATIONS

<i>TTS</i>	Text-to-speech Synthesis
<i>ASR</i>	Automatic Speech Recognition
<i>DNN</i>	Deep Neural Network
<i>MOS</i>	Mean Opinion Score
<i>US</i>	United States
<i>HMM</i>	Hidden Markov Model
<i>EGIDS</i>	Expanded Graded Intergenerational Disruption Scale
<i>AI</i>	Artificial Intelligence
<i>ISO</i>	International Organization for Standardization
<i>ASCII</i>	American Standard Code for Information Interchange
<i>IPA</i>	International Phonetic Alphabet
<i>POS</i>	Part of Speech
<i>COVID-19</i>	Coronavirus Disease 2019
<i>RDBMS</i>	Relational Database Management System
<i>ERD</i>	Entity Relationships Diagram
<i>R.O.C</i>	Republic of China
<i>OPS</i>	Open Syllables
<i>NAC</i>	Nasal Consonants
<i>NAV</i>	Nasal Vowels
<i>NaN</i>	Not Available
<i>STC</i>	Stop Consonants
<i>LTR</i>	Literary Pronunciation
<i>CLQ</i>	Colloquial Pronunciation
<i>KYM</i>	Kunyomi Pronunciation
<i>ATJ</i>	Ateji Pronunciation
<i>UCT</i>	Uncategorized Pronunciation
<i>UTF</i>	Unicode Transformation Format
<i>CJK</i>	Chinese Japanese Korean

CHAPTER 1 Introduction

1.1 Introduction

Text-to-speech Synthesis (TTS) Technology literally means technology that realizes conversion process of normal language text into speech or in other words – “a talking machine or a talking computer” [1], [2]. TTS has similar technology which is Automatic Speech Recognition (ASR) which realizes conversion process of human speeches into text and both of these technologies are under a field of study, Speech Synthesis [1]. There are vast applications of TTS but the first intention of inventors to develop TTS was to use it in reading system (convert text from book into speech) for visually impaired persons [1], [3], [4]. This can be proven by the first commercial TTS system – Kurzweil reading machine for blind was invented and produced in 1976 [5]. Other than that, [1] stressed that TTS technologies are also used in call-centre automation and news stories reading, weather reporting, and others. Additionally, well known physicist Stephen Hawking used TTS as speech prosthesis which developed by Intel [4], [6]. Moreover, verification of orthography correctness of written word and listening of personalized text which contains unknown pronunciation in second-language learning as well as automated audio books and computer games narratives are also implementing TTS technologies [4]. These applications manifest the high implementation and significance of TTS in human daily life. However, such manifesting and useful technology must have various of tedious requirements.

In order to build a speech synthesis for a language, identification of language type is vital to ensure the best approaches used in the speech synthesis development towards its quality. There are few major types of languages; for instance, tonal language, intonation language, pitch-accent language which are language type of Chinese Mandarin, English, and Japanese respectively. Google’s paper that introduced WaveNet – a deep neural network (DNN) technology that generate raw audio waveform made comparison of voice quality between DNN, concatenative technique, parametric technique and human speech by using Mean Opinion Score (MOS) towards US English and Mandarin Chinese. By only comparing the results using concatenative technique and parametric technique, US English showed advantageous results on concatenative whereas Mandarin Chinese shows advantageous results on parametric technique [7]. As mentioned previously, before

opting the best speech synthesis approach, identification of language type should be done to ensure its quality. On the other hand, a well-developed TTS synthesizer postulate a numerous type of resources. As instances. Text and speech databases, linguistic analysing tools, dictionaries, signal processing, and grammars are needed during the development process [4]. Resources selection placed the most important part as it directly affects the accuracy and the quality of the speech synthesis; hence, [4] suggested that during the corpus construction, professional speakers or voice talents should be invited during the recording of natural speech. Moreover, text resources should be obtained from the assorted text genres representable sources to ensure obtainable of the target language's optimal statistical distribution coverage [4]. Futhermore, language's pronunciation dictionary, lexical stress, syllable boudaries as well as morphological structure are also required in the development progress [4]. [4] also asserted that languages with nonalphabetical writing systems require systematic transcription (standardized romanization in orthography) for further realizing TTS conversion; Mandarin Chinese with Pinyin is one of the successful examples. However, there are lots of unwritten languages that exist only in verbal, entire culture without writers in entire globe; there are only 3,982 languages have developed writing system and approximately 3,135 are likely unwritten [8]. Without developed writing system with well documentation, it is impossible to convert text to speech. This makes TTS in an awkward and unsatisfying state as there are only few languages and dialects are supported by TTS [4].

In this research, a likely unwritten language in Malaysia, Penang Hokkien is taken as case study for this research. The most obvious problems for Penang Hokkien which are caused by weak documentation. Although the language exists since few hundred years ago in Malaysia, the language is lack of documentation on the tones, texts, phrases, idioms, and others. Moreover, Penang Hokkien is lack of literatures where Hokkien Language Association of Penang had dug up old books and documents to prove that Penang Hokkien is a language not a dialect of Mandarin [9]. These would be the main factors that causing the corpora is not complete. Unstandardized writing system and pronunciation system would be another problem that currently faced by Penang Hokkien. Thence, this research will standardize the orthography and pronunciation of Penang Hokkien by collecting resources from high

mutual intelligibility documented languages that has same progenitor with Penang Hokkien as a first step before proceeding into future construction of Text Corpora and Speech Corpora for Penang Hokkien.

1.2 Problem Statement

1. No standard orthography in Penang Hokkien. There 2 writing system and 2 romanization system in Penang Hokkien Language. Traditional Chinese character as graphemes of the language is suggested by Hokkien Association of Penang with its Romanised using Penang Hokkien Spelling System that modified and referred to Tâi-lô (Tâi-uân Lô-má-jī Phing-im Hong-àn) [10]. Timothy Tye suggests fully Romanised writing system as graphemes for Penang Hokkien which is called Taiji System [11].

2. Lack of formal pronunciation dictionary in Penang Hokkien. Timothy Tye developed an online dictionary for Penang Hokkien with Taiji System as main listed entries followed by definitions in English, Malay, et cetera [12]. However, the Taiji System is developed only by Timothy Tye himself without any other linguistic experts involved during the developing process and criticized by an Southern Min expert towards Tye's unprofessional transliteration and romanization on Penang Hokkien pronunciation [13].

1.3 Motivation

The motivation of this project is to model pronunciation of Penang Hokkien Language for Text-to_Speech System by standardize the written text for Penang Hokkien Script and the romanization of the text and collect audio guidelines of each phoneme as first step of TTS system development of Penang Hokkien. The standardization methods will be discussed in the Chapter 3. Since the language is facing extinction in future 40 years as [9] reported prediction of a language expert, Catherine Churchman, palliation of the threat of extinction should be taken for rescuing the language; hence, this project is vital as one part of the processes of rescuing Penang Hokkien Language.

1.4 Research Objectives

1. To collect large number of phrases and words of Penang Hokkien for text corpus construction and standardising the language orthographies including the romanization and Chinese characters and collect each morphemes audio sample as pronunciation guidelines.
2. To collect large number of sentences and articles that similar to Penang Hokkien as a future resource in selecting the most suitable sentences for studio recording in future research project.

1.5 Research Background

1.5.1 Speech Synthesis Techniques

Text-to-Speech Synthesis System as mentioned in the introduction could be one of the approaches to rescuing the endangered unwritten languages. It converts language text into speech. There are several techniques currently exist in this world which are Concatenative TTS, Formant Synthesis, Parametric TTS, Articulatory Synthesis, HMM-based Synthesis, Sinewave Synthesis and Hybrid (Deep Learning) Approaches. TTS synthesis system is evaluated in aspects of intelligibility, naturalness, preference, and comprehensibility [14].

Concatenative TTS required high-quality of audio recordings for combination to form the speeches. Labelling and segmentations of the range of speech units that recorded from the voice actors are done by the linguists from phones to phrases and sentences to build a gigantic database. When synthesizing the speech, Text-to-Speech engine concatenates the speech units that matches with the inputted text to produce an audio file by searching them from the databases [14].

Formant Synthesis techniques generates artificial signals based on a set of specified rules imitating the formant structure and other unique properties of natural speech to produce speech segments. It produces the synthesized speech by using an additive synthesis and an acoustic model that include parameters such as voicing, fundamental frequency, noise level and others [14].

Parametric TTS was invented to address the limitations of the concatenative TTS by combining the parameters such as fundamental frequency, magnitude spectrum and

others to process them into speech. In this method, extraction of linguistic features like phonemes, duration and others from the texts is required then will be moving into the next steps which is extraction of vocoder features (human speech characteristics): spectrogram, cepstra, fundamental frequency and others to use in audio processing in the future. Vocoder will be involved in this technique by transforming the features of the generated waveform and estimating the parameters of speech [14].

1.5.2 Min Languages Origin

Yue Retreated to Min, Fusion of Min and Yue, Min Languages Progenitors

The earliest recorded history of Fujian Province was after the Spring and Autumn Warring States period, royal family of Yue States retreated to Fujian area that had mountainous and riverine terrains, integrated with the natives, and founded Minyue where the relics and ruins of the royal family can be found in Northern Min region. In the consensus of the linguistics world, ancient Minyue Language family have certain blood relationships with Kra-Dai Language family and its lexical elements are reserved in various modern Min Languages [15].

Han Annexed Minyue, Huge Migration to North, Garrisoned Mother Tongues as High-Influencing Progenitors

Huge migration from Minyue to Yangtze occurred with result of the successful invasion of Han Dynasty towards Minyue in 110 BC. Han militaries from Jiangdong (Wu people) and Jiangxi (Chu people) brought their mother tongues, Ancient Wu and Ancient Chu Languages respectively integrated with Min Yue Language and became Proto-Min Language during their garrisoning period in Fujian; In the modern Min Languages, the lexical elements of Ancient Wu and Ancient Chu Languages are preserved [15].

Fall of Jin, Chaos Caused Migration into Min, Shaping of Min Languages

There were two main events that has huge influence on the shaping of the Min Languages. “Chaos of Yongjia”, was the first event that was a result from the event of Five Barbarians Upheaving in northern China during Western Jin Dynasty. The chaos forces huge migration from northern China to Fujian where the mother tongues were brought into the region too, this second event named as “Eight Great Surnames Entering Min” [15].

Two Waves of Migrations, Phonology System Influences, Shaped and Distinctive Min Languages

During Tang Dynasty, there were two waves of migrations from Central Plains to Fujian. The first migration was occurred in Early Tang where peoples were mainly arrived at Zhangzhou and Quanzhou. The noticeable events after their arrival were Tan brothers' achievements: (1) Pacified She Rebellions, (2) Developed Zhangzhou, and (3) Increased Quanzhou Population. These events shaped the Southern Min Languages that are predecessors of Malaysian Hokkien, Amoy Hokkien, Taiwanese Hokkien et cetera. While the second waves of the migration occurred during the Late Tang where peoples mainly migrated to Fuzhou. The noticeable event after their arrival was Uong father and son successfully occupied Fuzhou by military and founded Min State. Min Languages were fully shaped during Tang Dynasty with contribution of imperial examination system that brought Qieyun phonology system into Fujian. Moreover, the languages in Fujian had distinction due to the huge migrations and mountainous and riverine terrains which finally categorized Min Languages mainly into Puxian Min, Northern Min, Eastern Min, Southern Min, and Central Min [15].

1.5.3 Penang Hokkien Language

Penang Hokkien Language is the mother-tongue of the initial group of Chinese settlers in Malaysia came to Georgetown, Penang during 1786, which was the year of British Port establishment in Georgetown by Francis Light. After that, Hokkien community started to develop in Georgetown in 1800s. Penang Hokkien Language has been used as the major teaching language among the native Chinese Hokkien for more than a century even before Francis Light founded Penang. Unfortunately, during 20th century, Dr. Sun Yat-sen initialized Chinese Revolution for promoting Mandarin to gathering all Chinese subethnic [16]. Research in 2009 proved that Chinese education system has been influenced by the Chinese Revolution by establishing new school and reading clubs in big cities and small towns replacing small-scale private education; the first modern school was established in Penang in 1904 [17]. Enclosed with Malaysianization and Globalization, Penang Hokkien boasted with multi-lingual by speaking British English, Mandarin, varieties of Chinese Provincial Dialects

(Cantonese, Teochew, Hock chew, Hakka, Hainanese et cetera.), and Bahasa Malaysia [16].

1.5.4 Penang Hokkien Facing Extinction

Mok (2016) reported that a language expert, Catherine Churchman said that Penang Hokkien Language will face extinction if Penangites keep replacing it with English or Mandarin. The Penang Hokkien communities is now facing a huge problem which is fractured communities due to language barrier, and this key problem for this is majority of younger generation of Penangites are not cognizant and not conversant in Penang Hokkien [9]. This could be proved with the research result towards the Penang Chinese teenagers in 2020 where the Penang Hokkien vitality falls at Expanded Graded Intergenerational Disruption Scale (EGIDS) Level 6b [18]. EGIDS estimate languages' overall development against endangerment [8], [19]; EGIDS Level 6b, the result of Ting and Teng (2021) research means Penang Hokkien Language currently is threatened and it is only used during face-to-face communication by all generation with losing users. Mok (2020) also reported that the Hokkien Language Association of Penang is chronicling the slow decline of Southern Chinese languages by holding an exhibition in attempt to revive the use of languages among the varieties of sub-ethnic Chinese groups. Sim Tze Wei, President of Hokkien Language Association of Penang said that they dug up old documents and books to prove that Penang Hokkien is a language not a dialect of Mandarin [20]. These could show that the weaken of Penang Hokkien Language in communities could be affected by weak documentation on the corpus, dictionary, and speaking system. Moreover, the domination of Mandarin enclosed with globalization causes rapid improvement on complexity of the written text and adapting technologies such as Artificial Intelligence (AI), TTS, and others. Thus, other Chinese languages which could not catch up with new technologies would facing extinction. These problems should be tackled to palliate the extinction of Penang Hokkien and have a higher chance to develop and spread the language globally.

1.6 Project Scope

The title of this project is Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System. The scope for this project is to model the Penang Hokkien pronunciation that act as the first step that contributes to future development of

Penang Hokkien Speech Synthesis System. The project will be collecting a large amount of lexical resource of Penang Hokkien Language included: 1. Phonemes/Morphemes, 2. Tone Rules, 3. Graphemes, 4. Transcriptions, 5. Phrases 6. Penang Hokkien Written Sentences, 7. Most Similar Language Articles, and 8. Audio Guideline Generated by Most Similar Language TTS System. After the collection of the lexical resources, the Phonemes/Morphemes and Graphemes will be further arranged into designed database. The rest of the collected resources will be needed in the future research towards the Penang Hokkien TTS System.

1.7 Contributions

This research is placed at higher vitality towards the efforts of Penang Hokkien Language revitalization. Since Penang Hokkien had weakened due to domination of Chinese Mandarin via education and other channels, problem palliation should be taken to avoid extinction of the language. With the current available resources, ongoing revitalization efforts of Penang Hokkien communities as well as the sophisticated technologies, this research is significant to standardize the orthographies and pronunciations of Penang Hokkien as well as gathering the resources that will be needed in the future developments. Whereby this project act as the first step that contributes to the effort of developing Penang Hokkien Synthesis System in part of revitalization of Penang Hokkien Language.

1.8 Report Organization

The details of the research are sectioned into following chapters. Chapter 2 reviews related orthography systems and corpora constructions. Furthermore, flow of the research, Database design, Database construction flow, and dictionaries and software used are categorized in Chapter 3. Chapter 4 presents phonemes, tones of Penang Hokkien, database construction with problem faced and article as well as audio guidelines in this research. Chapter 5 evaluate the Chapter 4 pre-implementation of tables in the database and presents simple data analysis towards the processed data tables. Moreover, solutions to tackle problem faced in Chapter 4, redesign of ERD and Data Dictionaries and Reimplementation of the Database are also presented in Chapter 5. Lastly, Chapter 6 concludes the entire report of this research.

CHAPTER 2 Literature Reviews

This literature review will be separated into Current Existing Penang Hokkien Language Writing and Pronunciation System, Non-Similar Language Corpus Construction Review, and Similar Language Corpus Construction Review.

2.1 Penang Hokkien Language Writing and Pronunciation System

2.1.1 Traditional Chinese Characters with 9 Tones

Hokkien Language Association of Penang promotes Traditional Chinese Characters as writing system for Penang Hokkien which is recommended by the Ministry of Education of Republic of China [21]. Moreover, they proposed Penang Hokkien Spelling System as the romanization system for Penang Hokkien Language as the system is referenced from the *Tâi-lô* – official romanization system for Taiwanese Hokkien that is also promoted by the Ministry of Education of Republic of China [10]. This system consists of 9 tones where the second and sixth tones have no difference, and the third and seventh tones are also having no difference leaving the tones become seven; sixth tone and ninth tone are used on the loanwords [10].

Table: 臺灣教育部閩南語推薦用字 : Recommended Characters for the Hokkien Language

編號 No.	建議用字 Recommended Character	音讀 Pronunciations	又音 Alternative Pronunciations	對應華語 Mandarin Equivalent
1	阿	a		阿
2	阿姘	a-kīm		舅媽
3	阿姆	a-m̄		伯母
4	仔	á		仔、子
5	壓霸	ah-pà		霸道
6	曷	áh		何須、哪
7	抑是	áh-sī	iah-sī/ah-sī/á-sī/iah-sī	或是、或者
8	愛	ài		喜歡、想要、愛
9	偑	āinn	iāng	背 (人)
10	沃	ak		澆、淋
11	泔	ám		稀的、米湯

Figure 2.1 Traditional Chinese Character and Romanized Form

Figure 2.1 shows example of Traditional Chinese Character and Romanized form that suggested by the Ministry of Education of Republic of China to use in writing Hokkien Language which currently adapted by Hokkien Language Association of Penang.

2.1.2 Taiji System with 4 Tones

Timothy Tye, a Penang Hokkien native speaker promotes Romanized form of writing system called Taiji System or TJ System [22]. In dealing with varieties of pronunciation of the Penang Hokkien, he simplified it and create Taiji Romanization for Hokkien Speakers which is used as the writing system; there are total only 4 tones proposed in this romanization system which are 1, 2, 3 or 33 and 4 and these tones are corresponded to the four tones in Mandarin [11].

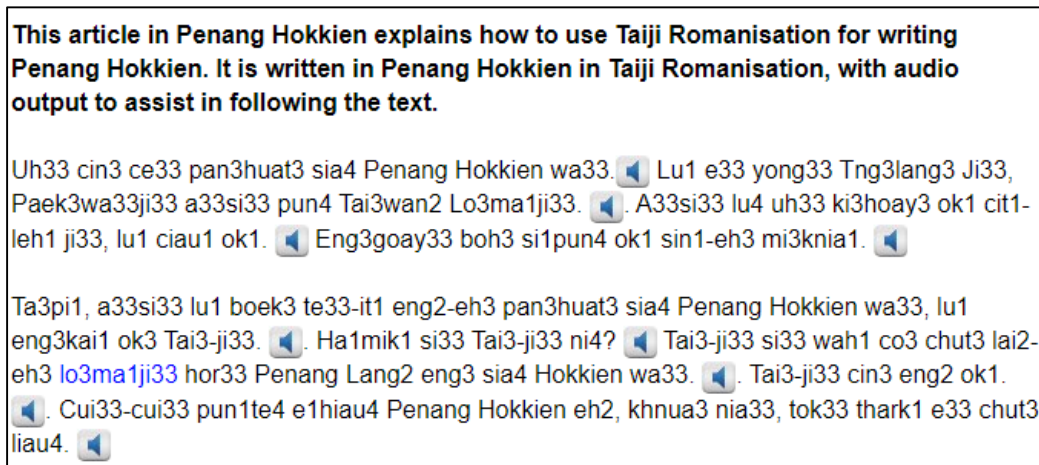


Figure 2.2 Taiji Writing and Romanized System

Figure 2.2 shows article written in Taiji Writing and Romanized System which created and promoted by Timothy Tye by combining tone numbers with Romanized form.

2.1.3 Comparison of Both Systems

Penang Hokkien Language which under Southern Min family has Chinese as macrolanguage as stated in ISO 639-3 [8]. Hence, romanization of the text of the language should be avoided to preserve the culture of Chinese elements. But the romanization should be using as a transliteration not as main writing system of the language. On the other hands, the 9 tones from the Penang Hokkien Spelling System are most suitable to use as it will increase the accuracy of the speech in the future TTS system developing research.

2.2 Non-Similar Language Corpus Construction Review

2.2.1 Cantonese Corpora Construction

Cantonese is one of the Southern Chinese Languages which is one of the major languages used in Southern China, Hong Kong, and others. Few researchers proposed that Cantonese is a monosyllabic language as the meaningful words, sentences and phrases are constructed from the inventory of 1600 syllables, and this scenario was perplexing as same syllable might referred to different characters which carrying totally different meaning [23].

The research from [23] explained that Cantonese is also a tonal language which means that each tone carry a lexical function. For instance, syllables are constructed from fusion of varieties of phonological segments, and these syllables mean differently and corresponded to different of characters which carrying different tones. Cantonese have 9 lexical tones, and they are classified into six main tones as three of them is entering tones (Law et al., 2017). Lo, Lee and Ching [23] implemented six-tone system in Cantonese phonemic transcriptions; they had also separated the speech data by different linguistic levels into corpora: syllable – CUSYL, word – CUWORD, sentence – CUSENT, and passage – CUPASS. Moreover, CUDIGIT – continuous Cantonese digit strings read speech corpus and CUCMD – small task specific word corpus was also included in the research [23].

The data collection for the Cantonese corpora by [23] was performed in a closed silent recording room for good quality data collection. Moreover, high quality microphones were used in the recording process. Throughout the corpora verification process, two stages were done by trained assistant (first stage in verifying syllabus corpus; second stage in verifying CUDIGIT and CUCMD) and phonetics experts (second stage in verifying word, sentence, passage corpora) to resolve all the highlighted problems and mistakes. The data organization were done by allocating the obtained data from same speaker within a same corpus will be placed under the same directory and the directory name was determined using corpus code, speaker code and gender [23].

By highlighting the weaknesses and the strengths of this Cantonese corpora construction, they could be summarized as below:

Weaknesses:

The data collection for speech using recording was costly as this method was using high quality microphones, and the venue taken for the process was performed in a closed silent recording room. These might increase the cost of researching. Moreover, during the stages of verification process, trained assistants and phonetics experts were involved and this might increase the cost of research.

Strengths:

The data were classified precisely and specifically into different corpora which made retrieving process of data easier. Furthermore, implementing tones system of the language had increased the accurate rate of each phrase during recording process. The involvement of expertise in the corpora verification processes ensured the quality of the corpora. Lastly, the obtained data were organized appropriately according to the speakers' directories which determined using gender, speaker code, and corpus code.

2.2.2 Czech Corpora Construction

In the text collection process, [24] selected phonetically rich text sentences as it follows the desired distribution of phonetic units which their algorithm obeyed that strategy. For instance, sentences that containing all phonetic events with maximum occurrence of uniform distribution will be selected as it corresponded with another strategy which is naturally balanced sentences selection that contain phonetics events corresponded with the frequency in natural speech [24].

After the researchers collected the text materials, they were ready to be recorded; a special environment which was based on SoundForge digital editing software was built for the recording purposes [24]. During the recording phase, recording session management took a very important responsibility to control the quality of the recording. [24] reported that the entire recording for the corpora usually takes several weeks with several thousand sentences. When dealing with non-professional speakers, it was understandable that they could unconsciously vary their speech style or voice which might causing unnatural glitches during concatenation of speech units sequences varying in voice colour or temp; thus, before the recording phases start, the speakers should experience three stages (warming stage, tune-training stage, and tune-checking stage) to ensure the quality of the recordings [24].

Checking module will be taken after the recording phase was done. [24] reported that “pluggable check module architecture” were equipped on the special environment that stated previously was used for checking the recorded audio data and rejecting them if the signals did not match the expected conditions.

[24] also reported that two annotation methods were used after the checking module was accomplished. The first annotation method used was orthographic annotation by using the software – Transcriber with assistance of two skilled annotators. During orthographic annotation, two phases were done: 1st phase – transcribe of all recordings; 2nd phase – revise the transcription and do correction if found error [24]. Phonetic annotation was the second method used, it was done in fully automated way based on the acoustic signals and revised annotations of the recordings; then, HMM-based speech segmentation process was used to supplement every recorded speech signal with predictions of boundaries between phone [24].

By highlighting the weaknesses and the strengths of this Czech corpora construction, they could be summarized as below:

Weaknesses:

Before the recording started, this corpus construction method needs to build a special environment to record and manage the recording which is time-wasting and costly. Besides, during annotation of the recording, skilled annotators were required to transcribe although third party software was involved; thus, this corpus construction was demand high cost and expertise.

Strengths:

The built of special environment could decrease the workloads of recording individuals as it managed the recording phases and handle the checking module after recording phases ended. Besides, two annotations methods were used during the corpora construction for maintaining the quality and accuracy of the recordings.

2.3 Similar Language Corpus Construction Review

2.3.1 Singapore Hokkien Corpus Construction

[25] reported that Singapore Hokkien did not have speech corpora during their effort in building limited-vocabulary speech recognition for conversational voice agent in Singapore Hokkien. They also pointed out that Singapore Hokkien is much different than Mandarin and Minnan so that these languages of corpora could not be directly used as it borrows words from other languages [25].

During the construction of corpora, [25] proposed 37 token phoneset with 19 consonants and 19 vowels for Singapore Hokkien but they abandoned tone markings for speeding up lexicon construction. Furthermore, in construction of lexical resources for Singapore Hokkien, a dictionary for common words has built which each lexical entry includes: 1. Romanized Orthography, 2. Equivalent English Meaning, 3. Equivalent Chinese Meaning, and 4. Phonetic Spelling [25]. ASCII phoneset symbols was used to concatenate with Romanized orthography for each phoneme in a syllable [25]. In order to expand lexicon, two methods were used; 1. Expanding from lists of English lexical entries; 2. Transcribe elicited natural spoken translations of English to Hokkien sentences from Singapore Hokkien speakers following with new lexical entries mining.

In speech collection, two phases were performed by [25]. First phase was conducted by generation and collection of 300 sentences 100 of them were obtained from Chinese language newspapers, another 100 of them were obtained from generation of date/time and sequence grammar, and the last 100 of them were elicited naturally from day-to-day uses of language in Singapore by imagination; in this phase language experts were recruited in assistance of Singapore Hokkien transliteration [25]. In the next phase, 52 speakers aged between 18 and 55 were recruited for recording session of the sentences and this phase collected 44.6 hours of speeches [25].

By highlighting the weaknesses and the strengths of this Singapore Hokkien corpora construction, they could be summarized as below:

Weaknesses:

Singapore Hokkien as a tonal language most likely required tone marking for getting more accurate results. However, this corpora construction did not perform it, this

might affect the result of the orthography romanization. Moreover, language experts were involved in this construction which increase the cost.

Strengths:

The concatenation of ASCII phoneset symbols with Romanized Orthography for each phenome in syllable might give better results on the romanization of the language. Transliteration processes performed by the language experts would give the better results as accuracy would increase by excluding the worse cases during the recording sessions.

2.3.2 Minnan Child Speech Corpus Construction

[26] reported that data collection for the Minnan Corpus took around three years between August 1997 and July 2000 under assistance and supports from National Science Council in Taiwan. As this corpus is mainly for children, participants in the speech recording were aged between 1 year 2 months and approximately 6 years old [26]. The recording sessions were done by visiting the participants' home regularly every two/three weeks and each recording sessions took 40 – 60 minutes; recorded sound files were separated according to the recording session [26].

Transcription of the sound files were into text files using orthographic transcription and phonetic transcription in International Phonetic Alphabet (IPA) [26]. [26] also proposed that under orthographic transcription, logographic orthography and spelling-based romanization system for Minnan were used; hence, the transcription of the texts was divided into Hanzi (Chinese Character) text file and Minnan text file. Orthographic transcription in Chinese Characters had performed first as this written form was the closest to majority of the native speakers' intuition [26]. Minnan do not have conventionalized orthography which Mandarin has; thus, there are no consistent way to write several words in Minnan [26]. Therefore, 7 Minnan dictionaries were consulted during the transcription process [26]. After the previous transcription was done, orthographic transcription in Minnan Pinyin was continued by selecting Taiwan Southern Min Phonetic Alphabetic as Minan Pinyin System [26]. [26] had used part of speech (POS) annotations and discourse annotations in construction process.

CHAPTER 2

Lexical Bank was constructed during the corpus construction, and it contains logographic orthography, spelling-based orthography, part-of-speech, and alternative forms [26].

By highlighting the weaknesses and the strengths of this Minnan Child Speech Corpora construction, they could be summarized as below:

Strengths:

This corpora construction method would be the most suitable method among the reviewed methods as 2 orthography transcriptions were done for separating the character and the romanization effectively with higher accuracy.

CHAPTER 3 System Model

3.1 Research Flow Chart

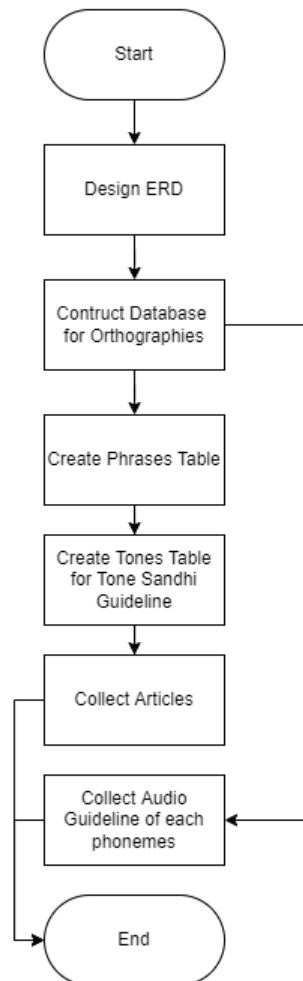


Figure 3.1 Flow Chart of Research

In Figure 3.1, this research started with ER-Diagram design for the Penang Hokkien Database that store the Phonemes, Graphemes, and the Mapped Phonemes and Graphemes. Next, Phrases table was created that gather phrases of Penang Hokkien from Speak Hokkien Campaign Facebook Page. After that, tones table was created by standardizing the researched tones and tone sandhi. Articles that have high mutual intelligibility with Penang Hokkien were collected was the next steps and continued with audio guideline collection for phonemes from the database. The research progress was ended with end of audio guideline collection.

3.2 Penang Hokkien Database Design

3.2.1 ER Diagram

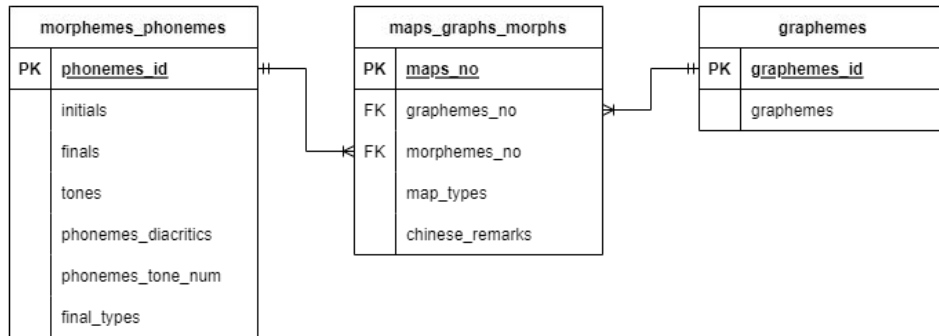


Figure 3.2 Penang Hokkien ER Diagram

The ER diagram in the Figure 3.2.1-1 shows relations between three tables: morphemes_phonemes, maps_graphs_morphs, and graphemes. The maps_graphs_morphs is the table that solves many-to-many relations between morphemes_phonemes and graphemes tables, and it also act as the mapping table for both tables. Foreign key “graphemes_no” in maps_graphs_morphs table is referencing graphemes_id in graphemes table and morphemes_no table is referencing morphemes_id in morphemes table.

3.2.2 Data Dictionary

morphemes_phonemes			
Field Name	Data Type	Nullable	Description
phonemes_id	Integer	N	Unique ID number for each phoneme of initial and final combination.
initials	Text	N	Initial of Penang Hokkien Romanized phoneme.
finals	Text	N	Finals of Penang Hokkien Romanized phoneme.
tones	Integer	N	Tones numbers of Penang Hokkien.
phonemes_diacritics	Text	N	Romanized phonemes with diacritics symbol on the Latin alphabets. Contain Special

			characters.
phonemes_tone_num	Text	N	Romanized phonemes with tone number.
final_types	Text	N	Abbreviation of the final types.

Table 3.1 Data Dictionary for morphemes_phonemes

graphemes			
Field Name	Data Type	Nullable	Description
graphemes_id	Integer	N	Unique ID number for each grapheme.
graphemes	Text	N	Chinese Japanese Korean (CJK) Unified Ideographs Extensions texts.

Table 3.2 Data Dictionary for graphemes

maps_graphs_morphs			
Field Name	Data Type	Nullable	Description
maps_no	Integer	N	Unique ID number for each mapped graphemes, morphemes, and map types.
graphemes_no	Integer	N	ID number for each grapheme.
morphemes_no	Integer	N	ID number for each morpheme.
map_types	Text	N	Abbreviation of the map types.
chinese_remarks	Text	N	Abbreviation of the map types in Chinese.

Table 3.3 Data Dictionary for maps_graphs_morphs

3.3 Penang Hokkien Database Construction

3.3.1 Database Construction Flow Chart

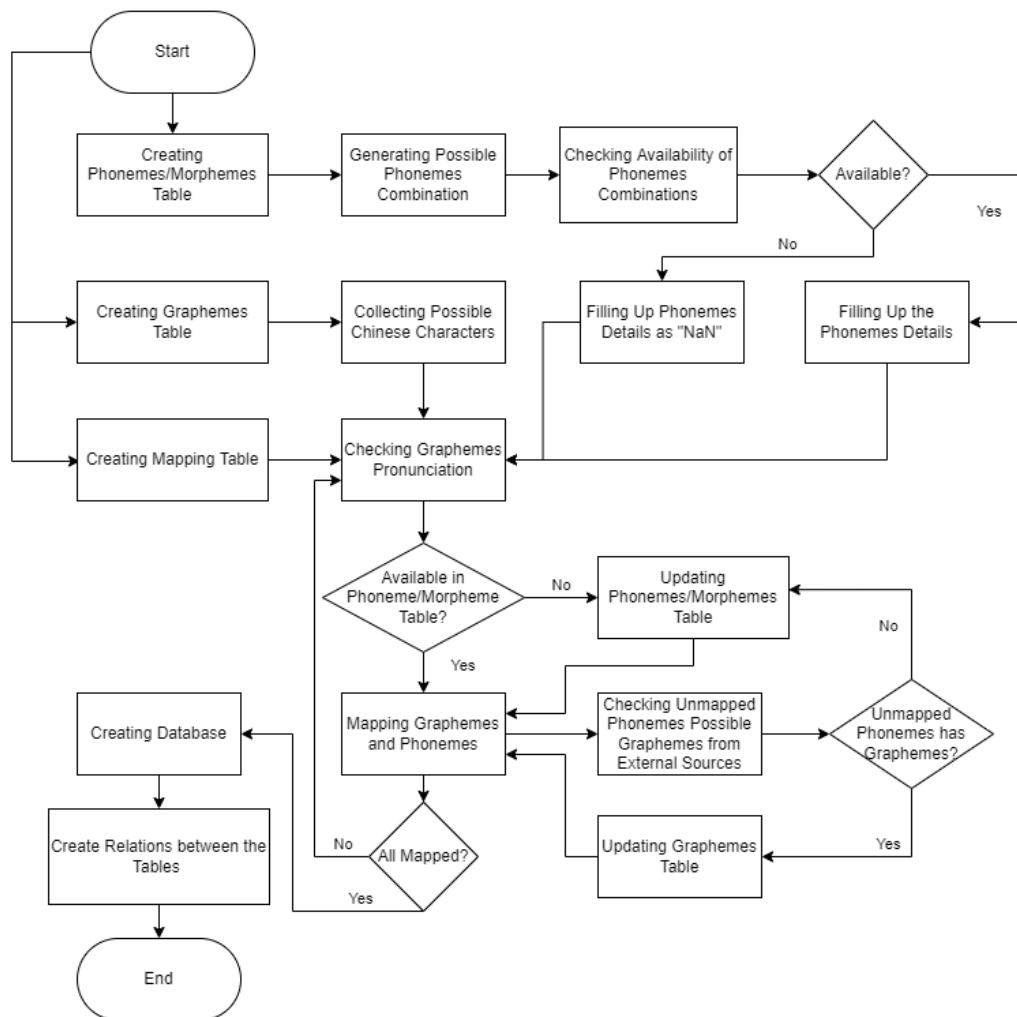


Figure 3.3 Flow Chart of Database Construction

In Figure 3.3.1-1, the database construction process for this research started by creating multiple tables that stores different types of data. The first table is Phonemes/Morphemes Table, where after this table was created, possible phonemes combination will be generated. After that, the availability of the combinations was checked with dictionaries by manually. The phonemes details were filled up in the table if the phonemes combinations are available, else all the details filled up became “NaN”. From the second created table – Graphemes, literally stores graphemes of the Penang Hokkien which are Chinese Characters. Possible Chinese Characters were collected and stored in the tables.

The third created table is mapping table that maps the graphemes and phonemes. After Mapping Table was created, the pronunciations of graphemes from the Graphemes Table were checked using dictionaries by manually. If the pronunciations of the graphemes were available in the Phonemes/Morphemes Table, the graphemes and phonemes were mapped by filling up the information needed in the mapping table; else, the Phonemes/Morphemes Table was updated by adding the phonemes that were checked as unavailable. After that, mapping processes were continued.

During the mapping process, when there were phonemes that were unmapped, the phonemes were checked in dictionaries to investigate the missed-out graphemes during the data collection and they were updated in the Graphemes Table when the graphemes were available, else, the phonemes that checked has no graphemes were updated in the Phonemes/Morphemes Table with all the details filled up became “NaN”. The mapping processes were stopped until all the graphemes and phonemes were all mapped, then a database was created. The database construction progress ended with created relations between the 3 tables that stated in the Figure 3.3.1-1.

3.4 Software and Dictionaries Used

1. Amoy Hokkien Dictionary by Scottish Missionary - Carstairs Douglas [27]
2. Taiwanese Hokkien Dictionary by Ministry of Education, R.O.C. [28]
3. zdic.net Online Chinese Dictionary
4. Microsoft Excel
5. MySQL
6. BabelStone Han [29]

CHAPTER 4 EXPERIMENT

4.1 Phonemes of Penang Hokkien

There are total of 118 phonemes in Penang Hokkien Language, and it is categorized into initials and finals. Finals phonemes are further categorized into normal, nasals, and stops.

4.1.1 Initials

Table 4.1 Penang Hokkien Initials Phonemes

p-	l-	j-
h-	k-	d-
ph-	kh-	f-
b-	g-	r-
m-	ng-	sh-
t-	ts-	w-
th-	tsh-	y-
n-	s-	NaN

- Total Initials = 23 + 1, as there are phonemes without initials
- Loanwords Initials (Highlighted in Grey) = 8

4.1.2 Finals

Table 4.2 Penang Hokkien Finals Phonemes

Normal			Nasals				Stops		
-a	-y	-oi	-m	-uinn	-eng	-ern	-ah	-ioh	-at
-e	-ai	-ou	-ann	-iaunn	-eeng	-in	-eh	-uah	-et
-ee	-au	-ua	-enn	-uainn	-ing	-on	-eeh	-ueh	-ert
-er	-ei	-ue	-inn	-am	-ong	-un	-erh	-ak	-it
-i	-ia	-ui	-onnn	-em	-ung	-yn	-ih	-ek	-ot
-o	-io	-eoi	-ainn	-im	-iang	-ian	-oh	-eek	-ut
-oo	-ioo	-iau	-iann	-om	-iong	-uan	-ooh	-ik	-iat
-u	-iu	-uai	-ionnn	-um	-uang	-ng	-uh	-ok	-uat
			-oinn	-iam	-an		-aih	-uk	-ap

	-uann	-ang	-en		-auh	-iak	-ip
					-iah	-iok	-iap

- Total Finals = 95
 - Normal Finals = 24
 - Nasals Finals = 38
 - Stops Finals = 33
- Total Loanwords Finals (Highlighted in Grey) = 16
 - Normal Loanwords Finals = 2
 - Nasal Loanwords Finals = 10
 - Stop Loanwords Finals = 4

4.2 Tones of Penang Hokkien

4.2.1 Nine Tones of Penang Hokkien

Table 4.3 Tones and Diacritic Marks of Penang Hokkien

Tone	1 st	2 nd	3 rd	4 th	5 th	6 th	7 th	8 th	9 th
Diacritic	○	◌́	◌̀	○	◌̂	◌̃	◌̄	◌̇	◌̈́

Table 4.4 Tone Naming

	Level	Rising	Departing	Entering
Dark	1	2	3	4
Light	5	6	7	8
				9

As mentioned in previous chapters, there are nine tones in Penang Hokkien due to increase of loanwords. By referring to Table 4.3 and Table 4.4, First tone is Dark Level tone with no diacritic mark, second tone is a Dark Rising tone with Acute diacritic mark, third tone is Dark Departing tone with Grave diacritic mark, fourth tone named as Dark Entering tone with no diacritic mark, fifth tone named as Light Level tone with Circumflex diacritic mark, sixth tone is Light Rising tone with Caron diacritic mark, seventh tone named as Light Departing tone with Macron diacritic mark, eighth tone is Light Entering tone with Vertical Line Above diacritic mark, and the ninth tone is High Entering tone with Double Acute mark [30]. From both Table

4.3 and Table 4.4, the first tone and fourth tone do not have diacritic marks. In order to differentiate the first and fourth tones, Table 4.4 had explained all, fourth tone is an entering tone so it only appears on stop finals which can be referred in Table 4.2. Moreover, the sixth tone and ninth tone which are in bold, and italic means they are tones for loanwords. Sixth tone from Table 4.3, sixth tone that highlighted in black means, this tone is not applicable in Penang Hokkien where during the research, there were no phrases found with sixth tones.

4.2.2 Tones Letter and Tones Contour

Table 4.5 Tones Letter and Tones Contour Table

Tones Number	Tones Name	Tones Letter			Contours Number		
1	Dark Level	ㄐ		ㄒ	33		44
2	Dark Rising	ㄐ	ㄑ	ㄒ	445	53	51
3	Dark Departing	ㄐ		ㄒ	21		22
4	Dark Entering	ㄐ?			3		
5	Light Level	ㄐ		ㄒ	23		24
6	Light Rising	-			-		
7	Light Departing	ㄐ			21		
8	Light Entering	ㄐ?			4		
9	High Entering	ㄐ?			32		

The Table 4.5 shows each tone number with their corresponded tone name, tone letter and contour numbers. Tone letters are represented by using IPA Tone Letter invented by Yuan Ren Chao [31]. According to research results of [32] in late 2013, older generations have different tones as compared with younger generations of Penang Hokkien Speakers, [10] has summarized the tone sandhi of the tones, but the tone letters and tone contours for each tone is not specify. The details on different tone contours of the 1st, 2nd, 3rd and 4th tones will be discussed on 4.2.5 Tone Sandhi.

4.2.3 Tones Graph

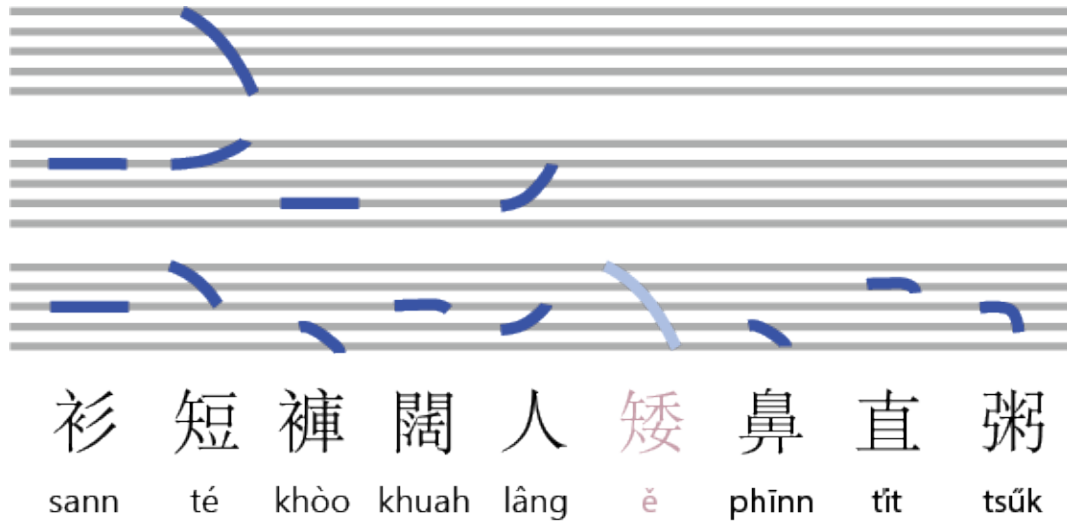


Figure 4.1 Tones Graph with examples of Graphemes and Romanizations

The Figure 4.1 shows graph marking on 3 five-line staff for each of the nine tones in ascending order from first tone to ninth tone. First two five-line staffs are the alternate tone contours of the third. Detailed explanation will be explained in 4.2.5 Tone Sandhi. Word examples are presented at below of the five-line staff with its pronunciation orthography. The translation of each word from left to right are clothes, short, pants, wide, person, short, nose, straight, congee.

4.2.4 Tones Marking Rules

Table 4.6 Tone Marking Priority

a	>	e	>	o	>	y	>	i
								u
High Priority				→				Low Priority

Table 4.6 shows priority of tone marking with high priority start from left to low priority end to right. By referring the rules in Penang Hokkien Spelling System created by Hokkien Association of Penang in the appendix, ‘i’ and ‘u’ have same priority but if both of these vowels appeared in the same final like ‘iu’ and ‘ui’, the last vowel will be tone marked. The rules from the appendix also stressed that if ‘oo’, ‘ee’ and ‘ng’ are finals, first vowels will be tone marked.

4.2.5 Tone Sandhi

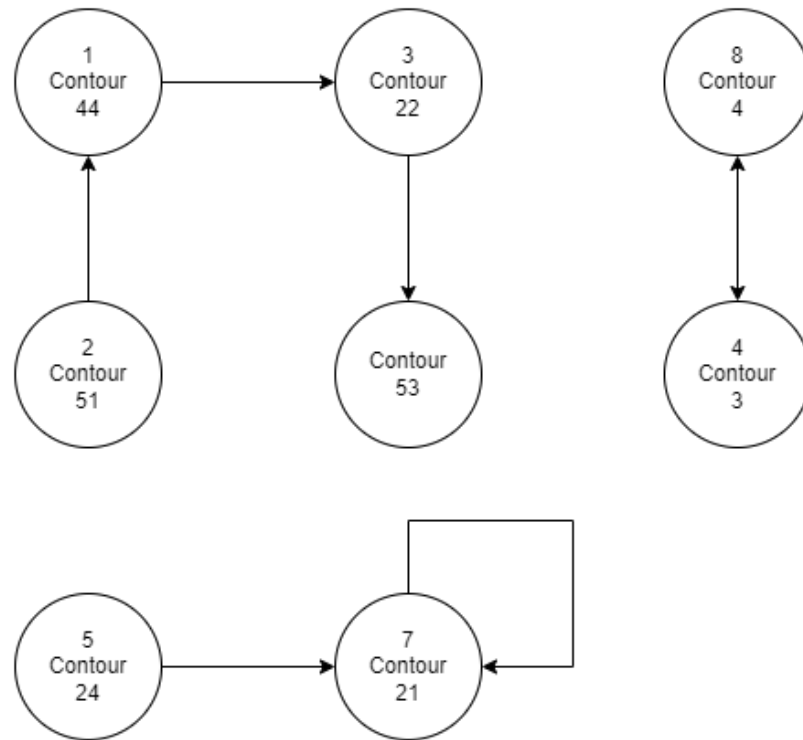


Figure 4.2 60 to 80 Age Group Penang Hokkien Speaker Tone Sandhi

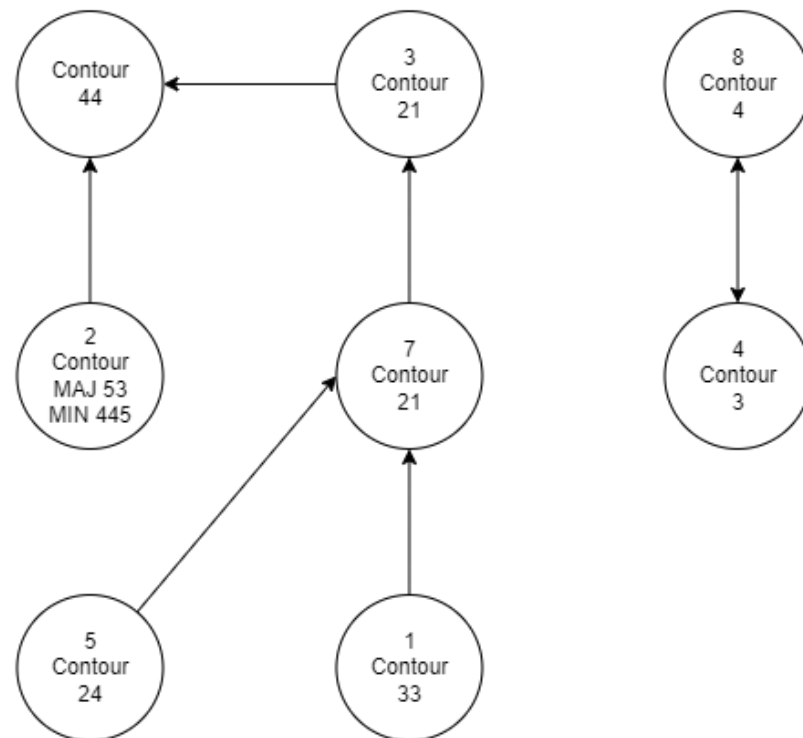


Figure 4.3 40 to 60 Age Group Penang Hokkien Speaker Tone Sandhi

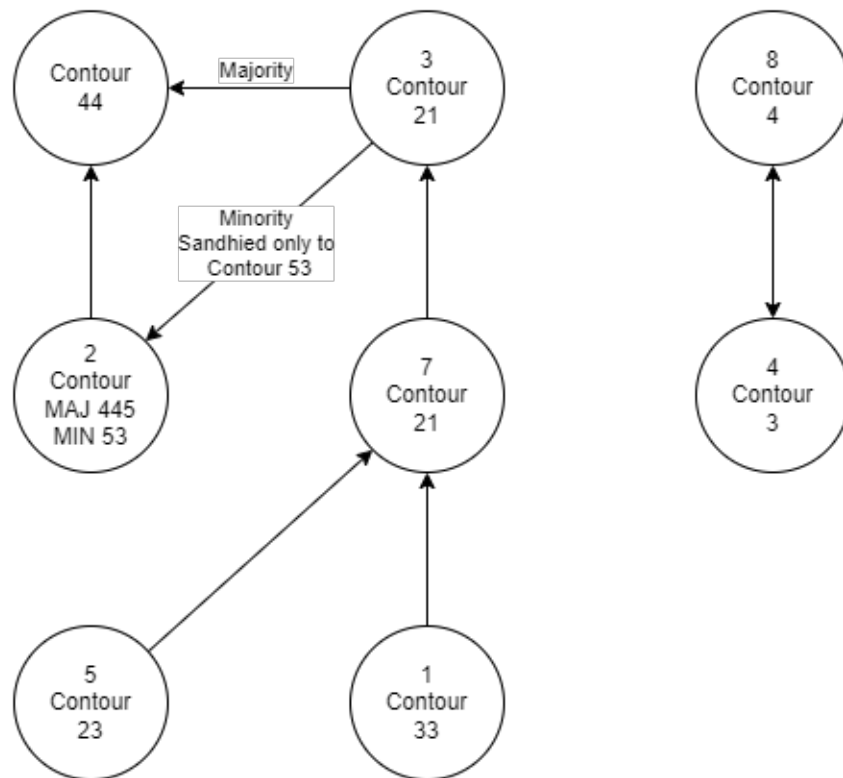


Figure 4.4 20 to 40 Age Group Penang Hokkien Speaker Tone Sandhi

By conducting the research on the Penang Hokkien tone sandhi towards [32], Penang Hokkien users mainly categorized into 60 to 80, 40 to 80 and 20 to 40 age groups as of 2013. Figure 4.2, 4.3, 4.4, are the tone sandhi that summarized from [32] research results. By referring to the Figure 4.2, age group 60 to 80 has very different tone sandhi than the younger generation. The most obvious one is the tone 3 and tone 7, tone 3 is sandhied to Contour 53 which is not tone contour any original tones of 60 to 80 group age, but it became tone contour of tone 2 of younger generations as referring to Figure 4.3 and 4.4; for tone 7, it is not sandhied, so it remains at 7th tone. This gave a foreshadowing of younger generation tone sandhi of tone 7 to tone 3 where both tones have same tone contour which is Contour 21. For the ton sandhi from tone 2 (Contour 51) to tone 1(Contour 44) also gave a foreshadowing of younger generation tone sandhi from tone 2 to Contour 44 where their contour of tone 1 is Contour 33. Tone sandhi of tone 4 and tone 8 are same throughout the generations.

By referring to Figure 4.3, tone 2 of age group 40 to 60 has two tone contours which are Contour 53 for majorities and Contour 445 for minorities. The tone 7 is sandhied to tone 3 which has same tone contour as stated previously and the tone 3

sandhied to Contour 44 which is the tone 1 of the older generation but Contour 33 born in younger generations.

In Figure 4.4, there are distinction from previous older generations where tone 3 can be sandhied in two ways, tone 3 to Contour 44 for majorities and tone 3 to tone 2 with Contour 5 for minorities. Apart from that, the group age 20 to 40 have two tone contours for tone 2, Contour 445 for majorities and Contour 53 for minorities. Moreover, this group age has their tone 5 contour changed from Contour 24 to Contour 23. These explained why in the previous sub chapters tone contours and tone letters for tone 1, tone2, tone 3 and tone 5 have more than one and explained how [10] standardized the Penang Hokkien tone sandhi. The standardization for the Penang Hokkien tone sandhi is visualized in Figure 4.5 below.

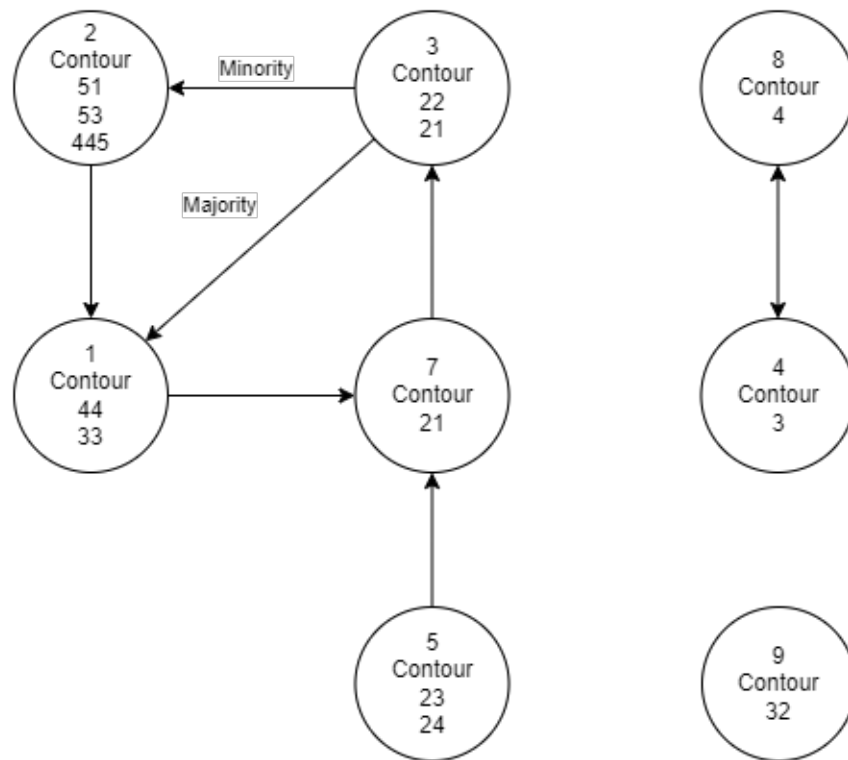


Figure 4.5 Standardized Penang Hokkien Tone Sandhi with Tone Contours

4.2.6 Tone Table Creation as Guidelines for Future Research

A tone table that referred to Figure 4.5 was created in Workbench Excel file during the research.

The attributes in the tables are listed at below with description:

1. tones_id: Tone number of each tone.
2. tones_name_zh: Tone Name in Chinese Characters.
3. tones_name_en: Tones Name in English.
4. tones_letter: IPA Tone letters for each tone.
5. alt_tones_letter1: First alternate tone letter for each tone.
6. alt_tones_letter2: Second alternate tone letter for each tone.
7. contours_no: Tone contours for each tone.
8. alt_contours_no1: First alternate tone contours for each tone.
9. alt_contours_no2: Second alternate tone contours for each tone.
10. sdh_tones_id: Sandhied tone number for each tone.
11. alt_sdh_tones_id: Alternate sandhied tone number for each tone.

4.3 Database Construction for Penang Hokkien Morphemes and Phonemes

4.3.1 Morphemes_Phonemes Table Creation

Based on the previous sub chapters, 4.1.1 Initials and 4.1.2 Finals, and 4.2.1 Nine Tones of Penang Hokkien, possible combination of initials, finals and tones can be calculated.

The equation is very simple, all we need is just multiply them as below:

$$I \times F \times T = \text{Possible Phonemes}$$

I = Total initials

F = Total Finals

T = Total Tones

Total Initials = 24 (Included NaN)

Total Finals = 95

Total Tones = 9

$24 * 95 * 9 = 20520$

Possible Phonemes Combination = **20520**

Table 4.7 Sample Phoneme_Morpheme Table from Excel

No	Initial	Final	Tones	Phoneme_ Diacritics	Phoneme_Tone Num	Final_ Types
1	p	a	1	pa	pa1	OPS
2			2	pá	pa2	OPS
.....
20520	NaN	iap	9	NaN	NaN	NaN

Table 4.7 shows sample Phoneme_Morpheme table taken from the Excel file. “No” in the table above is represented as the table primary key in the database where the values are unique. These values were created by using simple function “=ROW-1”.

CHAPTER 4

The initials were taken from Table 4.1 which starts with “p” and ends with “NaN” where “NaN” is to represent the phonemes that only have finals. Each initial has 1*95*9 records, total of 855 records. The finals were taken from Table 4.2 which starts by “a” and ends with “iap” for each initial. The tones in the table are represented in integer from 1 to 9. Values in Phoneme_Diacritics are represented by combination of initials and finals with the diacritics as shown in Table 4.3 while values in Phoneme_ToneNum are represented by combining initials and finals as well as the tone number that shown in sample from Excel in Table 4.7 above. These two attributes have value “NaN” to represent records that have no combination of initials, finals and tones. Final_Types is the attributes that restrict the value entered to ensure the consistency of value types.

The value that allowed in Final_Types are:

1. OPS – Open Syllables. Eg: Finals without consonants except -er to differentiate with -e and -ee.
2. NAC – Nasal Consonants. Eg: Finals end with -n and -m.
3. STC – Stop Consonants. Eg: Finals end with -p, -t, -k, -h.
4. NAV – Nasal Vowels. Eg: Finals end with mn, -ann, -enn.
5. NaN – Used this when Phonemes_Diacritics and Phonemes_ToneNum is NaN.

The possible combinations of the data were determined by checking with the two dictionaries mentioned in 3.4 and phrases collected from Speak Hokkien Campaign Facebook Posts as well as Wiktionary that the phonemes were rechecked with the dictionaries to ensure the data reliability.



Figure 4.6 Phonemes and Tones Indexing of Online Taiwanese Dictionary

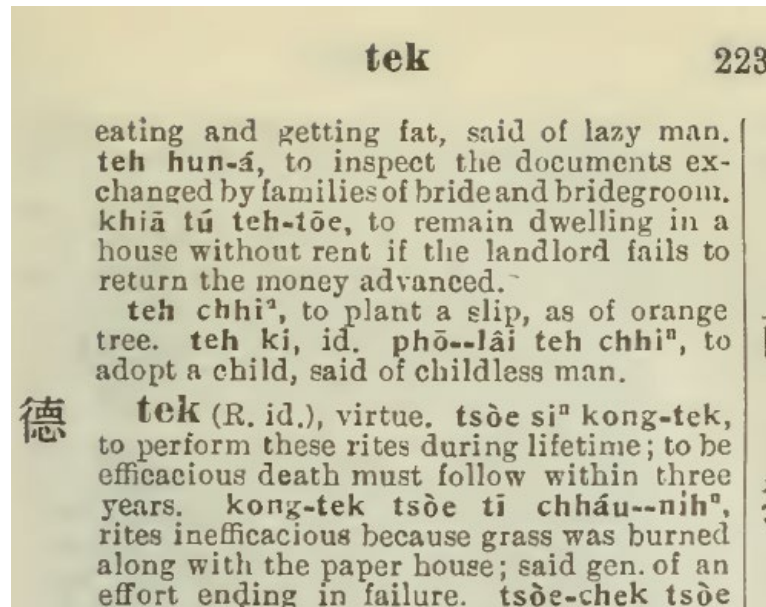


Figure 4.7 Amoy Hokkien Dictionary Page 223 Cropped Image

By referring to Figure 4.6, the Online Taiwanese Hokkien Dictionary provides indexing by initials, finals, and tones. The listed-out indexes of possible combination of initials, finals and tones were taken as data in the Morphemes_Phonemes Table. There are distinctions of finals used in Penang Hokkien and Taiwanese Hokkien, “-ik” and “-ek” are the examples from lots of finals distinctions in both languages. Taiwanese Hokkien uses “-ik” but Penang Hokkien and Amoy Hokkien use “-ek”[10], [27], [28]. In order to utilize both dictionaries, a basic knowledge of Tâi-lô and Pêh-oē-jī are needed to distinct both romanization system. Both systems look similar, but the finals vowels used are different and the diacritics marking priority on the vowels. As an example, for final vowel, Tâi-lô prefer “-ua” but Pêh-oē-jī prefer “-oa”. Moreover, the nasal vowels are Romanised differently, Tâi-lô prefer “-(vowel)nn” but Pêh-oē-jī prefer “-(vowel)ṃ”. As an instance of distinction of both romanization system, when a diphthong only exists “i” and “u”, Pêh-oē-jī will choose to mark on “u” but Tâi-lô will choose to mark on last vowel which can be referred in subchapter 4.2.4 as Penang Hokkien use Tâi-lô as romanization system.

4.3.2 Graphemes Table Creation

The Graphemes or Chinese Characters were taken from dictionaries that mentioned in previous subchapters which are Online Taiwanese Hokkien Dictionary by Ministry of Education Republic of China and Amoy Hokkien Dictionary by Scottish Missionary,

Carstairs Douglas, to create the Graphemes Table. Apart from that, Wiktionary on Southern Min Languages was also taken as sources with evaluation by checking with two dictionaries as mentioned above. Moreover, the graphemes are also taken from Speak Hokkien Campaign Facebook posts and from its previous Facebook page “Write Penang Hokkien” Facebook posts.

Table 4.8 Sample Graphemes Table from Excel

No	Grapheme
1	一
.....
4137	攏
4138	跔

Table 4.8 is a sample from the excel file that contains the graphemes or CJK Unified Ideographs. “No” in the table above is also represented as the table primary key in the database where the values are unique which were also created by using simple excel function “=ROW-1”. The Graphemes on each record are unique with the usage of data validation “=COUNTIF(INDIRECT("Graphemes[Graphemes]"),B2)<2” which prevents redundancies of the graphemes.

4.3.3 Mapping Table Creation

After creation of the Phonemes_Morphemes Table and Graphemes Table, Mapping Table was created. The table meant to map the Phonemes/Morphemes with Graphemes and classify them. A grapheme could have more than one phoneme or morpheme, so the in the previous subchapter, this mapping table is created to solve the many-to-many relations between Phonemes_Morphemes table and Graphemes Table.

Table 4.9 Sample Mapping Table from Excel

No	Grapheme_No	Morpheme_No	Map_Type	Chinese_Remarks
1	1	20452	UCT	無
2	1	11906	KYM	訓
....

Table 4.9 shows a sample Mapping Table from Excel with 5 attributes. “No” same as previous tables were also using the same function to generate the values and the “Grapheme_No” was restricted to input values “No” from Reference_Table_Grapheme Table that duplicated from Graphemes Table to save as backup during the evaluation that need to edit the original table. While “Morpheme_No” was restricted to input values “No” from Reference_Table_Morpheme Table that duplicated from Morpheme_Phonemes Table that excluding “NaN” Records to save as backup during the evaluation that need to edit the original table.

The data validation for both attributes are shown below:

1. Grapheme_No:

=Reference_Table_Morpheme!\$A\$2:\$A\$2111

2. Morpheme_No:

=Reference_Table_Grapheme!\$A\$2:\$B\$4251

“Map_Type” that shown in the Table 4.9 restricts the values that only can be chosen from the selection list:

- **LTR (Literary Pronunciation):** Literary pronunciations were brought from Tang Dynasty when imperial examination system introduced Qieyun phonology system into Fujian as well as mother-tongue brought from Central Plain of China during that dynasty as previously mentioned in the subchapter 1.5.2 Min Languages Origin [15]. According to [33], the Literary pronunciations of Southern Min are found by using Fanqie method towards Middle Chinese pronunciations that frequently used in writing during Middle

Chinese. The pronunciations could also be found in orally and it corresponded neatly with Middle Chinese pronunciations.

- **CLQ (Colloquial Pronunciation):** Colloquial pronunciations were also brought from Central Plains during the Tang Dynasty and can also be founded using Fanqie methods towards the Middle Chinese pronunciations [15], [33]. However, according to [33], there has differences in different phonetics level overlapping that increases the complexity level in the language and different areas have different colloquial pronunciation towards the same grapheme.
- **KYM (Kunyomi Pronunciation):** Kunyomi pronunciation is derived from Japanese which means Japanese native reading on the Chinese character that have same meaning with the Japanese native pronunciation [34]. In Southern Min, there already have similar categorization system towards the graphemes which are LTR and CLQ. However, by understanding the history of the Min Languages that mentioned in the previous subchapter, the Min language have a basic level of the lexical elements from Minyue Language and integrated with Old Wu and Old Chu Languages, and then integrated with Old Chinese Language. In Tang Dynasty, the Middle Chinese with phonology system were brought to Fujian and integrated with the language and became the current modern Min Languages. With these integrations of lots of languages, there is a huge problem in Southern Min Languages which is morphemes without graphemes. In order to solve this problem, KYM is introduced to categorize the graphemes that pronounced in natively as a temporarily solution for linguists and archeolinguists for them to trace back the actual word for the forgotten graphemes.
- **ATJ (Ateji Pronunciation):** Ateji pronunciation is also derived from Japanese which means the Chinese character is used only to represent the sound but not the meaning of the character. In Southern Min languages have lots of phonetic-loaned graphemes that are not standardized due the same problem as explained in the KYM. Hence, this categorization is introduced in this research to categorize the graphemes that used for representing the sound but not for its meaning. The categorization is introduced to solve the weak graphemes categorization of Online Taiwanese Hokkien Dictionary by Ministry of Education that suggested newly created graphemes and phonetic-

loaned graphemes are categorized into alternate words category. This categorization is problematic as the KYM graphemes are also categorized into the alternate words category by that dictionary. This will confuse the users whether the graphemes used for the morphemes are based on the graphemes meaning or based on the phonemes of the graphemes. Thence, ATJ is introduced as one of the categorization methods for the Penang Hokkien mapped graphemes.

- **UCT (Uncategorized Pronunciation):** According to [33], if there are graphemes that are categorized neither in LTR nor CLQ and only has one morpheme, those graphemes are uncategorized. Hence, the graphemes that mapped with morphemes that has no categorization are marked with UCT as these mapped graphemes could be taken from contemporary Mandarin; hence, those graphemes should remain uncategorized.

The Chinese Remarks are Penang Hokkien Translation of the abbreviations above which written in one Traditional Chinese Character for each abbreviation. Table 4.10 below shows the Penang Hokkien Translation towards the Abbreviation used in the Mapping Table.

Table 4.10 Map Types Abbreviation and Penang Hokkien Translation

Abbreviation	Penang Hokkien Translation in Tâi-lô
LTR	Bûn-thak-im
CLQ	Pêeh-thak-im
KYM	Hùn-thak-im
ATJ	Uan-jī-im
UCT	Bô-kui-luī

4.3.4 Implementation of Tables into the Database

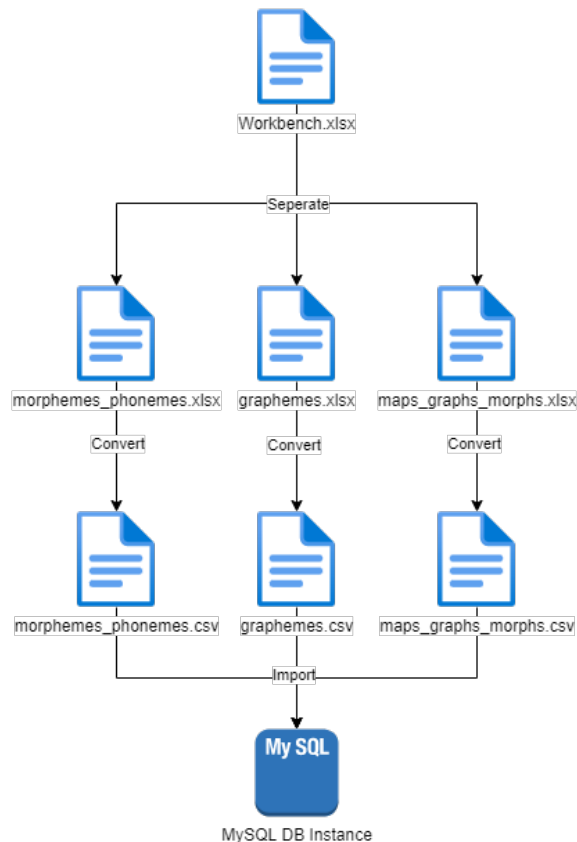


Figure 4.8 Database Tables Importing Process Diagram

By referring to Figure 4.8, the tables are needed to be separated into the different files. There are totals of 3 tables needed in the database, morphemes_phonemes, graphemes, and maps_graphs_morphs, which corresponded with Morpheme_Phonemes Table, Graphemes Table, and Mapping Table in the Workbench Excel file that gathered all needed data. The separated excel files were needed to change the attributes names and files names according to the ER Diagram design in subchapter 3.2.1. After the naming progress was done, the files were converted into the .csv file UTF-8 due to special characters in the table which the default .csv file is not supported.

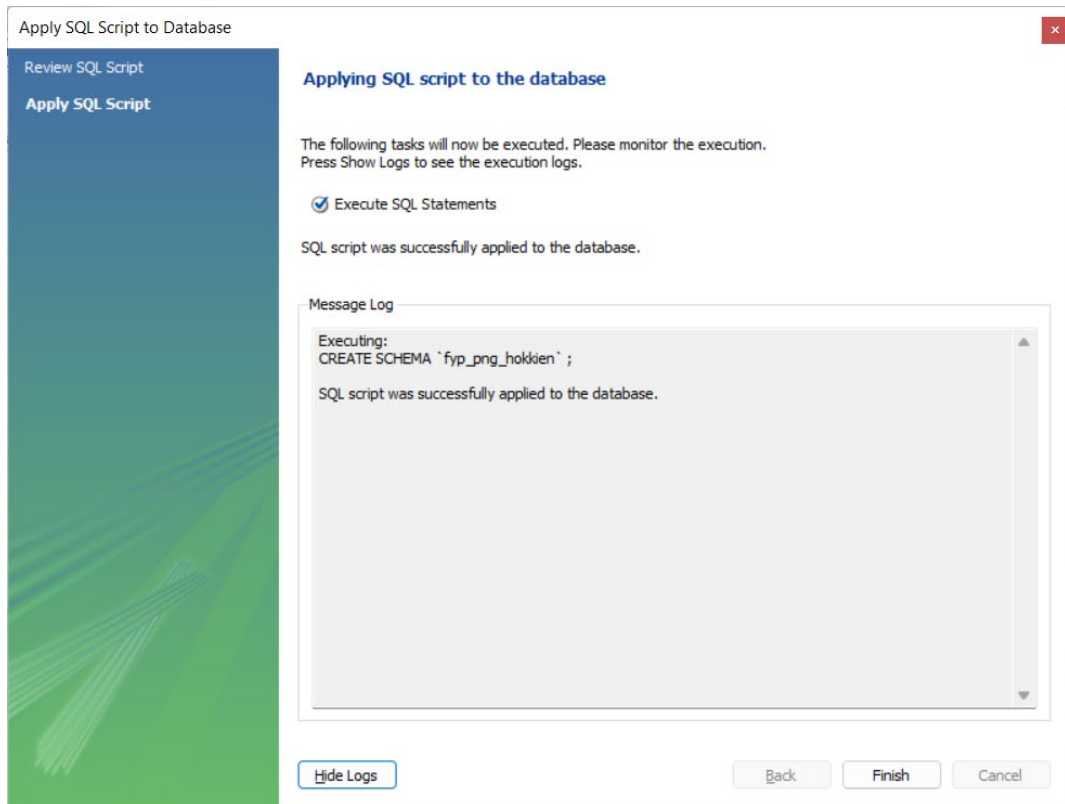


Figure 4.9 Step1: SQL Script Creation

First steps as shown in Figure 4.9, an empty SQL Script was created into the database.

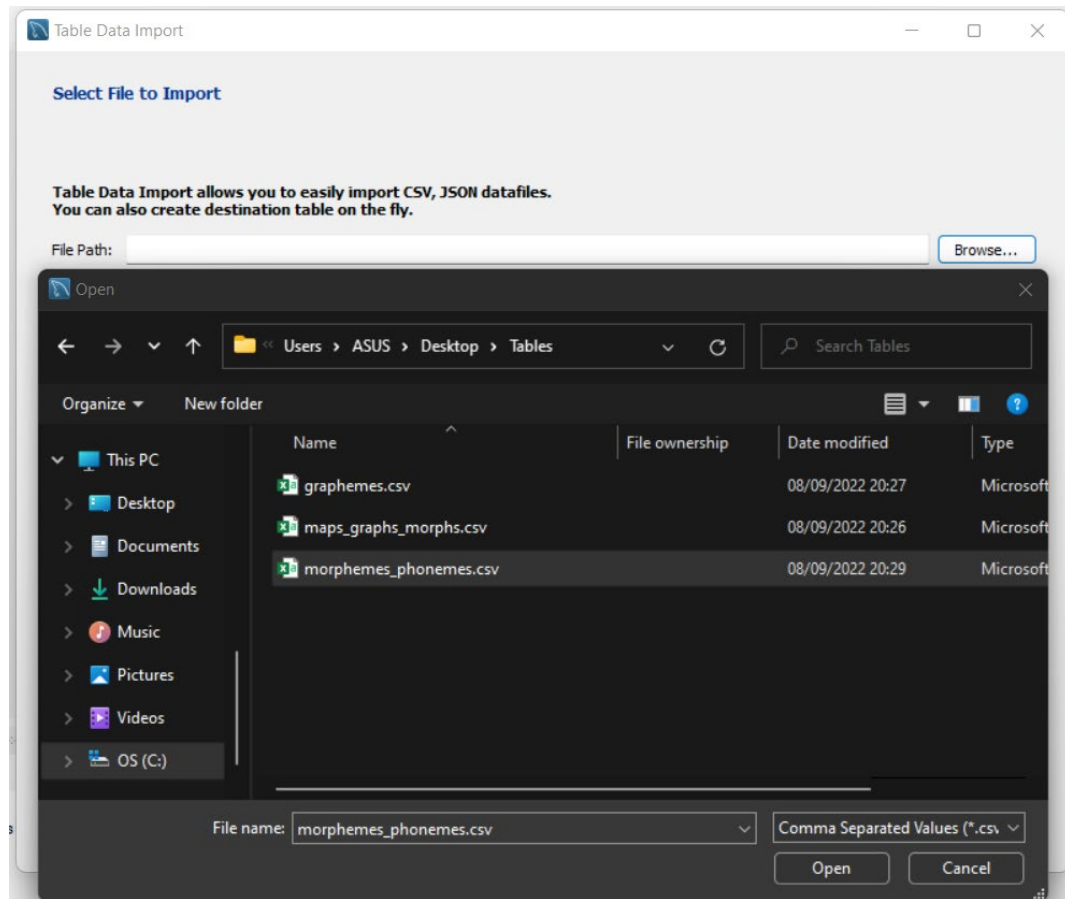


Figure 4.10 Step 2: Import Selected File

Figure 4.10 shows second steps of the implementation where file was selected to import into the database.

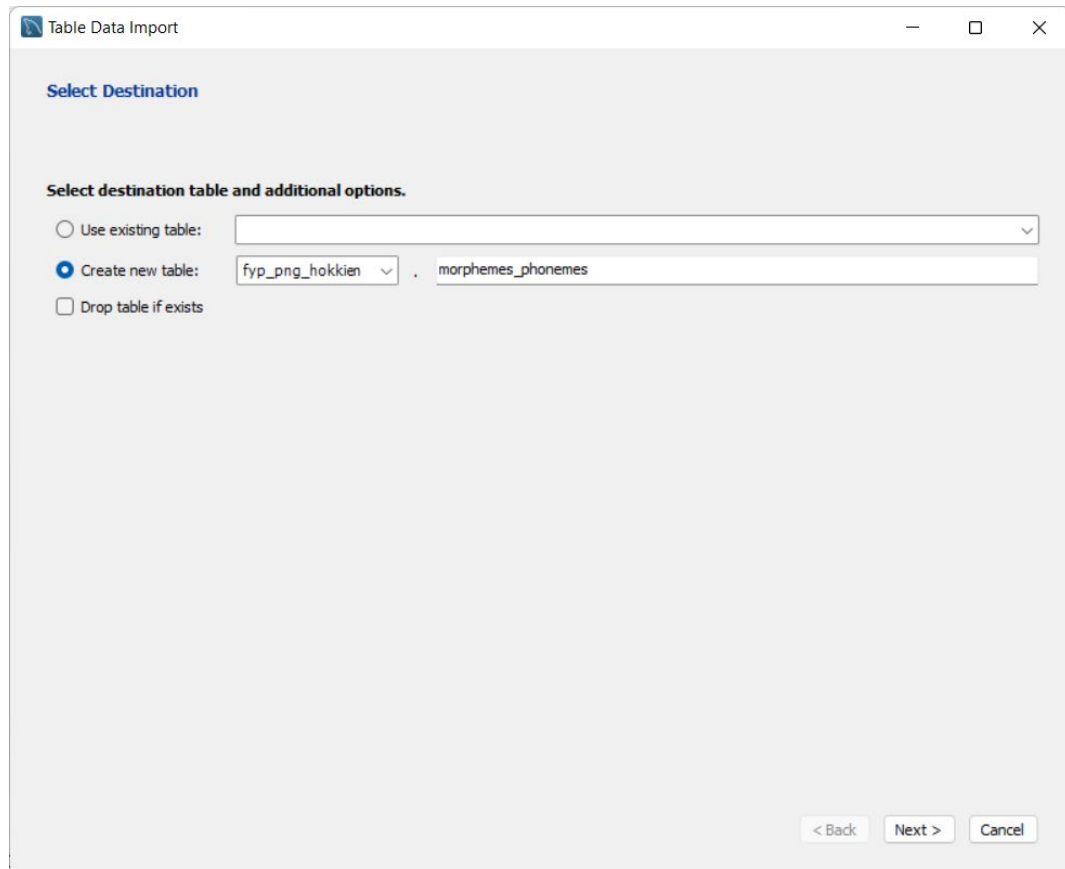


Figure 4.11 Step 3: Select Import Destination

Since the script file was newly created, the all the tables that need to import in the database were newly created. From Figure 4.11, radio button “Create new table” was selected and the import destination was selected as “fyp_png_hokkien” with table name that previously set before the table importing which made the import process easier.

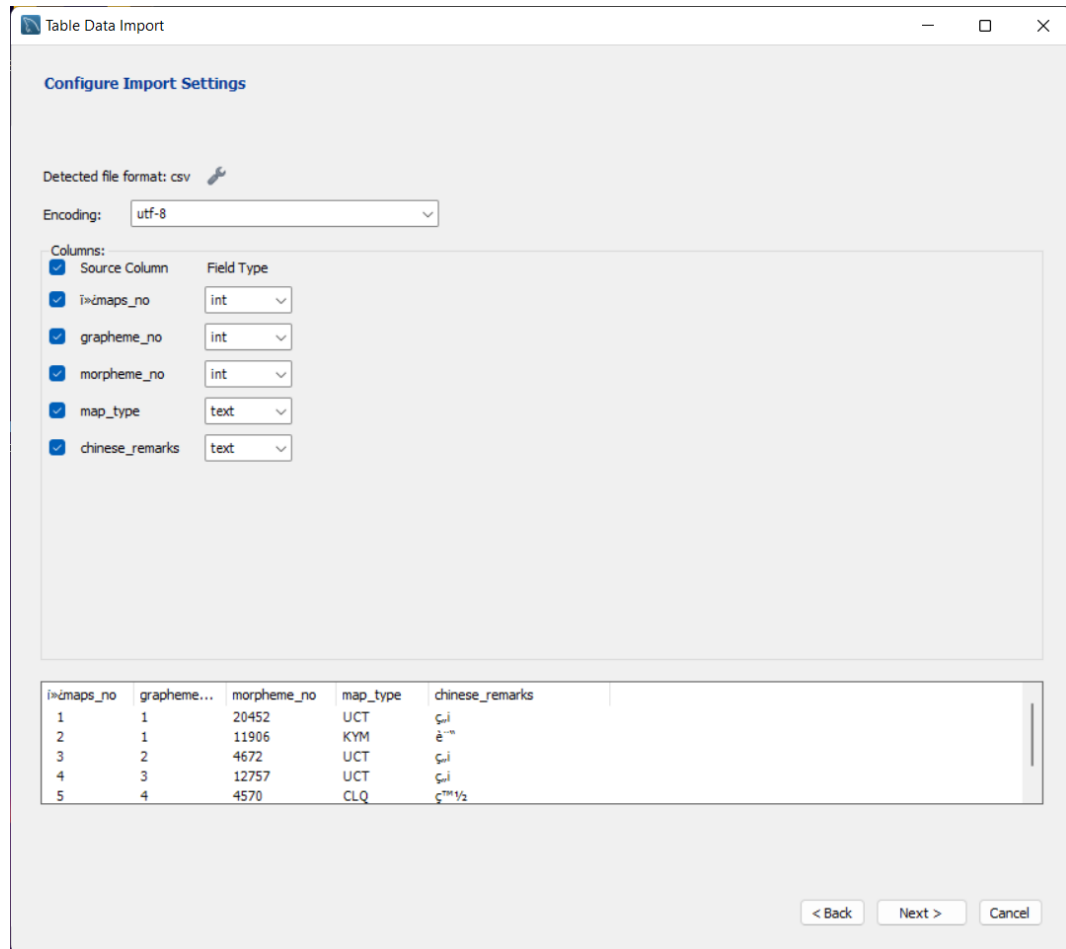


Figure 4.12 Step 4: Select Encoding Type and Attributes Field Type

Figure 4.12 shows the step 4 of the tables implementation, in this step, encoding type for the table values was selected, however, the preview showed that the chinese_remarks value which should be in Chinese characters became unreadable texts, further details are discussed in 4.3.5 Implementation Issues and Challenges. Moreover, in this step, the Field Types of the attributes were selected according to the Data Dictionaries in subchapter 3.2.2. After that, pressed next button to proceed to data import process. The implementation process is not over yet due to the errors were found. The further implementation processes will be discussed in Chapter 5 System Evaluation and Discussion.

4.3.5 Database Implementation Issues

As mentioned in previous subchapter 4.3.4, errors were encountered during the implementation process. The implementation errors snapshots were attached at below:

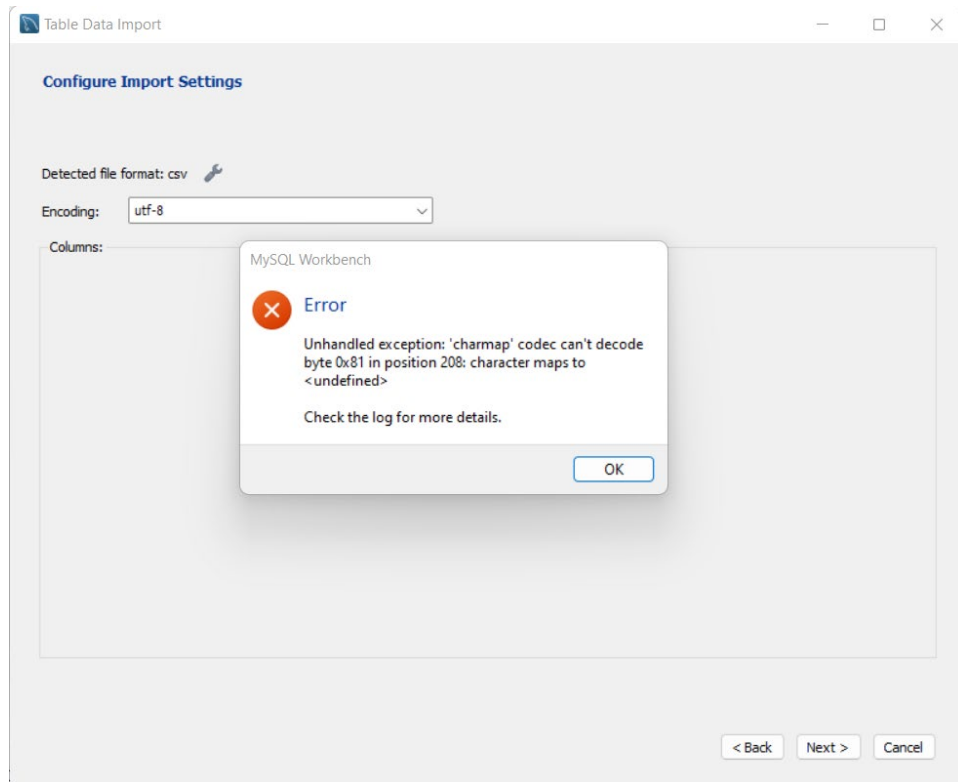


Figure 4.13 Error Snapshot 1: morphemes_phonemes.csv

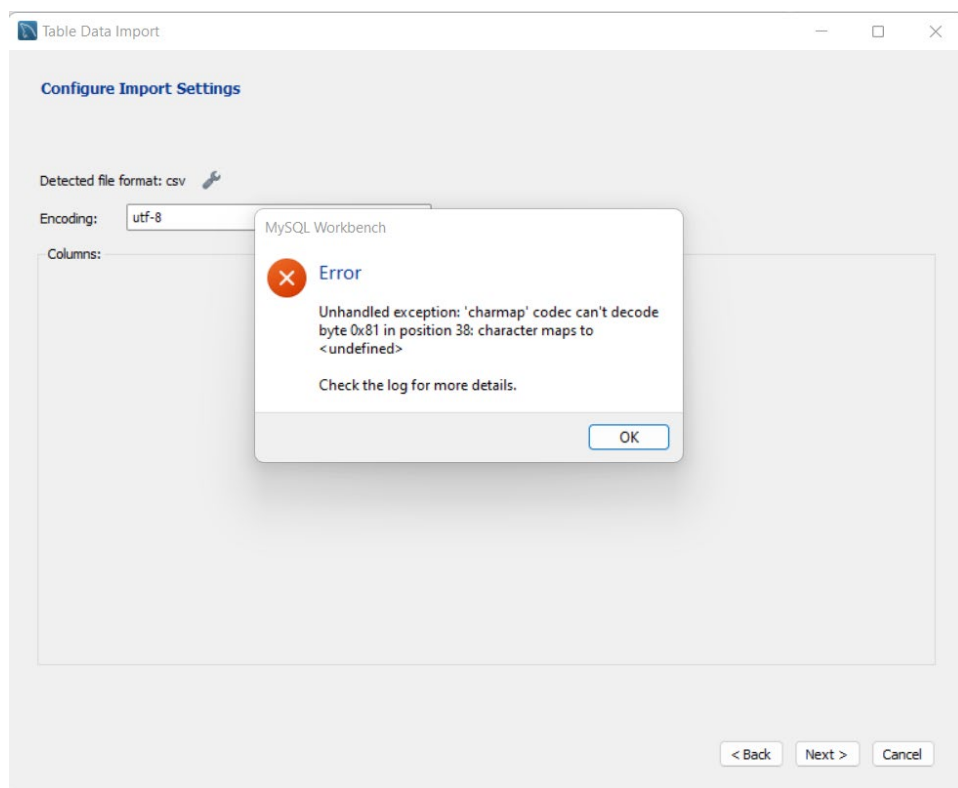


Figure 4.14 Error Snapshot 2: graphemes.csv

CHAPTER 4

By combining the issues that encountered during the implementation process, the root of the problems could be the Chinese Texts and Romanization Diacritics used as values in the tables.

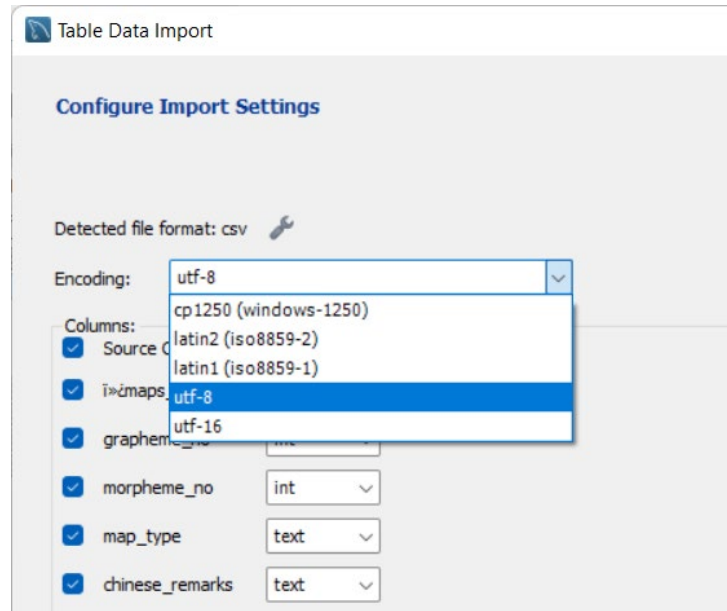


Figure 4.15 Encoding Types in MySQL

By referring to Figure 4.15, MySQL only have 5 encoding types; hence, the Chinese Characters and Romanization Diacritics could not be decoded in MySQL. Moreover, there is a Chinese Character are included in the CJK Unified Ideographs Extension E where not many fonts that can support this character, BabelStone Han font type used in this project during the graphemes table creation was to solve this problem. It was very disappointing that current technologies are not friendly to utilize when researching languages that are not using Latin characters as their written scripts. The solution for these errors will be discussed in Chapter 5.

4.4 Articles and Sentences Collection

4.4.1 Taiwanese Hokkien Articles

The Taiwanese Hokkien articles were taken from **R.O.C Ministry of Education Language Achievement Network** which were published on past e-Newsletter named “Reading to understand Hakka and Hokkien languages”. The articles taken were published in Year 102 until Year 104 of R.O.C Calendar (2013 - 2015). Due to limited time of this research, articles until latest year were not taken. These articles were written in Taiwanese Hokkien with graphemes that suggested by their Ministry of Education which increased the source reliability for future research on the Penang Hokkien as Penang Hokkien and Taiwanese Hokkien have same language progenitors; hence both languages have higher mutual intelligibility. There is another articles source which known as **Taiwanese Hokkien Communication BONG Newspaper**, however the articles from this source were not taken as the articles were written with mixing of Chinese Characters and romanization with diacritics.

Total 150 of articles were collected from the R.O.C Ministry of Education Language Achievement Network as shown in Figure 4.16 below:

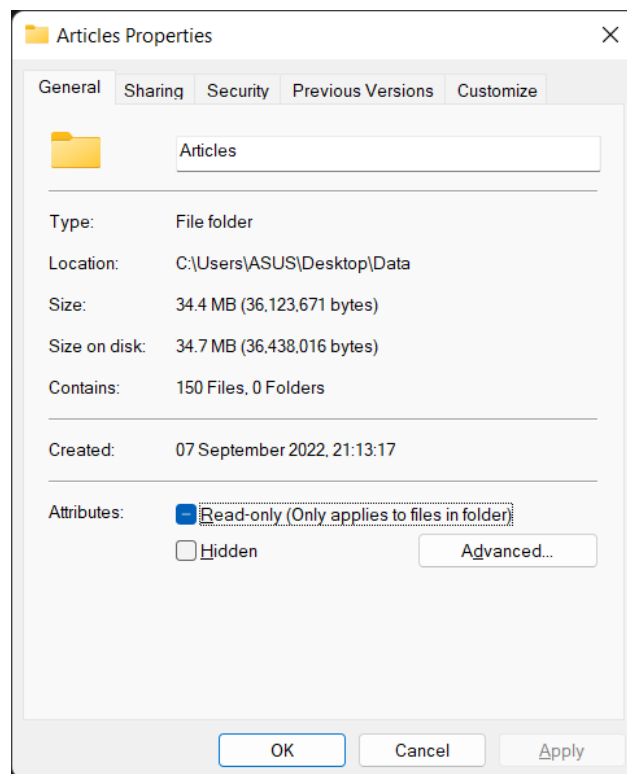


Figure 4.16 Article Folder Properties Window

Moreover, file naming standardization was suggested and implemented when renaming the file due to unreadable file name as shown in Figure 4.17 below:

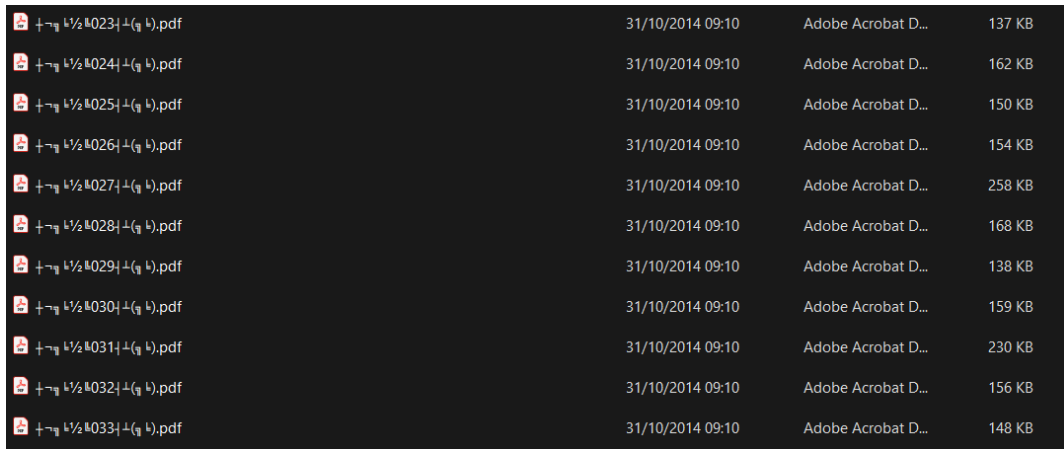


Figure 4.17 Articles Unreadable Filename Lists

The only recognizable text in the filenames were the number which are representing the Special Column Number the articles published in the e-Newsletter as stated previously with Figure4.18 as evidence.

102	世界一的臺灣Hinoki	陳憲國	閩	新知	2	567	023
102	關門開門	王秀容	閩	懷想	1	568	024
102	刳豬公無相請	邱文錫	閩	文史	2	569	025
102	佈稻仔	藍春瑞	閩	文史	2	570	026
102	平溪放天燈	陳憲國	閩	文史	1	571	027
102	未食五日節粽	王秀容	閩	懷想	1	572	028
102	鹿港八郊	邱文錫	閩	文史	1	573	029
102	鯪仔水俗鯪仔標	藍春瑞	閩	文史	3	574	030
102	馬偕牧師來臺灣宣教	陳憲國	閩	人物	1	575	031
102	西瓜皮	王秀容	閩	懷想	1	576	032
102	越南迴心適代	邱文錫	閩	懷想	2	577	033

Figure 4.18 Article Lists Snapshot

With the list provided in the Figure 4.18, 1st, 2nd, 3rd, 5th, and last columns data were taken to rename the file. The list was moved to excel file and create a new column for the file name. The filename format was standardized as below:

Excel Function (Translated into English):

$$=[@e-Newsletter\ Special\ Column\ Number]&"-"&[@Year]&"-"&[@Article\ Title]&"-"&[@Author]&"-"&[@Theme]$$

Outputs:

文件名字
2-102-嘉南大圳之父—八田與一-陳憲國-人物
3-102-甜粿-藍春瑞-文史
4-102-阿姑轉外家-邱文錫-文史
6-102-恩典-王秀容-懷想
7-102-罣礙-藍春瑞-懷想
8-102-語言內底的鬨-邱文錫-文史
10-102-同窗-王秀容-懷想
11-102-千算萬算，毋值天一劃-陳憲國-懷想
12-102-燦光寮山-藍春瑞-景點
14-102-阿母逐日-王秀容-懷想
15-102-竹仔開花-陳憲國-新知

4.4.2 Penang Hokkien Sentences

The Penang Hokkien Sentences were collected from two Facebook Pages, **Speak Hokkien Campaign**, and its predecessor page, **Write Penang Hokkien**.

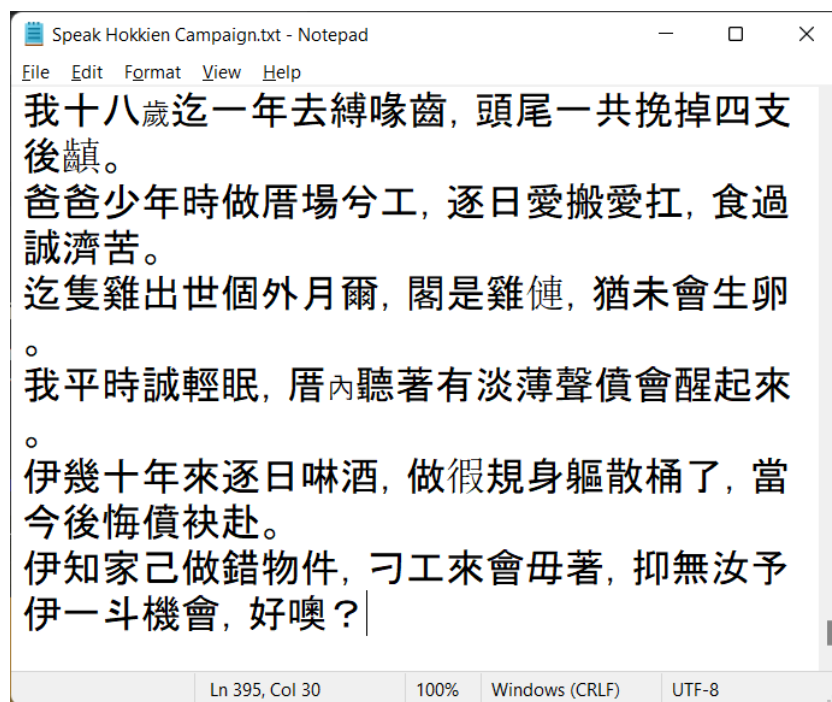


Figure 4.19 Sentences Collected from Speak Hokkien Campaign Facebook Page

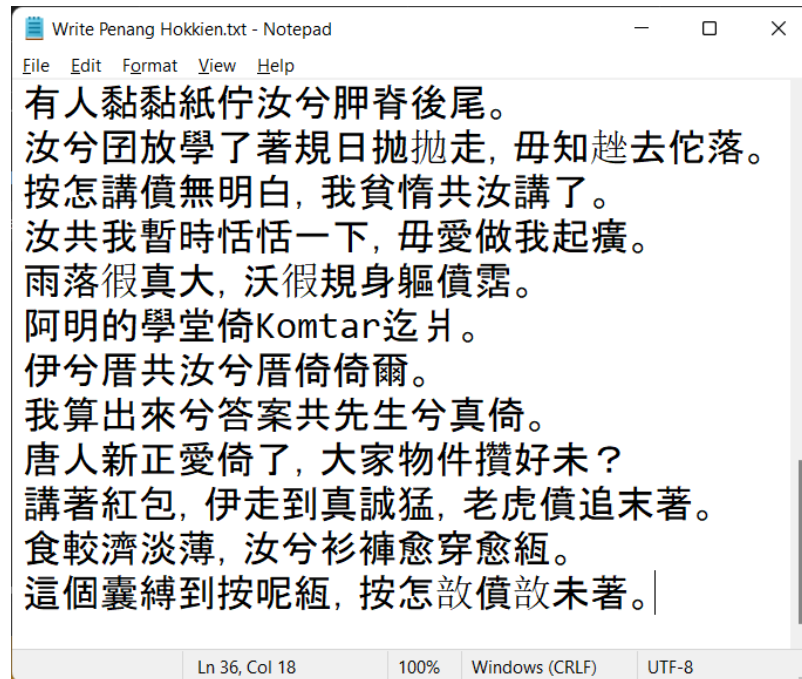


Figure 4.20 Sentences Collected from Write Penang Hokkien Facebook Page

From Figure 4.19 and Figure 4.20 above, 395 sentences were collected from Speak Hokkien Campaign Facebook Page and 36 Sentences Collected from Write Penang Hokkien Facebook Page.

4.5 Audio Guidelines Collection

The collection of audio guidelines was meant to decrease the load of the future research. These collected audio files are suggested to be used as guideline for future research when especially when encountering absents of linguist expert. The audio files were taken from Taiwanese Online TTS System that developed by National Yang Ming Chiao Tung University and National Taipei University of Technology. However, due to limited time for this research, only 60 audio files were collected and renamed as shown in Figure 4.21.

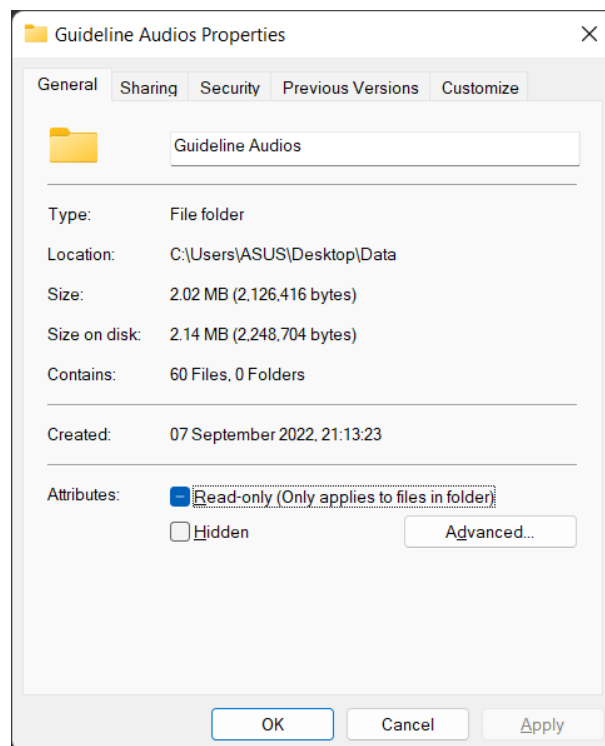


Figure 4.21 Guideline Audios Folder Properties

The audio files were named with the format below:

“phonemes_id”-“phonemes_tone_num”-“Speaker Region”-“Speaker Gender”

Results:

1-pa1-Kaohsiung-female.wav	05/09/2022 22:57	WAV File	44 KB
2-pa2-Kaohsiung-female.wav	05/09/2022 23:07	WAV File	36 KB
3-pa3-Kaohsiung-female.wav	05/09/2022 23:10	WAV File	36 KB

Figure 4.22 Phonemes Guideline Audios Filename Lists

4.6 Concluding Remarks

As a summarization of this chapter, three tables: (1) Morpheme_Phonemes Table, (2) Graphemes Table, and (3) Mapping Table were created for implementation into the database; however, due to the encoding problem on the diacritics and Chinese characters, the implementation of table into the database was failed. The solution will be discussed in the next chapter. Moreover, tone sandhi rules had been researched and standardized with given tone letters and tone contours for each tone except the tone 6 that diachronically merged with tone 2 and tone 7 [32]. The 150 Taiwanese Hokkien articles and 431 Penang Hokkien sentences were collected. Finally, due to limited research time, only 60 audio guidelines that generated by Taiwanese Hokkien TTS System for Penang Hokkien were collected.

CHAPTER 5 SYSTEM EVALUATION AND DISCUSSION

5.1 Data Refining on Pre-Implementation Database Tables

5.1.1 Tables of Pre-refining Process

The checking process towards the Morpheme_Phonemes Table and Graphemes Table was taken during the mapping process of the phonemes or morphemes with graphemes from respective tables in Mapping Table. Before the mapping process was taken, there were total of 2102 records in Morpheme_Phonemes Table as shown in Figure 5.1, and total of 4138 records in Graphemes Table as shown in Figure 5.2.

20493	20492	NaN	uat	8	uát
20498	20497	NaN	ap	4	ap
20502	20501	NaN	ap	8	áp
20507	20506	NaN	ip	4	ip
20520	20519	NaN	iap	8	iáp
20522					

Ready 2102 of 20520 records found Accessibility: Investigate

Figure 5.1 Total Number of Phonemes Before Mapping

4135	4134	黠
4136	4135	滂
4137	4136	嚮
4138	4137	捩
4139	4138	躅

Graphemes Suggestion1 Suggestion2 Refe

Figure 5.2 Total Number of Graphemes Before Mapping

5.1.2 Data Mapping and Refining Process

The refining process of data was taken along with mapping of phonemes with graphemes. The mapping and refining process were supported by dictionaries mentioned in the previous subchapter 3.4. Graphemes were the main data that focused during the mapping process. When the graphemes mapped with phonemes, map types were selected from the list. As previously mentioned on the abbreviation of the map

types, Taiwanese Hokkien Online Dictionary categorized the phonetic-loaned graphemes and KYM graphemes as “Alternate Words”. In order to differentiate these while select the map types for the mapping results, Online Chinese Dictionary was used to check on the graphemes meaning. If the searched grapheme meaning is same with Taiwanese Hokkien Dictionary but in Taiwanese Hokkien Online Dictionary marked it as “Alternate words”, then that grapheme is categorized as KYM. If a grapheme in the Taiwanese Hokkien Online Dictionary marked as “Alternate words” and searched in Online Chinese Dictionary has different meaning, it is marked as “ATJ”. These rules applied to the graphemes collected from Speak Hokkien Campaign and Write Penang Hokkien Facebook posts as these graphemes provided come along with its meaning. This mapping process had tracked graphemes and phonemes that missed out and extra phonemes that added in the corresponded tables. The logs are attached at Appendix B.

5.1.3 Results

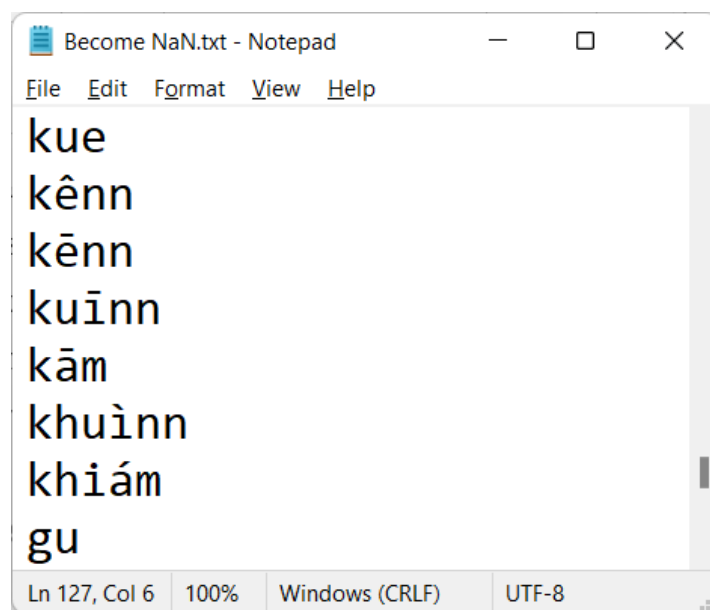


Figure 5.3 Total Number of Extra Phonemes

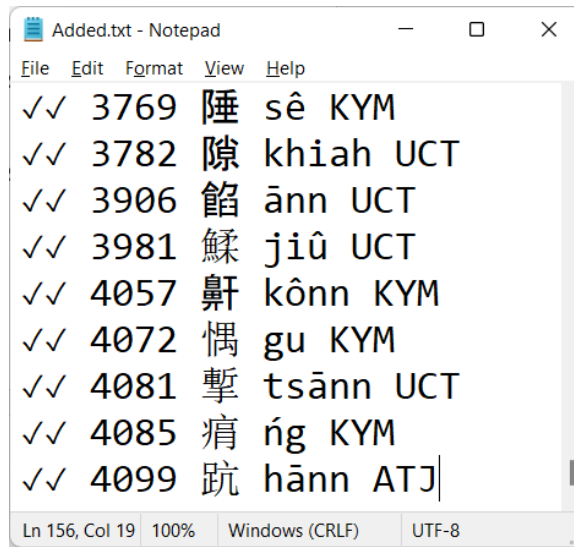


Figure 5.4 Total Numbers of Unadded Graphemes and Phonemes

4246	4245	飾		
4247	4246	適		
4248	4247	釋		
4249	4248	席		
4250	4249	毫		
4251	4250	旄		

Figure 5.5 Total Numbers of Graphemes after Data Refining

51	50	p	o	5	pô
53	52	p	o	7	pō
56	55	p	oo	1	poo
57	56	p	oo	2	póo

Figure 5.6 Total Numbers of Phonemes after Data Refining

From Figure 5.3, there were 127 phonemes added as extra that were not available in Penang Hokkien and were removed from the Morpheme_Phonemes Table. From 5.4, there were 156 unmapped graphemes due to phonemes were not added in phonemes tables. By comparing the total graphemes from Pre-Refining Graphemes Table with the total graphemes from Post-Refining Graphemes Table, there are total of $4250 - 4138 = 112$ graphemes that added into the Post-Refining Graphemes Table as shown

in Figure and 5.5. The refined Morpheme_Phonemes table has its records increased from 2102 to 2110 as shown in Figure 5.1 and 5.6.

5759	5758	4246	13522	UCT	無
5760	5759	4247	13522	UCT	無
5761	5760	4248	13526	UCT	無
5762	5761	4249	3481	UCT	無
5763	5762	4250	3481	KYM	訓

Figure 5.7 Total Mapped Records

Lastly, there were total of 5,762 records mapped in Mapping Table as shown in Figure 5.7.

5.2 Simple Data Analysis Towards Pre-Implementation Database Tables

5.2.1 Morphemes_Phonemes Table

Count of Initial Column Labels	Row Labels	NaN	OPS	NAC	STC	NAV	Grand Total
b		779	33	25	18		855
d		854		1			855
f		854	1				855
g		782	30	28	15		855
h		691	63	50	32	19	855
j		815	16	15	9		855
k		677	68	56	29	25	855
kh		729	49	43	27	7	855
l		731	49	48	27		855
m		811	32	5	7		855
n		821	29	3	2		855
NaN		667	66	62	34	26	855
ng		835	15	2	3		855
p		717	53	39	29	17	855
ph		758	40	26	19	12	855
r		854		1			855
s		677	64	57	36	21	855
sh		854		1			855
t		667	62	62	35	29	855
th		743	43	34	23	12	855
ts		667	65	60	37	26	855
tsh		720	50	48	22	15	855
w		853	1	1			855
y		854		1			855
Grand Total		18410	829	668	404	209	20520

Figure 5.8 Initials and Final Types Analysis

% distribution of 'Final_Types'

Row Labels	Count of Final_Types
OPS	39.29%
NAC	31.66%
STC	19.15%
NAV	9.91%
Grand Total	100.00%

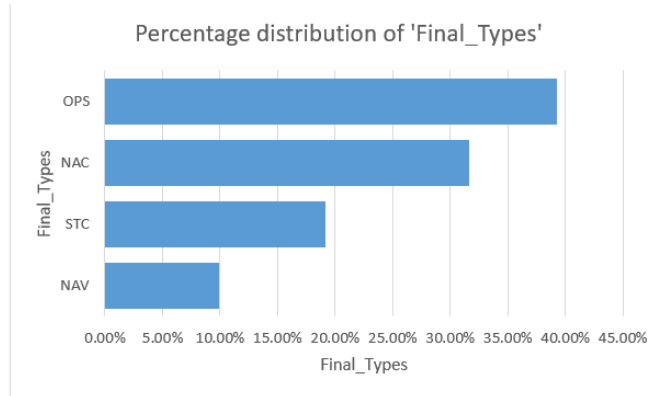


Figure 5.9 Percentage Distribution of Final_Types

Figure 5.7 showed 18,410/20,520 possible phonemes combinations are invalid. There are 829 OPS, 668 NAC, 404 STC, and 209 NAV which give the total sum of 2,110 valid phonemes combinations for the Penang Hokkien. Among the 2,110 valid phonemes combinations, 39.29% are OPS, 31.66% are NAC, 19.15% are STC and 9.91% are NAV as shown in Figure 5.8.

5.2.2 Mapping Table

Percentage Distribution of 'Map_Type'.

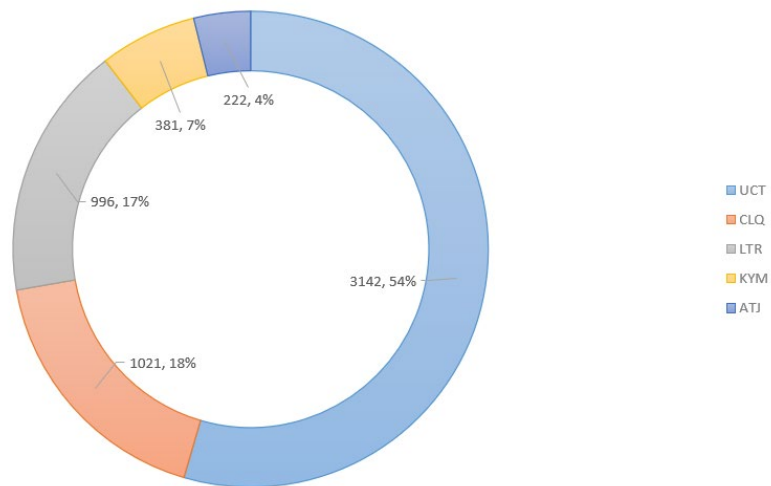


Figure 5.10 Percentage Distribution of Map_Type

From the Figure 5.9, with Grand Total of 5,762 Mapped Graphemes and Phonemes, 3,142 are UCT with distribution of 54%, 1,021 and 996 are CLQ and LTR respectively with distribution of 18% and 17% in the Map_Type. Moreover, KYM

has 381 records with 7% distributed and ATJ is the least distributed map type with only 222 records and 4% of distribution.

5.3 Solutions on Errors During Database Implementation

5.3.1 morphemes_phonemes.xlsx

In morphemes_phonemes.xlsx, the solution was deleting the phoneme_diacritics column as the special diacritics were the root problems of importing the table into the database.

5.3.2 graphemes.xlsx

In graphemes.xlsx, the solution was creating two new attributes that represent the Chinese characters as they are the root problem of importing the table into the database. The first attribute created was unicodes_dec, which means Unicode number of the Chinese characters in decimal format. The excel function used was “=UNICODE([@[graphemes_chars]])”. The second attribute created was unicodes_cjk_hex, which means unicodes in CJK Ideographs Hex format. The excel function used was “=“U+”&DEC2HEX([@[unicodes_dec]])”. After that, this excel file was saved as .csv file then the column “graphemes” was deleted.

5.3.3 maps_graphs_morphs.xlsx

In maps_graphs_morphs.xlsx, the solution was deleting the chinese_remarks column as the Chinese characters were the root problems of importing the table into the database.

5.3.4 ER Diagram Redesign with Data Dictionaries

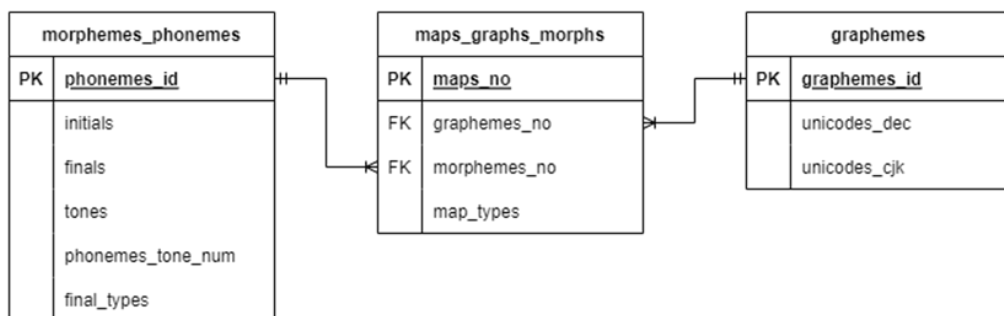


Figure 5.11 Redesigned ER Diagram for Penang Hokkien

Table 5.1 Redesigned Data Dictionary for morphemes_phonemes

morphemes_phonemes			
Field Name	Data Type	Nullable	Description
phonemes_id	Integer	N	Unique ID number for each phoneme of initial and final combination.
initials	Text	N	Initial of Penang Hokkien Romanized phoneme.
finals	Text	N	Finals of Penang Hokkien Romanized phoneme.
tones	Integer	N	Tones numbers of Penang Hokkien.
phonemes_diacritics	Text	N	Romanized phonemes with diacritics symbol on the Latin alphabets. Contain Special characters.
phonemes_tone_num	Text	N	Romanized phonemes with tone number.
final_types	Text	N	Abbreviation of the final types.

Table 5.2 Redesigned Data Dictionary for graphemes

graphemes			
Field Name	Data Type	Nullable	Description
graphemes_id	Integer	N	Unique ID number for each grapheme.
graphemes	Text	N	Chinese Japanese Korean (CJK) Unified Ideographs Extensions texts.

Table 5.3 Redesigned Data Dictionary for maps_graphs_morphs

maps_graphs_morphs			
Field Name	Data Type	Nullable	Description
maps_no	Integer	N	Unique ID number for each mapped graphemes, morphemes, and map types.
graphemes_no	Integer	N	ID number for each grapheme.
morphemes_no	Integer	N	ID number for each morpheme.
map_types	Text	N	Abbreviation of the map types.
chinese_remarks	Text	N	Abbreviation of the map types in Chinese.

5.3.5 Reimplementation of Tables in Database

The steps for importing the tables into the database were same as in the previous subchapter 4.3.4.

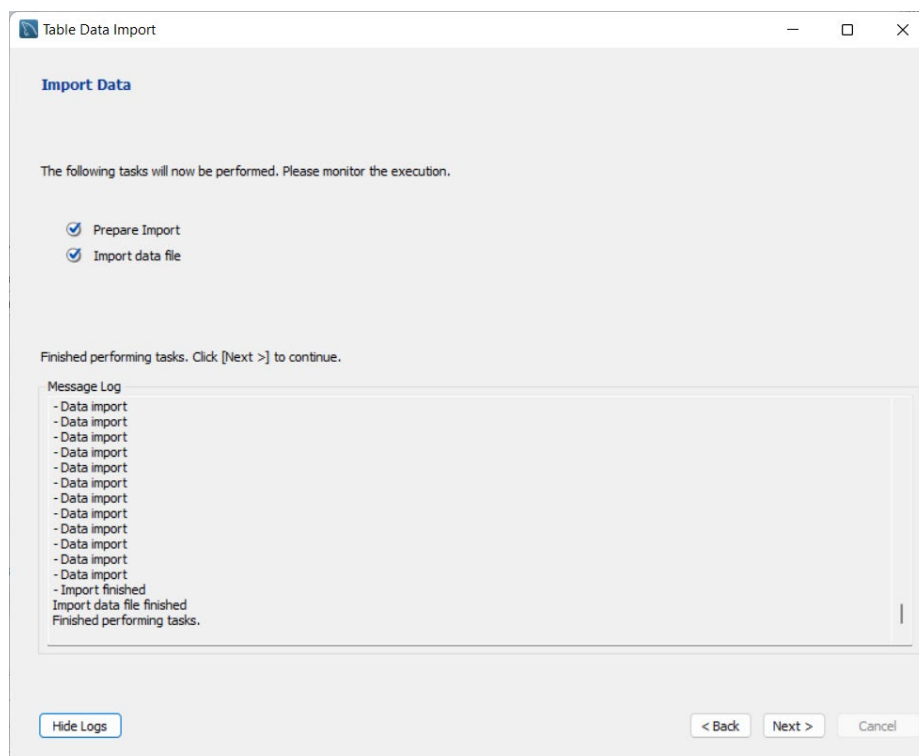


Figure 5.12 Successful Import Logs

After the tables were successfully imported into the database where their successful import logs were similar as shown in Figure 5.11, the next step was to set up the primary keys and foreign keys in the tables by following the redesigned ER Diagram and Data Dictionaries in the previous chapter 5.3.4 as shown in Figure 5.12, 5.13, and 5.14. Next, the foreign keys were set up in maps_graphs_morphs as shown in Figure 5.15 with successful set up result shown in Figure 5.16 below.

Reverse Engineer Function in MySQL were used to visualize the Entity Relation of the Tables as shown in Figure 5.17 below which is same as the Redesigned ER Diagram in subchapter 5.3.4.

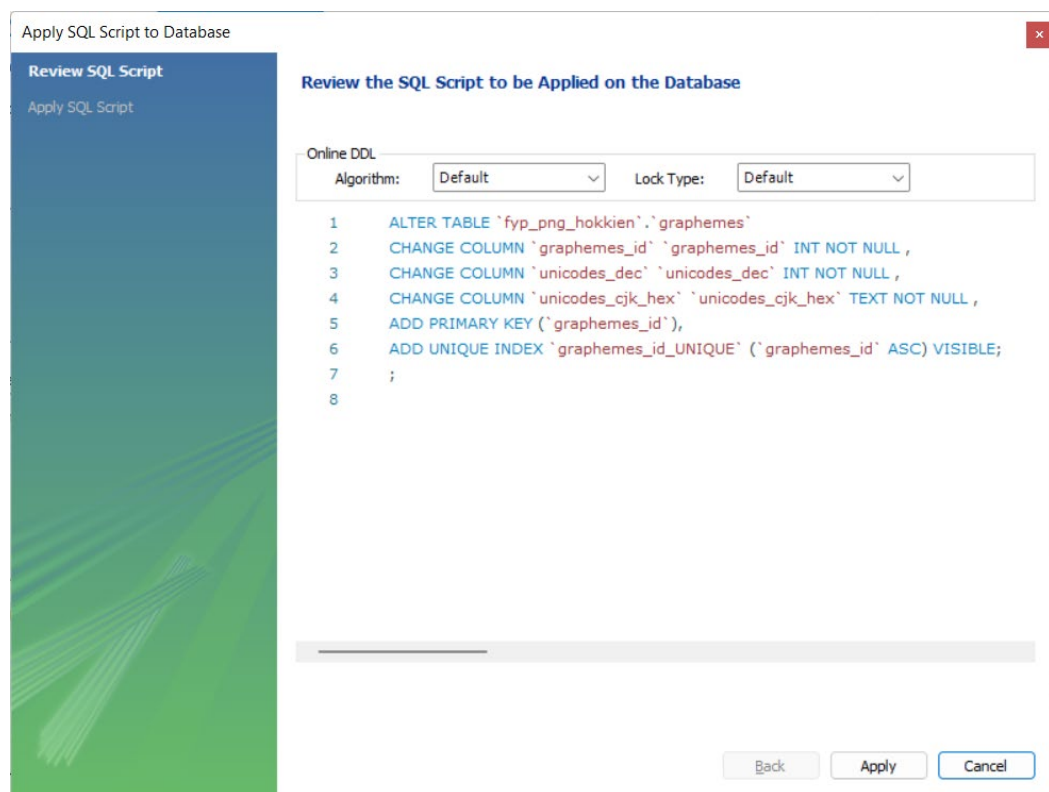


Figure 5.13 Primary Key and Not Null Set Up for graphemes

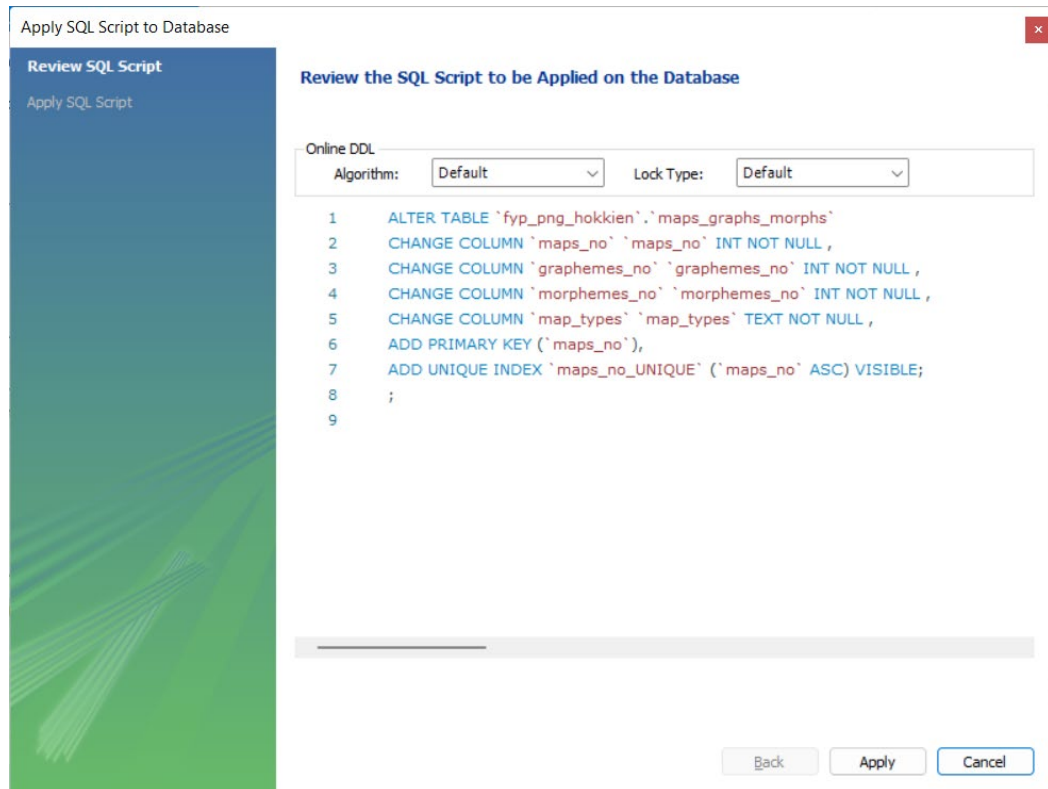


Figure 5.14 Primary Key and Not Null Set Up for maps_graphs_morphs

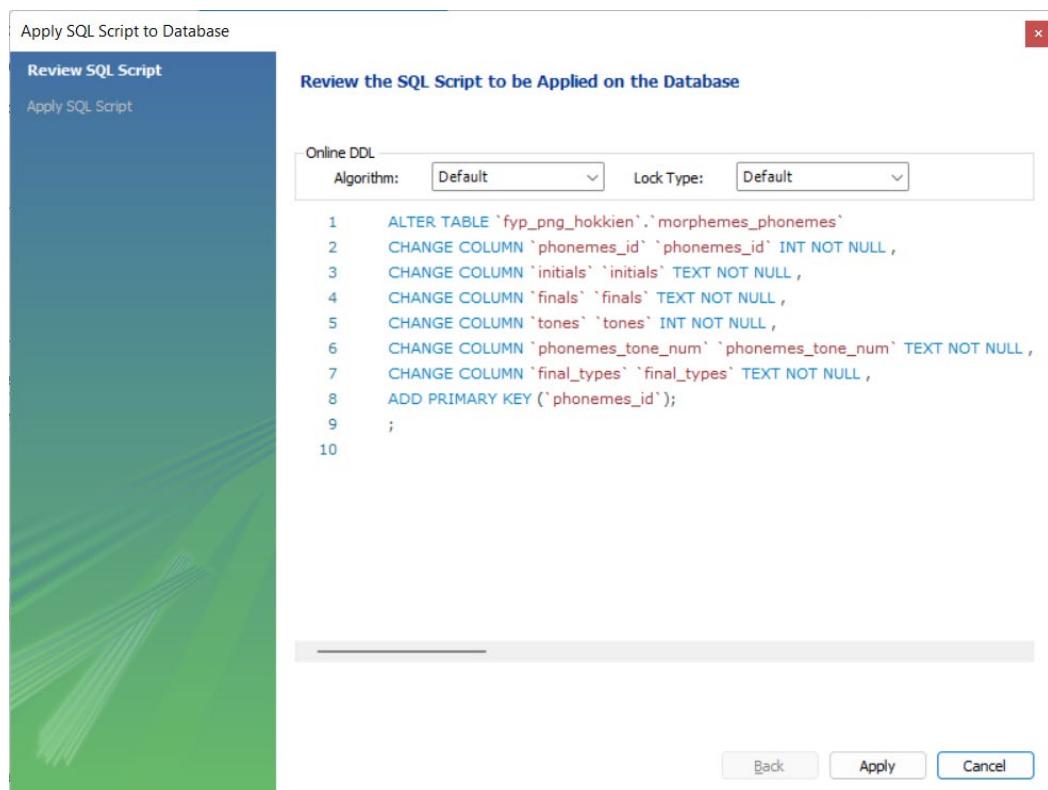


Figure 5.15 Primary Key and Not Null Set Up for morphemes_phonemes

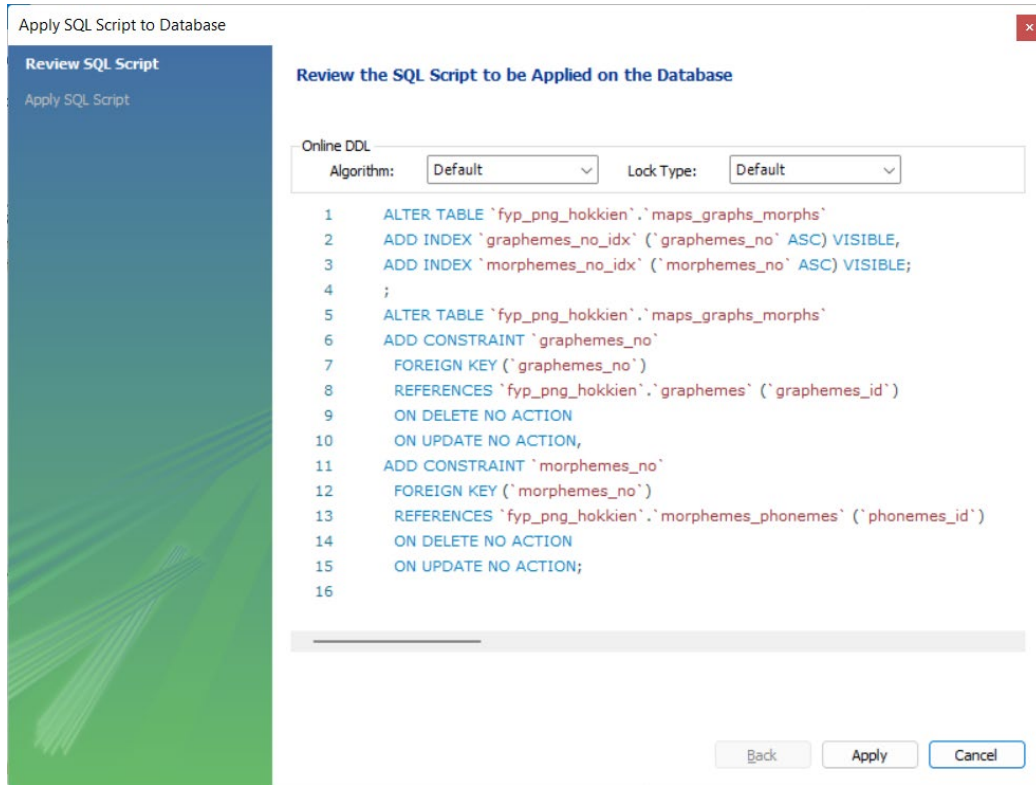


Figure 5.16 Foreign Keys Set Up for maps_graphs_morphs

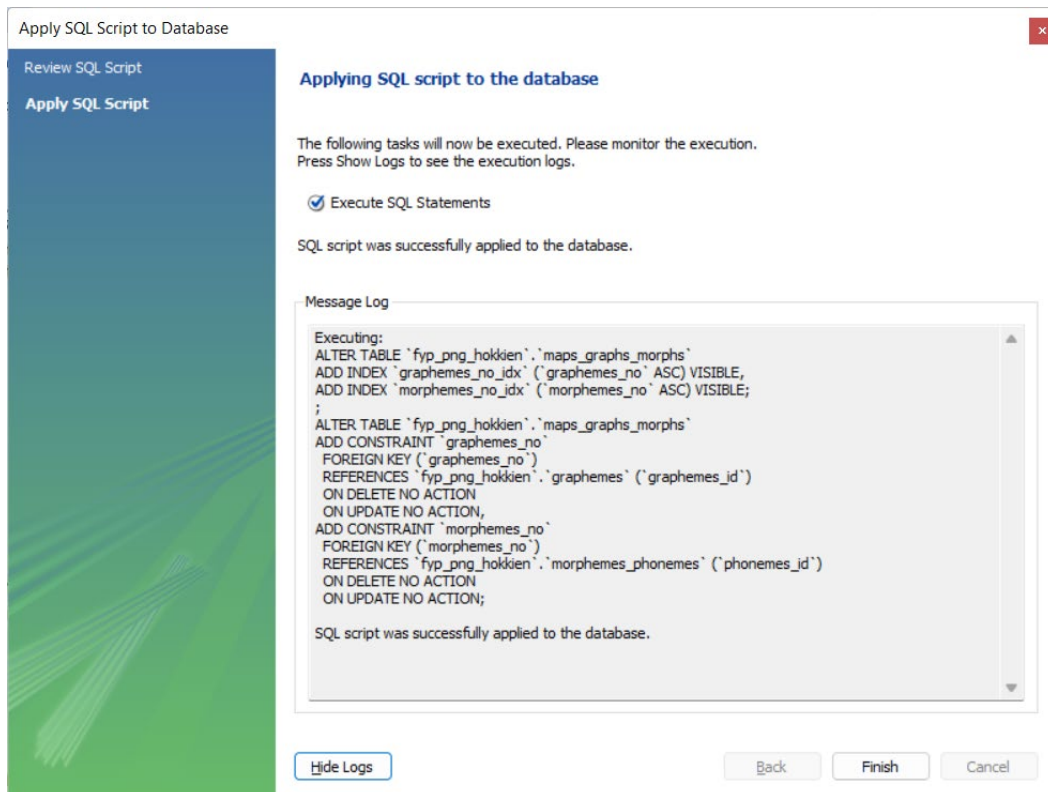


Figure 5.17 Foreign Keys Set Up Results for maps_graphs_morphs

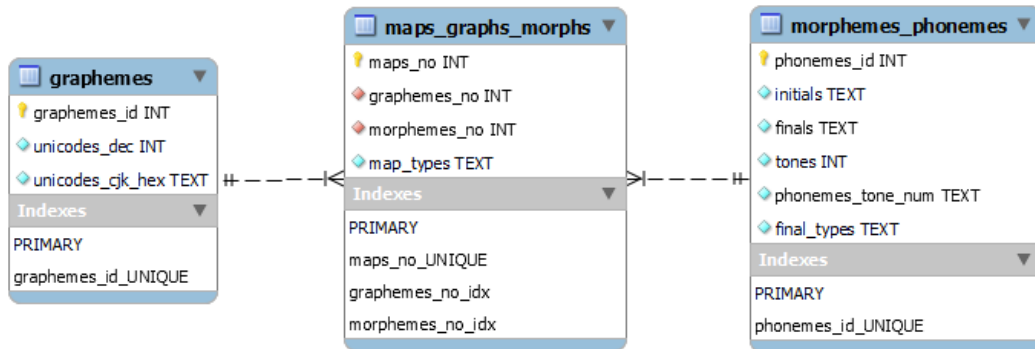
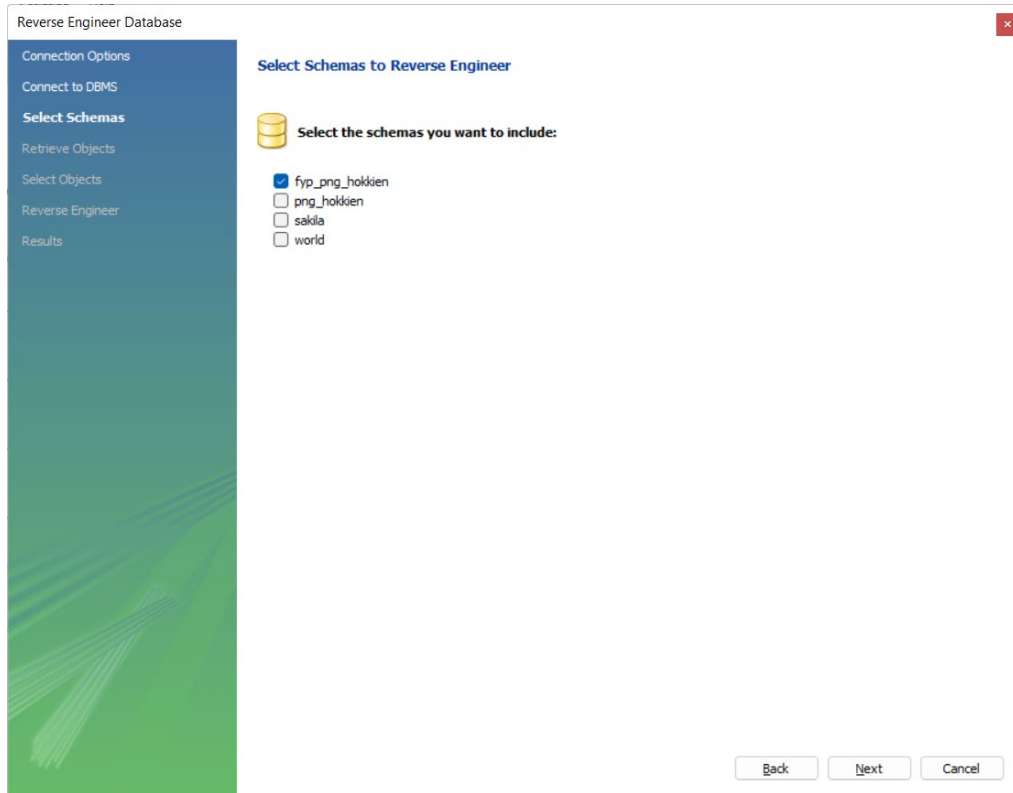


Figure 5.18 ERD Visualization with Reverse Engineering

5.4 Project Challenges

The challenges in this research project were the limited time given and resources as well as limited human sources. I had contacted with the Speak Hokkien Campaign authority to request on the data that needed in the research; however, no replies were received. Hence, I decided to collect the data all by myself and spent several months to rush on the research. When I further contact the Speak Hokkien Campaign authority to ask on the diacritics symbol used, they replied to me and answered my

questions and from that conversation, I found that there are only two persons that managing on the Speak Hokkien Campaign Facebook page although they are under the Persatuan Bahasa Hokkien Pulau Pinang.

The most disappointed challenge was the current technologies that could not support the written scripts that are not written in English alphabet or Latin alphabets.

5.5 Concluding Remarks

As a summary in this chapter, data refining process was conducted correspondingly with graphemes and morphemes mapping progress in the Mapping Table. Simply data analysis that summarized the total number of data collected and distributions by Final Types and Map Types were visualized in graphs. Moreover, solutions to tackle the database importing errors were given and conducted by eliminating the special characters that database could not decoded. Hence, ER-diagram was redesigned together with the related data dictionaries. Re-implementation of the tables into the database was successful and using the MySQL Re-engineering function to generate ER Diagram of the tables in the database is the same with the redesigned ER Diagram.

CHAPTER 6 Conclusion and Recommendation

6.1 Conclusion

As an encapsulation for the entire research report, the title of this research is Pronunciation Modelling of Penang Hokkien Dialect for Text-to-Speech System. The scope of this project to model the Penang Hokkien pronunciation that act as the first step that contributes to future development of Penang Hokkien Speech Synthesis System. Penang Hokkien have potential to be developed into speech synthesis however, this language has obvious problems which are unstandardized orthography, unstandardized text, et cetera due to poor documentation of lexical resources as mentioned previously that Hokkien Association of Penang dug up the old documents and books to prove Penang Hokkien is a language not a dialect [20]. These reasons hinder Penang Hokkien in effort of develop a speech synthesis system for it. Hence, this research has a responsibility 1. to collect large number of phrases and words of Penang Hokkien for text corpus construction and standardising the language orthographies including the romanization and Chinese characters and collect each morphemes audio sample as pronunciation guidelines., and 2. collect large number of sentences and articles that similar to Penang Hokkien as a future resource in selecting the most suitable sentences for studio recording in future research project. As stressed in previous chapters, Penang Hokkien currently is being threatened and would be extinct approximately 40 years from now; thence, this gives motivation on this project to take part in efforts of revitalizing Penang Hokkien Language with implementation of current technologies. After reviewing numerous related resources, Penang Hokkien Spelling System that created by Hokkien Association of Penang and Traditional Chinese Characters are selected as standards of the pronunciation and orthography for Penang Hokkien Language. During the research, an ER Diagram along with related Data Dictionaries were designed. Three dictionaries were involved in this research with the assistance of MySQL Database and Microsoft Excel as well as BabelStone Han font style throughout the research progress. There are three tables were created to implemented into the database however due to special characters in the tables, the implementation was failed. Apart from that, tone sandhi rules had been standardized during the research. Taiwanese Hokkien articles and Penang Hokkien sentences was collected for contributing to future research purpose. Audio guidelines that planned to

collect from Taiwanese Hokkien TTS System was halted due to limited time of the research progress. Data refining process was taken correspondingly with Mapping Table creation process and a simple data analysis was conducted. Furthermore, solution by eliminating special characters in the tables solved the implementation problems. The research was ended with successful Re-implementation of tables into database that match with the redesigned ERD and Data Dictionaries.

6.2 Recommendations

This research project was the first steps of TTS System Development of Penang Hokkien. Due to limited time and scarcity of human resources, lots of early planned research objectives were halted and abandoned. The high complexity of Penang Hokkien Dialects under categorization of Southern Min Language was one of the obstacles in this research project. However, this research increased the sanity capacity of the researchers. As first recommendation towards the future research of related research, knowledge of different romanization of Southern Min Languages must be acquired due to current unstandardized romanization methods were invented. This could assist in the data obtaining process where more data can be acquired. Furthermore, expertise of old and middle Chinese phonology and graphemes should be participated in the research to trace the forbidden graphemes. Due to unsuitable graphemes used to replace the forbidden graphemes by Taiwanese Ministry of Education in their Online Dictionaries, there are lots of controversies towards the Ministry, the graphemes taken from them are needed to compare with other Southern Min linguists that are not working under Taiwanese Ministry of Education.

REFERENCE

- [1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.
- [2] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, 1st ed. Springer Netherlands, 1997. doi: 10.1007/978-94-011-5730-8.
- [3] National Academy of Sciences, *Voice Communication Between Humans and Machines*. Washington, DC: The National Academies Press, 1994. doi: 10.17226/2308.
- [4] J. Trouvain and B. Möbius, “Speech Synthesis: Text-To-Speech Conversion and Artificial Voices,” in *Handbook of the Changing World Language Map*, S. D. Brunn and R. Kehrein, Eds. Springer, Cham, 2020, pp. 3837–3851. doi: 10.1007/978-3-030-02438-3_168.
- [5] D. H. Klatt, “Review of text-to-speech conversion for English,” *Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987, doi: 10.1121/1.395275.
- [6] K. Cagle, “A Digital Voice For The Mute,” *Bbntimes.com*, 2020. <https://www.bbntimes.com/technology/a-digital-voice-for-the-mute> (accessed Aug. 22, 2021).
- [7] D. Mwiti, “A 2019 Guide to Speech Synthesis with Deep Learning,” *Heartbeat*, 2019. <https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd#70c3> (accessed Aug. 23, 2021).
- [8] D. M. Eberhard, G. F. Simons, and C. D. Fennig, “Ethnologue: Languages of the World,” *SIL International*, 2021. <http://www.ethnologue.com>
- [9] O. Mok, “Penang Hokkien Will Be ‘Dead’ In 40 Years If People Stop Using It, Says Language Expert,” *Malaymail.com*, 2016. <https://www.malaymail.com/news/malaysia/2016/08/02/penang-hokkien-will-be-%09dead-in-40-years-if-people-stop-using-it-says-langua/1174401> (accessed Aug. 23, 2021).

REFERENCE

- [10] Persatuan Bahasa Hokkien Pulau Pinang, “Learn To Read And Write Hokkien,” *Speak Hokkien Campaign*, 2021. <https://www.speakhokkien.org/learn-to-read-and-write-hokkien> (accessed Aug. 24, 2021).
- [11] T. Tye, “Taiji Romanisation for Hokkien Speakers,” *Penang Travel Tips*, 2009. <https://www.penang-traveltips.com/taiji-romanisation-for-hokkien-speakers.htm> (accessed Aug. 24, 2021).
- [12] T. Tye, “Penang Hokkien Dictionary,” *Penang Travel Tips*, 2013. <https://www.penang-traveltips.com/dictionary/index.htm> (accessed Aug. 27, 2021).
- [13] T. Tye, “Taiji Romanisation of Penang Hokkien,” *Penang Travel Tips*, 2009. <https://www.penang-traveltips.com/timothy-tye-penang-hokkien-romanisation.htm> (accessed Sep. 06, 2022).
- [14] Sciforce, “Text-to-Speech Synthesis: an Overview,” *Sciforce Blog*, 2020. <https://medium.com/sciforce/text-to-speech-synthesis-an-overview-641c18fcd35f> (accessed Aug. 24, 2021).
- [15] J. Y. Hou, “Minyu [Min Languages],” in *Xiandai Hanyufangyan gai lun*, Shanghai: Shanghai Educational Publishing House, 2002, pp. 207–209.
- [16] C. J. & Teh and Y. L. Lim, “An Alternative Architectural Strategy to Preserve the Living Heritage and Identity of Penang Hokkien Language in Malaysia,” *Int J Humanit Soc Sci*, vol. 4, no. 3, pp. 242–247, 2014.
- [17] Y. F. Lee, “Chinese Education and Identity in Singapore,” *International Journal of History Education*, vol. 10, no. 2, pp. 11–25, 2009, doi: 10.17509/historia.v10i2.12218.
- [18] S.-H. Ting and J. Z.-M. Teng, “Chinese teenagers’ perceptions of vitality of Hokkien Chinese in Penang, Malaysia,” *Int J Soc Lang*, 2021, doi: 10.1515/ijsl-2020-0024.
- [19] M. P. Lewis and G. F. Simons, “Assessing endangerment: Expanding Fishman’s GIDS,” *Revue Roumaine de Linguistique*, vol. 55, no. 2, pp. 103–120, 2010.

REFERENCE

- [20] O. Mok, “Exhibition in George Town chronicles the decline of the Hokkien language in bid to save it,” *Malaymail.com*, 2020. <https://www.malaymail.com/news/life/2020/08/23/exhibition-in-george-town-%09chronicles-the-decline-of-the-hokkien-language-in/1896270> (accessed Aug. 24, 2021).
- [21] Persatuan Bahasa Hokkien Pulau Pinang, “Recommended Characters,” *Speak Hokkien Campaign*, 2021. <https://www.speakhokkien.org/hokkien-characters> (accessed Aug. 24, 2021).
- [22] T. Tye, “Introduction to Learning to Write Penang Hokkien,” *Penang Travel Tips*, 2015. <https://www.penang-traveltips.com/hokkien/learn-to-write-penang-hokkien-01.htm> (accessed Aug. 24, 2021).
- [23] W. K. Lo, T. Lee, and P. C. Ching, “Development of Cantonese spoken language corpora for speech applications,” in *International Symposium on Chinese Spoken Language Processing*, 1998, pp. 102–107.
- [24] J. Matoušek, D. Tihelka, and J. Romportl, “Building of a speech corpus optimised for unit selection tts synthesis,” in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008, pp. 1296–1299.
- [25] V. Lim, H. S. Ang, E. Lee, and B. P. Lim, “Towards an interactive voice agent for Singapore Hokkien,” in *HAI 2016 - Proceedings of the 4th International Conference on Human Agent Interaction*, 2016, pp. 249–252. doi: 10.1145/2974804.2980495.
- [26] J. S. Tsay, “Construction and automatization of a Minnan Child Speech Corpus with some research findings,” *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 12, no. 4, pp. 411–442, 2007.
- [27] C. Douglas, *Chinese-English dictionary of the vernacular or spoken language of Amoy: with the principal variations of the Chang-Chew and Chin-Chew dialects*. London: Trübner, 1873.
- [28] Ministry of Education Republic of China, “Dictionary of Frequently-Used Taiwan Minnan.”

REFERENCE

- [29] A. C. West, “BabelStone Han,” *BabelStone Fonts*. <https://www.babelstone.co.uk/Fonts/Han.html> (accessed Sep. 08, 2022).
- [30] Logith, “Běi mǎ, bīnláng fújiàn huà bīng shēngdiào[Penang Hokkien Tones],” *banlam.tawa.asia*, Oct. 28, 2012. <https://archive.ph/NdBBP#selection-295.1-299.11> (accessed Sep. 07, 2022).
- [31] Y.-R. Chao, “ə sistim əv "toun-letəz [A system of ‘tone-letters’],” *Le Maître Phonétique*, 1930, Accessed: Sep. 07, 2022. [Online]. Available: <https://www.jstor.org/stable/44704341>
- [32] C. Chuang, Y.-C. Chang, and F. Hsieh, “Complete and not-so-complete tonal neutralization in Penang Hokkien,” Dec. 2013.
- [33] Ministry of Education R.O.C, “Jiaoyubu Taiwan Minnanyu Changyongci Cidian - Bianji Shuoming [Ministry of Education Taiwan Southern Min Frequent Used Phrases Dictionary - Editing Instructions - Phonetics Instructions].” https://twblg.dict.edu.tw/holodict_new/compile1_3_9_3.jsp (accessed Sep. 08, 2022).
- [34] K. Dexter, “On’yomi And Kun’yomi in Kanji: What’s the Difference?,” *Tofugu*, Sep. 05, 2017. <https://www.tofugu.com/japanese/onyomi-kunyomi/> (accessed Sep. 08, 2022).

APPENDIX A Penang Hokkien Spelling System

Penang Hokkien Spelling System



spelling	p-	h-	ph-	b-	m- ^a	t-	th-	n-	l-	k-	kh-	g-	ng- ^a	ts-	tsh-	s-	j-
meaning	fat	good	scent	none	porridge	(a surname)	listen	egg	you	dog	leg	cow	hard	run	hand	three	word
example	pūi	hó	phang	bó	múi	tán	thiam	nú	lú	káu	kha	gú	ngé	tsáu	tshú	sann	ji
d- ^b		f- ^b	r- ^b	sh- ^b		w- ^b	y- ^{a,b}										
midwife	bi-dan	sofa	ring-git	ku-shén	wé / us	wá-làng	sa-yang										

^a used in loanwords.
^b also used in Finals. See below.

Initials

	a	e	ee	er	i	o	oo	u	y ^b
dry	small	tea	steal	poem	treasure	rain	fish	Choe Cheong Fun	
ta	sè	téa	sêr-kh	sí	pó	hoo	hú	tsy-tshióng-fán	
write	invite	chief	flower	unlucky	open	open	'easy-peasy'	bird	obedient
siá	tsio	tai-lóu	hua	sue	khui	sáp--sáp--séoi		tsiáu	kuai
mother	mother	keep	keep	keep	keep	keep	keep	keep	keep
nioá	niu	niu	niu	niu	niu	niu	niu	niu	niu
nioá	niu	niu	niu	niu	niu	niu	niu	niu	niu

Finals

-ai	-au	-ei ^a	-ia	-io	-ioo ^c	-iu	-oi ^a	-ou ^a	-ua	-ue	-ui	-oo	-u	-y ^b
know	package	non sequitur	write	invite	mother	keep	bottle	chief	flower	poem	treasure	rain	fish	Choe Cheong Fun
tsai	pau	mou-lai-thau	siá	tsio	nioá	niu	bót-toi	tai-lóu	hua	sue	khui	sáp--sáp--séoi		tsiáu

^a a variant of -io, used only with n- and h-.
^b a final borrowed from Cantonese /eoy/, pronounced by some as -ue.
^c a variant of -io, used only with n- and h-.

Nasal

	m ^b	-ann	-enn	-inn	-onn	-ainn	-iann	-ionn	-oinn	-uann	-uinn	-iaunn ^a	-uainn
(negation)	dress	fight	sky	interjection	finger	fear	piece	sleep	hill	bright	incense	lock up	kuainn
m̄	sann	tsenn	thinn	hónn	tsáinn	kiann	tiann	óinn-óinn	suann	kuinn	hiann	kuainn	kuainn
nap' (sound)	hiám	dislike	dream	success	Perang	ring-git	wind	village	same	hurt	turnip		
tám tsit-siam	hiám	hiám	bàng	séng	Pr-iéng	ring-git	hóng	kám-pung	siàng	sióng	bàng-kuang		
	-an	-en ^a	-en ^a	-enn ^a	-in	-on ^a	-un	-yn ^a	-ian	-uan	-ng ^b		
	slow	unneeded	unneeded	unneeded	root	dry-tossed noodles	boil	sudden	build	bottle	ice		
	bán	mén	mén	mén	kin	kón-lo-mi	kún	mou-tyñ-tyñ	kián	kuán	sng		

^a a variant of -tonn.
^b assimilated diphthongs:
-ian > -en
-iat > -et

-ah	-eh	-eoh	-eth	-ih	-oh	-ooh	-uh	-ah	-auh	-lah	-loh
hit	eight	book	burp	fold	rope	oh	Pierce	water pipe	plump	wall	borrow
phah	peh	tsheeh	phah-erh	tsih	soh	ooh	tuh	tsui-path	phauh	plah	tsiah
	-uah	-ueh	-ueh	-ak	hom	cheque	Tic-Tac	tsék	tsúik	smash	tsiak
	kuah	kuah	kuah	kek	kek	tsék	tsék	tsék	tsék	tsék	tsék

Stops

-lok	-at	-et ^a	-ert ^a
bless	kill	set	stabbed
tsiook	sat	set	sé-lerit

^a 3 and 7 are identical before sandhi.
^b Some speakers change from 3 to 1.
^c 6 and 9 are loan syllables therefore do not change tones.

Tones

Level	Rising	Departing	Entering
1	2 /	3 \	4
Dak	dress	trousers	wide
5	6	7	8
Light	person	short/meh	nose
làng	é	phinn	ti
			9

According to the sandhi rules, a syllable in the spoken language generally changes its tone if it precedes another syllable. However, in the written language syllables are marked in their original tones. Hyphens (-) connect syllables to form words. No diacritic mark is needed for tones 1 and 4.

www.speakhokkien.org

APPENDIX B EVALUATION LOGS

Phonemes Needed to be NaN

pe	puánn	mē	sen	hèn	tshên	hàm	thà	lam	khíám
pé	puānn	kē	en	phèn	sên	hiah	thènn	lím	gu
pê	png	uinn	hén	lèn	ên	phiánn	thiánn	līm	giō
pu	pén	tsng	bén	kèn	pēn	phék	niû	liám	geng
piann	pèh	thuīnn	tén	gèn	hēn	bui	nah	lín	gûn
puánn	pueh	hen	lén	sèn	tēn	biāu	ngèh	liah	tsínn
puānn	ne	phen	kén	èn	lēn	bòng	jiám	kiā	tsiánn
png	le	ten	khén	hên	kēn	bēn	ā	kue	tsuánn
pe	mé	then	gén	bên	gēn	bèh	ínn	kēnn	tsuánn
pé	ngé	ken	tshén	tên	sēn	bak	iànn	kēnn	tsám
pê	mê	khen	sén	lên	hiâ	méh	iaunn	kuīnn	tshā
pu	khê	pen	én	khên	hainn	tooh	iàm	kām	tshôo
piann	gê	tshen	pèn	gên	huánn	tiap	uèh	khuīnn	tshèeh
tshuh	sau	siò	sáng	siàng					

Phonemes and Graphemes Needed to be Added

✓✓ 92 伯 phek	✓✓ 2203 琶 pē	✓✓ 4102 賴 phué	✓✓ 2009 潮 tiô
✓✓ 204 債 tsèe	✓✓ 2216 瑪 bée	✓✓ 4102 賴 phé	✓✓ 2132 牒 tiáp
✓✓ 404 叉 tshee	✓✓ 2411 砂 see	✓✓ 4149 魄 phek	✓✓ 2159 欸 gīn KYM
✓✓ 504 唆 sō	✓✓ 2438 碧 phek	✓✓ 4150 辟 phek	✓✓ 2170 獠 ngiáu KYM
✓✓ 505 唇 tún	✓✓ 2443 碼 bée	✓✓ 4151 劈 phek	✓✓ 2235 甌 bót change to KYM
✓✓ 521 問 muī	✓✓ 2664 紗 see	✓✓ 4152 霹 phek	✓✓ 2268 畫 uā
✓✓ 528 啉 tshiunn -> tshiù	✓✓ 2771 罵 mēe	✓✓ 575 嚙 ngiáu KYM	✓✓ 2290 疾 tsék UCT
✓✓ 529 喂 uê	✓✓ 2810 耙 pē	✓✓ 596 囧 buê	✓✓ 2319 癩 khuê
✓✓ 542 啲 iō	✓✓ 2810 耙 pē	✓✓ 653 場 tiáunn	✓✓ 2436 牒 tiáp
✓✓ 543 喻 jū	✓✓ 3163 蝦 hē	✓✓ 749 姆 m	✓✓ 2652 糶 thiò
✓✓ 711 夾 ngeeh	✓✓ 3226 裳 see	✓✓ 852 寂 tsék UCT	✓✓ 2874 脍 tshuinn KYM
✓✓ 712 命 phānn	✓✓ 3470 躡 phīn	✓✓ 890 就 tsiū UCT	✓✓ 2899 膜 mòoh
✓✓ 750 姊 tsé	✓✓ 3492 艘 nng	✓✓ 909 脛 ban	✓✓ 2952 艾 ngāi
✓✓ 1224 扒 pee	✓✓ 3527 辦 pīnn	✓✓ 945 川 tshuinn KYM	✓✓ 2964 糠 khng
✓✓ 1224 扒 pē	✓✓ 3723 鑽 tsuinn	✓✓ 1008 康 khng	✓✓ 3112 藹 ái
✓✓ 1244 把 pé	✓✓ 3793 雅 ngéc	✓✓ 1018 廠	✓✓ 3166 蝶 tiáp

APPENDIX

		tshiáng LTR	
✓✓ 1270 拍 phək	✓✓ 3820 霞 hêe	✓✓ 1038 張 tiaunn	✓✓ 3189 蠕 lùn
✓✓ 1615 杷 pèe	✓✓ 3832 靶 pée	✓✓ 1238 扶 ío	✓✓ 3223 裙 gun KYM
✓✓ 1629 枷 kêe	✓✓ 3885 飄 phiau	✓✓ 1297 挑 thio	✓✓ 3295 話 uā
✓✓ 1741 標 phiau	✓✓ 3886 飄 phiau	✓✓ 1321 掬 kák ATJ	✓✓ 3325 謀 tiáp
✓✓ 1866 沙 see	✓✓ 3916 馬 bée	✓✓ 1382 搶 tshiáng LTR	✓✓ 3407 賊 tsék LTR
✓✓ 1866 沙 sée	✓✓ 4060 齋 tsee	✓✓ 1403 擱 nuá	✓✓ 3460 蕘 seh UCT
✓✓ 1879 泊 phək	✓✓ 4074 拐 khêe	✓✓ 1407 撓 ngiáu KYM	✓✓ 3475 蹠 tshiáng KYM
✓✓ 1947 渣 tsee	✓✓ 4102 賴 phué	✓✓ 1554 暖 juán	✓✓ 3489 賬 lò KYM
✓✓ 1988 漂 phiau	✓✓ 4102 賴 phé	✓✓ 1684 械 kāi UTC	✓✓ 3586 遮 jia UCT
✓✓ 2117 爬 pèe	✓✓ 4149 魄 phək	✓✓ 1718 楹 ênn CLQ	✓✓ 3606 郎 nng CLQ
✓✓ 2122 爸 pèe	✓✓ 4150 辟 phək	✓✓ 1941 淺 khín	✓✓ 3674 缺 giap KYM
✓✓ 2161 猛 mée	✓✓ 4151 劈 phək	✓✓ 1976 潑 kō ATJ	✓✓ 3726 長 tíaunn
✓✓ 3902 餒 lué UCT	✓✓ 964 帕 phèe UCT	✓✓ 1720 榆 jiū UCT	✓✓ 3348 諱 hooh KYM
✓✓ 3943 髓 tshué CLQ	✓✓ 980 幔 mua KYM	✓✓ 1857 沁 tshim UCT	✓✓ 3441 趙 tiō UCT
✓✓ 3995 烏 ngiáu ATJ	✓✓ 1202 戎 luang KYM	✓✓ 1977 滓 tsáinn UCT	✓✓ 3537 远 hānn ATJ
✓✓ 4128 飭 khiū ATJ	✓✓ 1272 拈 khinn UCT	✓✓ 2326 癩 tiò KYM	✓✓ 3603 邵 siō UCT
✓✓ 4133 滲 gàn UCT	✓✓ 1354 揉 jiū UCT	✓✓ 2427 硯 hīnn UCT	✓✓ 3648 釣 tiò CLQ
✓✓ 4183 慄 lek	✓✓ 1364 擲 tshih ATJ	✓✓ 2653 亂 khiú ATJ	✓✓ 3660 鉗 khinn CLQ
✓✓ 4212 籍 tsék UCT	✓✓ 1375 摸 khiú ATJ	✓✓ 2685 絢 kù ATJ	✓✓ 3728 門 tshuànn UCT
✓✓ 4217 斜 tshiá CLQ	✓✓ 1440 擲 tshih KYM	✓✓ 2728 縞 gue ATJ	✓✓ 3769 睡 sè KYM
✓✓ 4227 嘎 sà UCT	✓✓ 1497 斬 tsánn CLQ	✓✓ 2812 耳 hīnn	✓✓ 3782 隙 khiah UCT
✓✓ 844 宰 tsái UCT	✓✓ 1528 昕 hin UCT	✓✓ 2961 苳 mún ATJ	✓✓ 3906 餡 ānn UCT

APPENDIX

✓✓ 864 寤 út ATJ	✓✓ 1637 柔 jiû UCT	✓✓ 3073 蔣 tsiáunn UCT	✓✓ 3981 繇 jiû UCT
✓✓ 3981 繇 jiû UCT	✓✓ 4072 惆 gu KYM	✓✓ 4085 瘡 íg KYM	
✓✓ 4057 𪗇 kônn KYM	✓✓ 4081 擊 tsānn UCT	✓✓ 4099 𪗇 hānn ATJ	


APPENDIX C FYP 2 Poster

PRONUNCIATION MODELLING OF PENANG HOKKIEN DIALECT FOR TEXT-TO-SPEECH SYSTEM

NAME: LIM KANG JIE 18ACB02259
SUPERVISOR: DR JASMINA KHAW YEN MIN

檳城閩南語
文字轉語音
系統兮發音
建模研究

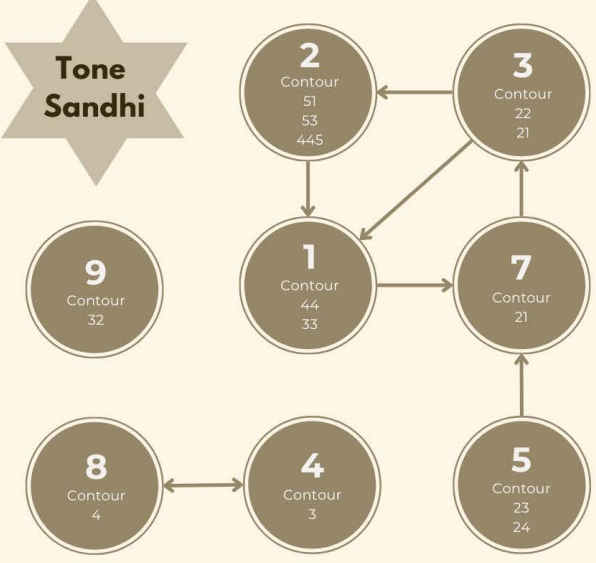
PENANG HOKKIEN




STANDARDIZE
ORTHOGRAPHIES
PRONUNCIATIONS
TONES

COLLECT
PHONEMES
GRAPHEMES
ARTICLES
AUDIOS

Tone Sandhi



Tones Graph




衫 短 褲 闊 人 矮 鼻 直 粥
sann té khò khuah lâng é phinn tit tsúk

Dictionaries


廈英大辭典
Chinese English dictionary of Vernacular or Spoken Language of Amoy
教育部臺灣閩南語常用詞辭典
Ministry of Education of Republic of China
Taiwanese Hokkien Online Dictionary

RESULTS IN DATABASE

- 2110 Categorized Phonemes/Morphemes
- 4250 Graphemes
- 5762 Categorized Mapped Graphemes with Phonemes



Revitalization of our mother tongues is our duty



✨ ✨ 甲乙丙丁戊己庚辛壬癸子丑寅卯辰巳午未申酉戌亥 ✨ ✨

APPENDIX D

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Year 4 Trimester 1	Study week no.: Week 4
Student Name & ID: Lim Kang Jie 18ACB02259	
Supervisor: Dr Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System	

1. WORK DONE Generate Possible combination of phonemes and graphemes
2. WORK TO BE DONE Check the reliability of the combination
3. PROBLEMS ENCOUNTERED No
4. SELF EVALUATION OF THE PROGRESS 3 out of 5



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Year 4 Trimester 1	Study week no.: Week 6
Student Name & ID: Lim Kang Jie 18ACB02259	
Supervisor: Dr Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System	

1. WORK DONE Check the reliability of the combination
2. WORK TO BE DONE Research on the Tone Sandhi Collect Graphemes
3. PROBLEMS ENCOUNTERED No
4. SELF EVALUATION OF THE PROGRESS 3 out of 5



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Year 4 Trimester 1	Study week no.: Week 8
Student Name & ID: Lim Kang Jie 18ACB02259	
Supervisor: Dr Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System	

1. WORK DONE Research on the Tone Sandhi Collect Graphemes
2. WORK TO BE DONE Mapping Graphemes with Phonemes/Morphemes
3. PROBLEMS ENCOUNTERED Huge amounts of records need to map manually
4. SELF EVALUATION OF THE PROGRESS 1 out of 5



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Year 4 Trimester 1	Study week no.: Week 13
Student Name & ID: Lim Kang Jie 18ACB02259	
Supervisor: Dr Jasmina Khaw Yen Min	
Project Title: Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System	

1. WORK DONE Mapping Graphemes with Phonemes/Morphemes
2. WORK TO BE DONE Import Tables into Datab
3. PROBLEMS ENCOUNTERED Huge amount of records need to map manually
4. SELF EVALUATION OF THE PROGRESS 1 out of 5



Supervisor's signature



Student's signature

PLAGIARISM CHECK RESULT

PRONUNCIATION MODELLING OF PENANG HOKKIEN DIALECT FOR TEXT-TO-SPEECH SYSTEM

ORIGINALITY REPORT

3 %	3 %	1 %	1 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.papercamp.com Internet Source	1 %
2	penanghokkien.com Internet Source	1 %
3	Vanessa Lim, Hui Shan Ang, Estelle Lee, Boon Pang Lim. "Towards an Interactive Voice Agent for Singapore Hokkien", Proceedings of the Fourth International Conference on Human Agent Interaction, 2016 Publication	1 %
4	medium.com Internet Source	<1 %
5	www.ee.cuhk.edu.hk Internet Source	<1 %
6	en.wikipedia.org Internet Source	<1 %
7	Submitted to University of Malaya Student Paper	<1 %

PLAGIARISM CHECK RESULT

8	umpir.ump.edu.my Internet Source	<1 %
9	Submitted to Indira Gandhi Delhi Technical University for Women Student Paper	<1 %
10	esg.tsmc.com Internet Source	<1 %
11	newnaratif.com Internet Source	<1 %
12	www.arcademachines.xyz Internet Source	<1 %

Exclude quotes On

Exclude matches < 8 words

Exclude bibliography On

Form Title: Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1 of 1



FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	LIM KANG JIE
ID Number(s)	18ACB02259
Programme / Course	IA
Title of Final Year Project	Pronunciation Modelling of Penang Hokkien Dialect for Text-To-Speech System

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceed the limits approved by UTAR)
Overall similarity index: <u> 3 </u> % Similarity by source Internet Sources: <u> 3 </u> % Publications: <u> 1 </u> % Student Papers: <u> 1 </u> %	
Number of individual sources listed of more than 3% similarity: <u> 0 </u>	
Parameters of originality required, and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note: Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Jasmina Khaw Yen Min

Date: 9 September 2022

Signature of Co-Supervisor

Name: _____

Date: _____


FYP 2 CHECKLIST**UNIVERSITI TUNKU ABDUL RAHMAN****FACULTY OF INFORMATION & COMMUNICATION
TECHNOLOGY (KAMPAR CAMPUS)****CHECKLIST FOR FYP2 THESIS SUBMISSION**

Student Id	18ACB02259
Student Name	LIM KANG JIE
Supervisor Name	DR JASMINA KHAW YEN MIN

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Front Plastic Cover (for hardcopy)
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.



(Signature of Student)

Date: 09/09/2022