

國立交通大學

資訊科學與工程研究所

碩士論文

漢語間統計式機器翻譯語料處理—用臺灣
閩南語示範

Corpus Preprocessing for Statistical Machine
Translation between the Chinese Languages - Using
Taiwan Southern Min as Examples

研究生： 薛丞宏

指導教授： 張智星教授

易志偉教授

中華民國103年11月

踏話頭

臺灣是_ㄟ一_ㄟ個_ㄟ世_ㄟ多_ㄟ元_ㄟ民_ㄟ族_ㄟ，多_ㄟ元_ㄟ語_ㄟ言_ㄟ的_ㄟ國_ㄟ家_ㄟ
tâi uân sí tsit ê to guân bìn tsok to guân gí giân ê kok ka

。講_ㄟ母_ㄟ語_ㄟ，使_ㄟ用_ㄟ母_ㄟ語_ㄟ是_ㄟ上_ㄟ基_ㄟ本_ㄟ的_ㄟ權_ㄟ利_ㄟ，毋_ㄟ
kóng bó gí sú iōng bó gí sī siāng ki pún ê kuân lī m̄

過_ㄟ母_ㄟ語_ㄟ的_ㄟ電_ㄟ腦_ㄟ相_ㄟ關_ㄟ應_ㄟ用_ㄟ煞_ㄟ誠_ㄟ少_ㄟ，需_ㄟ要_ㄟ加_ㄟ
koh bó gí ê tiān náu siong kuan òng iōng suah tsiānn tsíó - su iàu ka

強_ㄟ自_ㄟ然_ㄟ語_ㄟ言_ㄟ處_ㄟ理_ㄟ的_ㄟ研_ㄟ究_ㄟ恰_ㄟ語_ㄟ料_ㄟ收_ㄟ集_ㄟ整_ㄟ理_ㄟ
kiōng tsū lián gí giân tshú lí ê giân kiù kap gí liáu siu tsip tsing lí

。臺_ㄟ灣_ㄟ本_ㄟ土_ㄟ語_ㄟ言_ㄟ百_ㄟ百_ㄟ種_ㄟ，本_ㄟ論_ㄟ文_ㄟ是_ㄟ針_ㄟ對_ㄟ閩_ㄟ
tâi uân pún thóo gí giân pah pah tsiong pún lūn bûn sī tsiam tui bân

南_ㄟ語_ㄟ，研_ㄟ究_ㄟ伊_ㄟ翻_ㄟ譯_ㄟ語_ㄟ料_ㄟ的_ㄟ特_ㄟ性_ㄟ。除_ㄟ了_ㄟ閩_ㄟ南_ㄟ
lām gí giân kiù i huan ik gí liáu ê tik sing tui liáu bân lām

語_ㄟ本_ㄟ身_ㄟ以_ㄟ外_ㄟ，嘛_ㄟ希_ㄟ望_ㄟ研_ㄟ究_ㄟ結_ㄟ果_ㄟ對_ㄟ別_ㄟ的_ㄟ本_ㄟ
gí pún sī í guā mā hī bāng giân kiù kiāt kó tui pat ê pún

土_ㄟ語_ㄟ言_ㄟ有_ㄟ幫_ㄟ助_ㄟ。
thóo gí giân ū pang tsōo

本_ㄟ論_ㄟ文_ㄟ提_ㄟ出_ㄟ一_ㄟ個_ㄟ自_ㄟ動_ㄟ整_ㄟ理_ㄟ漢_ㄟ語_ㄟ語_ㄟ料_ㄟ的_ㄟ
pún lūn bûn thê tshut tsit ê tsū tōng tsing lí hàn gí gí liáu ê

方_ㄟ法_ㄟ，予_ㄟ資_ㄟ訊_ㄟ無_ㄟ完_ㄟ整_ㄟ的_ㄟ語_ㄟ料_ㄟ庫_ㄟ補_ㄟ足_ㄟ資_ㄟ訊_ㄟ，
hong huat hōo tsu sīn bō uân tsing ê gí liáu khòo póo tsio̍k tsu sīn

發_ㄟ揮_ㄟ上_ㄟ大_ㄟ的_ㄟ價_ㄟ值_ㄟ，BLEU 分_ㄟ數_ㄟ對_ㄟ9.30 換_ㄟ到_ㄟ13.82。另_ㄟ
huat hui siāng tuā ê kè tat hun sòo tui giú kàu līng

外_ㄟ閣_ㄟ用_ㄟ實_ㄟ驗_ㄟ證_ㄟ明_ㄟ平_ㄟ行_ㄟ語_ㄟ料_ㄟ數_ㄟ量_ㄟ無_ㄟ到_ㄟ十_ㄟ萬_ㄟ
guā koh òng sít giām tsing bing ping hing gí liáu sòo liōng bó kàu tsap bân

句_ㄟ的_ㄟ時_ㄟ，加_ㄟ語_ㄟ料_ㄟ對_ㄟ翻_ㄟ譯_ㄟ的_ㄟ效_ㄟ果_ㄟ影_ㄟ響_ㄟ非_ㄟ常_ㄟ
kù ê sī ka gí liáu tui huan ik ê hāu kó íng hióng hui siōng

大_ㄟ，原_ㄟ本_ㄟ64121 句_ㄟ加_ㄟ到_ㄟ99147 句_ㄟ了_ㄟ後_ㄟ，BLEU 分_ㄟ數_ㄟ對_ㄟ
tuā guān pún kù ka kàu kù liáu āu hun sòo tui

13.82 提_{ㄊㄧˊ}昇_{ㄕㄨㄥˊ}到_{ㄉㄠˋ} 19.33 。

關_{ㄍㄨㄢˊ}鍵_{ㄑㄩㄢˊ}字_{ㄗˋ}：臺_{ㄊㄞˊ}灣_{ㄨㄢˊ}閩_{ㄇㄢˊ}南_{ㄋㄢˊ}語_{ㄩˊ}、華_{ㄏㄨㄚˊ}語_{ㄩˊ}、翻_{ㄈㄢˊ}譯_{ㄩˊ}、語_{ㄩˊ}料_{ㄌㄠˊ}、

斷_{ㄊㄨㄢˊ}詞_{ㄘㄨˊ}、語_{ㄩˊ}言_{ㄍㄢˊ}分_{ㄈㄢˊ}類_{ㄌㄟˊ}



摘要

臺灣是一個多元文化、多元語言的國家。講母語、使用母語是最基本的權利，不過母語的電腦相關應用卻很少，需要加強自然語言處理的研究和語料收集整理。臺灣本土語言很多種，本論文是針對閩南語，研究閩南語翻譯語料的特性，除了閩南語本身以外，也希望研究結果對別的本土語言有幫助。

本論文提出一個自動整理漢語語料的方法，讓資訊不完整的語料庫補足資訊，發揮最大的價值，BLEU 分數從 9.30 拉到 13.82。另外證明平行語料數量不到十萬句的時候，增加語料對翻譯的效果影響非常大，原本 64121 句加到 99147 句之後，BLEU 分數從 13.82 提昇到 19.33。

關鍵字：臺灣閩南語、華語、翻譯、語料、斷詞、語言分類

Abstract

Taiwan is a multi-culture and multi-language country. Speaking in mother tongues is a basic human right, but there are few computer applications for mother languages. The applications are supported by corpus and research of natural language processing. There are many local languages in Taiwan. This thesis focuses on Southern Min Taiwanese, is major local language in Taiwan. It contains research into corpus preprocessing to get good performance in statistical machine translation. We wish it can help the computational linguistic research of other local language of Taiwan.

This thesis introduces a method to preprocess the corpus whose information is lacking. After refining, the BLEU score is raised from 9.30 to 13.82. Experiments in this thesis show that translation performance is sensitive to the amount of parallel corpus when the amount of parallel corpus sentences is less than 100,000. The BLEU score raises from 13.82 to 19.33 as the amount of sentences increased from 64121 to 99147.

Keyword: Southern Min, Taiwanese, Mandarin, Chinese, Translation, Corpus, Segmentation, Language Identification

勞力

佇遮就愛先感謝我厝內的人，支持我做臺灣母語的研究，我拄著問題時敲電話轉去員林，個攏誠有耐性教我閩南語。蔡文莉除了關心我的研究以外，閣陪我學閩南語、客話佮 Tayal，予我佇這條路頂有一个伴。

多謝指導老師易志偉佮張智星教授，支持我研究臺灣母語，予我家已決定研究的方向。陳孟彰、高明達、呂仁園佮楊允言老師，感謝個予我建議，而且提供我閩南語的語料佮資源。多謝張俊盛老師佮劉昭宏教我翻譯的做法，林政源、邱祈添、陳江村，林奇嶽學長教我語音的技術，予我一息仔就有法度入去這兩個領域，對自然語言處理閣較了解。

上尾感謝劉勝權老師、張光宇老師佮莊淨婷學姐，予我開始思考研究臺灣母語，嘛教我有關漢語佮語言學的智識，這篇論文才有遮爾濟養份。

閣有捌鬥相共的親朋好友，丞宏佇遮共恁說多謝！！

目錄

踏話頭	i
摘要	iii
Abstract	iv
勞力	v
目錄	vi
圖目錄	xi
表目錄	xii
演算法目錄	xiv
1 研究背景	1
1.1 研究目的	3
1.2 語料狀況	4
1.3 論文貢獻	5
1.4 論文架構	5
2 相關研究	7
2.1 音標系統	7

2.1.1	記音系統	7
2.1.2	拼音系統	7
2.2	語音學	9
2.3	音韻學	11
2.3.1	共時音韻學	12
2.3.2	歷史音韻學	14
2.4	漢語聲韻學	16
2.4.1	韻冊	16
2.4.2	古早中國語音紀錄	16
2.5	機器學習	17
2.5.1	分類器	17
2.5.2	隱性馬可夫模型	18
2.6	語音辨識	18
2.6.1	HTK	18
2.6.2	Kaldi	19
2.7	語音合成	19
2.7.1	模型合成	19
2.7.2	接音合成	20
2.7.3	變調恰重音預測	20
2.7.4	相關系統	21
2.8	斷詞	22

2.8.1	長詞優先斷詞	22
2.8.2	評分方式	23
2.8.3	相關系統	23
2.9	剖析	24
2.9.1	相關系統	24
2.10	翻譯	24
2.10.1	對齊模型	25
2.10.2	語言模型	25
2.10.3	解碼器	26
2.10.4	評分方式	26
2.11	語料收集整理	27
2.11.1	語言分類	27
2.11.2	平行語料語句對齊	27
2.12	語料庫	27
2.12.1	閩南語語料種類	28
2.12.2	閩南語語料—教育部辭典	29
2.12.3	閩南語語料—新聞語料庫	29
2.12.4	閩南語語料—臺文典藏	30
2.12.5	閩南語語料—TGB 通訊臺灣組合	31
3	研究介紹	33
3.1	閩南語斷詞	33

3.2	未知詞問題	34
3.3	整理語料	34
3.4	語言分類	35
4	研究方法	36
4.1	挂好長度斷詞	36
4.2	未知詞另外翻譯	37
4.3	漢羅全羅對齊	39
4.4	補全漢俗全羅	40
4.5	語言分類特徵	40
5	實驗結果	43
5.1	閩南語斷詞實驗	43
5.2	語料整理實驗	44
5.3	分類語言實驗	46
5.4	加入 TGB 語料庫實驗	47
5.5	斷詞樣式俗斷字樣式的翻譯結果實驗	48
6	結論俗未來發展	51
6.1	結論	51
6.2	機器校對	52
6.3	斷詞	53
6.4	字幕辨識	53



圖目錄

2.1	逐門自然語言處理研究項目對應的相關語言學	8
4.1	未知詞另外翻譯流程	38
4.2	判斷語言流程	42
5.1	互相整理流程	46
5.2	無仝特徵詞數量，分類 3741 段閩南語華語	47
5.3	斷字俗斷詞語料的翻譯效果比較—BLEU 用詞拍分數	48
5.4	斷字俗斷詞語料的翻譯效果比較—BLEU 用字拍分數	49

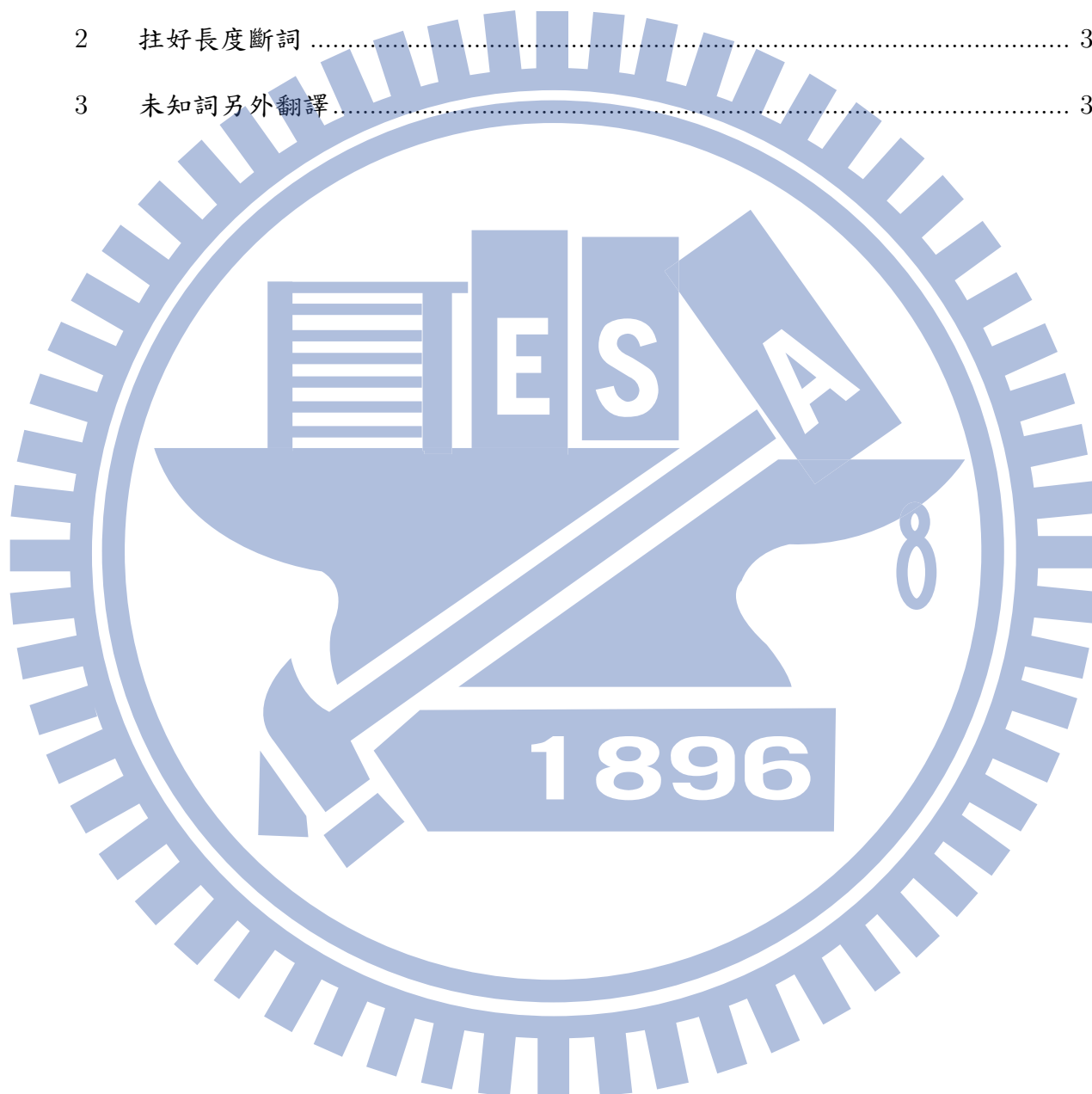
表目錄

2.1	元音表	10
2.2	音節分析	10
2.3	閩南語上細配對	11
2.4	閩南語濁聲母無鼻化的選擇	13
2.5	閩南語濁聲母有鼻化的選擇	13
2.6	閩南語低元音提昇過程 [1]	15
2.7	定用的分類器優缺點	17
2.8	臺灣母語語音合成相關研究、系統	21
2.9	召回率、精確率、F- 測量的公式	23
2.10	翻譯結果一、翻譯結果二對答案的 BLEU 分數	26
2.11	自然語言處理技術需要的語料庫	28
2.12	閩南語語料種類比較表	28
2.13	教育部辭典例句	29
2.14	臺文典藏語料漢羅、全羅對照	31
2.15	TGB 通訊語料狀況	32
3.1	未知詞問題範例	34
3.2	語料庫狀況	35
3.3	語言分類範例	35

4.1	長詞優先毋著的情形	36
4.2	挂好長度成本	38
4.3	長詞優先比挂好長度斷詞較好的狀況	38
4.4	$n=7000$ 、 $m=3000$ 的頭前九個定用詞俗特徵詞	41
5.1	實驗工具版本	43
5.2	閩南語斷詞的效果	44
5.3	新聞語料庫俗臺文典藏互相整理的實驗	44
5.4	新聞語料庫佇互相整理的變化	45
5.5	臺文典藏佇互相整理的變化	45
5.6	加入 TGB 語料的翻譯效果	47
6.1	字幕辨識問題分析	53

演算法目錄

1	長詞優先斷詞方法	22
2	拄好長度斷詞	37
3	未知詞另外翻譯	39



第一章 研究背景

臺灣是多元民族的國家，逐個民族攏有家己的文化，嘛無仝，有人講閩南語，有人講泰雅語，有人講客話……最近二十幾年來閣有越南話、印尼話……新住民¹的語言。臺灣語言主要會當分做南島語（Austronesian）佻漢語（Chinese）兩種²。

南島語是全世界分布上闊的語族 [3]，上北到臺灣、南到紐西蘭，東到復活節島，西到馬達加斯加。民族語言網 [4] 共南島語分做十個族群，臺灣摻蘭嶼這十個族群攏有，語系變化誠大，代表臺灣佇南島語的研究內底非常重要，所以臺灣的原住民語系變化誠大。

個閣比漢人較早到臺灣，才會予人叫做原住民。佇歷史的文書記錄面頂，十七世紀荷蘭佻西班牙來臺灣進前就有一個幾仔族原住民組合起來的大肚王國 [5]，個保護個家己的土地，抵抗荷蘭人，鄭成功政權，清國的統治，到大甲西社事件才滅國。到今仔日，臺灣的原住民除了中華民國政府的原民會認定的十六族以外，閣有誠濟猶未認定的族，親像臺南的西拉雅佻埔里的噶哈巫……攏總有二三十族以上。

臺灣用的漢語主要會當分做三大種，閩南語、客家話佻官話 [6]。閩南語佻客家話是對四百外年前開始，佇明國、清國的福建人因為枵腹肚蹣袂落，姑不而將駛船渡過烏水溝³，來臺灣趁食。佇清國的時陣透過政治佻經濟的力量一步一步食掉原住民的土地，上尾佇臺灣變做上大的族群之一。

講著閩南語的文字，定定有人問：「閩南語是欲按怎寫！？」除了用拼音的方式寫出來以

¹佇 2013 年提著中華民國國籍 5004 人中，對越南來的 3855 人，對印尼來的 566 人 [2]

²新住民的越南語是南亞語系，泰國語是狀侖語系

³又叫臺灣海峽

外，閩南語有九成以上是有漢字的 [7]，鶴佬人⁴祖先搬去閩越地區時，就佢遐的壯苗族原住民通婚，因為原住民人濟，所以予鶴佬人同化的時留落來這一寡毋是漢語的詞，董忠司 [8] 認為「查甫」佢「查某」的「查」，「大家」佢「大官」的「大」是狀侖族的詞頭。雖然「查」大部份人攞讀「tsa」，毋過佇臺灣漢語辭典 [9] 有記錄「ta」的音，佢「大」全音。

張光宇 [10] 研究歷史音韻學佢聲韻學⁵認為，閩南語的漢語部份是西晉後尾動亂的兩擺大移民、唐朝陳元光派兵鎮壓狀苗族原住民，佢南宋時期文教影響，攞總四擺移民，佢政府制度影響，造成四層漢語語言層的閩南語。頭前三層語言層的語音叫做白話音，上尾第四擺後南宋音號做文讀音，親像「石」這個字，有「石^ㄕ頭^ㄊ」⁶、「石^ㄕ榴^ㄌ」⁶、「藥^ㄩ石^ㄕ」⁷三種語音，佇語言是規律變化的假設之下，這個「石」字就代表上少有三層語言層。

客家話這馬佇臺灣較大腔口有「四海大平安」⁸。鶴佬人佢客人毋管佇亞洲大陸抑是臺灣，生活攞無遠，誠濟詞的用法攞相像，親像閩南語講「頭^ㄊ前^ㄑ」，四縣客話講「頭^ㄊ前^ㄑ」⁹，閩南語講「好^ㄏ勢^ㄕ」，四縣客話講「好^ㄏ勢^ㄕ」。

佇臺灣的鶴佬人佢客人來臺灣佢原住民通婚，生活中嘛濫著原住民的用詞，親像「臺灣」是西拉雅語「Taian」或「Tayan」對外地人的稱呼 [13]，咱定定食著的「菝仔」嘛是平埔族話 [14]。

毋但原住民話，閩南語佢客話嘛有佢別的语言交插，像是的「六甲地」的「甲」是從荷蘭「akker」來的 [15]。受過日本的統治，閩南語佢客話攞有濫著日語，親像「臭柿仔」，

⁴閩南人。漢字有爭議，教育部無訂落來，採用洪惟仁教授的用字慣勢

⁵相關理論請看2.3節佢2.4節

⁶一種果子

⁷方藥與砒石

⁸四縣腔、海陸腔、大埔腔、饒平腔、詔安腔

⁹客話拼音會當看客話拼音 [11]，意思會當查客話辭典 [12]

日語唸「ト マ ト」，閩南語閣會當講「ト マ ト」¹⁰，客話講「ト マ ト」，「オートバイ」
to ma to *kha7 *ma1 *tooh4 toˊ ma do oo to bai

閩南語講「オート バイ」俗「機車」，客話講「オートバイ」。會當講是臺灣的歷史藏佇臺灣的語言內底。
*oo1 *too7 *bai2 o do bai

上尾一个漢語官話是中華民國政府拍輸中華人民共和國矣，恁誠濟中國人來臺灣，因為逐个的故鄉攏無全 [6]，民國政府就繼續用佇亞洲大陸的規範，共北京官話的語音、白話文的用法當做基礎，利用政府機關，教育認知¹¹ …等手法佇臺灣揀，所以華語這馬是佇臺灣是上有政治優勢的語言。毋過這馬佇臺灣實際用的官話閣恰北京用的官話無全款，有加入臺灣本土的元素，是北京官話的次方言之一。除了這三種漢語以外，佇二次大戰了中華民國政府恁來臺灣的人，個的母語嘛誠濟種 [6]，毋過這馬攏已經消失甲差不多矣。

為著方便起見，本論文下跤的閩南語恰客話攏是講佇臺灣用的閩南語恰客話，官話就照中國海外華人的慣勢，稱呼佇臺灣使用的北京官話次方言做華語。本論文主要用閩南語寫，毋過會用著一寡華語恰客話。為著格式一致，予人有法度一看就知影是啥物語言，閩南語會用臺羅拼音 [16] 恰方言音符號 [17]，客話會用客話拼音 [11]，華語會用注音符號 [17]。引用的閩南語的羅馬拼音攏會轉臺羅，毋過人名、文章名、冊名保持原樣。

1.1 研究目的

最近十幾冬政府開始注重人權，講母語是人上基本的權利，毋但國校仔國中攏開始有鄉土語言的課矣¹²，逐家嘛較重視研究母語¹³，毋過定定會拄著文字、語料、教材數量無

¹⁰ 音標有 * 代表外來詞，頭前兩音節 kha7、ma1 免閣變調

¹¹ 本人阿姨有佇學校講母語予人罰過錢

¹² 除了本土語言以外，閣有新住民語言

¹³ 請看2章

夠的問題。

親像電視台想欲製做母語的新聞，毋過大部份攞是華語的材料；學生想欲知影華語一句話，母語按怎講，這時陣就會當利用華語資料攞誠濟的優勢，共華語翻譯做母語，按呢就會解決這幾項問題。

這馬翻譯的技術已經發展到一個坎站矣，毋過華語翻譯做母語的效果攞無蓋好，主要是母語的語料數量無夠，本論文就針對華語到閩南語的翻譯，研究按怎處理數量有限的閩南語語料，予翻譯閣較好。予後壁的人會使利用這個成果，繼續研究閩南語，抑是會當利用這篇文章的經驗，推廣到別的臺灣語言，上直接的，就是會當予母語使用者有閣較方便的數位環境。

1.2 語料狀況

佇處理閩南語語料進前，都愛先了解閩南語語料的歷史。

上早的閩南語文體是明國時代流傳落來的荔鏡記戲文 [18]，清國時代有閣較濟的字典、歌仔冊、教會詩歌，日本時代開始有人感受閩南語消失的威脅，產生出臺灣話文論戰，到中日戰爭開始，日本人禁止用漢文。中華民國政府來臺灣隨就二二八，臺灣文學因為白色恐怖，到最近二十幾冬，閩南語才開始有大量的文章。

因為歷史誠久長，閩南語的書寫方式有誠濟種，大約會當分做三種。第一種全部用漢字，叫做全漢，就全部用漢字表達閩南語，因為閩南語有部份毋是漢語，拄著這種情形逐家用的漢字攞無全，到最近幾冬，才有教育部以官方單位規範用字¹⁴。

第二種是用拼音，佇清國時期，基督教的傳教士為著學閩南語，向鶴佬人傳教，定一

¹⁴臺灣閩南語推薦用字 700 字表 [19]，佇 96 ～ 99 年公佈修正

套閩南語的羅馬拼音，一般號做「教會羅馬拼音」¹⁵。日本時期，日本政府嘛是為著統治原因，用日本的假名來記錄閩南語 [20][21]。到中華民國時期，有模仿華語注音符號的方言音符號。最近幾十年，有主張佢英文教學系統較倚的通用拼音。上尾教育部改教羅羅馬拼音的缺點¹⁶，號做「臺灣羅馬字拼音」，後壁號做「臺羅」。若規篇文章攞用拼音，就號做「全羅」。

面頂兩種表示法攞有缺點，上深的漢字抑是全篇的拼音對無學過拼音的人來講較歹接受，所以第三種就是共漢字佢拼音濫做伙，主要是漢字，拄著揣無漢字本字的，就寫拼音。

1.3 論文貢獻

本論文主要有四個對翻譯語料的貢獻，第一個是本論文提出「拄好長度斷詞」的演算法。第二個貢獻是比較漢語語料樣式對翻譯的影響，比較斷詞佢斷字翻譯模型，佢按怎的組合之下效果上好。第三個是提出一個自動整理漢語語料的方法，予資訊無完整的語料庫補足資訊，發揮上大的價值。上尾一個是提出分類兩種漢語的方法，免用傷濟特徵詞就會當得著袂稔的分類效果，佢網路頂掠落來的資料就有法度分類。

1.4 論文架構

第2章介紹語言學研究佢技術原理。第3章定義本篇論文愛處理的問題，分別是斷詞方法的比較、翻譯的未知詞問題、補語料庫資訊佢語言分類。第4章提出拄好長度斷詞、未

¹⁵幾仔个版本，一般是講打馬字的版本

¹⁶毋是一擺就改好，中央閣有 TLPA 拼音

知詞另外翻譯、補全漢全羅佻增加特徵詞，解決頂一章翻譯語料的問題。第5章做實驗而且驗證結果。上尾的結論佻以後會當發展的方向攏記佇第6章。



第二章 相關研究

華語到閩南語的翻譯是自然語言處理（Natural Language Processing）¹的一部份，佇處理自然語言的時陣就需要目標俗語言學相關的智識，會當簡單整理做圖2.1，無仝的自然語言研究方向，愛知影的語言學智識嘛無仝。繼落來就介紹語言學俗自然語言處理的研究文獻俗母語的研究狀況：

2.1 音標系統

音標系統有兩種，一種是研究語音用的記音系統，一種是予一般人拼寫用的拼音系統。

2.1.1 記音系統

研究一个語言，愛先了解伊的語音，這時陣就需要一个標準化的記音符號。這馬上時行的是國際語言學會（International Phonetic Association）制定的國際音標（International Phonetic Alphabet, IPA）[22]，毋管記錄的語言有仝款無，只要語音的特徵仝款，就會用仝款的符號。

2.1.2 拼音系統

記音的音標符號較濟，對指定的自然語言，誠濟符號根本用袂著，用起來無方便。實際寫文章、編教材大部份會用另外的音標系統。下跤照歷史年代簡單介紹閩南語主要三種拼音系統：

¹人講的話攞是自然語言

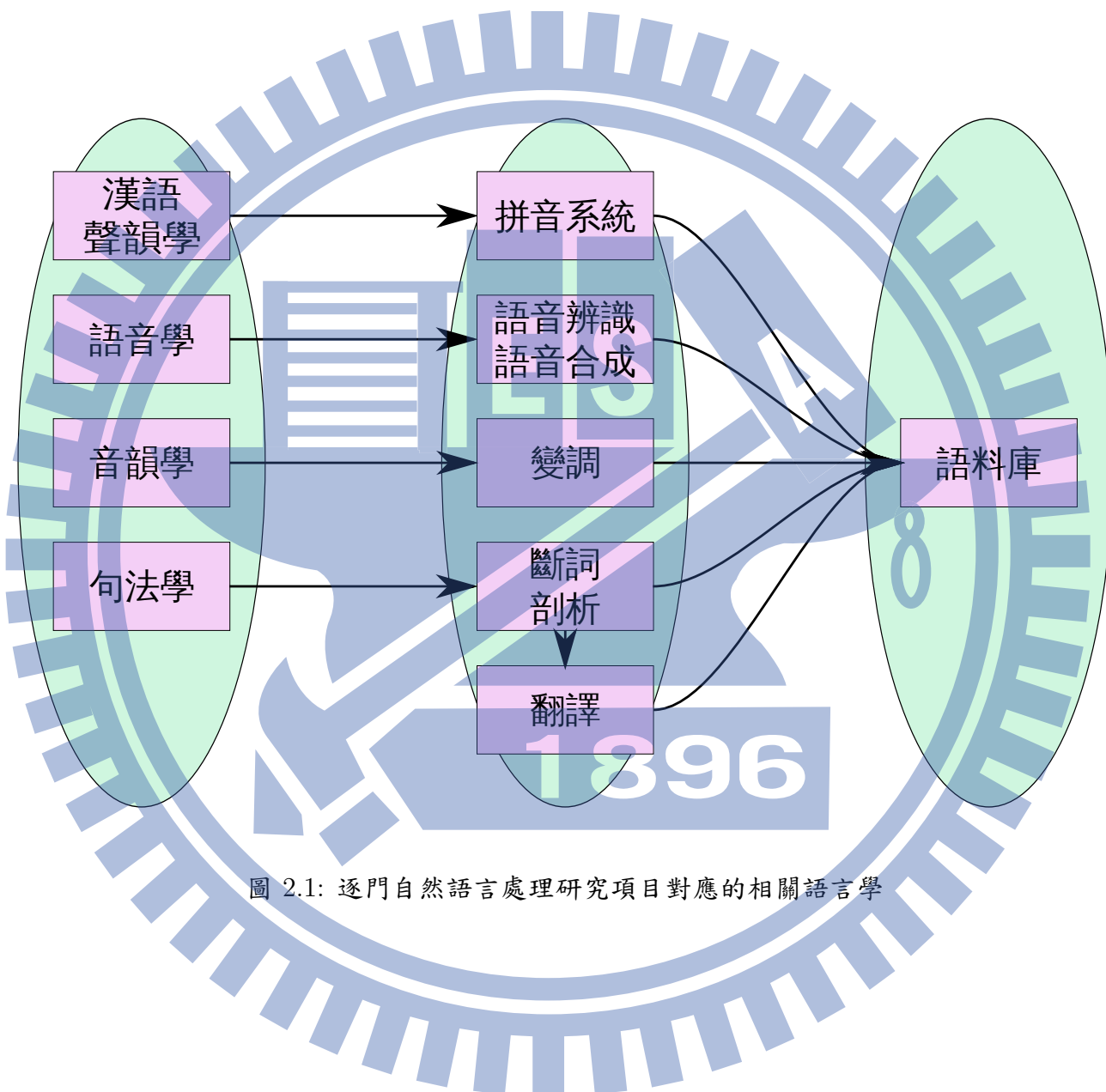


圖 2.1: 逐門自然語言處理研究項目對應的相關語言學

臺灣閩南語羅馬字拼音

臺灣閩南語羅馬字拼音是中華民國教育部佇 2008 年發佈的拼音方案，簡稱做臺羅。伊的前身是教會羅馬字（白話字）[23] 佉臺灣語言音標（Taiwan Language Phonetic Alphabet, TLPA）[24]，這馬猶原相容教會羅馬字。

方音符號

方音符號，又稱方言音符號 [17] 將華語的注音符號閣加一寡聲母、韻母佉聲調符號來標注閩南語。是 1946 年朱兆祥教授設計，佇 1998 年時，中華民國教育部嘛捌公告使用過。

通用拼音

余伯泉教授魚頭所發展的拼音系統，想欲予華語、閩南語、客語佉原住民語攏通用的拼音系統。中華民國教育部佇 2002 年到 2008 年規定華語通用拼音當做羅馬字譯音標準。

2.2 語音學

語音學（Phonetics）主要討論喙舌的運動方式佉語音的物理性質。

前國際語言學會會長 John Ohala 捌講過：「語音變化是連續的（Continuous），為著研究只好假設做離散特徵（Discrete）的音素（Phoneme）。」可比講「狗_{kau2}」，音標寫做 [kau]，實際上 [k] 佉 [a]、[a] 佉 [u] 中央有誠濟過渡的音，毋過過渡的音實在傷濟，無法度一个一个寫出來，只好用 [k]、[a] 佉 [u] 三个符號代表「狗_{kau2}」的音標。

表 2.1: 元音表

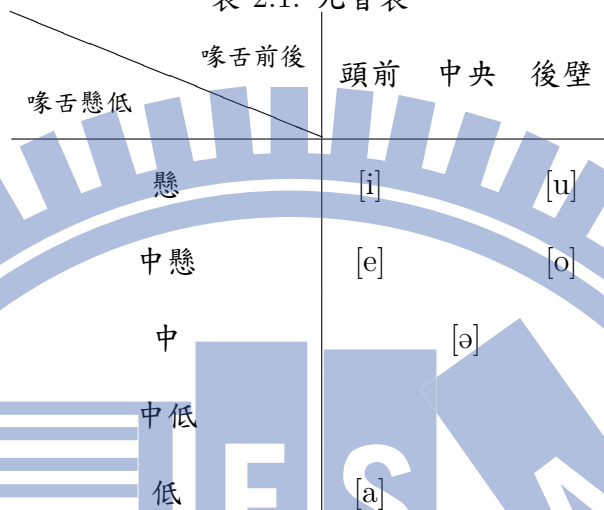


表 2.2: 音節分析

字	聲母	韻母	聲調		
	介音	主要元音	韻尾		
良 <small>カニ</small> liang5	[l]	[j]	[o]	[ŋ]	5
嬌 <small>カニ</small> sui2	[s]	[u]	[i]		2
遠 <small>カニ</small> hng7	[h]		[ŋ]		7
意 <small>カニ</small> i3		[i]			3

表 2.3: 閩南語上細配對

	喙唇	中央
元音、鼻化音	[i] (異 ⁻) i7	[ĩ] (院 ⁻) inn7
配合濁聲母	[bi] (味 ⁻) bi7	[mĩ] (麵 ⁻) mi7

語音學為著方便討論，語音會當用氣流有順無，大略仔分做元音 (Vowel) 佢輔音 (Consonant)。表2.1是幾仔个閩南語定用的元音，會當看著 [i] 佢 [u] 攞是喙舌較懸發的音，而且喙舌佇 i 發音時比 u 發音閣較頭前，這嘛影響著個的頻譜 (Frequency Spectrum)，[i] 佢 [u] 共振鋒 (Formant) 嘛小可無仝。[i] 佢 [u] 喙型嘛無仝，唸「ㄨ」的時陣喙尖尖，唸「一」的時陣喙較平，嘛影響著語音的變化。

佇分析漢語的時，會親像表2.2仝款分析，共音節拆做聲母、韻母佢聲調來看，韻母閣會當分做介音、主要元音佢韻尾。毋過愛注意，這個分析只是方便研究，聲母、韻母的元素佢聲調猶原會互相影響。

若了解語音學，就會當知影語音變化的道理，支持音韻學的理論。

2.3 音韻學

現代的音韻學 (Phonology) 是對 Ferdinand de Saussure[25] 提出語言學研究，必須分做共時 (Synchronic Phonology) 佢歷史 (Diachronic Phonology) 兩種音韻學。

2.3.1 共時音韻學

共時音韻學就是對一個時間的一個語言，討論「頭殼底所想的音」到「唸出來的聲音」的語音變化。佇討論語音進前，阮就愛先訂出共時的語音單位，定用的語音單位是音素。音素代表一個人「頭殼底」，對一個語言的「語音單位」。愛判斷兩個音是毋是全一個音素，會當來揣「上細配對」。可比講表2.3第一逝，一般元音恰鼻化元音會當分出無全的字，所以 [i] 恰 [ĩ] 對人來講是辨識語音的單位，是無全的兩個音素。

$$B \rightarrow \begin{cases} [b], & \text{佇一般元音頭前} \\ [m], & \text{佇一般鼻化音頭前} \end{cases} \quad (2.1)$$

毋過看第二逝，濁輔音佇 [i] 頭前是塞音 [b]，佇 [ĩ] 頭前是變化鼻音 [m]，因為無對比通證明 [b] 恰 [m] 有分辨字的能力，所以 [b] 恰 [m] 對人來講「可能」是全一個語音單位，會當歸類做一個音素 B^5 ，只是 B 佇無全的所在會唸無全的音。

音韻學家就想欲揣出「頭殼底所想的音」到「唸出來的聲音」的關係，討論為啥物面頂的 B 有時唸 [b]，有時陣唸 [m]，目前較大的有衍生音韻學（Generative Phonology）恰優選理論（Optimality Theory）兩大派。

²佇優選理論內底，← 代表上好的選擇

³佇優選理論內底，* 代表違反限制，! 代表出局

⁴佇優選理論內底，殍色底代表選擇出局，毋免比

⁵這個符號用 m、b 攞會使，伊只是代表一個音素

表 2.4: 閩南語濁聲母無鼻化的選擇

Bi+ 無鼻化	元音符合鼻音要求	規个音節鼻化一致	上好的選擇
[bi]			← 是 ²
[bĩ]	*! 毋是 ³	* 毋是 ⁴	
[mi]		*! 毋是	
[mĩ]	*! 毋是		

表 2.5: 閩南語濁聲母有鼻化的選擇

Bi+ 有鼻化	元音符合鼻音要求	規个音節鼻化一致	上好的選擇
[bi]	* 毋是		
[bĩ]		* 毋是	
[mi]	* 毋是	* 毋是	
[mĩ]			← 是

衍生音韻學

衍生音韻學是希望揣出的對應音韻規則 (Phonological Rule)，親像面頂 [b] 恰 [m] 的問題會當看閩南語濁聲母變化規則2.1，*B* 若後壁是一般元音，就變做 [b]，若後壁是鼻化音，就變做 [m]。

優選理論

第二派優選理論是揣出一寡人類語言通用的現象，而且共這現象，照重要程度排先後，去揀人為啥物愛唸這個音。

可比講 [b] 恰 [m] 的問題，既有「元音符合鼻音要求」恰「規个音節鼻化一致」[26] 兩個現象，第一個現象「元音符合鼻音要求」是希望主要元音愛符合頭殼內有鼻音無鼻音的條件，第二個現象「規个音節鼻化一致」是希望音節全部的音素，個鼻化狀況是全款的，而且第一個現象比第二個現象優先，若第一個現象無過，就免比第二個現象。

親像表2.4，假設頭殼底想的是「Bi+ 無鼻化」，阮先產生 [bi]、[bĩ]、[mi]、[mĩ] 四個選擇⁶，其中 [bĩ] 恰 [mĩ] 違反第一個現象，[bĩ] 恰 [mĩ] 予人揀掉，[bĩ]、[mi] 違反第二個現象，毋過因為 [bĩ] 早就違反第一個現象，無需要閣判斷第二個現象，所以第二個現象揀掉 [mi] 上尾賭 [bi]，就是上好的選擇。表2.5是「Bi+ 有鼻化」的例，伊佇第一個現象揀掉 [bi]、[mi]，第二個現象揀掉 [bĩ]，上尾賭 [mĩ] 是上好的選擇。

2.3.2 歷史音韻學

⁶選擇其實閣有 pi、pĩ …無限濟个，個會用別的現象揀掉。為著簡單說明就無寫出來。

⁷[i] 的介音寫做 [j]

表 2.6: 閩南語低元音提昇過程 [1]

階段	音值	來源
Stage 1	jan/jat ⁷	Doty (1853), dictionary of Amoy dialect
Stage 2	jan/jat	Luo & Zhou's fieldwork in 1930 (L&Z 1975)
Stage 3	jen/jet	Taiwanese and some Southern Min dialects
Stage 4	en/et	New forms among young Taiwanese speakers

$$[jaC] \rightarrow [jeC] \rightarrow [eC], \text{ 若 } C \in \{n, t\} \quad (2.2)$$

$$[jaC] \nrightarrow [jeC], \text{ 若 } C \notin \{n, t\} \quad (2.3)$$

歷史音韻學是討論語言長期的變化恰語言互相的影響，這個變化無一定是講話的人講無清楚，嘛有可能是因為音相倚，聽話的人聽毋著，一緣一緣的人沓沓仔改變的。

親像表2.6記錄閩南語「先^{ㄒㄩㄢˊ}」_{sian1} 恰「節^{ㄗㄧㄝˊ}」_{tsiat4} 韻母這兩百冬的變化，所以阮會當共這變化整理做規則2.2。除了寫出規則以外，閣愛需要用語音學解釋為啥物會按呢生，是因為頭前有介音 [j]，喺舌佇較懸較頭前的所在⁸，後壁 [n] 恰 [t] 是舌尖音，喺舌嘛是佇較懸較頭前的所在⁹，予原本喺舌較低的 [a]，變做喺舌較懸較頭前一寡的 [e]。

毋過觀察別的韻煞無這種變化，親像規則2.3，「閃^{ㄕㄩㄢˊ}」_{siam2}、「雙^{ㄕㄩㄢˊ}」_{siang1} 猶原是 [jam] 恰 [jan]。用語音學的角度來看，干焦頭前介音 [j]，無後壁的韻尾配合，無法度予 [a] 變懸變做 [e]。

⁸ 元音的所在會當看表2.1

⁹ 舌尖音恰 [i]、[j] 的位較倚，讀者會當唸看覓 [in]、[en] 恰 [an]，觀察喺舌的變化

2.4 漢語聲韻學

聲韻學是研究漢語的歷史語言學，參考韻冊恰現代漢語方言，討論方言自古到現代的語音變化恰互相影響。

2.4.1 韻冊

聲韻學的主要研究材料就是韻冊，韻冊會提供逐個字的聲母、韻母恰聲調資訊，這就是聲韻學家會當提來推測古早漢語的發音系統的原因。

中國自三國南北朝時就有地區方言的韻冊¹⁰，到宋國官方綜合中國北方恰南方的語音系統的廣韻 [27][28]，廣韻是綜合各地的系統，它的聲母、韻母、聲調紀錄佇無仝的漢語方言攞會當用。

2.4.2 古早中國語音紀錄

反切

反切是古代中國用的記音方式，記錄佇韻冊內底。反切記音需要記兩個字，一字代表聲母，一字代表韻母恰聲調，親像閩南語的「東^{ㄉㄨㄥ}」會用「端^{ㄊㄨㄢ}」「通^{ㄊㄨㄥ}」表示。

綜合各地的韻冊，個記錄的反切嘛會當佇無仝的漢語方言用，除了閩南語的「東」會當切做「端通」華語的「東^{ㄉㄨㄥ}」嘛會當反切做「端^{ㄊㄨㄢ}」「通^{ㄊㄨㄥ}」。

¹⁰ 李登《聲類》、呂靜《韻集》

表 2.7: 定用的分類器優缺點

分類器	優點	缺點
高斯混合模型	知影「結果正確」的機率	訓練的過程無一定收斂
決策樹	輸入會當是字串、符號	訓練資料的答案數量愛平均分配
支持向量機	效果好，效果閣穩定	答案的種類無法度傷濟
深層學習	效果上好	訓練資料愛非常濟，使用門檻非常懸

2.5 機器學習

自然語言的問題定非常複雜，有的時陣無法度揣著規則來解決。機器學習 (Machine Learning) 這個時陣就會當分析資料，揣出適合的參數來模擬資料的行為。

這節就簡單介紹分類器 (Classifier) 恰隱性馬可夫模型 (Hidden Markov Model, HMM):

2.5.1 分類器

分類器是機器學習的一類研究，先決定分類器的模型生啥款，模型內底有誠濟參數，予伊訓練資料，訓練資料內底有一堆問題恰對應的答案，分類器會用輸入訓練資料的問題，揣出上適合的參數，予輸入的問題會當用這參數對應到訓練資料的答案。等待後壁有新的問題入來時，分類器會用伊訓練出來的參數來算答案。

定用的分類器模型有高斯混合模型 (Gaussian mixture model, GMM)、決策樹 (Decision Tree, DT)、支持向量機 (Support Vector Machine, SVM) 恰深層學習 (Deep learning)，逐個分類器的專長無全，應用的所在嘛無全，分類器優缺點會當看表2.7。

2.5.2 隱性馬可夫模型

隱性馬可夫模型 [29] 是針對有狀態轉移的問題，毋過咱看袂著「狀態」本身，咱只會當看著佢狀態有關係的「現象」。隱性馬可夫模型就是想欲用咱看著的現象，去推測實際的狀態倒底是按怎變化。

2.6 語音辨識

語音辨識 (Speech Recognition) 就是共語音轉做文字，會當用佇語音指令佢問答系統¹¹。

語音辨識主要的做法是揣出語音佢音標的對應，共語音轉做一个一个 MFCC 聲學特徵，用分類器去判斷是佇一个音標。

因為語音訊號連續閣無固定長度，為著解決這個問題，就假設語音變化是狀態的轉移，用隱性馬可夫模型來模擬語音狀態的變化。

這方面的開源工具有 HTK 佢 Kaldi：

2.6.1 HTK

HTK 全名號做 Hidden Markov Model Toolkit[30]，發展的時間較早¹²，伊主要用的高斯混合模型當做分類器，而且用決策樹合併相倚的高斯模型。

¹¹親像蘋果公司的 Siri

¹²對 1989 年到最近上新的 2009 年版本

2.6.2 Kaldi

Kaldi[31] 是較新的工具，除了訓練一開始嘛是恰 HTK 全款用高斯混合模型以外，伊訓練後壁閣加入深層學習恰其他的演算法，效果比 HTK 閣較好。

2.7 語音合成

語音合成 (Speech Synthesis) 是恰文字轉做聲音，恰語音辨識顛倒反。親像車站廣播，有聲冊攏是語音合成的應用。這馬時行的做法有模型合成 (Model-based Speech Synthesis) 恰接音合成 (Corpus-based Speech Synthesis) 兩種：

2.7.1 模型合成

這個方法揣出音標¹³恰語音的對應關係，啥物音標會唸啥物音。頭一步是用合成器 (Vocoder) 共語音訊號轉做一個一個的頻譜恰頻率，才用隱性馬可夫模型恰分類器，共音標當做輸入，頻譜恰頻率當作輸出訓練隱性馬可夫模型恰分類器的參數。等欲合聲音時，才閣照模型的特徵，用合成器合聲音出來。

伊的輸出語音，韻律攏誠自然，毋過聲音的品質比原本語料的音質較稔一寡，因為訓練的時，聲音先用合成器轉做特徵，合成的時，特徵閣用合成器轉語音，兩擺轉換造成音質變稔。

¹³用人工抑是語音辨識軟體標記

HTS

這方面開源軟體有 HTS (HMM-based Speech Synthesis System) [32]，伊是對 HTK 修改來的，全款用隱性馬可夫模型、高斯混合模型佢決策樹。

HTS 有 3000 ~ 7000 句的訓練語料，就會當得著袂稔的效果，缺點歹揣出聲音效果稔的原因。

2.7.2 接音合成

模型合成有音質的問題，為著保持音質，接音合成共原本的音檔庫切做一個一個細音檔，合成的時陣，提細音檔來接起來。

對漢語來講，只需要共逐種音錄起來就好矣，毋過一字一字聽起來無順無自然。

為著改善接起來韻律無自然的問題，接音合成會配合模型合成，用模型產生的韻律，揀接起來較自然的細音檔。按呢做的優點是聲音品質誠好，缺點是語料愛有夠，因為需要誠濟的細音檔來配合韻律，若拄著合成品質無好的語句，就錄音增加音檔庫就好矣。

2.7.3 變調佢重音預測

佇書寫的音標佢實際唸的發音定定是無全款的，上大的差別就是聲調 (Tone) 佢重音 (Stress)。漢語有聲調，聲調到發音之間會變調 (Tone Sandhi)。南島語有重音，標註語句的重音 (Stress Prediction) 嘛是一個音韻學的研究問題。

用閩南語做例，準做想欲共閩南語的音標變做閩南語的聲音，毋過實際的音標到唸的發音閣無全，後且閩南語的變調傷過複雜的，需要專門的處理。

楊允言教授就有做過閩南語的變調系統 [33]，用斷詞、詞性佢句型的資訊，配合 20 類

表 2.8: 臺灣母語語音合成相關研究、系統

1999	林川傑	閩南語翻譯俗語音合成系統 [34]
2002	李雪貞	客家語語音合成之初步研究 [35]
2005	楊允言、劉杰岳、李盛安	台語羅馬字發音試驗系統 [36]
2008	陳信宏、余秀敏、羅烈師	客語文句轉語音及語音辨認之研究 [37]
2010	蔡依玲	基於隱藏式馬可夫模型之客語文句轉語音系統 [38]
2013	薛丞宏	意傳文化科技 [39]

音韻規則來變調，得著 88.90% 的準確率。音韻規則的優點是語料無蓋濟的時，就會當得著八九成的正確率，毋過若數量一濟，規則的先後會較歹處理。

變調嘛會當用分類器來做，共規則式用著的特徵當做參數下入去，看佢一个分類模型會較好。伊上大的好處就是管理方便，拄著新語料，重訓練模型就好矣，缺點是訓練資料需要誠濟。

2.7.4 相關系統

目前臺灣母語的語音合成只有閩南語俗客話，親像表2.8的系統，頭前三个攞是用接音合成，配合字的錄音第四第五个是接音合成配合韻律模型上尾一个主要是訓練閩南語 HTS 模型。

因為南島語的語料較少，嘛無法度親像漢語幾字仔錄音就有基本的效果，就需要請專人來錄音、整理，這是咱閣愛拍拼的所在！

2.8 斷詞

斷詞 (Word Segmentation) 是共語句照一个詞一个詞分開的技術，親像漢語佻日語的文字定定是一字一字無分開，就看袂出來倒底佻一字佻佻一字是一个詞，若愛提著詞的資訊，就需要程式來斷詞。斷詞的效果嘛會影響著翻譯、變調佻其他技術的效果。

閩南語的變調、翻譯...一寡語言現象會受著詞的資訊影響，毋過漢語語句的文本攞無詞的資訊，所以就需要斷詞，共詞佻詞分開，後壁的應用才有法度繼續落去。

2.8.1 長詞優先斷詞

定看著的斷詞方法有長詞優先 (Maximum Matching)，伊的做法是自語句後頭開始看¹⁴，往頭前幾個字是毋是會當揣著一个佇辭典的詞，若會使，就揀上長的彼个，希望詞愈長愈好，會當看演算法1。

演算法 1 長詞優先斷詞方法

輸入: 辭典上長的詞字數 k , 無斷詞的語句 $[j_1, j_2, \dots, j_m]$

輸出: 斷詞的語句 $[s_1, s_2, \dots, s_n]$

揣一个上細的 i ，予 $j_{i+1}, j_{i+2}, \dots, j_m$ 是辭典的一个詞，而且 $m-k \leq i \leq m-1$

斷詞的語句加入 $s = j_{i+1}, j_{i+2}, \dots, j_m$

$m=i$ ，重做第 1 步揣 i ，到 $i=0$ 為止

¹⁴ 華語實驗的結果，自後頭開始的效果比對頭前的閣較好

表 2.9: 召回率、精確率佅 F- 測量的公式

$$\text{召回率} = \frac{\text{答案的斷詞數量}}{\text{斷著的斷詞數量}}$$

$$\text{精確率} = \frac{\text{答案的斷詞數量}}{\text{結果的斷詞數量}}$$

$$F\text{-測量} = \frac{2 \times \text{召回率} \times \text{精確率}}{\text{召回率} + \text{精確率}}$$

2.8.2 評分方式

斷詞的評分方式主要會當分做召回率、精確率佅 F- 測量三種 [40]。召回率是比較斷著的斷詞佅標準答案斷詞的比率，精確率是比較斷著的斷詞佅程式輸出結果的比率，F- 測量是綜合召回率佅精確率，這三个評分方式的公式會當看表2.9。

因為的召回率佅精確率意義無全，有時陣無法度比較，就需要參考 F- 測量。準若有兩個方法，第一個召回率比第二個方法懸，第一個精確率比第二個方法低，這時陣就會參考 F- 測量的數值。

2.8.3 相關系統

閩南語斷詞的標準，自誠早以前就有人討論矣 [41]，教育部嘛有出「臺灣閩南語羅馬字拼音方案連字符使用原則」 [16]，定義連字符的標準。

華語這方面有中研院中文斷詞系統 (CKIP) [42]。

2.9 剖析

剖析 (Parsing) 是了解語句內底詞的關係，分析語句的句型。剖析會當用佇翻譯俗語意分析，較定看的就是問答系統¹⁵。

做剖析需要句法學的知識，閣愛有一致的人工檢查，這是一個誠大的工程。

2.9.1 相關系統

楊允言教授捌做過閩南語的剖析樹 [43]，毋過數量猶原無夠，需要閣整理。

一開始的語料歹收集，有一个辦法是借用華語的剖析，先共閩南語翻譯做華語，才閣共華語語句的剖析樹 [44] 對應去閩南語語句，按呢就有初步的剖析樹矣。訂好剖析樹規則了後，就會使整理初步的剖析樹，等待資料有夠濟，就會提剖析樹的語料來訓練一个閩南語的剖析器 [45]。

2.10 翻譯

這馬電腦時行的翻譯方式是統計式機器翻譯 (Statistical Machine Translation)，這是對 1993 年 Brown 提出數學模型 [46] 開始，一直發展到這馬。統計式機器翻譯會當分做對齊模型 (Alignment Model)、語言模型 (Language Model) 佢解碼器 (Decoder) 三个部份：

¹⁵親像蘋果公司的 Siri

2.10.1 對齊模型

對齊模型的功能是予解碼器知影詞愛按怎翻譯，可比講是一個雙語的辭典。

對齊模型有分斷詞對齊恰剖析樹對齊兩種，下跤用斷詞對齊說明伊的原理。

先準備一組一組的華語閩南語平行語料，親像「我 要 吃飯」和「我 欲 食飯」，繼落來產生語詞對照表。華語詞的「要」，會對應到「我」、「欲」、「食」、「飯」閩南語詞，經過大量的平行語料，上尾知影華語的「要」定定對應著閩南語的「欲」，也就是共對應頻率懸的組合留落來。

開源工具 GIZA++[47] 實作 Brown 1993 的演算法，而且這馬嘛有支援多核心的 MGIZA[48]。

2.10.2 語言模型

第二部份是語言模型，伊是會當用來判斷一句話是好像是。

伊的做法是去記錄逐個詞後壁定定會接啥物詞，若有一句話是「…欲 食…」，有「欲」恰「食」兩個詞，咱知影「…欲 食」的後壁接「飯」比「…欲 食」的後壁接「湯」的機率較大，也就是講「欲 食 飯」連繼詞比「欲 食 湯」連繼詞機率大，若語言模型一擺看「欲 食 飯」三個詞，就是三連繼詞模型 (3-grams model)。語言模型判斷一句話，伊出現的機率有佻大，就是看這句話伊內底連繼詞的機率是佻大。

這方面的工具有 IRSTLM[49]、SRILM[50] 恰 KenLM[51]，其中 IRSTLM 恰 KenLM 是 LGPL 開放授權，SRILM 是學術授權。

表 2.10: 翻譯結果一佢翻譯結果二對答案的 BLEU 分數

翻著的數量	一連繼詞	兩連繼詞	三連繼詞	四連繼詞	BLEU 分數
結果一	5/6	3/5	2/4	1/3	53.73
結果二	6/6	3/5	1/4	0/4	0.00

2.10.3 解碼器

上尾一部份是解碼器，提面頂講的對齊模型、語言模型，來翻譯華語到閩南語。

因為翻譯的問題毋是多項式時間（NP problem）會當解出來的，所以解碼器袂使硬算全部的可能，必須用有效率的演算法來翻譯。

上有名的開源程式就是 Moses[52]，伊整合對齊模型佢語言模型的介面，閣有專工的訓練包通使用 [53]。

2.10.4 評分方式

翻譯大部份攞用 BLEU（Bilingual Evaluation Understudy）來評分，伊用連繼詞的概念來評分， $BLEU = 100 \times e^{\max(0, \frac{\text{結果}-\text{答案長度}}{\text{結果長度}})} \times \sum_{n=1}^4 (n \text{ 連繼詞})^{\frac{1}{4}}$ [54]。

準若翻譯的答案是「這 幾 工 寒流 閣再 展威」，咱有兩個翻譯的結果，翻譯結果一「這 幾 工 寒流 有 展威」佢翻譯結果二「寒流 這 幾 工 閣再 展威」，個的分數請看表2.10。答案有「這 幾 工」、「幾 工 寒流」、「工 寒流 閣再」佢「寒流 閣再 展威」4 个三連繼詞，翻譯結果一有出現 2 个，所以翻譯結果一的三連繼詞分數是 2/4，翻譯結果二有出現 1 个，分數是 1/4。因為翻譯結果二無對應的四連繼詞，翻譯結果二的分數都比翻譯結果一低。

2.11 語料收集整理

2.11.1 語言分類

語言分類 (Language Identification) 是輸入一句話，判斷是佢一種語言。這馬上時行的方法是以字元為單位，語言模型算分數 [55]

南島語主要嘛是拼音文字，所以會使用這個方法，南島語佢漢語的語言分類較簡單，就算漢語用羅馬拼音，拼音的種類嘛差誠濟，只要檢驗有聲調抑是拼音規則就會使判斷是南島語抑是漢語。

2.11.2 平行語料語句對齊

翻譯語料的平行語料需要一句一句對齊，若原本的語料是一篇一篇對應的，就需要平行語料語句對齊 (Parallel Corpus Sentences Alignment)。

語句對齊的方法有誠濟種，有照字元數量對齊的 Gale and Church 算法 [56]，嘛有用翻譯結果輔助的 Bleualign[57]。

2.12 語料庫

自然語言處理需要語料才有法度訓練模型，就需要語料庫 (Corpus) 共語料存起來。

有的語料庫是純文字的資料庫，記錄雙語語料，抑是斷詞佢剖析資料。有的是語音語料庫，記錄語音佢伊的文字內容。語料庫的設計愛看需求，定看著的技術會當參考表2.11。

表 2.11: 自然語言處理技術需要的語料庫

技術	語料樣式
變調	原始文本、本調音標俗變調音標
語音辨識	濟人音檔俗對應文本
語音合成	孤人音檔俗對應文本
斷詞	斷詞的語料
剖析	剖析樹
翻譯	兩種語言的平行語料

表 2.12: 閩南語語料種類比較表

種類	範例	備註
全漢	我欲食飯	全部漢字
全羅	gua2 beh4 tsiah8-png7	全部羅馬拼音，有斷詞資訊
漢羅	我 beh4 食飯	漢字拼音濫用

2.12.1 閩南語語料種類

閩南語是漢語的一支方言，大部份的字擺會當揣著漢字，毋過閩南語嘛毋是純漢語，有的字無對應的漢字。

有的人慣勢全部用羅馬拼音創作，這種寫法號做「全羅」。嘛有人慣勢全部用漢字創作，號做「全漢」毋過有的字無法度揣著對應的漢字，全漢實際上的拄著造字、揣字的困難，為著創作方便，知影的字用漢字寫，賸的用音標寫落來，號做「漢羅」。詳細會當看表2.12的範例。

表 2.13: 教育部辭典例句

全漢	彼个查某囡仔真嬌。
全羅	Hit ê tsa-bóo gín-á tsin suí.
華語	那個女孩子很漂亮。

2.12.2 閩南語語料—教育部辭典

教育部辭典全名「臺灣閩南語常用詞辭典」[58]，正式版是 100 年上線，伊有 25892 的詞條¹⁶，內底誠濟生活的用語，大部份詞條攏有漢字、音標、解釋、例句佮翻譯。

這個辭典是教育部編的，當然漢字有照教育部家己的規範¹⁷來寫，所以伊內部的用字前後有一致，為著翻譯的效果佮使用者的方便，就共教育部辭典的用字當做標準，若有用字佮教育部的用字無仝的，就改做教育部的用字。

伊的例句，除了閩南語漢字佮音標以外，閣有攸華語的翻譯，親像表 2.13 的例。漢字佮音標的對應會使提去做用字參考的字典，音標的部份會當提來斷詞，斷詞了會當訓練語言模型，閩南語佮華語的對應會使提來做平行語料。

除了一般的詞條以外，教育部辭典嘛有收一寡俗語、臆謎猜，攏總 388 句做附錄句。因為附錄句干焦提供解釋，所以無法度提來做平行語料，毋過會使提來訓練語言模型。

2.12.3 閩南語語料—新聞語料庫

iCorpus 臺華平行新聞語料庫（後壁用「新聞語料庫」稱呼）[59] 是中央研究院資訊所陳孟彰老師主持，內底的文章主要是何澤政翻譯的。

¹⁶1021230 的資料

¹⁷臺灣閩南語推薦用字 700 字表，佇 96 ～ 99 年公佈修正

何澤政¹⁸對民國九十七年十一月開始，逐工揣兩篇華語新聞，先人工斷詞斷句，後尾翻譯做閩南語教會全羅，罕得改變用詞的先後。

親像原本的新聞「這幾天寒流再度發威」，翻譯做「tsit4-kui2-kang han5-liu5 koh-t sai3 tian2-ui」（這幾工寒流閣再展威）¹⁹。

澤政佇語料內底用「捫^{ㄅㄣˊ}揀^{ㄓㄢˊ}」²⁰、「作^{ㄗㄨㄛˊ}孽^{ㄋㄧㄝˊ}」²¹本土的詞以外，伊嘛會配合這馬發生的代誌，用較時行的閩南語，親像「喙^{ㄅㄣˊ}罨^{ㄢˊ}」²²、「心^{ㄒㄩㄢˊ}肌^{ㄇㄩㄣˊ}梗^{ㄍㄥˊ}窒^{ㄓˋ}」²³、「自^{ㄗㄣˊ}來^{ㄌㄞˊ}水^{ㄨㄟˊ}」²⁴。

而且除了現代閩南語，澤政伊閣會去查台華線頂辭典 [60] 選擇較古典的用詞²⁵，親像「瘡^{ㄘㄨㄞˊ}篤^{ㄉㄨˊ}篤^{ㄉㄨˊ}」、「鬥^{ㄉㄡˊ}贊^{ㄗㄢˊ}手^{ㄕㄨˊ}」。拄著外來詞，澤政嘛會選擇保留原文，拄著華語的「歐巴馬」恰「西藏」，會翻轉去英文「Obama」、「Tibet」。

因為新聞語料庫源頭無全漢，有人工本論文用的新聞語料庫有閣半自動半人工補充全漢語料。

2.12.4 閩南語語料—臺文典藏

台語文數位典藏資料庫（下跤號做臺文典藏）[61] 是國家臺灣文學館收集 1885 到 2006 年的語料。臺文典藏的語料來源百百款，攞總 2167 篇，照時代分做清國時期 170 篇、日

¹⁸ 一九七零年代出世，臺中烏日人

¹⁹ 原文是教會羅馬字，為著文章一致，以教育部的臺羅書寫

²⁰ 捫揀意思共物件擲掉、放揀，嘛就是華語的「丟棄」

²¹

²² 華語的「口罩」

²³ 華語的「心肌梗塞」

²⁴ 閩南語較古典的用法，會號做「水道水」

²⁵ 台華線頂辭典是古早語料，一個詞若台華線頂辭典查有，教育部辭典查無，就當做伊是較古典的詞

表 2.14: 臺文典藏語料漢羅、全羅對照

漢羅	Koh m7 知 u7 危險.....,
全羅 ²⁷	Koh m7-tsai u7 gui5-hiam2.....,

治時代 490 篇，民國統治 1507 篇。內底嘛有照語料形式分做四類，有詩 387 條、散文 1127 篇、小說 387 篇、劇本 49 篇。

因為伊是百外冬的語料庫，較古早的語料，用詞就較古典。格式嘛無全，有的是漢羅，有的是全羅。臺文館因為著格式統一，就替原底是全漢抑是漢羅的語料補全羅拼音²⁶。若原本是全羅，就請人拍漢羅，會當看表2.14的語料，拍漢羅的時，個若知影漢字，就會拍漢字，賸的揣無漢字，抑是外來語的語詞就會用音標來拍。伊有的語料是一句對齊一句，有的是一段對齊一段。

2.12.5 閩南語語料— TGB 通訊臺灣組合

TGB 通訊 [62] 是學生台灣語文促進會對 1999 年 10 月開始 [63] 一個月一期的刊物。頭前 60 期以閩南語為主，61 期後有提供華語對照，有時陣閩南語句內底會濫一寡華語詞，形式較無固定，親像表2.15的語料。

²⁶有時陣劇本邊仔的解釋會用漢字，親像「(福哥仔出場)」

表 2.15: TGB 通訊語料狀況

平行語料	範例
閩南語漢羅	Gín-á beh 講 sián-mih 款語言 kám 是 gín-á ka-tī 決定 -ê? Iah 是大人提供 ê 選擇?
華語漢字	孩子要講什麼語言是孩子自己決定的嗎? 還是大人提供的選擇?
濫做伙語料	範例
語料一	「糟了, 是工地火燒厝, 緊轉去打火! 」建設公司的工地主任從手機接到消息, 通話結束後就帶著那群混混先離開了。
語料二	『聽說妳最近遇到什麼問題, 是不是? 怎麼了? 』好性地 ê QA 繼續問 -落-去。

第三章 研究介紹

阮的目標是予閩南語的翻譯，效果閣較好，效果的好穰是看 BLEU 拍的分數¹。

佇遮希望預處理語料，予翻譯效果變好。統計式機器翻譯的效果決定佇統計的模型，若愛翻譯翻較好，有兩個大方向通做：第一個方向是予語料的形式相像，若語料的形式愈全款，翻譯的統計機率會閣較好。第二個方向是資料愈濟愈好，加新的語料庫了後，翻譯模型揀著好的語詞來翻譯的機會愈大。

本章第3.1摻3.2節針對第一個方向做，因為閩南語目前閣無剖析的程式，本論文對齊模型用的是斷詞翻譯²，語料的樣式就是以斷詞為主。3.1節討論佗一種斷詞方法，對數量無濟的閩南語語料上好。3.2節說明斷詞的語料翻譯，一寡詞會翻袂出來的問題。第二個方向是第3.3和3.4節，3.3節講摻樣式無全的語料庫時，會發生啥物問題。3.4節對網路頂掠落來的資料，愛按怎共語料照語言分類。

3.1 閩南語斷詞

愛用斷詞的形式來翻譯，華語有中研院的中文斷詞系統，閩南語無現成的系統。所以頭一個問題就是閩南語欲按怎斷詞，而且比較無全的斷詞方法，個的效果分別是按怎。

¹請看2.10.4節的紹介

²請看2.10節的紹介

表 3.1: 未知詞問題範例

訓練語料 1	自稱一輩子離不開預報工作的吳德榮, tsu7-tshing1 tsit8-si3-lang5 li5-be7-khui1...
訓練語料 2	開始傷愁了, khai1-si2 siong1-tshiu5 ah4 ,
...	
試驗輸入	一輩子吃穿都不用憂愁了
翻譯結果	tsit8-si3-lang5 tsiah8 tshing7 long2 m7-bian2 憂愁

3.2 未知詞問題

用斷詞做語料單位，親像有的華語詞無出現佇語料過，翻譯模型毋知伊愛對應到佗一个閩南語詞，就會翻袂出來。因為訓練語料無可能有全部的華語，親像表3.1的例就無「憂愁」華語詞的翻譯方法，就愛想一个辦法，處理有詞翻袂出來的情形。

3.3 整理語料

完整的閩南語語料應該有全漢、全羅俗斷詞三種資訊，毋過誠少有語料庫三種資訊攏有，而且逐个語料庫資訊狀況攏無全。

第三个問題就是討論按怎利用手頭有的語料庫，補好全漢、全羅俗斷詞三種資訊，狀況親像表3.2，希望整理了的語料庫，會當予翻譯的效果閣較好。

³ 因為新聞語料斷詞無規範

⁴ 臺文典藏少數無全羅

表 3.2: 語料庫狀況

	全漢	全羅	斷詞
教育部辭典	O	O	O
新聞語料庫	O	O	X ³
臺文典藏	X	O ⁴	O

表 3.3: 語言分類範例

語句	語言
聽人講 khah 早有出現過『小蜜蜂』	閩南語
我 beh tng 來種作! ——記 0312 Truku 反亞泥・還我土地運動	閩南語
有台灣味 ê 繪本——《我和我的腳踏車》。	閩南語
「糟了, 是工地火燒厝, 緊轉去打火! 」建設公司的工地主任從手機接到消息, 通話結束後就帶著那群混混先離開了。	華語
去越南胡志明市 4 工/越南胡志明市四日行 @Giok-hōng	華語

3.4 語言分類

增加語料庫的一个方法就是去網路頂掠閩南語的資料, 毋過網頁頂的閩南語定定佻華語濫做伙, 所以愛揣一个方法分類這兩個語言, 可比講表3.3按呢。先前大部份語言分類的研究攏是拼音文字, 用字元的語言模型去判斷, 毋過閩南語佻華語誠濟用詞是全款的, 用字的語言模型去判斷效果無好, 所以第四個問題就是閣欲加揣啥物款的特徵, 會當來幫助語言分類。

第四章 研究方法

4.1 拄好長度斷詞

3.1節講愛比較斷詞的方法，定用的斷詞方法有2.8.1節的長詞優先，佇遮本論文提出一个「拄好長度斷詞」的方法。

因為長詞優先有時陣會揀著無好的組合，親像表4.1內底的例，長詞優先有的情形會斷毋著。觀察這個情形，第一組例是斷詞結果的詞數比答案閣加一个，斷出來的詞傷濟矣。第二組例是斷詞結果恰答案攏是斷詞兩個詞，因為斷詞結果共四字攏分配做一字詞恰三字詞，字數分配無齊勻。

綜合這兩個觀察，本論文提出「拄好長度斷詞」。成本函式親像公式4.1是訂做一字詞 1 分、兩字詞 $1/2$ 分、三字詞 $1/3$ 分、四字詞 $1/4$ 分……斷詞的方法是演算法2配合維特比 (Viterbi) 算法揣出成本上低的斷詞法。用面頂表的例，算出來的成本會當看表4.2

而且愛注意拄好長度斷詞毋是全部的語料攏會比長詞優先斷詞好，親像表4.3的例，雖

表 4.1: 長詞優先毋著的情形

方法	結果
長詞優先 (對頭前)	猶 掠做 唱歌 仔 戲 真 簡單
答案	猶 掠做 唱 歌仔戲 真簡單
長詞優先 (對後壁)	甚至 和 國 小學生 嘛 想 袂 開
答案	甚至 和 國小 學生 嘛 想 袂 開

然挂好長度的總詞數比長詞優先閣較少，毋過佢答案相比，這個例猶原是長詞優先閣較好淡薄仔。

$$\text{成本函式}(n) = \frac{1}{n} \quad (4.1)$$

演算法 2 挂好長度斷詞

輸入：需要斷詞字數 m , 辭典上長的詞字數 k , 無斷詞的語句 $[j_1, j_2, \dots, j_m]$

輸出：斷詞成本, 斷詞的語句

if $n == 0$ then

 回傳 $(0, \emptyset)$

end if

$i \leftarrow \underset{i}{\operatorname{argmin}} \{ \frac{1}{m-i} + (\text{挂好長度斷詞}(i, k) \text{ 的成本}) \}$, 其中 $j_{i+1}, j_{i+2}, \dots, j_m$ 是辭典的一个詞,

而且 $0 \leq m-k \leq i \leq m-1$

頂一層成本 c , 頂一層斷詞的語句 S

成本 $c' \leftarrow$ 頂一層成本 $c + \frac{1}{m-i}$

斷詞的語句 $S' \leftarrow$ 頂一層斷詞的語句 S 加入詞 $(j_{i+1}, j_{i+2}, \dots, j_m)$

回傳 (c', S')

4.2 未知詞另外翻譯

對3.2節來看，用斷詞翻譯會挂著未知詞的問題。佇遮阮提出一个演算法3，準若阮用斷詞翻譯模型，挂著未知詞的時陣，這個未知詞會使提予斷字翻譯模型去翻譯。

表 4.2: 挂好長度成本

斷詞結果	挂好長度成本
猶 掠做 唱歌 仔 戲 真 簡單	$\dots + \frac{1}{2} + \frac{1}{1} + \frac{1}{1} + \dots$
猶 掠做 唱 歌仔戲 真 簡單	$\dots + \frac{1}{1} + \frac{1}{3} + \dots$
甚至 和 國 小學生 嘛 想 袂 開	$\dots + \frac{1}{1} + \frac{1}{3} + \dots$
甚至 和 國小 學生 嘛 想 袂 開	$\dots + \frac{1}{2} + \frac{1}{2} + \dots$

表 4.3: 長詞優先比挂好長度斷詞較好的狀況

答案	七月半 鴨仔 毋 知 死活
挂好長度	七 月半 鴨仔 毋知死 活
長詞優先	七 月半 鴨仔 毋 知 死活

未知詞：一百五十項、...

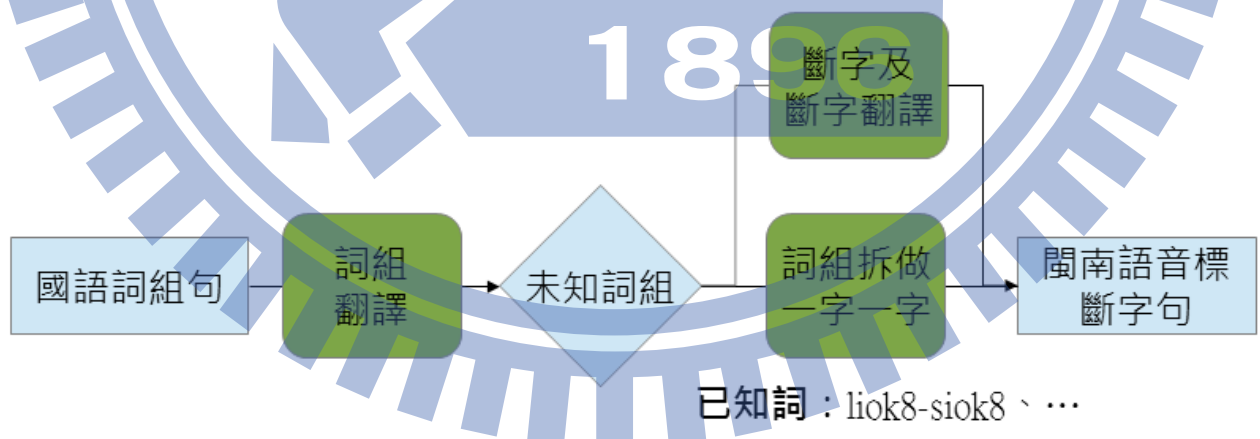


圖 4.1: 未知詞另外翻譯流程

可比講「陸續 開放 一百五十項 的 規費」提予斷詞組模型翻譯，得著「liok8-siok8 khai1-hong3 一百五十項 e5 規費」，閣來共「一百五十項」佢「規費」這兩個詞組切做斷字「一 百 五 十 項」佢「規 費」，閣擲去斷字模型翻譯，流程會當看圖4.1。

演算法 3 未知詞另外翻譯

輸入：原本華語句 $H = [h_1, h_2, \dots]$

輸出：翻譯閩南語句 $M = [m_1, m_2, \dots, m_n]$

$M = [m_1, m_2, \dots, m_n] \leftarrow$ 斷詞翻譯 H 的結果

while 存在 $B = [m_i, m_{i+1}, \dots, m_j]$ 攞是未知詞， m_{i-1}, m_{j+1} 是已知詞 **do**

$T \leftarrow B$ 提去斷字翻譯

共 M 內的 B 換做 T

end while

4.3 漢羅全羅對齊

佇2.12.4節有講著，臺文典藏是提供漢羅佢全羅的對照，因為阮的翻譯需要一個漢字對一個音標的一對一，所以愛共臺文典藏伊原本一段對齊一段的語料改做一字對一字。而且愛注意臺文典藏佇2006年完成，教育部的漢字規範佇2007年才公佈，所以個兩個的用字規範是無完全全款的。毋過臺文典藏的語料倩人整理的時陣內部有訂標準，伊的漢字有一半以上攞是會用得。本論文用字以教育部的為主，對齊的做法是共漢羅逐字攞去對看覓全羅，看漢羅字一字佢全羅一字的對應有佇字典內底無，揣出上長的對應組合。

4.4 補全漢佻全羅

佇3.3節有講著，閩南語語料的完整資訊有全漢、全羅、斷詞三項，若全部的語料攞有這三種資訊，翻譯效果會閣較好。

斷詞的部份佇4.1節有討論矣，這節的重點是下佇按怎自動整理語料，予個有完整的全漢佻全羅資訊，也就是替個補起哩欠的漢字佻羅馬音標。

因為有的詞可能一字音標、一字是漢字，親像「彰化」，寫做「tsiong-化」。本論文提出一个做法，整理語料的時猶原用斷詞方法，毋過用的辭典愛小可仔改變，逐個詞的全部形式攞愛加佇辭典內底，閣愛加「彰化」、「彰^{ㄓㄨㄥˋ}huà」、「彰化^{ㄓㄨㄥˋ}huà」、「tsiong化」、「tsionghuà」、「tsiong化」、「彰^{ㄓㄨㄥˋ}化」、「彰^{ㄓㄨㄥˋ}huà」、「彰^{ㄓㄨㄥˋ}化^{ㄓㄨㄥˋ}huà」攞總九種¹。

決定斷詞斷佇佢位了後，逐個斷詞的所在可能有超過一个的候選詞，上尾閣用語言模型，配合維特比算法，揀出機率上懸的語句。

4.5 語言分類特徵

3.4節有講，閩南語佻華語以字為單位的語言模型效果無好，為著予電腦會當分別閩南語佻華語，阮就愛準備幾項閩南語佻華語無仝的特徵。本論文提出一个以斷詞資訊做判斷特徵的方法，除了以斷詞算語言模型以外，閣加入斷詞了一字詞、兩字詞、三字詞、四字詞的詞數，毋過按呢猶原無夠。

閩南語佻華語上大差別就是用詞無仝，閩南語寫「食飯」、「無法度」，華語寫「吃飯」、「沒辦法」，所以阮揀定用詞出來，當作阮的特徵之一。

¹一个字的資訊可能是「漢字」、「音標」、「漢字音標攞有」三種其中一種。兩字，攞總 $3^2 = 9$ 種

表 4.4: $n=7000$ 、 $m=3000$ 的頭前九個定用詞俾特徵詞

第幾個	1	2	3	4	5	6	7	8	9
閩南語定用詞	的 _ê	伊 _i	有 _ū	是 _{sī}	我 _{guá}	人 _{lâng}	無 _{bô}	講 _{kóng}	佇 _{tī} ...
華語定用詞	的 _{de}	是 _{shì}	在 _{zài}	一 _{yí}	有 _{yǒu}	了 _{le}	不 _{bù}	我 _{wǒ}	個 _{ge} ...
閩南語特徵詞	佇 _{tī}	個 _ê	閣 _{koh}	攞 _{lóng}	恰 _{kap}	個 _{in}	咧 _{teh}	咱 _{lán}	彼 _{hit} ...
華語特徵詞	我 _{wǒ}	們 _{men}	很 _{hěn}	她 _{tā}	沒 _{méi}	有 _{yǒu}	或 _{huò}	他 _{tā}	們 _{men} 更 _{gèng} 則 _{zé} 把 _{bǎ} ...

毋過閩南語俾華語有誠濟共同詞，親像「火車」、「電腦」，個寫法是全款的，阮袂使直接提定用詞來做，因為內底會有共同詞，所以阮愛揀出無共同詞的「特徵詞」。

選特徵詞的方法是先統計閩南語語料俾華語語料，分別揣出 n 個定用詞²，了後揀出頭前 m 個閩南語定用詞，而且這 m 個閩南語定用詞無出現佇華語 n 個定用詞，這 m 個詞阮就號做閩南語特徵詞。華語部份嘛全款，揀出頭前 m 個華語定用詞，這 m 個詞無出現佇閩南語的 n 個定用詞，這 m 個詞就是華語的特徵詞。

佇遮阮設 定用詞數量是 $n = 7000$ ，特徵詞數量是 $m = 3000$ 來揣華語閩南語的特徵詞，頭幾個定用詞俾特徵詞佇表4.4，「的_ê」俾「伊_i」攞是閩南語的定用詞，毋過這兩個詞「的_{de}」俾「伊_{yí}」華語攞會用著，所以袂使做閩南語的特徵詞。親像「佇_{tī}」華語就罕得用著「佇_{zài}」，就會使當做閩南語的特徵詞。

有斷詞、語言模型俾特徵詞的資訊，就會當親像圖4.2全款，交予分類器做語言分類。

²有算標點符號

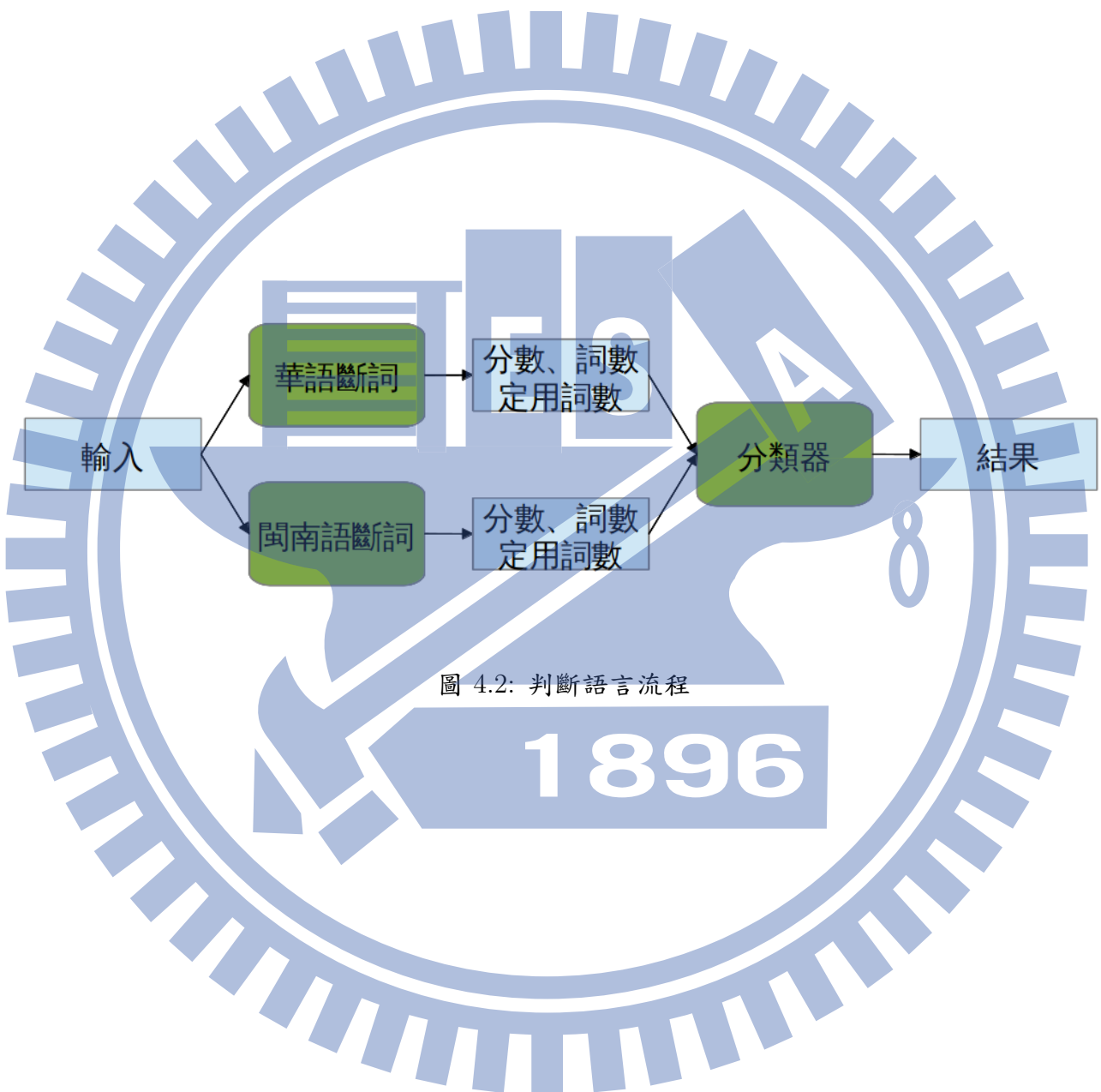


圖 4.2: 判斷語言流程

第五章 實驗結果

為著逐家後壁研究的方便，本論文研究的程式份結果，全部公開佇網路頂的專案 [65]。開發工具的版本會當看表5.1，詳細的參數，可比講語言模型用 Witten-Bell 加 discounting 的算法，翻譯模型用預設的訓練包攏會當佇內底的設定看著。

5.1 閩南語斷詞實驗

本實驗是提教育部辭典的 35130 個詞條當做訓練語料，試驗語料是教育部辭典例句 8027 句。共試驗語料的斷詞資訊提掉了後，用拄好長度斷詞份長詞優先去斷詞，才閣份原本例句的斷詞比較，得著表5.2的結果。

會當看著拄好長度斷詞的分數有比長詞優先閣較好淡薄，毋過無明顯的進步。這兩種斷詞方法的精確率攏比召回率低誠濟，對表2.9的公式來看，斷詞斷出來的詞數比答案的詞數閣較濟，代表有的詞無揣出來，有一部份原因可能是辭典內底收的詞閣無夠濟。

表 5.1: 實驗工具版本

工具	版本
臺灣言語工具 [64]	0.5.0
Moses[52]	commit 40c819d285cdeb40c0b8cc428bfde2fcb531b655
GIZA++[47]	1.0.7
SRILM[50]	1.7.0

表 5.2: 閩南語斷詞的效果

斷詞方法	召回率	精確率	F 測量
拄好長度斷詞	91.1	85.1	88.0
長詞優先斷詞 (對頭前)	91.0	84.9	87.9
長詞優先斷詞 (對後壁)	91.1	85.0	88.0

表 5.3: 新聞語料庫恰臺文典藏互相整理的實驗

整理幾擺	原始語料	1	2	3	4	5
BLEU 分數	9.30	14.72	13.77	13.82	13.82	13.82

5.2 語料整理實驗

完整的閩南語語料愛有全漢、全羅恰斷詞資訊。因為新聞語料庫有全漢、全羅，無斷詞，臺文典藏有斷詞毋過無完整的全漢恰全羅。

本實驗是提教育部辭典的 35130 個詞條恰附錄 388 句當做標準語料，用拄好長度斷詞來整理新聞平行語料 64121 句恰臺文典藏 329476 句。因為逐個資料庫有的資訊攏無全，會使做親像圖 5.1 全款，用教育部辭典恰臺文典藏共新聞語料庫斷詞，閣共斷好的新聞語料庫恰教育部辭典提來標臺文典藏的全漢，做幾仔擺的整理。

表 5.3 是整理的結果，分數是用詞為單位拍的。整理的結果一息仔就收斂，會當看著整理了的分數比猶未整理前好欲一半。毋過整理第二擺了後，分數有降一寡，看整理了的結果，是因為新聞恰臺文典藏內底的攏有一寡錯誤，所以第二擺用著遮的資料，會影響著整理的結果。

表 5.4: 新聞語料庫佇互相整理的變化

原始語料	指 ^ㄓ tsí	用 ^ㄩ īng	二 ^ㄉ jī	十 ^ㄕ tsap	三 ^ㄙ sann	ㄗ é	字 ^ㄗ jī	母 ^ㄇ bú	ㄗ tiānn	ㄗ tiānn
整理 1 擺	指 ^ㄓ tsí	用 ^ㄩ īng	二 ^ㄉ jī	十 ^ㄕ tsap	三 ^ㄙ sann	个 ^ㄗ é	字 ^ㄗ jī	母 ^ㄇ bú	定 ^ㄗ tiānn	定 ^ㄗ tiānn
整理 2 擺	指 ^ㄓ tsí	用 ^ㄩ īng	二 ^ㄉ jī	十 ^ㄕ tsap	三 ^ㄙ sann	的 ^ㄗ é	字 ^ㄗ jī	母 ^ㄇ bú	定 ^ㄗ tiānn	定 ^ㄗ tiānn
...										

表 5.5: 臺文典藏佇互相整理的變化

原始語料	佇 ^ㄗ tī	已 ^ㄧ í	經 ^ㄕ king	開 ^ㄕ khui	出 ^ㄗ tshut	的 ^ㄗ é	選 ^ㄗ suán	票 ^ㄗ phiò	中 ^ㄗ tiong
整理 1 擺	佇 ^ㄗ tī	已 ^ㄧ í	經 ^ㄕ king	開 ^ㄕ khui	出 ^ㄗ tshut	的 ^ㄗ é	選 ^ㄗ suán	票 ^ㄗ phiò	中 ^ㄗ tiong
整理 2 擺	佇 ^ㄗ tī	已 ^ㄧ í	經 ^ㄕ king	開 ^ㄕ khui	出 ^ㄗ tshut	的 ^ㄗ é	選 ^ㄗ suán	票 ^ㄗ phiò	中 ^ㄗ tiong
...									

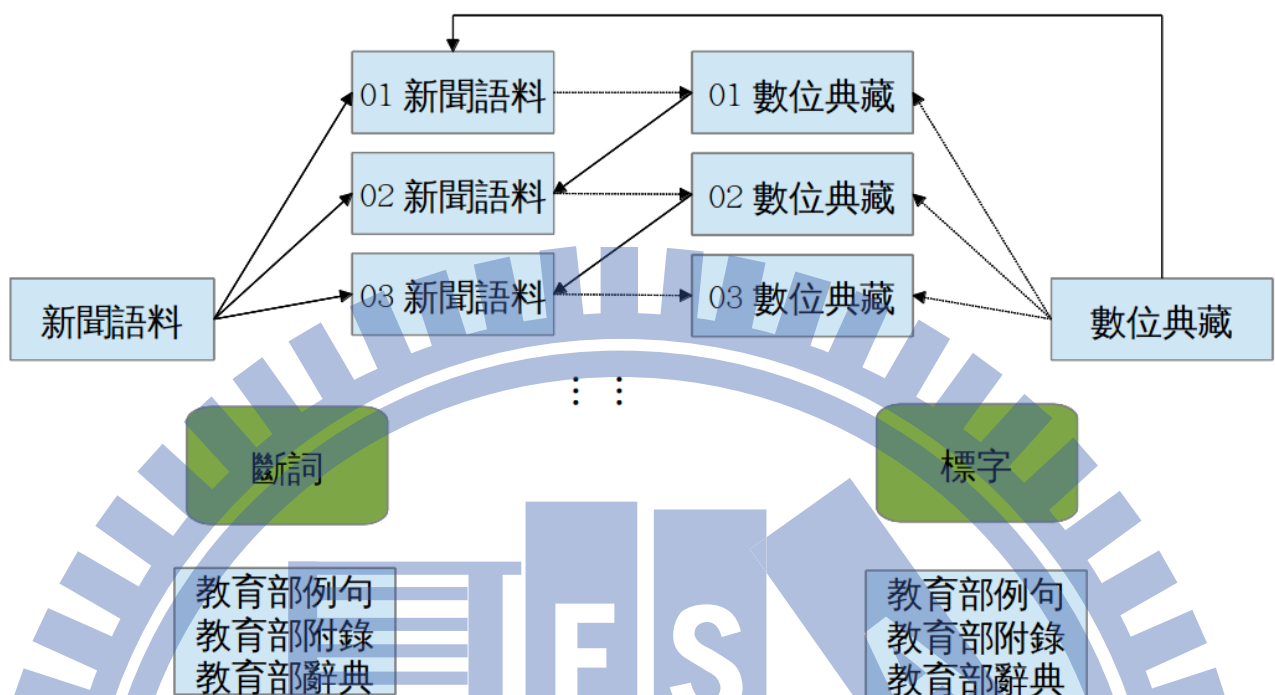


圖 5.1: 互相整理流程

可比講佇表5.4的例，一開始「ㄜ」無漢字，頭一擺整理得著正確的「个ㄜ」，毋過第二擺煞得著錯誤的「的ㄜ」。阮揣出這種情形的原因，是因為頭一擺整理時，辭典內底無「二十三」這個詞，所以斷詞斷做「二十 三」揀出「三 个」，第二擺整理時，辭典內底有「二十三」這個詞，毋過語言模型無「二十三 个」的出現頻率，就選出「三 的」。表5.5的例是因為有語料是「是 對 海關 出的」，斷詞的時陣就斷毋著矣。

5.3 分類語言實驗

這節實驗的語料是對 TGB 通訊創刊開始，到 2014 年 6 月 12 日為止攞總 177 期 1179 篇文章，提出頭前 1000 篇做訓練語料，閩南語有 9368 段 488844 詞，華語有 8519 段

圖 5.2: 無仝特徵詞數量，分類 3741 段閩南語華語

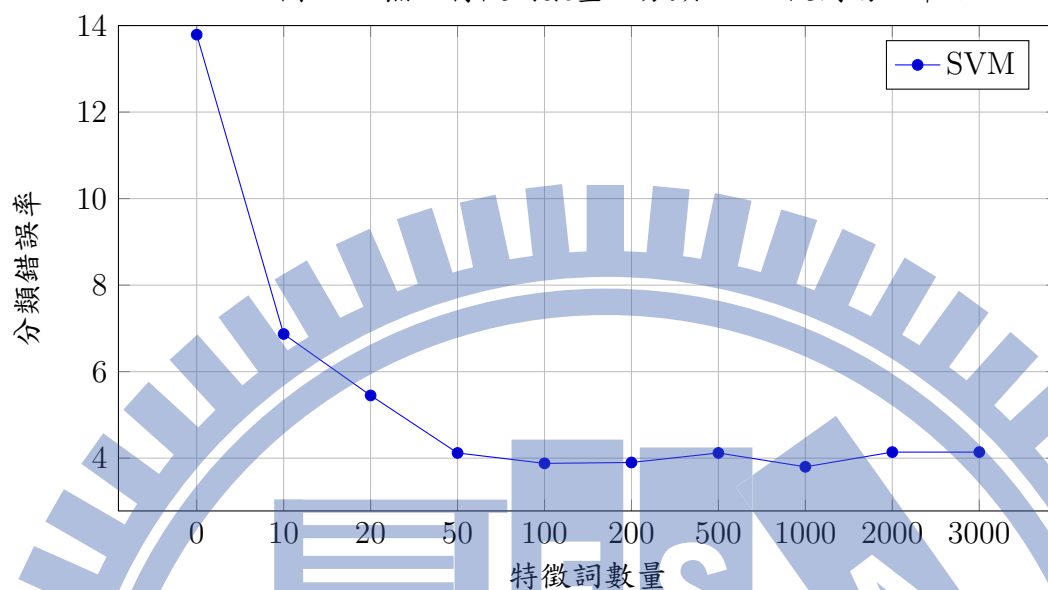


表 5.6: 加入 TGB 語料的翻譯效果

	加 TGB 語料前	加 TGB 語料後
平行語料句數	64121 句	99146 句
BLEU 分數	13.82	19.33

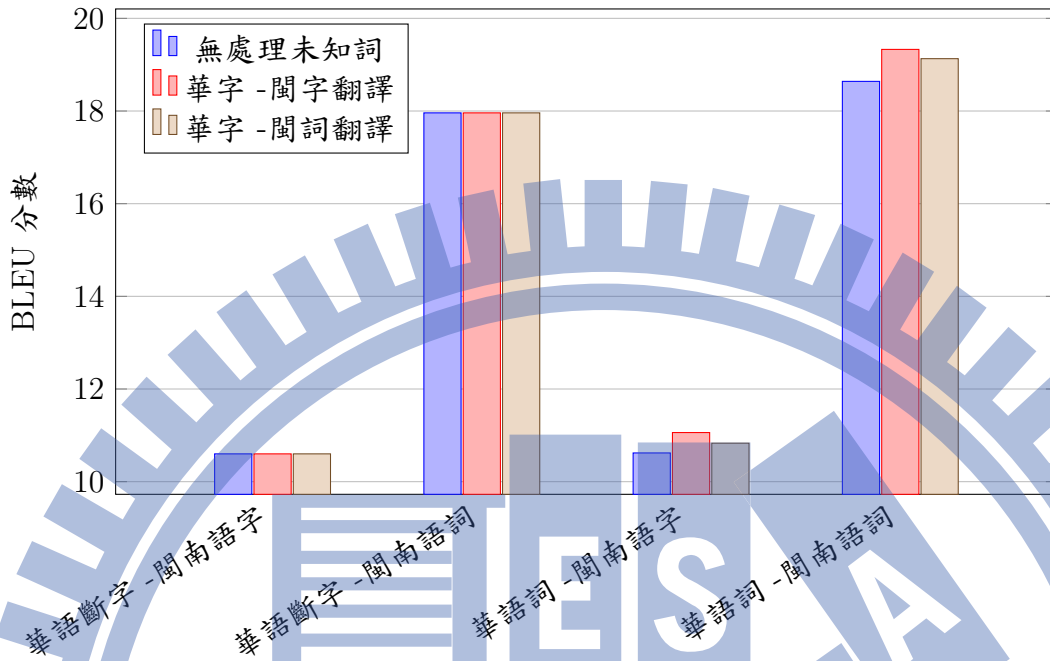
439436 詞；後壁 179 篇做試驗語料，閩南語有 1344 段 75282 詞，華語有 2397 段 114901 詞。以段做辨識單位，提來予支援向量機分類。

毋過 6012 个特徵實在是傷濟矣，所以阮試看覓共 3000 特徵詞減少，看會影響著辨識率無。實驗結果佇圖 5.2，佇 50 ~ 100 个特徵詞分類效果就收斂矣，加閣較濟的特徵詞，無啥影響著辨識的效果。

5.4 加入 TGB 語料庫實驗

頂一節做分類語言的實驗，繼落來就是共 TGB 語料摻入來翻譯語料。

圖 5.3: 斷字俚斷詞語料的翻譯效果比較—BLEU 用詞拍分數



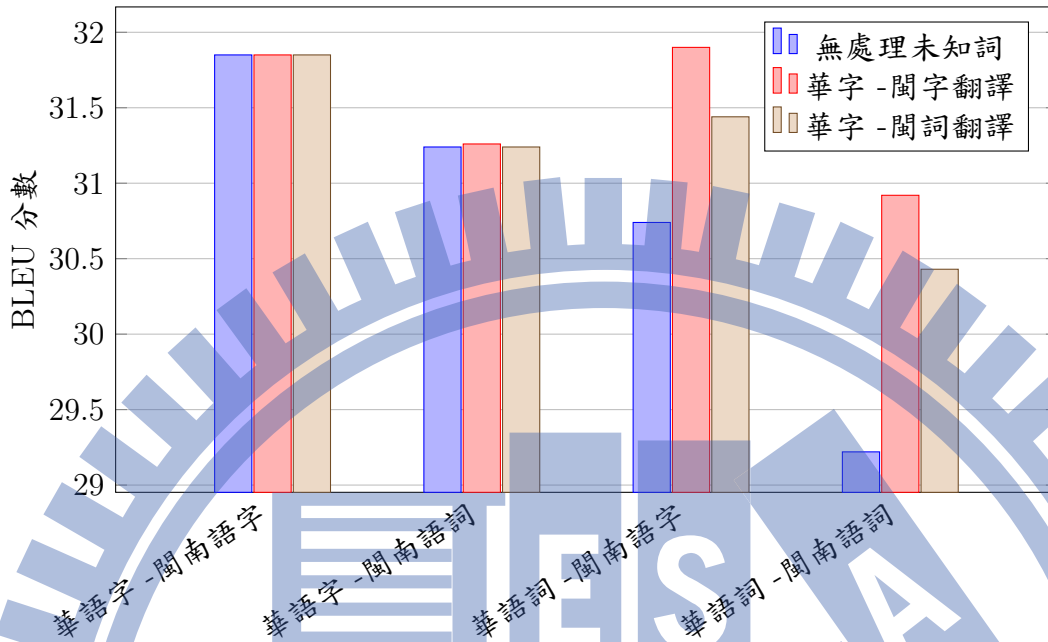
先提教育部詞條俚附錄句和 5.2 節整理了的新聞俚典藏語料來整理 TGB 語料，閣用 Bleualign 來對齊，就會使提著 35025 句 TGB 平行語料。

實驗訓練語料除了 5.2 節的語料外，閣加入 TGB 平行語料，除了教育部附錄句、典藏千焦會使做語言模型以外，原本平行語料千焦是新聞語料庫 64121 句，加入 TGB 語料 35025 句了後變做 99146 句，對表 5.6 會當看著進步誠濟，代表平行語料閣無夠，閣佇加語料就會使幫助翻譯的階段。

5.5 斷詞樣式俚斷字樣式的翻譯結果實驗

為著愛無全形式的語料對翻譯效果的影響，就共全部的狀況試一擺。華語俚閩南語兩種語言，摻字俚詞兩種形式，攏總有「華語字-閩南語字」、「華語字-閩南語詞」、「華語詞-閩南語字」，「華語詞-閩南語詞」四個狀況。

圖 5.4: 斷字恰斷詞語料的翻譯效果比較—BLEU 用字拍分數



因為華語用詞做單位，會出現3.2節的未知詞問題，若翻譯袂出來，就用「華語字-閩南語字」抑是「華語字-閩南語詞」的翻譯模型去處理，結果會當看圖5.3，這個實驗的訓練、試驗語料恰頂一个實驗是全款的。

因為上尾比較是用詞做單位，所以「華語字-閩南語字」恰「華語詞-閩南語字」的結果愛閣重斷工，會當看著「華語詞-閩南語詞」的翻譯效果上好，毋閣若用字做單位，佇圖5.4的實驗，「華語字-閩南語詞」分數上懸。

這個實驗會當看著幾件代誌，第一个是用詞拍分數的情形下「華語詞-閩南語詞」效果上好，因為翻譯了的文本若需要語音合成抑是再利用，攞需要斷詞的資訊，代表講用「華語字-閩南語詞」是上好的。雖然用字做單位拍分數時伊分數毋是上懸，毋過別的組合需要重斷詞，會當看著重斷詞影響著上尾的分數誠濟。

第二个是未知詞若有處理，對翻譯有幫助，而且未知詞用「華語字-閩南語字」的效果

比「華語字 - 閩南語詞」閣較好，可能是因為未知詞大部份攞是需要一字一字照翻的，嘛有可能是語料無夠濟，斷字對斷字的統計數量較濟。

上尾第三件代誌是用字拍分數的角度來看，華語的斷詞對翻譯有幫助，毋過閩南語的斷詞煞無。可能是閩南語的斷詞閣無夠準，造成翻譯效果變糗。



第六章 結論佢未來發展

目前臺灣的本土語言佇沓沓仔流失，需要逐家用心來注意，這嘛是這篇論文研究華語翻譯到閩南語的上主要目的。下跤會盡量共經驗分享予其他母語，希望閩南語以外，客話佢咱的原住民南島語，嘛會使利用這個研究成果。

6.1 結論

本論文提出「拄好長度斷詞」的方法，試圖改善「長詞優先」的缺點，毋過佇5.1節的實驗內底，對無濟訓練語料來講，兩種斷詞方法的效果略略仔爾。

佇5.2節佢5.4節的實驗，會當看著整理閩南語的效果，對原本的 9.30 分加到 19.33 分，代表對閩南語來講，這馬無夠十萬句的語料，是翻譯效果穰的一大限制。對臺灣別的本土語言來講，語料數量是一個上大的問題。

分類華語佢閩南語兩種語言佇5.3節嘛有 96% 正確的成績，若愛一擺分類三種以上 n 个漢語，會使簡化做兩種漢語的分類，先共 n 个語言，兩個兩個語言彼此之間揣出特徵詞，訓練出 $\frac{n \times (n-1)}{2}$ 个分類模型，上尾投票，看佢一个語言贏較濟，就當作是彼个語言。

可比講這馬有閩南語、客家四縣話、客家饒平話、佢華語四種語言愛分類，先訓練「閩南語 - 客家四縣話」、「閩南語 - 客家饒平話」、「閩南語 - 華語」、「客家四縣話 - 客家饒平話」、「客家四縣話 - 華語」、「客家饒平話 - 華語」六个模型，閣來共試驗語料提入來予六个模型判斷，分別判斷出「客家四縣話」、「客家饒平話」、「閩南語」、「客家饒平話」、「客家四縣話」、「客家饒平話」，看這六个模型分類內底「客家饒平話」上濟，就判斷做「客家饒平話」的語料。準做有需要分類南島語，會當看2.11.1節的說明。

繼落來講兩個會當加強翻譯的方法，恰一個翻譯模型會當閣利用的所在：

6.2 機器校對

5.2節的實驗結果共阮講，就算阮提著的語料母是蓋完整，阮嘛會當補足伊的資訊，予翻譯變好。這嘛表示語料正確度對翻譯有影響，用機器整理語料有伊的限制，若愛予語料正確，一定就需要人工校對。

毋過人工校對是開錢開時間開氣力的代誌，準做有機器校對系統，予資料的正確率閣較懸，就會減少人工檢查的負擔，嘛會當閃避無彩工，一直改全款的語料錯誤。

所以做一个即時更新(Online)的機器校對系統就是重要的問題，這個問題準做有 n 組人工檢查前的錯誤語料 t_i 恰人工檢查了的標準語料 p_i ，決定一開始的訓練語料數量 m 組，希望繼落來第 $m+1, m+2, \dots, n$ 組攞會當用進前校對過的資料來做機器校對，予機器校對的結果 p'_i 改做 p_i 的人工校對功夫上少，會當定義寫做公式6.1。

$$\sum_{i=m+1}^n \{p'_i \text{編輯到 } p_i \text{ 的人工校對成本}\}, \quad (6.1)$$

其中 p'_i 是 t_i 機器校對的結果，

機器校對是 $(t_1, p_1), (t_2, p_2), \dots, (t_{i-1}, p_{i-1})$ 訓練的

發展技術恰照顧語料對發展臺灣母語的研究來講平平仔重要，絕對袂使重視技術煞袂記得語料。

表 6.1: 字幕辨識問題分析

	輸入	輸出
語音辨識	臺灣母語的語音	臺灣母語的字幕
翻譯	華語的字幕	臺灣母語的字幕
字幕辨識	臺灣母語的語音 華語的字幕	臺灣母語的字幕

6.3 斷詞

華語有夠濟的斷詞標記語料，所以華語斷詞是一個發展足完整的技術。毋過閩南語的語料無華語遐爾濟的標記，就會影響著斷詞效果。

本論文是用對「長度優先」的算法改做「挂好長度斷詞」，除了這個方法以外，嘛會使閣用統計方法看斷詞的效果會較好無。用統計的方法會使用翻譯工具來做，一字一字斷字的輸入配合一詞一詞斷詞的輸出，提去訓練翻譯模型，按呢就有一個統計的斷詞工具。

到底佢一個方法對閩南語、客話這種只有幾萬、幾十萬句語料，效果會閣較好，就需要另外研究矣。

6.4 字幕辨識

臺灣母語的語音資料其實誠濟，親像電視劇、廣播攞有誠濟的語音資訊，只毋過遮的聲音語料大部份攞是配華語字幕。若有一個工具會當補母語字幕，阮就會當用這母語資訊來教學，抑是會來做語言學佢自然語言處理的研究。

字幕辨識的問題就是輸入「臺灣母語的語音」佢「華語的字幕」，想欲輸出「臺灣母語的字幕」。會當看表6.1，其實這個問題是綜合語音辨識佢翻譯的問題，語音辨識會當對「臺灣母語的語音」提供字詞佢先後的資訊，翻譯會當對「華語的字幕」提供詞的機率，語言模型會當提供語句的合理性，就需要另外的研究看這三个物件按怎用合理的數學方式鬥起來。



參考文獻

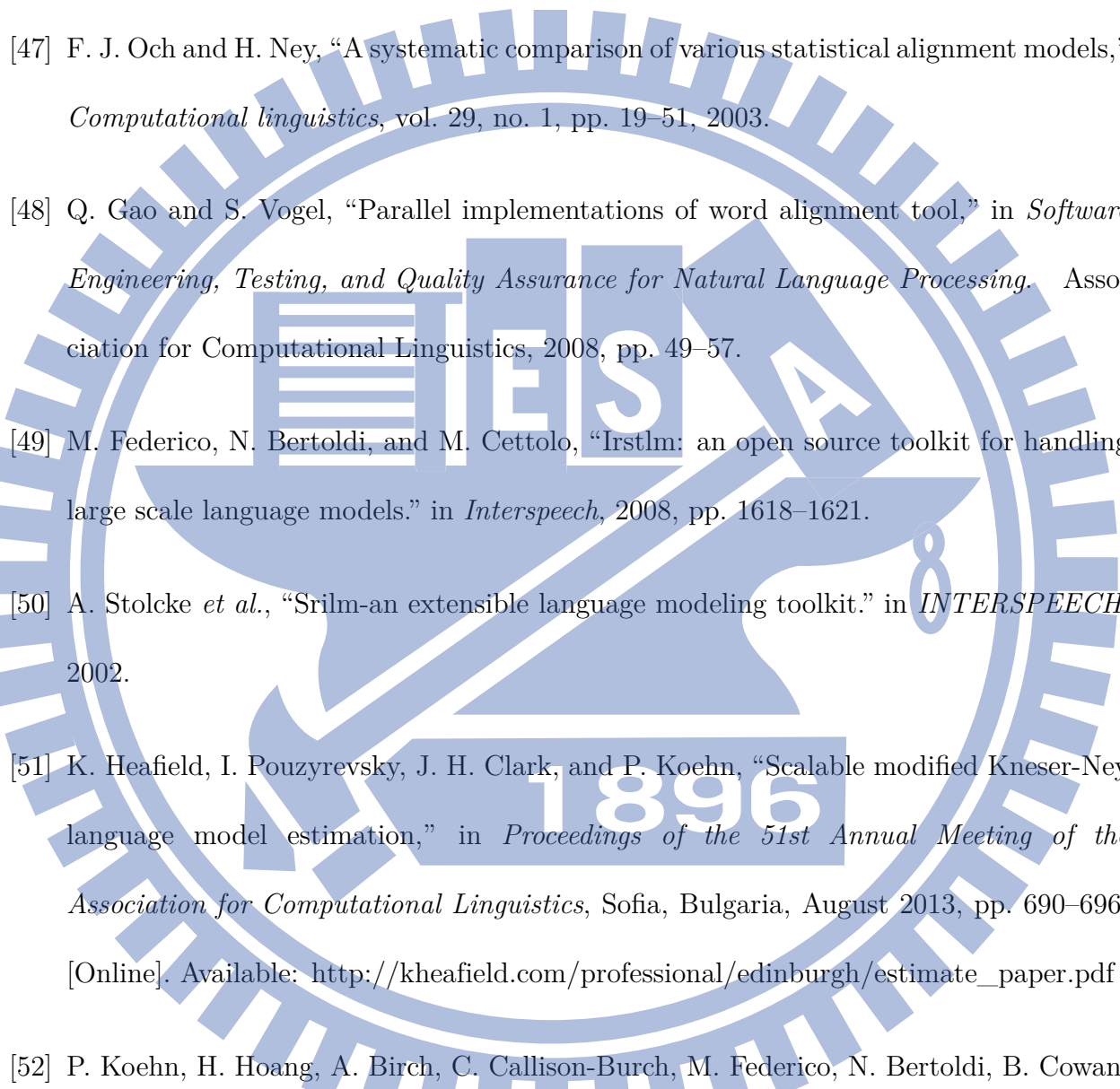
- [1] F.-f. Hsieh, "Low vowel raising in sinitic languages: Assimilation, reduction, or both?" *Language and Linguistics*, vol. 13, pp. 583–623, 2012.
- [2] 內政部戶政司, "國籍之歸化取得人數." [Online]. Available: <http://sowf.moi.gov.tw/stat/year/y02-06.xls>
- [3] 李壬癸, 臺灣原住民史—語言篇. 臺灣省文獻委員會, 1999.
- [4] G. F. S. Lewis, M. Paul and C. D. F. (eds.), *Ethnologue: Languages of the World*, seventeenth ed. Dallas, Texas: SIL International, 2014. [Online]. Available: <http://www.ethnologue.com>
- [5] 平埔文化資訊網. 中研院民族所數位典藏. [Online]. Available: <http://www.ianthro.tw/p/64>
- [6] 何萬順, 2008 年, "滅絕與新生：從多語言文化看外省族群的母語與國語," in 東吳大學外國語文學院 2008 年校際學術研討會:『多語言文化 -教學與研究』(2008.3.29), 三月二十九日 2008.
- [7] 洪惟仁, "台語不是沒有文字, 只是漢字不夠用," in 台語文學與台語文字. 前衛出版社, 1992.
- [8] 董忠司, "早期臺灣語裡的非漢語成分初探," in 『臺灣閩南語概論』講授資料彙編. 臺灣語文學會, 1996.

- [9] 許成章, 臺灣漢語辭典. 自立晚報, 1992.
- [10] 張光宇, 閩客方言史稿. 南天書局有限公司, 1996.
- [11] 教育部, 臺灣客家語羅馬字拼音方案使用手冊, 2012. [Online]. Available: http://www.edu.tw/FileUpload/3653-15592/Documents/hakka_pinyin3.pdf
- [12] —, 臺灣客家語常用詞辭典. [Online]. Available: <http://hakka.dict.edu.tw/hakkadict/>
- [13] 台灣名稱的由來. [Online]. Available: http://content.edu.tw/junior/co_tw/ch_yl/city/citaiwan.htm
- [14] 客語外來語: 含原閩客國(語)互借詞. 四縣腔, 初版 ed. 客委會, 2011, p. 94.
- [15] 甲(單位) - 維基百科, 自由的百科全書. [Online]. Available: <http://zh.wikipedia.org/wiki/%E5%8F%B0%E7%94%B2>
- [16] 教育部, 臺灣閩南語羅馬字拼音方案使用手冊, 2008. [Online]. Available: <http://www.edu.tw/FileUpload/3677-15601/Documents/tshiutsheh.pdf>
- [17] 吳守禮, 華、台語注音符號溯源. [Online]. Available: <http://olddoc.tmu.edu.tw/chiaushin/marker-0.htm>
- [18] —, 荔鏡記戲文研究——校勘篇. 從宜編輯室, 1961.
- [19] 教育部, 臺灣閩南語推薦用字 700 字表. [Online]. Available: <http://olddoc.tmu.edu.tw/chiaushin/marker-0.htm>

- [20] 小川尚義, 臺日大辭典. 臺灣總督府, 1931,1932.
- [21] ——, 日臺大辭典. 臺灣總督府, 1907.
- [22] 國際音標 - 維基百科, 自由的百科全書. [Online]. Available: <http://zh.wikipedia.org/wiki/%E5%9C%8B%E9%9A%9B%E9%9F%B3%E6%A8%99>
- [23] 白話字 - 維基百科, 自由的百科全書. [Online]. Available: <http://zh.wikipedia.org/wiki/%E7%99%BD%E8%A9%B1%E5%AD%97>
- [24] 臺灣語言音標方案 - 維基百科, 自由的百科全書. [Online]. Available: <http://zh.wikipedia.org/wiki/%E8%87%BA%E7%81%A3%E8%AA%9E%E8%A8%80%E9%9F%B3%E6%A8%99%E6%96%B9%E6%A1%88>
- [25] F. De Saussure, *Course in general linguistics*. Columbia University Press, 2011.
- [26] M. Yip, “Lexicon optimization in languages without alternations,” 1996.
- [27] 林明雄, 廣韻注漳州漢音. 太普公關, 2002. [Online]. Available: <http://books.google.com.tw/books?id=gW4sQwAACAAJ>
- [28] 余迺永, 新校互註宋本廣韻: (定稿本). 里仁書局發行, 2010. [Online]. Available: http://books.google.com.tw/books?id=R_0Z_gAACAAJ
- [29] 隱馬爾可夫模型 - 維基百科, 自由的百科全書. [Online]. Available: <http://zh.wikipedia.org/zh-tw/%E9%9A%90%E9%A9%AC%E5%B0%94%E5%8F%AF%E5%A4%AB%E6%A8%A1%E5%9E%8B>

- [30] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, 2006, vol. 3.4.
- [31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, IEEE Catalog No.: CFP11SRW-USB.
- [32] H. Zen, T. Nose, T. Masuko, A. W. Black, and K. Tokuda, “The hmm-based speech synthesis system (hts) version 2.0.” Citeseer.
- [33] U.-G. Iunn, S. an Li, K.-G. Lau, and C.-Y. Kao, “台語變調系統實作研究 (a study on implementation of taiwanese tone sandhi system) [in chinese],” in *ROCLING’05*, 2005, pp. –1–1.
- [34] 林川傑 and 陳信希, “中文到閩南語之線上翻譯及閩南語之語音合成,” in 數位博物館中之語文處理技術研討匯集刊, 南港, 1999, pp. 11.1–11.4.
- [35] 李雪貞, “客家語語音合成之初步研究,” 2002.
- [36] 楊允言, 劉杰岳, and 李盛安, 台語羅馬字發音試驗系統, 2005. [Online]. Available: <http://210.240.194.97/tts/tts.asp>
- [37] 陳信宏, 余秀敏, 羅烈師 *et al.*, “客語文句轉語音及語音辨認之研究,” 2008.

- [38] 蔡依玲, “基於隱藏式馬可夫模型之客語文句轉語音系統,” 交通大學電信工程系所學位論文, pp. 1–59, 2010.
- [39] 薛丞宏, 意傳文化科技, 2013. [Online]. Available: <http://意傳.台灣/>
- [40] R. X. Z.-s. Z. Kam-Fai Wong, Wenji Li, *Introduction to Chinese Natural Language Processing*. Morgan & Claypool Publishers, 2010.
- [41] 曾金金, “台語斷詞原則討論,” in 台灣文學出版物收集、目錄、選讀編輯計畫結案報告說明. 行政院文化建設委員會, 1997, pp. 45–72. [Online]. Available: <http://ip194097.ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm>
- [42] W.-Y. Ma and K.-J. Chen, “Introduction to CKIP chinese word segmentation system for the first international chinese word segmentation bakeoff,” in *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 168–171.
- [43] 楊允言, 張學謙 (共同), and 陳克健 (協同), 台語文語法結構樹建置研究成果報告, 國科會, 2009. [Online]. Available: <http://www.etpc.org.tw/jspui/handle/10537/19562;jsessionid=4902fc99c8f0245b328eede86732>
- [44] K.-J. Chen and Y.-M. Hsieh, “Chinese treebanks and grammar extraction,” in *Natural Language Processing-IJCNLP 2004*. Springer, 2005, pp. 655–663.
- [45] D. Klein and C. D. Manning, “Accurate unlexicalized parsing,” in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 423–430.

- 
- [46] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [47] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
- [48] Q. Gao and S. Vogel, “Parallel implementations of word alignment tool,” in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 49–57.
- [49] M. Federico, N. Bertoldi, and M. Cettolo, “Irstlm: an open source toolkit for handling large scale language models.” in *Interspeech*, 2008, pp. 1618–1621.
- [50] A. Stolcke *et al.*, “Srlm-an extensible language modeling toolkit.” in *INTERSPEECH*, 2002.
- [51] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, “Scalable modified Kneser-Ney language model estimation,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013, pp. 690–696.
- [Online]. Available: http://kheafield.com/professional/edinburgh/estimate_paper.pdf
- [52] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation,” in *Proceedings of the 45th*

- Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1557769.1557821>
- [53] *Moses - Moses/Baseline*. [Online]. Available: <http://www.statmt.org/moses/?n=moses.baseline>
- [54] *multi-bleu.perl*. [Online]. Available: <https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>
- [55] W. B. Cavnar, J. M. Trenkle *et al.*, “N-gram-based text categorization,” *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175.
- [56] W. A. Gale and K. W. Church, “A program for aligning sentences in bilingual corpora,” *Computational linguistics*, vol. 19, no. 1, pp. 75–102, 1993.
- [57] R. Sennrich and M. Volk, “Mt-based sentence alignment for ocr-generated parallel texts,” in *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, November 2010. [Online]. Available: <http://dx.doi.org/10.5167/uzh-38464>
- [58] 教育部, 臺灣閩南語常用詞辭典. [Online]. Available: <http://twblg.dict.edu.tw/>
- [59] 陳孟彰 and 何澤政, *iCorpus 臺華平行新聞語料庫*, 2014. [Online]. Available: <http://icorpus.iis.sinica.edu.tw/>

- [60] 楊允言, 台華線頂辭典. [Online]. Available: <http://ip194097.ntcu.edu.tw/iug/Ungian/SoannTeng/chil/Taihoa.asp>
- [61] 台語文數位典藏資料庫. [Online]. Available: <http://xcm.nmtl.gov.tw/dadwt/pbk.asp>
- [62] TGB 通訊. [Online]. Available: <http://taioan-chouhap.myweb.hinet.net/>
- [63] 關於 TGB / 關係 TGB @ 台灣組合:: 痞客邦 PIXNET ::. [Online]. Available: <http://taioanchouhap.pixnet.net/blog/post/32374696>
- [64] 薛丞宏, “臺灣言語工具,” 2014. [Online]. Available: https://github.com/sih4sing5hong5/tai5_uan5_gian5_gi2_kang1_ku7
- [65] ———, “翻譯研究,” 2014. [Online]. Available: https://github.com/sih4sing5hong5/huan1-ik8_gian5-kiu3