

漢語間統計式機器翻譯初探 用臺灣閩南語示範

Initial Study of Statistical Machine Translation
between the Chinese Languages
Using Taiwan Southern Min as Examples

103/10/07

國立交通大學 資訊工程與科學研究所

0156016 薛丞宏

指導教授：張智星教授

易志偉教授

目錄

- 第一節：研究介紹
- 第二節：背景智識
- 第三節：語料樣式探討
- 第四節：無仝語料整理
- 第五節：語言分類
- 第六節：結論俾未來發展
- 第七節：參考文獻
- 附錄

第一節：研究介紹

- 臺灣是多元語言的國家
 - 南島語
 - 阿美、泰雅、噶哈巫、西拉雅、...
 - 漢語
 - 閩南語、客話、華語（官話）、二戰後移民...
 - 其他
 - 越南（新住民）、...
- 講母語是人上基本的權利
 - 毋過臺灣母語消失誠緊
 - 學習資源無夠

研究方向

- 共華語翻譯做母語
 - 華語資料誠濟
 - 予欲學的人參考
 - 配合語音合成
- 針對漢語之間的翻譯
 - 主要處理翻譯語料
 - 用閩南語示範，結果嘛會使用佇客話
- 閩南語的資料
 - 資料傷少
 - 閩南語-幾十萬句 / 華語 - 百萬句
 - 有漢字佮音標兩種表示方法
 - 部分語料漢字無正規化
 - 大部份語對料音標標本調，有一部分標變調

第二節：背景智識

- 語料狀況
- 語料庫介紹
- 語料樣式
- 翻譯模型
- 評分方式

閩南語語料種類

種類	範例	語料庫	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	TGB通訊	漢字拼音濫用
全漢全羅對應	我欲食飯 gua2 beh4 tsiah8-png7	新聞語料庫 教育部辭典	漢字恰拼音擁有
漢羅全羅對應	我beh4食飯 gua2 beh4 tsiah8-png7	臺語文數位典藏	部份字有漢字

- 逐個語料庫攏無仝
 - 後壁愛處理的問題

語料庫－新聞語料庫

- 全名「臺華平行新聞語料庫」
 - 中研院資訊所陳孟彰老師主持，何澤政翻譯

- 有華語恰對應的閩南語全漢全羅

- 華語的新聞
- 閩南語全羅
 - 何澤政翻譯
- 閩南語全漢

華語	這幾天 寒流 再度 發威
全羅	tsit4-kui2-kang1 han5-liu5 koh4-tsai3 tian2-ui1
全漢	這幾工 寒流 閣再 展威

- 根據全羅，參考教育部用字補起哩的

- 翻譯時，罕得調整語詞先後
- 現代用語、古早用語攏有
- 97/11/06到103/3/14的文章
 - 2567篇文章、64121句
 - 359554華語詞組、366190閩南語詞組

語料庫－教育部辭典

○ 全名「臺灣閩南語常用詞辭典」

○ 較濟生活用語

○ 內部有規範

- 資料攏有全漢佮全羅
- 音標有斷詞

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。
----	-----------

○ 例句有華語翻譯

- 附錄句無華語翻譯

漢字	我欲去買雞卵。
----	---------

混合腔音標	Guá beh khì bé ke-nṅg.
-------	------------------------

偏泉腔音標	Guá beh khì bé kue-nṅg.
-------	-------------------------

○ 逐个相近詞攏有例句

- 為著驚無平衡，另外加的

語料庫－數位典藏

- 全名「台語文數位典藏」
 - 國家臺灣文學館收集1885～2006年的語料
- 攏總2167篇
 - 詩387條、散文1127篇、小說387篇、劇本49篇
 - 的語料
- 漢羅佮全羅對照
 - 原本只有一種，臺文館後來倩人拍字
 - 有的劇本全羅內底有漢字

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

語料庫—TGB通訊

- 「學生台灣語文促進會」主持
- 一個月一期的部落格
- 對1999年10月到這馬
 - 實驗的資料到2014年6月12日，1179篇文章

○形式真濟款

- 華語
- 閩南語
- 華語閩南語平行
- 華語閩南語濫做伙

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? lah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？

濫做伙範例

『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。

腔口無仝

- 閩南語有許多腔調
 - 偏漳腔、混合腔、偏泉腔
 - 混和腔有較濟偏漳腔的特色
- 語料狀況
 - 教育部資料
 - 有地方腔，鹿港「火her2」
 - 主要資料有記錄腔口，附錄句無
 - 新聞語料庫
 - 澤政是臺中烏日人，60年代出身
 - 偏漳腔，有時陣會濫著泉腔
 - 數位典藏
 - 四界收集來的，無記錄腔
 - TGB通訊
 - 無註明腔，干焦漢羅，無法度算
- 全部資料濫做伙訓練
 - 資料無逐个註明

教育部	偏漳	混合	偏泉	外來語
字/詞	36142	33756	46654	172
例句	10637	9829	14227	0

漳/泉	雞 ke1/kue1	近 kin7/kun7	火 hue2/he2
附錄句	16/0	5/0	2/0
新聞語料庫	284/13	710/45	1037/25
數位典藏	229/135	458/365	703/390

語料樣式

- 一般漢語文字無詞恰詞組的資訊
 - 因為字攏黏做伙無空白
 - 這個資訊翻譯用會著

- 語料樣式

- 斷詞組
 - 語料庫提供
 - 人工斷的
- 斷詞
 - 華語
 - 中研院中文斷詞系統 (CKIP)
 - 閩南語
 - 需要斷詞方法
- 斷字

語料樣式	範例
斷詞組	蔡崇名 細漢時 生活困苦
斷詞	蔡崇名 細漢 時 生活 困苦
斷字	蔡 崇 名 細 漢 時 生 活 困 苦

上長詞優先斷詞方法

○ 上長詞優先

- 上定看著的斷詞方法
- 對後壁開始看，希望詞愈長愈好
 - 華語的實驗結果，效果比對頭前閣較好

○ 做法

1. 對上後壁的字開始
2. 揣一个佇辭典的上長詞
3. 揣著詞了後，繼續對第1步做到結束

上長詞優先斷詞範例

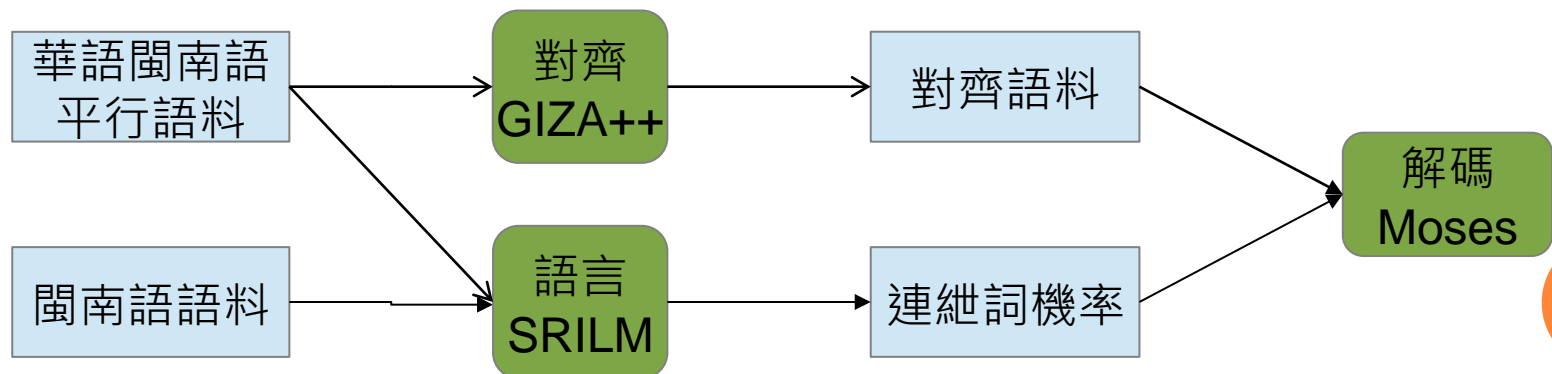
○ 範例 - 陸續開放一百五十項的規費

- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費
- 陸續開放一百五十項的規費

翻譯模型

○ 三个部份

- 對齊模型alignment model
 - 華語佾閩南語的詞對應機率
- 語言模型language model
 - 閩南語語句合理性
 - 連繼的詞接做伙的機率
- 解碼器decoder
 - 用對齊模型佻語言模型翻譯



對齊模型介紹

○ 目的

- 揣出華語佾閩南語的詞機率對應
- 功能親像雙語辭典
 - 對應資訊是對平行語料揣出來的

○ 做法

- 平行語料一句一句處理
- 記錄華語全部的詞佾閩南語全部的詞對應幾擺
- 對應較濟擺的對應組合就是翻譯選項之一

對齊模型範例

○ 輸入語料

1. 他 打 我/伊 共 我 拍

- 他-伊/他-共/他-我/他-拍
- 打-伊/打-共/打-我/打-拍
- 我-伊/我-共/我-我/我-拍

2. 打 鼓 很 好 玩/拍 鼓 誠 趣 味

- 打-拍/打-鼓/打- 誠/打-趣味
-

○ 「打」的對應結果

- 語料是整理過的

華-閩南對應	對應機率	華-閩南對應	對應機率
打-伊	1/8	打-鼓	1/8
打-共	1/8	打-誠	1/8
打-我	1/8	打-趣味	1/8
打-拍	2/8		

對齊模型種類無仝範例

○ 輸入語料

1. 他 打 我/伊 共 我 phah 單位

- 他-伊/他-共/他-我/他-phah
- 打-伊/打-共/打-我/打-phah
- 我-伊/我-共/我-我/我-phah

2. 打 鼓 很 好玩/拍 鼓 誠 趣味

- 打-拍/打-鼓/打- 誠/打-趣味
-

○ 「打」的對應結果

- 猶未整理的語料
- 後壁愛處理的問題

華-閩南對應	對應機率	華-閩南對應	對應機率
打-伊	1/8	打-鼓	1/8
打-共	1/8	打-誠	1/8
打-我	1/8	打-趣味	1/8
打-拍	1/8	打-phah	1/8

對齊模型樣式無仝範例

○ 斷詞組語料

1. 打鼓 很好玩/拍鼓 誠趣味
 - 打鼓-拍鼓/打鼓-誠趣味
2. 打電話 了沒/敲電話 矣未
 - 打電話-敲電話/打電話-矣未

○ 斷字語料

1. 打 鼓 很 好 玩/拍 鼓 誠 趣 味
 - 打-拍/打-鼓/打-誠/打-趣/打-味
2. 打 電 話 了 沒/敲 電 話 矣 未
 - 打-敲/打-電/打-話/.....

○ 樣式無仝影響統計

- 後壁愛處理的問題

華-閩南對應	對應機率	華-閩南對應	對應機率
打鼓-拍鼓	1/4	打-拍	1/10
打鼓-誠趣味	1/4	打-鼓	1/10
打電話-敲電話	1/4	打-敲	1/10
打電話-矣未	1/4	打-電	1/10

語言模型介紹

○ 目的

- 判斷語句合理性
- 語句分做細句，合理性就是逐個細句機率相乘

○ 訓練方法

- 先決定一個數字 n ，代表一攞看 n 個連繼的詞
- 共一句 k 個詞的句，產生 $k-n+1$ 組的 n 連繼詞
- 統計全部連繼詞的數量

○ 判斷方法

- 全款共試驗語句轉做 n 連繼詞
- 對逐個連繼詞，用頭前 $n-1$ 個詞，算第 n 個詞出現的條件機率
 - $p(\text{詞}_n | \text{詞}_1, \text{詞}_2, \dots, \text{詞}_{n-1})$
- 語句的合理性是全部 n 連繼詞的機率乘起來
 - $p(\text{一句話}) = p(\text{詞}_1, \text{詞}_2, \dots, \text{詞}_k)$
 - $\sim \prod_{i=1}^{k-n+1} p(\text{詞}_{i+n-1} | \text{詞}_i, \text{詞}_{i+1}, \dots, \text{詞}_{i+n-1})$

語言模型範例

- 假設 $n=2$ ，考慮二連繼詞
- 輸入語料
 1. 伊 共 我 拍
 2. 拍 鼓 誠 趣味
 3. 我 敲 電話 予 伊

連繼詞	對應機率	連繼詞	對應機率
$p([\text{句尾}] \text{拍})$	$1/2$	$p(\text{拍} \text{我})$	$1/2$
$p(\text{鼓} \text{拍})$	$1/2$	$p(\text{敲} \text{我})$	$1/2$
$p(\text{電話} \text{拍})$	10^{-10*}	$p([\text{句尾}] \text{電話})$	10^{-10*}
$p([\text{句尾}] \text{鼓})$	10^{-10*}		

*註：無出現過的機率，有專門的算法予機率加起來是1

語言模型範例

○ 假設 $n=2$ ，考慮二連繼詞

○ 語句機率

• 我 拍 鼓

○ $p(\text{拍}|\text{我}) \times p(\text{鼓}|\text{拍}) \times p([\text{句尾}]|\text{鼓})$

• 我 拍 電話

○ $p(\text{拍}|\text{我}) \times p(\text{電話}|\text{拍}) \times p([\text{句尾}]|\text{電話})$

連繼詞	對應機率	連繼詞	對應機率
$p([\text{句尾}] \text{拍})$	$1/2$	$p(\text{拍} \text{我})$	$1/2$
$p(\text{鼓} \text{拍})$	$1/2$	$p(\text{敲} \text{我})$	$1/2$
$p(\text{電話} \text{拍})$	10^{-10*}	$p([\text{句尾}] \text{電話})$	10^{-10*}
$p([\text{句尾}] \text{鼓})$	10^{-10*}		

*註：無出現過的機率，有專門的算法予機率加起來是1

BLEU評分

- Bilingual Evaluation Understudy
- 以連繼詞n-grams為單位
- $BLEU =$

$$100 \times e^{\max(0, \frac{\text{結果-答案長度}}{\text{結果長度}})} \left(\prod_{n=1}^4 \text{連繼詞}_n \right)^{\frac{1}{4}}$$

答案	這 幾 工 寒 流 閣 再 展 威
結果一	這 幾 工 寒 流 有 展 威
結果二	寒 流 這 幾 工 閣 再 展 威

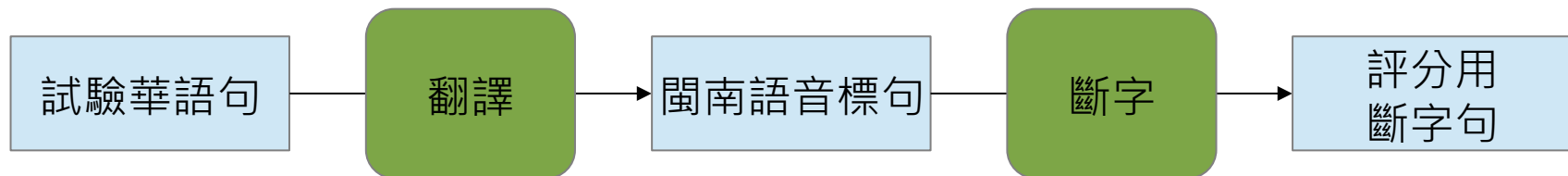
連繼詞 _n	n=1	n=2	n=3	n=4	BLEU
結果一	5/6	3/5	2/4	1/3	53.73
結果二	6/6	3/5	1/4	0/4	0.00

第三節：語料樣式探討

- 動機
 - 想欲改善翻譯效果
 - 改變語料樣式
- 問題
 - 探討它一種樣式較好
 - 改變翻譯的流程

原始斷詞組語料

- 使用新聞語料庫的原始斷詞組語料
 - 輸入華語漢字
 - 輸出對應閩南語的全漢全羅
 - 訓練語料2300篇新聞，57167句
 - 試驗語料267篇新聞，6954句
 - BLEU評分，用字為單位
 - 70.67分



未知詞問題

○ 原因

- 對齊模型、語言模型的單位攏是詞組
 - 因為語料是斷詞組
- 拄著無看過的詞組就會翻袂出來
 - 語料無全部的華語詞組

訓練語料1 動員了 一百五十位 志工 ，
liok8-siok8 siu1-tioh8 lak8-khin1 e5 soo2-tit4 kiau2-sue3-tuann1 。

訓練語料2 陸續 增加 好幾家 店面 。
liok8-siok8 tsing1-ka1 kui2-na7-king1 tiam3-bin7 。

訓練語料3 台南 空軍基地 要 在 十日 開放 參觀 ；
tong7-uan5-liau2 tsit8-pah4-goo7-tsap8-ui7 tsi3-kang1 ，

試驗輸入 陸續 開放 一百五十項 的 規費 ，

翻譯結果 liok8-siok8 khai1-hong3 一百五十項 e5 規費 ，

問題改善

現象研究

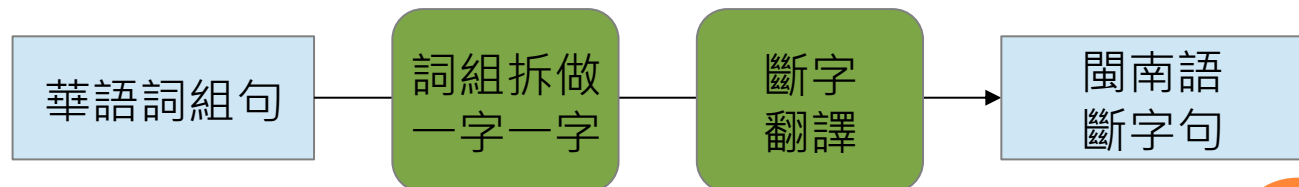
- 平行語料無法度充分利用
 - 訓練語料有「一百五十位」
 - 對「一百五十項」無法度用

解決方法

- 語料的單位對斷詞組改做斷字
 - 一个字做一个單位
 - 「一百五十位」的「一百五十」就會當充分利用

效果

- 82.94分



未知詞另外翻譯

○ 目的

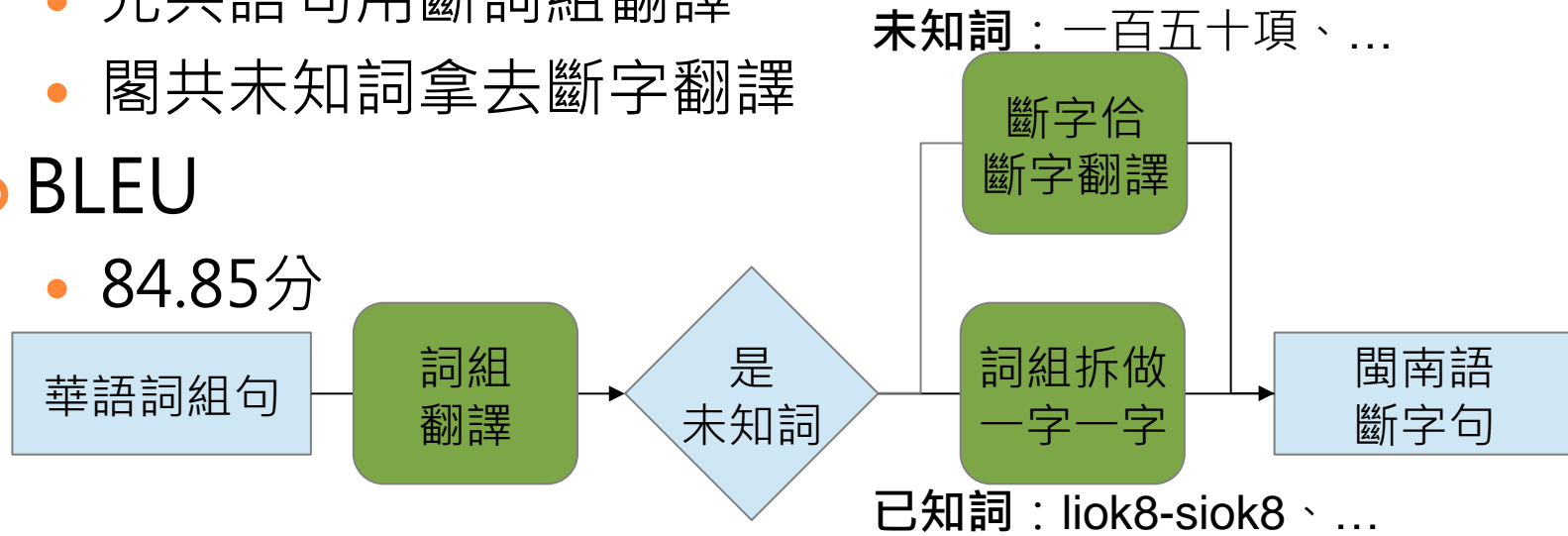
- 斷詞組翻譯有詞組資訊
- 斷字翻譯解決未知詞問題
- 揣一个方法綜合兩個方法的優點

○ 方法

- 先共語句用斷詞組翻譯
- 閣共未知詞拿去斷字翻譯

○ BLEU

- 84.85分



無仝樣式翻譯

○ 原因

- 來源的華語俚結果的閩南語形式會當無仝
- 試無仝樣式分別的效果

○ 華語樣式

- 原本斷詞組
- 中研院斷詞
- 斷字

○ 閩南語樣式

- 原本斷詞組
- 拄好長度斷詞
- 斷字

拄好長度斷詞方法

○ 目的

- 希望會當閃避上長詞優先的缺點
- 字平均分配到逐個詞
- 逐個詞愈長愈好

○ 方法

- 要求分數愈低愈好
- 對無全長度的詞分數無全
 - 一字詞1分
 - 兩字詞1/2分
 - 三字詞1/3分
 - ...
 - n字詞1/n分
- 用維特比 (Viterbi) 揣出分數上低的斷詞切法

上長詞優先 (後壁)

拄好長度斷詞

上長詞優先 (頭前)

拄好長度斷詞

甚至和國小學生嘛想袂開

甚至和國小學生嘛想袂開

猶掠做唱歌仔戲真簡單

猶掠做唱歌仔戲真簡單

拄好長度斷詞範例

○ 範例

- 頭前有一張椅仔
 - $1/2 + 1 + 1/2 + 1/2 = 2.5$ 分

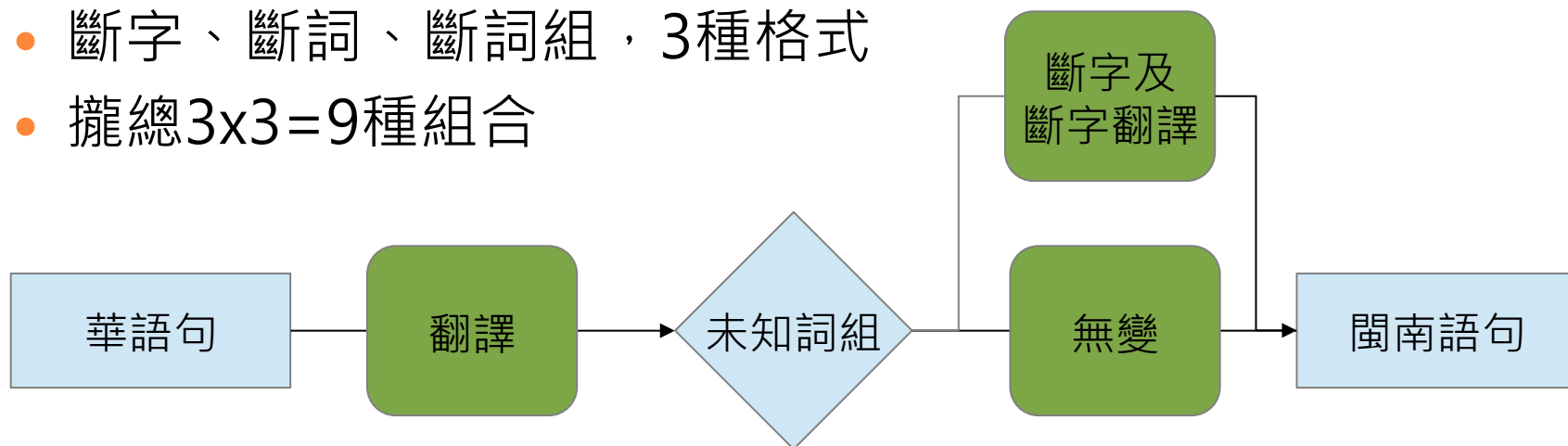
○ 缺點

- hoo7 i1 tsut4-khi3 sng2/予伊出去耍
- hoo7-i1 tsut4-khi3 sng2/雨衣出去耍

比較結果

○ 平行語料樣式

- 華語閩南語兩種語言
- 斷字、斷詞、斷詞組，3種格式
- 攏總 $3 \times 3 = 9$ 種組合



華語\閩南語樣式	斷字	上長詞優先*	拄好長度斷詞	斷詞組
斷字	82.94	82.74/82.78	82.81	80.23
斷詞	84.28	84.05/84.03	84.02	82.85
斷詞組	84.04	83.94/83.95	83.95	84.85

上長詞優先*：一個是對頭前斷詞，一個對後壁

第四節：語料整理

○ 目的

- 用加語料會當予翻譯的效果較好
- 翻譯語料樣式愛全款，較好
 - 原本逐個語料庫樣式攞無全款

○ 語料之間的問題

- 漢字用字無一致
- 無完整的全漢佢全羅
- 斷詞資訊

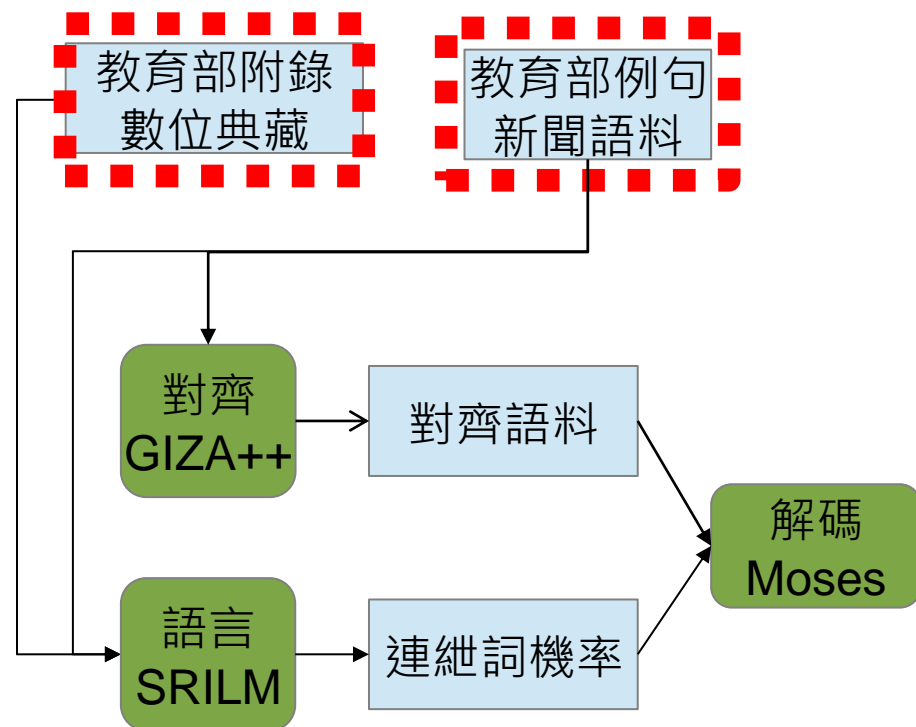
欲使用的語料

○ 對齊語料

- 教育部例句
 - 34693句、273619閩南詞
 - 377061字
- 新聞語料庫
 - 64121句、365855詞組
 - 759538字

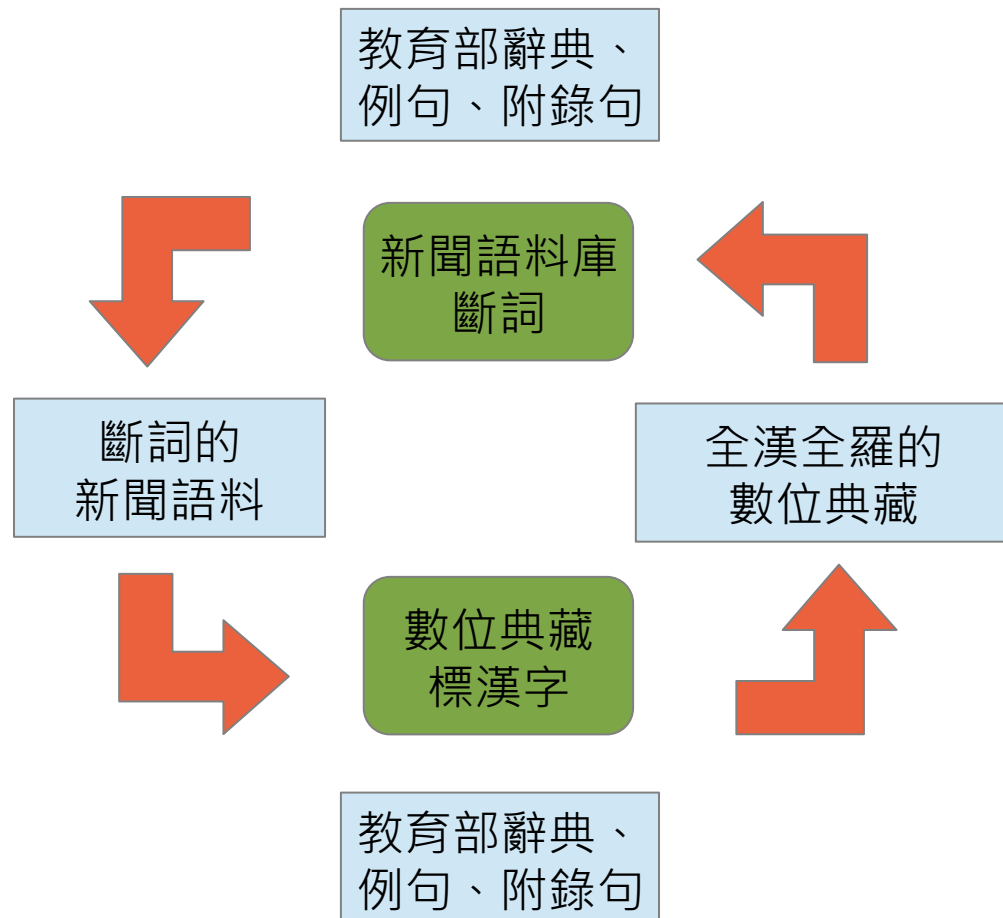
○ 語言模型

- 教育部附錄
 - 388句、3055詞、3795字
- 數位典藏
 - 2167篇、329476句
 - 加標點2250889詞、3027268字



語料無一致

- 教育部辭典
 - 斷詞
 - 全漢佖全羅
- 新聞語料庫
 - 斷詞組
 - 全漢佖全羅
- 數位典藏
 - 斷詞
 - 漢羅佖全羅



新聞語料庫斷詞

- 華語斷詞
 - 用中研院中文斷詞系統 (CKIP)
- 閩南語斷詞
 - 用教育部辭典、典藏的資料做辭典
 - 用拄好長度斷詞來斷新聞語料庫
 - 教育部辭典攏總116552詞

數位典藏標漢字

○ 問題

- 無全部的字擁有漢字
 - 典藏的是漢羅佮全羅
- 漢字用字佮教育部無仝

○ 方法

- 仝款用拄好長度斷詞
- 斷詞時查全漢、全羅、漢羅仝部形式
- 確定斷點後，選語言模型分數上懸的
- 查著了後，照原本全羅斷詞

頭家/thau5-ke1

頭-家

頭-ke1

頭-家ke1

thau5-家

thau5-ke1

thau5-家ke1

頭thau5-ke1

頭thau5-家

頭thau5-家ke1

標漢字流程

- 原本語料
 - 伊i1 ti7 tsiah8-飯png7-thiann1
- 斷詞了
 - 伊i1 佇ti7 食tsiah8-飯png7 廳thiann1
- 照原本斷詞
 - 伊i1 佇ti7 食tsiah8-飯png7-廳thiann1

整理流程圖 - 一開始

新聞語料

斷詞

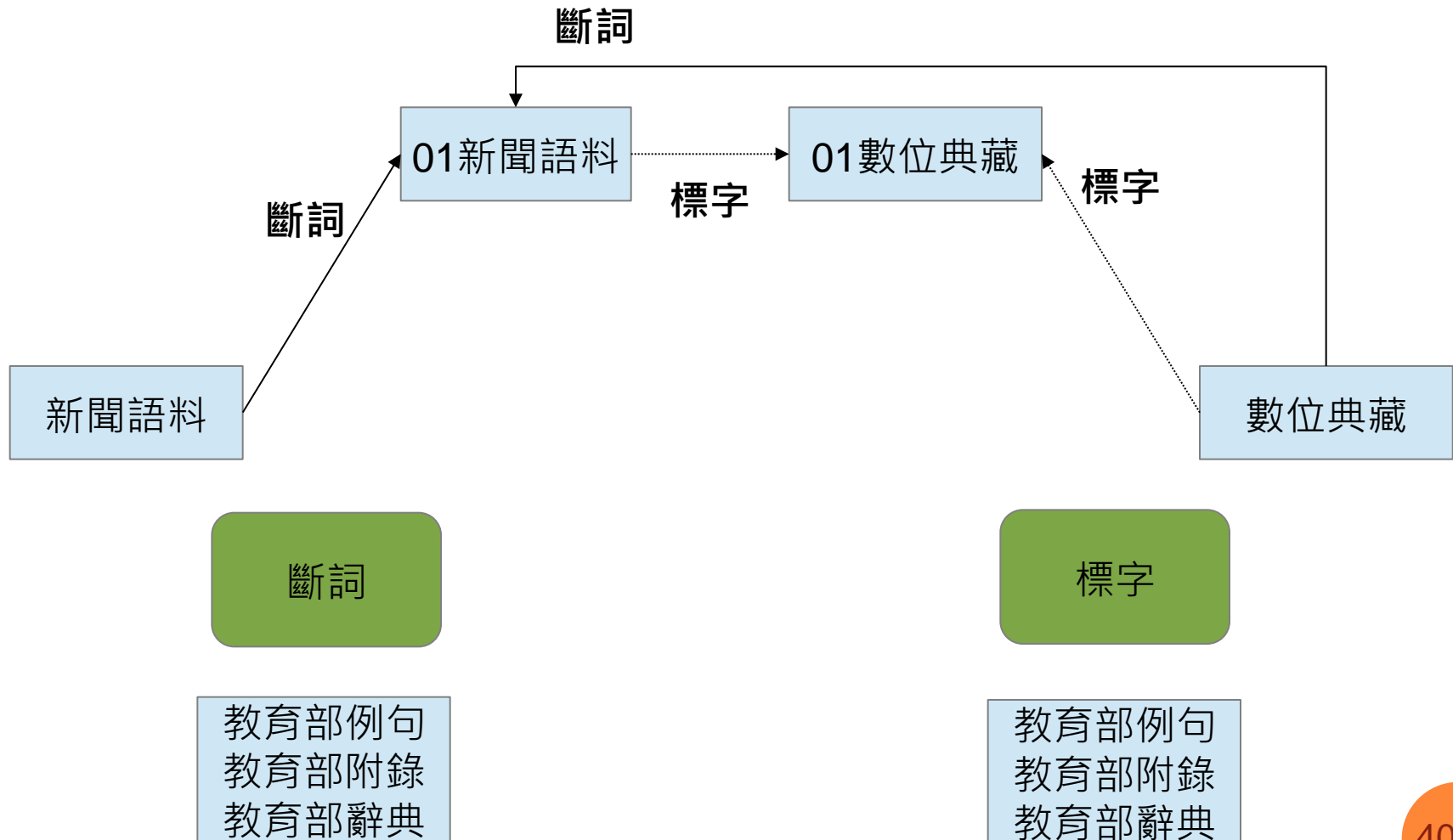
教育部例句
教育部附錄
教育部辭典

數位典藏

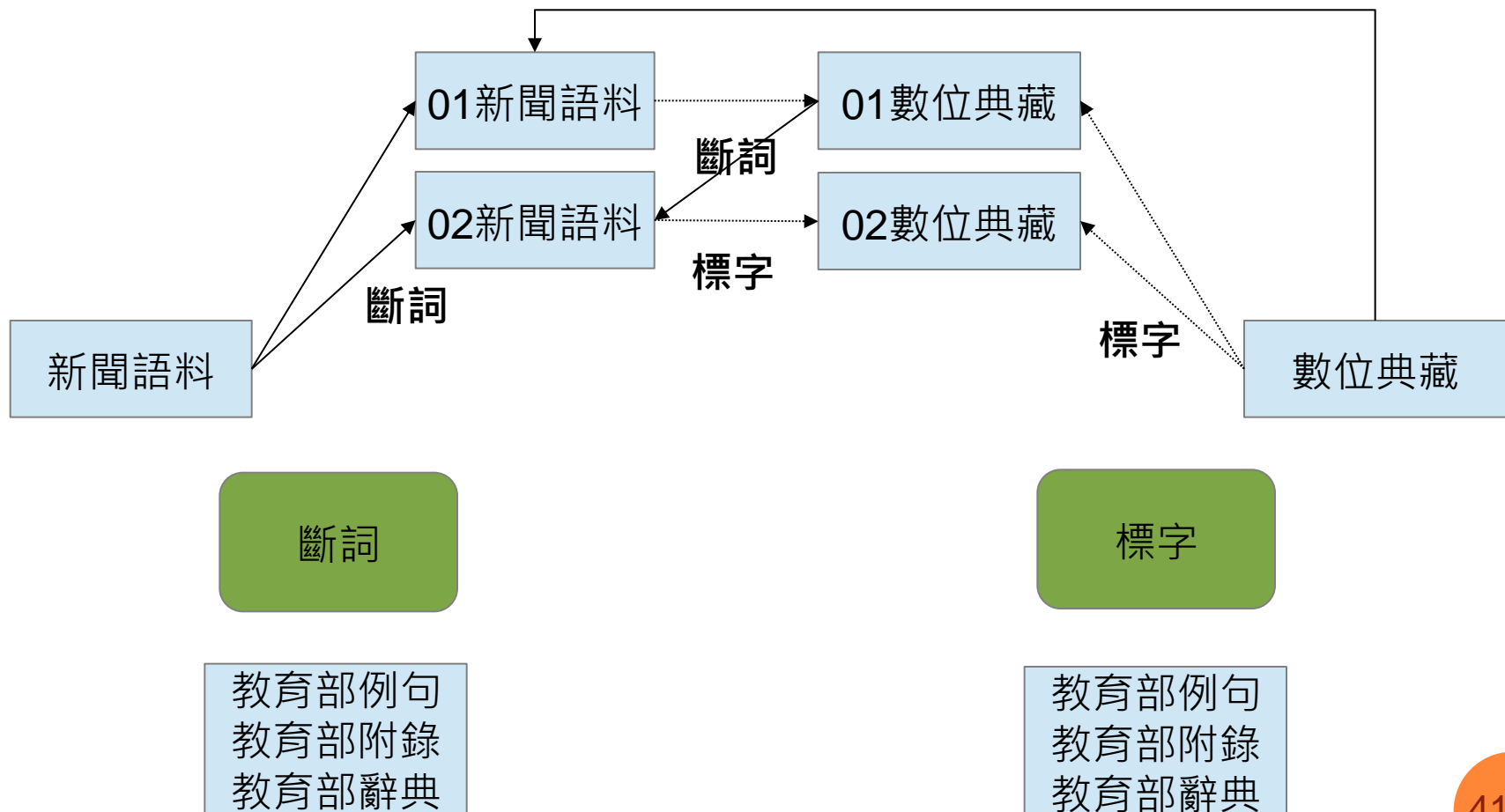
標字

教育部例句
教育部附錄
教育部辭典

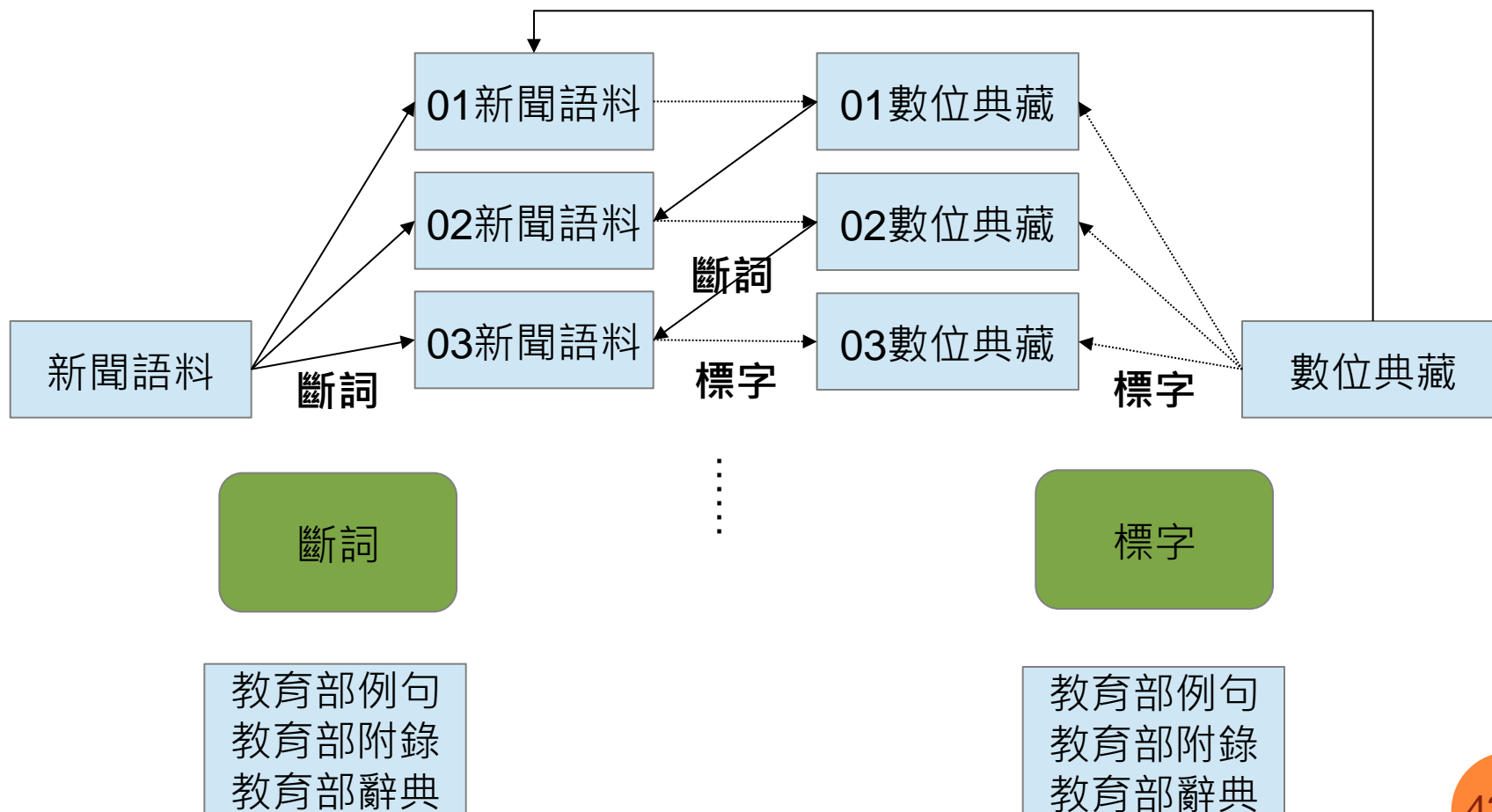
整理流程圖 - 第一擺



整理流程圖 - 第二擺



整理流程圖 - 第三擺



實驗結果

- 訓練語料
 - 新聞57167 佻教育部31200句
- 試驗語料
 - 新聞6954 佻教育部3493句
- 語言模型
 - 訓練語料佻教育部附錄句388句
- 分數以詞為單位

整理幾擺	0*	1	2	3
語言模型 無加典藏	50.57	53.91	53.91	53.91
語言模型 加典藏	49.50	51.94	51.97	51.97

0*：新聞只用教育部辭典斷詞

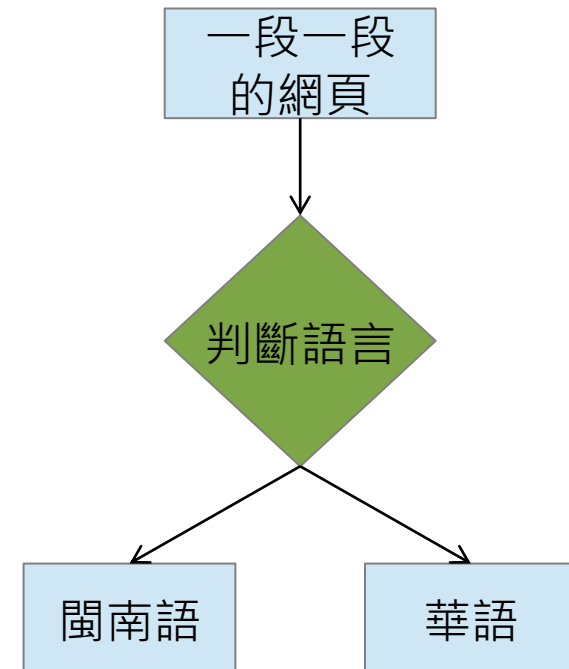
第五節：語言分類

○ 目的

- 想欲加閣較濟語料，予翻譯較穩定
 - 這馬閩南語語料四十萬句爾爾
- 對網路頂收集語料
 - 毋過網路頂的語料百百款
- 用TGB通訊語料庫

○ 問題

- 分類網頁中的閩南語佻華語
 - 閩南語網頁內底定定有華語



語言分類標準

○ 閩南語

- 以閩南語為主，有華語詞無要緊

○ 華語

- 有完整華語句
- 華語閩南語攏通的

閩南語

聽人講 khah 早有出現過『小蜜蜂』

我 beh tng 來種作！——記 0312 Truku 反亞泥 • 還我土地運動

有台灣味 ê 繪本——《我和我的腳踏車》。

華語

「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。

『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。

去越南胡志明市 4 工 / 越南胡志明市四日行 @Giok-hōng

判斷語言

○ 問題

- 以早判斷語言的研究
 - 對象是拼音文字為主
 - 用字元算語言模型分數
 - 無適合用佇分閩南語華語
 - 閩南語華語有誠濟共同詞

○ 解決方法

- 利用斷詞恰定用詞做特徵
 - 斷詞的詞數資訊
 - 定用詞愛提掉共同詞
 - 後壁號做「特徵詞」

特徵詞介紹

○ 特徵詞

- 無共同詞的定用詞

○ 方法

- 閩南語俚華語分別選 n 个上定用的詞
- 揣閩南語頭前 m 个無出現佇華語 n 个的定用詞
 - 就揣出閩南語 m 特徵詞
- 揣華語頭前 m 个無出現佇閩南語 n 个的定用詞
 - 就揣出華語 m 特徵詞

特徵詞範例

○ 閩南語統計來源

- 新聞語料庫、教育部辭典、數位典藏

○ 華語統計來源

- 中研院1000萬字平衡語料庫

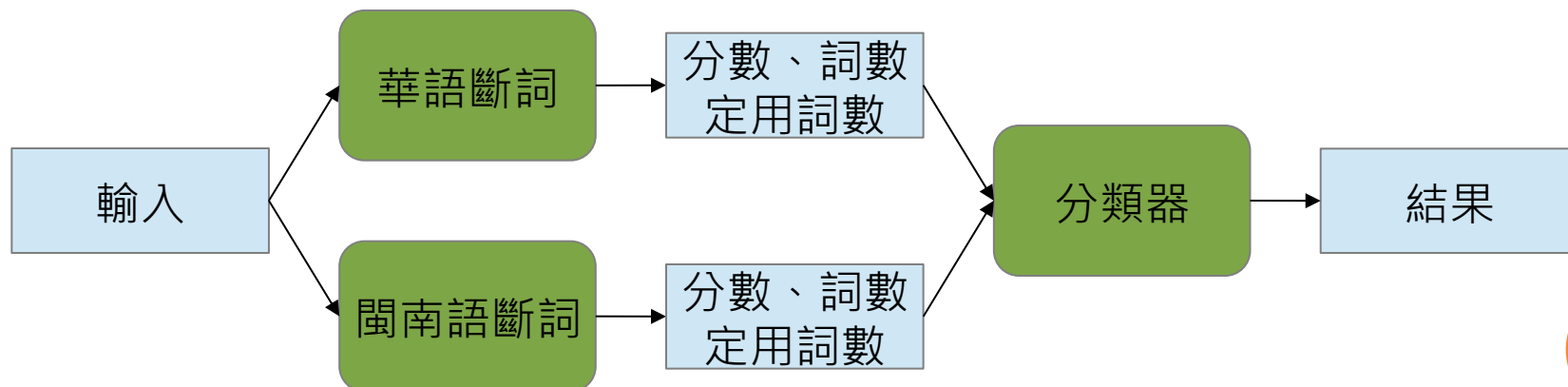
○ 定用詞數量 $n=7000$ ，特徵詞數量 $m=3000$

閩南語 定用詞	的 e5	伊 i1	有 u7	是 si7	我 gua2	人 lang5	無 bo5	講 kong2	佇 ti7	...
華語 定用詞	的	是	在	一	有	了	不	我	個	...
閩南語 特徵詞	佇 ti7	个 e5	閣 koh4	攏 long2	佱 kap4	□ in1	咧 teh4	咱 lan2	彼 hit4	...
華語 特徵詞	我們	很	她	沒有	或	他們	更	則	把	這些

判斷語言

○ 特徵

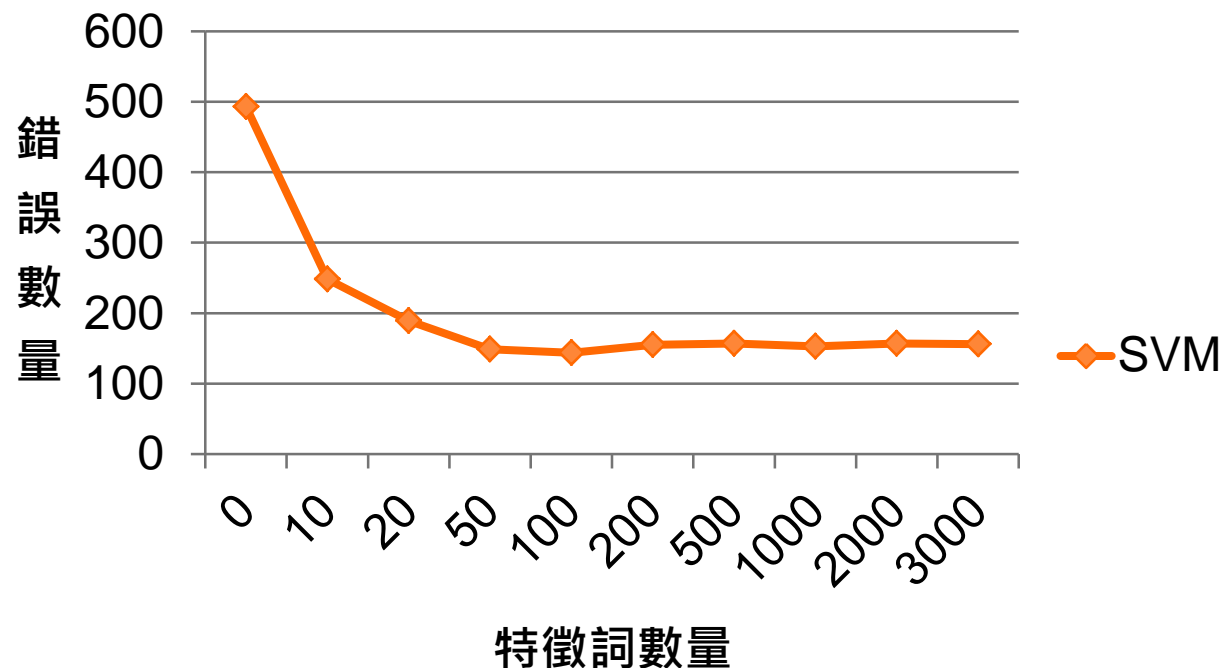
- 語言模型分數
- 斷詞詞數
- 1~4字詞分別數量
- 3000特徵詞
 - 3000个傷大，影響著速度



語言分類實驗結果

- 訓練語料
 - 1000篇
 - 華語8519段
 - 439436詞
 - 閩南語9368段
 - 488844詞
- 試驗語料
 - 179篇
 - 華語2397段
 - 114901詞
 - 閩南語1344段
 - 75282詞
 - 試驗攏總3741段
- 以段做辨識單位

分類結果



第六節：結論佢未來發展

○ 結論

- 語料樣式愛用斷詞
 - 雖然斷詞組佢好，毋過無算法通產生
- 語料庫經過整理了後，效果有較好
- 分類閩南語佢華語免傷濟特徵
 - 語言模型分數、斷詞資訊佢各50个特徵詞

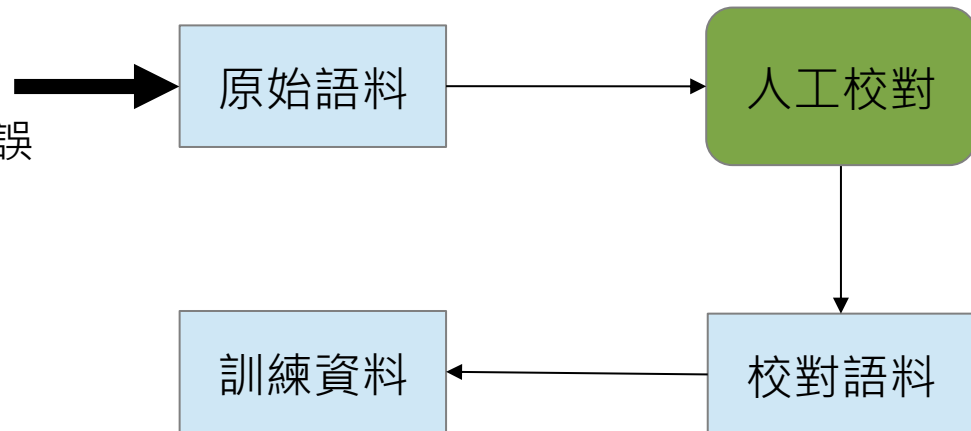
○ 未來發展

- 自動校對語料
- 加強斷詞
- 應用佇字幕辨識

未來發展—加強翻譯

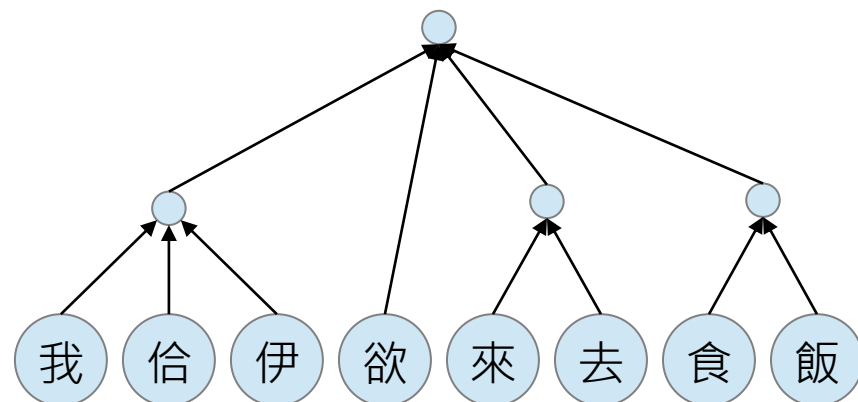
○ 自動校對語料

- 目的
 - 翻譯的訓練語料有誠濟錯誤
 - 語料需要人工校對
 - 減人工負擔
- 問題
 - 語料改錯字
- 方法
 - 用「原始語料」佢「校對語料」訓練翻譯模型



○ 加強斷詞

- 目的
 - 斷詞效果影響著翻譯效果
- 方法
 - 詞性斷詞
 - 剖析器
 - 我佢伊欲來去食飯



未來發展—應用

○ 字幕辨識

- 目的

- 聲音語料大部份母語發音、配華語字幕
 - 電視劇、廣播
- 補上母語字幕，學母語用字

- 問題

- 輸入母語聲音、華語字幕
- 輸出母語字幕

第七節：參考文獻

- Peter F. Brown , Vincent J. Della Pietra , Stephen A. Della Pietra , Robert L. Mercer, The mathematics of statistical machine translation: parameter estimation, Computational Linguistics, v.19 n.2, June 1993
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

多謝逐家

漢羅全羅對齊

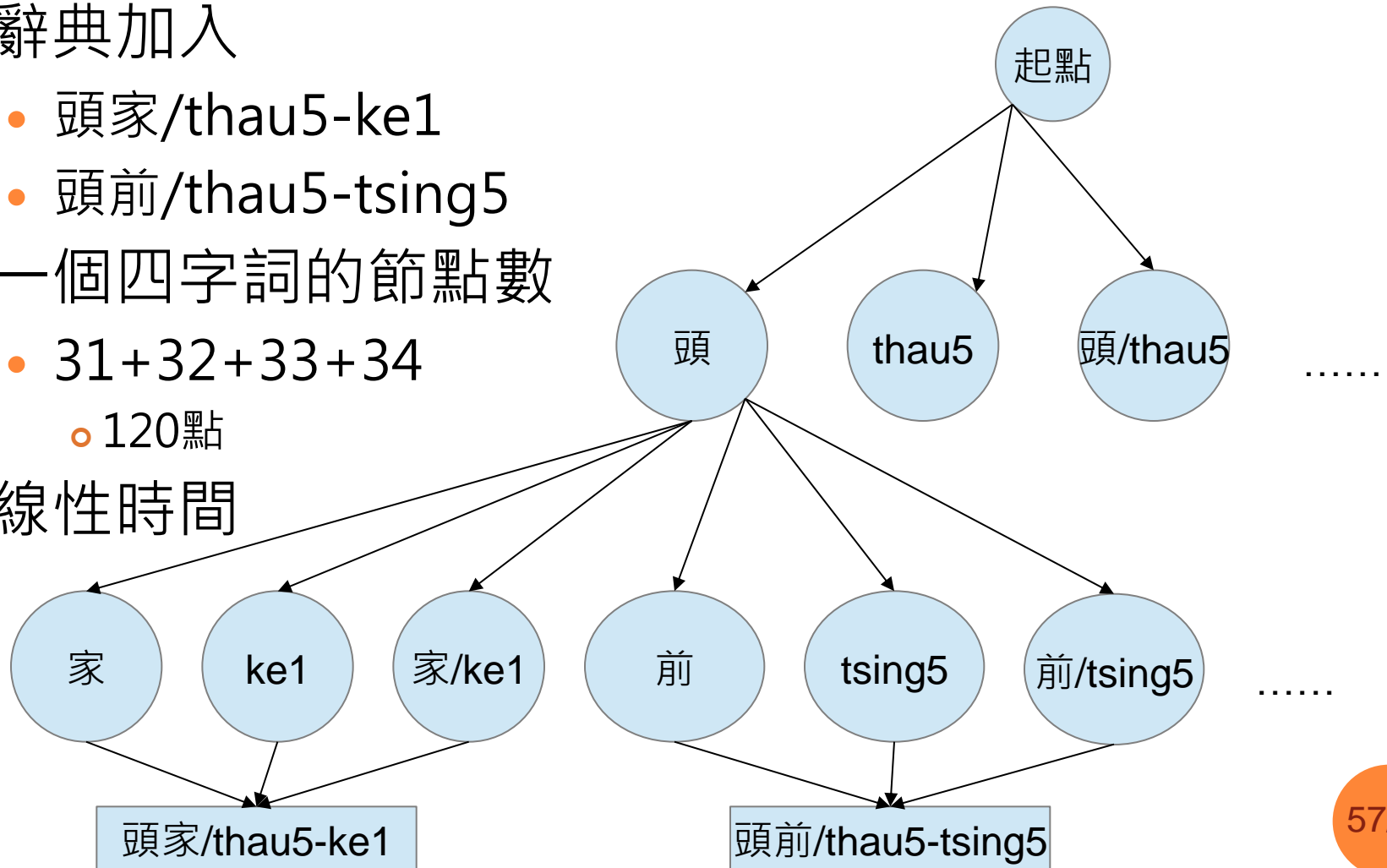
○ 做法

- 看字恰音有佇辭典無
- 揣上大的配對組合
 - Koh m7知u7危險
 - Koh m7-tsai u7 gui5-hiam2

配對	Koh	m7	知	u7	危	險
Koh	有					
m7		有				
tsai			有			
u7				有		
gui5					有	
hiam2						有

找候選詞

- 辭典加入
 - 頭家/thau5-ke1
 - 頭前/thau5-tsing5
- 一個四字詞的節點數
 - $31+32+33+34$
 - 120點
- 線性時間



附錄一：加臺華平行語料庫漢字

- 何澤政翻譯
 - 參考線頂辭典
- 由華語的新聞翻做閩南語全羅
 - 這幾天 寒流 再度 發威
 - tsit4-kui2-kang han5-liu5 koh-tsai3 tian2-ui
- 補上漢字變成一對一
 - 這幾工 寒流 閣再 展威
- 全部約37萬詞組
- 極少調動語句結構

補上漢字的方法

- 如果詞組數一樣，直接對齊
 - 這幾天 tsit4-kui2-kang1
 - 寒流 han5-liu5
 - 再度 koh-tsai3
 - 發威 tian2-ui
- 將詞一字一字對辭典，如果不符合要人工看過
 - 這幾天←要人工看，天不會唸kang1
 - 寒流←免檢查，「寒han5」「流liu5」字典都有
 - 37萬詞組約20萬詞組免檢查
- 人工看
 - 用教育部8000句做的語言模型攏會

實際狀況

- 從97/11/06到103/3/14的文章
 - 2567篇文章、64121句
 - 366190詞組
 - 依斷詞資訊看字音是否符合教育部規範
 - 199629詞組
 - 寒流 han5-liu5
 - 忽略斷詞資訊，由句子的華語，去找是否有符合音標的
 - 34434詞組
 - 民視 新聞報導/bin5-si7-sin1-bun5-po3-to7
 - 由資料庫處理無同音詞的詞
 - 37587由資料庫快速校對
 - 要人工看的
 - 94540 詞組

校對介面

○ 人工看

- 用教育部8000句做的語言模型猜漢字
 - 大部份只要按「確定」就好
- 一小時約可以檢查200詞
- 不夠快

型體	tng5-kenn1-penn7-inn7
音標	tng5-kenn1-penn7-inn7
參考語句型體	soo1-ti7-hun1 in1-ui7 siap8-hiam5 king2-bi2-khuan5-po2-kong1-si1 kah4 tng5-kenn1-penn7-inn7 hun5-lim5-hun1-inn7 siu1-ang5-pau1-pe3-an3 ,
參考語句音標	soo1-ti7-hun1 in1-ui7 siap8-hiam5 king2-bi2-khuan5-po2-kong1-si1 kah4 tng5-kenn1-penn7-inn7 hun5-lim5-hun1-inn7 siu1-ang5-pau1-pe3-an3 ,
參考國語	蘇治芬 因為 涉及 環美環保公司 及 長庚醫院 雲林分院 收賄弊案,
照國語字唸	soo1-ti7-hun1 in1-ui7 king2-bi2 khuan5-po2 kong1-si1 tiong5-king1 hun5-lim5-hun1-inn7 siu1-hue3-pe3-an3
建議結果	蘇治芬 因為 涉嫌 環美 環保 公司 恰 長 更 penn7 inn7 雲林分院 收紅包 弊案 ,

型體

長庚病院

音標

tng5-kenn1-penn7-inn7

有校對改過，請看結果

毋知，愛查，以後處理

外來詞袂曉處理

字典攏無收著的字，做標準

附錄二：教育部辭典處理

- 部份詞有地方腔
 - 辭典干焦一句混合腔例句
 - 地方全部套例句
- 寒著
 - 例句
 - 這馬去寒著矣乎
 - 三峽腔：冷著
 - 產生→這馬去冷著矣乎
 - 鹿港腔：寒著；感著
 - 產生→這馬去寒著矣乎
 - 產生→這馬去感著矣乎

日語外來詞

- 親像「蕃茄」
 - 有人號做「臭柿仔」
 - 閣有人叫「thoo7-ma1-tooh3」
 - 對日語「トマト」來
 - 日語就當做漢字トthoo7-マma1-トtooh3
- 親像「打火機」
 - lai2-tah4
 - 對日語「ライター」來
 - 長度無仝，用統一碼表意文字的符號
 - 標做「 ライ タ - 」
 - ライlai2- タ - tah4

暫存

辭典類

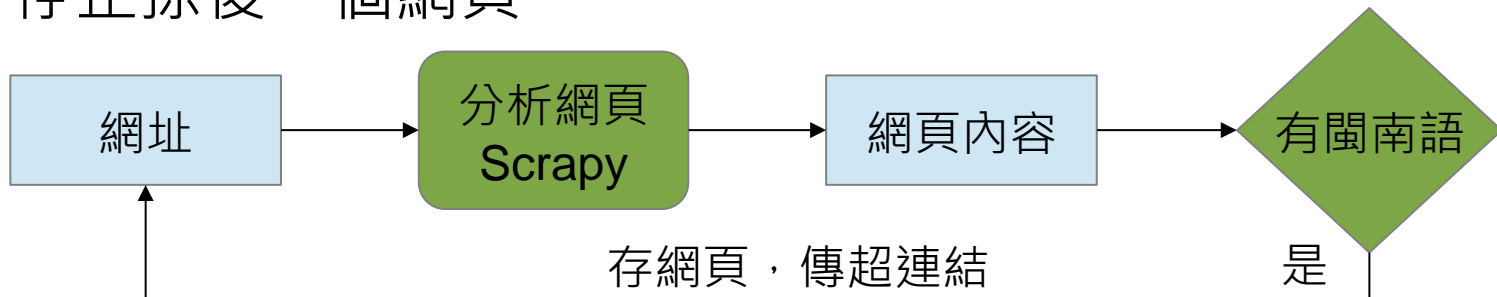
- 教育部辭典
- 陳孟彰老師的平行語料庫
- 信望愛語料庫
- 整理中的詞典01、03、05、06、07

閩南語文章語料

- 教育部辭典
- 陳孟彰老師的平行語料庫
- 數位典藏文本
- 中央研究院台語語音語料庫系統
- 張春鳳老師學生的翻譯
- 網路文章
 - TGB通訊
 - 台文通訊BONG報
 - 老刀烏白講 (Knife Says)

掠閩南語網頁

- 給一個網址
 - 看超連結，掠相關的網頁
- 不同網站收尋深度不同
 - 部落格
 - 歌詞網
 - 網路冊店
- 若這網頁無夠濟閩南語
 - 停止掠後一個網頁



網路文章

- 【技藝101】打鐵| PNN 公視新聞議題中心
 - <http://pnn.pts.org.tw/main/2013/07/15/%E3%80%90%E6%8A%80%E8%97%9D101%E3%80%91%E6%89%93%E9%90%B5/>
- 台灣歌是咱永遠ㄝ記憶: Taiwanese Lyrics
 - <http://enjoytpopmusic.blogspot.tw/search/label/Taiwanes%20Lyrics>
- 臺南市海東國小 - 閩南語教學網站
 - <http://web.https.tn.edu.tw/90ct/Default.htm>
- 第二集 少年好學顯壯志-節目音頻網路收聽-你好台灣網
 - http://www.hellotw.com/zthz/lzlds/ldsjpg/201311/t20131111_889380_4.htm
- [PDF]對鶯歌瓷仔用語來探討台灣常民文化來探討台灣常民文化
 - ir.lib.ntnu.edu.tw/retrieve/47577/metadata_02_06_s_05_0001.pdf