讓格書寫 以及 台華互譯 初探

LangGeh Orthography and an Initial Study of Statistical Translation Between Taiwanese and Mandarin

江永進 Yuang-Chin Chiang
清華大學 統計所
Institute of Statistics, National TsingHua University
jjchiang1@hotmail.com

楊佩琪
清華大學 統計所

林淑卿
清華大學 統計所

張春凰
清華大學 通識中心

高明達
中央研究院 資訊研究所

呂仁園
長庚大學 資訊研究所

陳孟彰
中央研究院 資訊研究所

摘要

讓格書寫 是 新近提議的 書寫方式 [1]， 主張 在 語句中 適當的地方 加上 空白字元， 適合 台客華語等 使用漢字ｅ 文字系統。 大體上講， 讓格書寫 是 分 簡短詞組， 對比的是 英語的 分詞書寫， 傳統華語的 分句書寫， 以及 語言技術的 分詞技術。 讓格書寫 有 減少模糊、 方便閱讀、 利益 語言技術 等 優點； 我們 甚至認爲， 空白字母的 地位 如同 數字系統的 零。 我們 使用 讓格書寫，製作 一套 主要是 台華語的 平行語料庫，各約 15 萬字，並且用來 初步探討 台華語 詞典製作、以及 台華語互譯 問題； 在 此過程 中， 我們 利用了 台華語的 二大類似： 共同詞多、 詞序類似。 比較 現時 詞組爲基礎的 統計式 翻譯 趨向， 讓格書寫 實質上 有讓 台語的 語言技術 站在 較佳的 基礎上。

Abstract

LangGeh orthography is a new writing style proposed by [1]. For Han family languages such Taiwanese or Mandarin that uses Chinese character, *LangGeh* proposes writing with spaces in-between, using simple short phrase as a unit. This is in contrast to word-based orthography in English and sentence-based orthography in traditional Mandarin. Easy to add spaces, *LangGeh* has the advangtages of reducing ambiguity, easier to read, and easier for text processing in Chinese characters. Using the LangGeh orthography, we produce a parallel corpus in Taiwanese and Madarin, about 150 thousand characters each. We then explore the extraction of "phrase dictionary" from the parallel corpus, and begin the study of statistical translation between Taiwanese and Mandarin[7][8].

關鍵詞：讓格書寫、 翻譯詞組、 詞組典、 統計式翻譯、 台語、 華語

Keywords: *LangGeh* orthography, pharse-based translation pair, statistical translation, Taiwanese, Mandarin

一、 介紹

江永進等人[1] 提議 「讓格書寫」的 新 書寫形式， 有 幫助閱讀、 減少模糊 等效果， 對 語言的 初學者、 外國人 有幫助； 不只 強勢語言 有利， 對 弱勢語言的 幫助， 更加明顯。

讓格書寫 採用 「四字左右 無模糊 原則」的 分 簡短詞組。 以台語爲例， 過去 受限於 斷詞正確率 不夠高， 台語 語言技術 常難以 做 進一步的 探討。 使用 讓格書寫的 台語， 多少 避開了 斷詞的門檻， 似乎 建立了 自動 語言處理 技術的 新平臺。

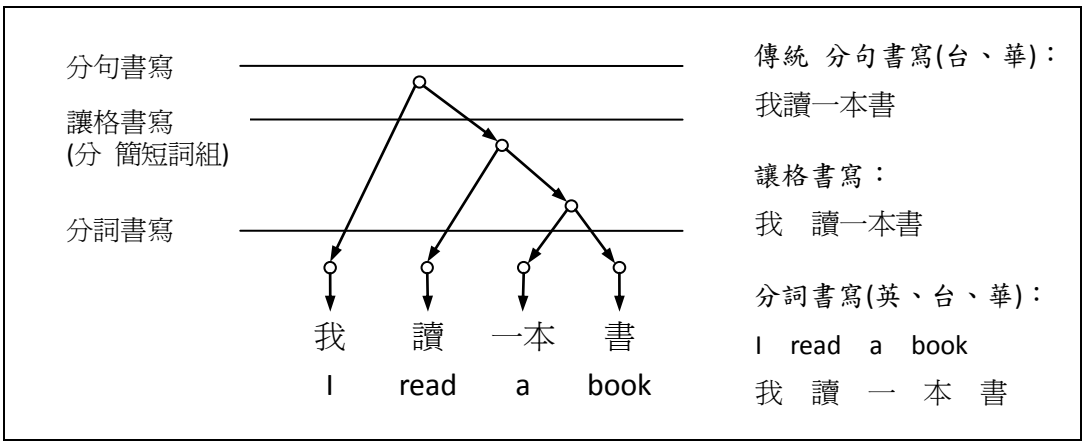
使用 讓格書寫， 本文 報告 「讓格 09 平行語料庫」。 這 是 一套 主要是 台華語的 讓格 • 平行語料庫。 同時 使用 此語料庫， 報告 台華語 對應詞組典的 抽取， 以及 使用 香腸詞組 針對 台華語對譯 做初步探討。

本文內容如下。 第二節 介紹 讓格書寫， 第三節 說明 「讓格 09 平行語料庫」的 製作過程，包括 翻譯、 讓格、 詞組對齊。 第四節 是 對應詞組典 抽取之後的 統計數字， 第五節 探討 香腸詞組 在 台華語對譯的 實作。 最後 是結論。

二、 讓格書寫 簡介

語句 有階層性， 文字書寫的 時候 將 此階層關係 轉換做 線性形式， 因此 多少 損失了 某些消息， 造成 語意模糊。 英語等 拼音文字 以詞 爲單位， 不妨叫 分詞書寫。 相對的， 華語 主要是分句， 以句 爲單位， 句之間 以標點符號(，。？！等) 隔開；華語 不妨叫 分句書寫。當然 華語的 部份 標點符號 也有 分詞功能 (如，

頓號、連音號等)， 但是 現在 華語書寫 主要是 分句。 江永進等[1] 提議使用 分 簡短詞組 書寫的 讓格書寫。 圖一 說明 之間差別。



圖一、 語言 階層性 以及 三種 書寫方式： 分詞書寫、 分句書寫、 讓格書寫(分簡短詞組)。 (讓格的 台語發音 是 ànggēh)

讓我們 簡要敘述 讓格書寫的 「發現」過程， 然後 簡要說明 讓格規則。

一開始， 我們 持續 觀察到， 語句 若加上 適當空白， 閱讀 可以 較簡單， 因此 主張過 “space as a optional punctuation” ([2], p.141)。

然後最近， 我們 無意中 閱讀過 西方拼音文字 約自 第二世紀 開始 使用的是「連續書寫」 (scriptura continua)， 直到 第七世紀 愛爾蘭神父 爲書寫 愛爾蘭文 才於 句中 加上空白，方便 弱勢的 愛爾蘭文 的 閱讀，再經過 約四百年的 傳播，「連續書寫」 漸漸爲 「加空白書寫」 所取代， 變做現在的 分詞書寫 (Saenger [3])。 同時發生的 是， 西方文字 由「朗讀」到「默讀」， 加空白 幫助掌握 詞的界線， 效率閱讀、 快速閱讀 才 變爲可能， Saenger [3] 注重默讀 對 閱讀效率的 論述。

第三個因素 是 我們的 台語 斷詞技術 一直 無法進展。使用語料庫的 斷詞系統需要 大量語料， 華語斷詞的 正確率 可以到達 95%以上。 但是 書寫方式 類似華語的 台語， 書寫人口 少， 語料庫 不足， 斷詞的正確率 一直停留在 85%左右， 受限於此， 語句分析、 語意分析 等 進一步的 研究 一直受到 很大的限制。 最近 我們 才警覺： 傳統的 教會白話字， 使用全拼音， 沿用 西方文字的 分詞， 因此 根本不用 斷詞， 斷詞的正確率 可以說是 100%正確！

第四個因素 是 嚴格的分詞 常常過度 瑣碎、 難行。 以 圖一的例， 嚴格斷詞的結果 可以是 「我 讀 一 本 書」; 這 過度瑣碎， 跟不分詞 差別不大。

因此， 讓格 採取 分 簡短詞組 的 策略。 讓格書寫的 主要原則 是「四字左右無模糊 原則」。 使用 四字左右 是 因爲 台華語的 雙字詞 很多， 二個詞 所合併的詞組， 模糊的機會 較小。 四字左右 只是原則， 不足四字 也可能 模糊， 超過四字也可能 不分詞組 較好：

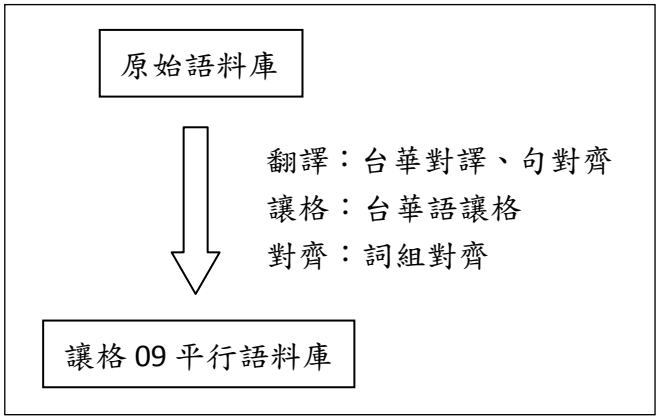
讀好書 有模糊， 最好 作者自己 分詞組；

國際性交易　　有模糊， 最好 作者自己 分詞組；
南港區公所　　二種分法 似乎 差別不大， 建議 合寫。
其他 讓格 建議規則， 請參考 [1]。

本文 是 以讓格 書寫， 我們的經驗 顯示 閱讀 較容易， 比較 中研院 詞庫小組 的 分詞書寫， 讓格書寫 使用 也較簡單。

三、 讓格 09 平行語料庫

根基 讓格書寫， 我們 收集、 製作 一套 主要是 台華語的 平行語料庫。 從 原始語料， 經過翻譯， 再 經過讓格， 最後 再執行 詞組對齊。 過程 如圖二。



圖二、 讓格 09 平行語料庫 製作程序

三・一、 原始語料

我們的 原始語料庫 主要是 利用 三組資料， 分別爲：

(1) 自由時報 《中英對照讀新聞》 中， 2008 年的新聞 (每天一則， 共 366 則)。 原始語料 是 英文、華文。 [4]
(2) 《發明的故事》 台譯本， 譯者 爲 游政榮(2006)， 原作 爲 Hendrik · Willem · van · Loon (房龍)。 原始語料 算是 台文。 [5]
(3) 《青鳥》台譯本， 譯者 爲 林慧婷等(2009)， 原作 爲 Maurice · Maetrlinck。 原始語料 算是 台文。 [6]

三・二、 翻譯

第二個 工作是 翻譯。

原始語料 中， 新聞語料 所使用的 書寫語言 爲 英語 與 華語， 而 後兩本 名

著譯本 則爲台語。 所以， 新聞語料 主要是 翻譯成 台語， 而 其他 兩本名著 則翻譯成 華語。

對於 台華對譯 而言， 翻譯華語 要比 翻譯台語 來得 簡單多了。 因爲 大多數人，「台語書寫」 困難過 「華語書寫」， 所能使用的 台語詞彙 要比華語 來得少。

在翻譯時， 是 一段台語 翻譯 一段華語， 在這同時， 我們 也要 對語料做「句對齊」的 翻譯。 這裡所說的「句」， 是指 將 台語段落 或 華語段落， 以「：，。；！？」此六種 標點符號， 分成 一句一句的 形式， 然後 翻譯時， 就 一句台語 翻譯 一句華語， 而 不要發生 二句台語 翻譯成 一句華語 等 情況。 如表一， 此例 是 語料庫中的 某一 台華對應的 平行段落，以及 其 對應的平行句，因爲 各句 各自對應，所以 叫 句對齊。

表一、 句對齊

台語段落	幾千年來，人 ganna 用空手去掠取活食，用空手 sa 起獵物，空手掠小動物 gah 飛禽，但是 suah ia 未想過可能 iau 有任何其他可行 e 辦法。	
翻譯成華語	幾千年來，人只有用空手去抓取活食，用空手拿起獵物，空手抓小動物和飛禽，但是卻也沒想過可能還有任何其他可行的辦法。	
句對齊	幾千年來，	幾千年來，
	人 ganna 用空手去掠取活食，	人只有用空手去抓取活食，
	用空手 sa 起獵物，	用空手拿起獵物，
	空手掠小動物 gah 飛禽，	空手抓小動物和飛禽，
	但是 suah ia 未想過可能 iau 有任何其他可行 e 辦法。	但是卻也沒想過可能還有任何其他可行的辦法。

做句對齊的 好處 就是， 可以方便 我們 在 學習語文 時， 清楚地 了解此句台語 就是 對應 此句華語， 而 不用再 費心思 去尋找； 也可以 幫助我們 做 之後的研究。 不過， 台華對譯的 句對齊 比較容易 實行， 若是 中英對譯 的話， 因爲 書寫的文法 不同， 比較 不容易 完成。

三・三、 讓格

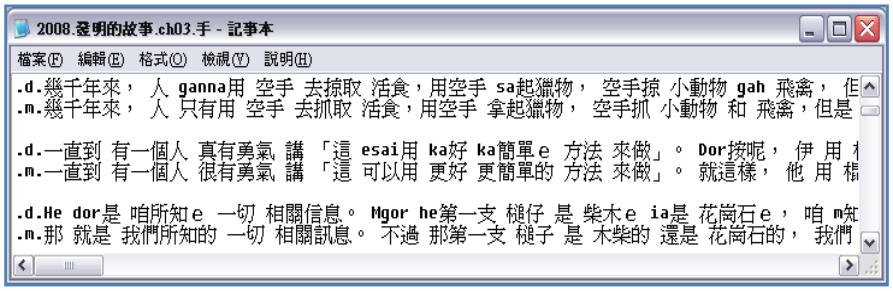
翻譯完成 後，我們 再對語料 做讓格。而 讓格規則 已在 第二章 概述過了。值得注意的是， 我們 在做 台華讓格 時， 並不需要 像句對齊 那樣， 強制做 讓格詞組的 依照順序 對齊，句中詞組的 順序 應依照 各自語言的 自然順序。只要服從 讓格規則， 台華語 獨立讓格。 如 台語句 與 華語句：

hong 迫去 面對 非傳統 網路競爭

被迫面臨 非傳統 網路競爭

華語句中的 ***被迫面臨***， 是 不用特地 爲了 配合台語， 而 讓格成 ***被迫 面臨***。

讓格之後的 結果， 我們 以 utf-8 編碼 儲存成 普通文字 檔案， 如圖三。



圖三、 讓格之後的 台華語 平行語料 格式。
其中 .d.表示 台語段落， .m. 表示 華語段落。

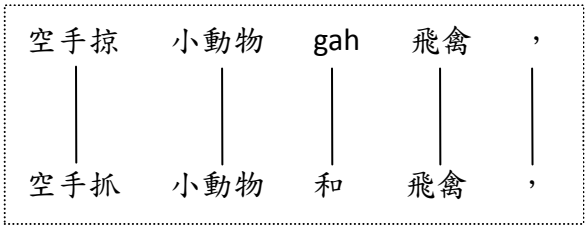
三・四、 對齊

所謂 「對齊」， 在此 是指「詞組對齊」， 簡稱「對齊」。

跟隨 Brown 等人[7][8]， 如圖四所示， 將 平行句 中 的 對應詞組 以 關聯線 連接， 每條 關聯線， 就稱爲 一個 連結 (connection)； 所謂 詞組對齊， 就是 這些連結的 集合。 圖四中 的 平行句， 總共有 5 個連結。 連結 可以用 符號、數字 表示， 方便 機器閱讀 自動處理。 以 圖四之例 爲例， 台華語 對應句的 對齊 可表示成：

空手掠[1] 小動物[2] gah[3] 飛禽[4] ，[5]
空手抓(1) 小動物(2) 和(3) 飛禽(4) ，(5)

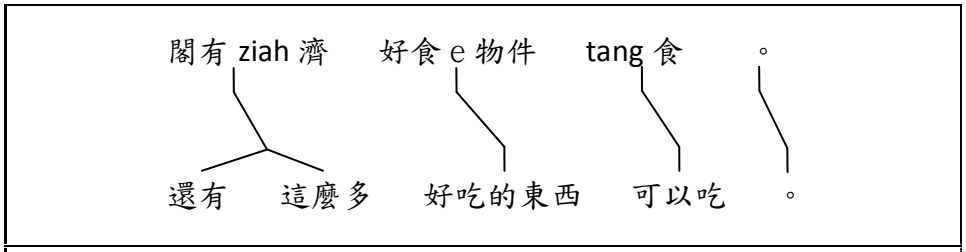
基本上 這 是[7]中的 記號， 我們 再加上 中括弧記號， 因爲 這 在 人工糾正 時，可立即 清楚知道， 華語詞組 所對應的 台語詞組 爲何者， 而 不用 再花時間 去數台語詞組的 位置。 簡言之， 前頭句的數字 代表「位置」， 後頭句的數字 代表「對齊」。



圖四、 台華讓格詞組的關聯線 以及 對齊的標記。

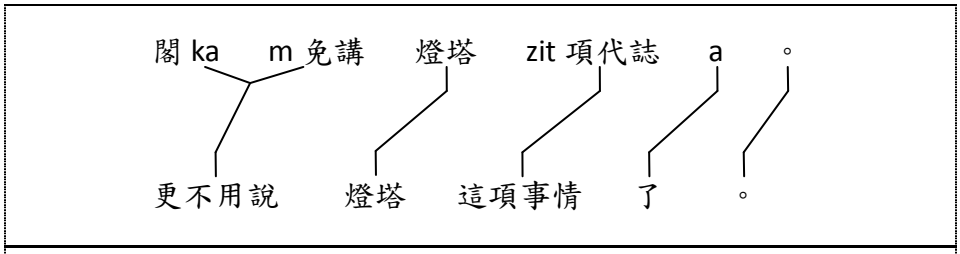
實務上， 詞組對齊的 標記 開始 是用 LCS(最長共同子序列) 爲基礎 先行自動標記， 然後 人工校正。 詳細 請見[9]。

圖四 是 對齊中 最簡單的 「1 對 1 對齊」， 圖五、六 是「1 對多對齊」、「多對 1 對齊」， 其他「多對多對齊」 以及 有時 無對應的 情形(「對應空詞組」)， 請見[10]。



閣有 ziah 濟[1] 好食 e 物件[2] tang 食[3] 。[4]
還有(1) 這麼多(1) 好吃的東西(2) 可以吃(3) 。(4)

圖五、「1 對多對齊」：台語詞組 1，華語詞組多

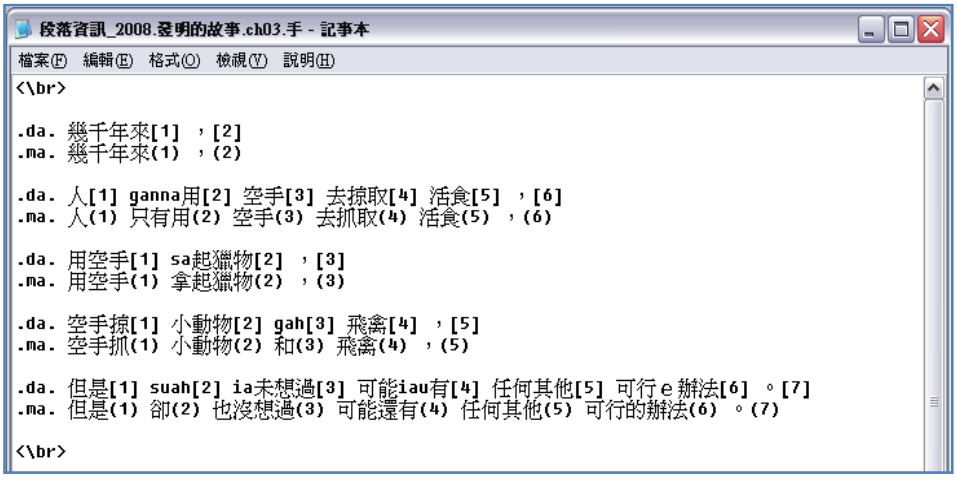


閣 ka[1] m 免講[2] 燈塔[3] zit 項代誌[4] a[5] 。[6]
更不用說(1,2) 燈塔(3) 這項事情(4) 了(5) 。(6)

圖六、「多對 1 對齊」：台語詞組多，華語詞組 1

將語料庫 做完翻譯、 讓格、 對齊 後， 即爲 讓格 09 平行語料庫。 此語料庫，是由 多人 合作完成。 利用 此語料庫， 我們 就可以 清楚知道， 台華對譯的 對應情形， 進而發展 後續的 研究工作， 如 讓格詞組典、 台華翻譯 實作。

讓格 09 平行語料庫 是 以 普通文字 檔案 儲存，只有加上 語言標記，如圖七；檔案編輯 可以使用 多樣的軟體。 儲存時， 使用 utf-8 編碼 ， 以便我們 使用 Python 程式[11] 做 讀取處理。 Python 3 全面使用 unicode， 處理文字 方便很多。



圖七、 讓格 09・平行語料庫 對齊後 格式。 .da. 表示 對齊的 台語句， .ma. 表示 對齊的 華語句， <\br> 表示 段落資訊。

四、 詞組對齊 與 台華語 對應詞組 抽取

有 讓格書寫的 平行語料 之後， 我們 可以 進一步 對齊詞組， 然後抽取 台華語 對應詞。

四・一、 台華讓格詞組典的抽取

讓格 09 平行語料庫 中 的 18363 組 對齊的句對， 容易抽取 對應的詞組對。 依照 這些詞組的 對齊情形 分類， 共 分成六類， 分別爲 「1 對 1 對齊」、 「1 對多對齊」、 「多對 1 對齊」、 「多對多對齊」、 「多種釋譯」、 「跳位對齊」， 如表二， 舉些例子， 以便了解 這六種 對齊方式。

表二、 台華對齊類型

對齊類型	台語詞組	對齊	華語詞組
1 對 1 對齊	會知影	→	會知道
1 對多對齊	iau 無重視	→	還沒有 重視
多對 1 對齊	edang 防止 日頭曝傷	→	可防曬傷
多對多對齊	敢 edang pah 開	→	可以打開 嗎
多種釋譯	edang・ho	→	可以讓、能讓、能使、能夠讓
跳位對齊	ga~講	→	告訴、告訴~說

前 四種分類 在前面 已介紹過，而「多種釋譯」是 台語詞組 可以 被翻譯成 多個 不同的 華語詞組。 基本上， 前 四類對齊 都是 單一釋譯，也就是 只有 一個 對應的 華語詞組； 而 「多種釋譯」 當中， 會有 前四類的情形 發生。 例如：

台語詞組：***一個月前***

對應的華語詞組：***一個月前***、***一個月 之前***

這當中 就有 1 對 1 對齊 (***一個月前→一個月前***)、 1 對多對齊 (***一個月前→一個月 之前***) 兩種情形。

而 「跳位對齊」 就是 台語詞組 或 華語詞組的 順序 是 不連貫的。 例如：

ga[1] zit 個警官[2] 講[3]

訴(1,3) 這名警官(2)

我們 會發現 華語詞組 ***告訴***， 所對應的 台語詞組 爲 ***ga~講*** 是 不連貫的； 又例如將 翻譯句子 改變：

ga[1] zit 個警官[2] 講[3]

告訴(1,3) 這名警官(2) 說(1,3)

則 台語詞組 和 華語詞組， 都有 不連貫的情形 (***ga~講→告訴~說***)。

然後 我們 去計算 這六種分類的 分佈情形， 如表三。 由 讓格 09 平行語料庫的 18363 組 句對齊 中， 我們 總共 得到了 108129 個 台語詞組， 其中 共有 37998 個

不同的 台語詞組。 由其分佈 可知， 大多數的 台華對齊 都是 1 對 1 的情形； 而 多種釋譯 雖然 只佔了 4.92 %， 但是 在 對齊類型 中 比例 是 第二高的， 所以 也是不容忽視的。 而 跳位對齊， 因 所佔的比例 只有 0.09 %， 所以 在做 後續的 台華翻譯 時， 我們 並沒有 針對 當 句子 發生跳位 時 的 特別處理，即 當 句子 發生跳位 時， 其 翻譯出來的 結果， 一定會 發生錯誤。

表三、 台華讓格詞組典的分佈

對齊類型	詞組數	比例
1 對 1 對齊	34594	91.04 %
1 對多對齊	540	1.42 %
多對 1 對齊	879	2.31 %
多對多對齊	80	0.21 %
多種釋譯	1869	4.92 %
跳位對齊	36	0.09 %
總和	37998	100.00 %

四・二、 華台讓格詞組典的抽取

華台讓格詞組典 的 製作， 與 台華讓格詞組典 相同， 只是要對 讓格 09 平行語料庫 的 詞組對齊 做轉換。 原本語料的 詞組 對齊方式 是

.da. 只 edang[1] 做出[2] 幾種有限 e [3] 動作[4] ，[5]

.ma. 只能做出(1,2) 幾種有限的(3) 動作(4) ，(5)

現要將之 轉換成

.ma. 只能做出[1] 幾種有限的[2] 動作[3] ，[4]

.da. 只 edang(1) 做出(1) 幾種有限 e (2) 動作(3) ，(4)

再利用 轉換後的形式， 以 相同方法， 來製作 華台讓格詞組典。

由 讓格 09 平行語料庫 的 18363 組 句對齊 中，我們 總共 得到了 97940 個 華語詞組， 其中 共有 37647 個 不同的 華語詞組。 而 其 六種 對齊方式 分佈 如表四。 與表三 做一對照， 其分佈 與 台華讓格詞組典 差異不大。

表四、 華台讓格詞組典的分佈

對齊類型	詞組數	比例
1 對 1 對齊	33943	90.16 %
1 對多對齊	879	2.33 %
多對 1 對齊	543	1.44 %
多對多對齊	81	0.22 %
多種釋譯	2166	5.75 %
跳位對齊	35	0.09 %
總和	37647	100.00 %

我們的 讓格詞組典， 主要是 由 這些不同的 對齊類型 所組成。 可發現，就 我們的語料庫 而言， 主要是「1 對 1 對齊」 爲多，「多種釋譯」 次之。而 當中的「多種釋譯」 可以 被應用 來做 一個 「同義詞詞典」， 提供 較豐富的詞彙。

五、 台華語 互譯 初探

爲了 台華對譯的 需求， 本章 仿效[7][8]的 統計式翻譯方法， 並且 提出 簡化的 「香腸詞組」 翻譯法， 並報告 初步的結果。

在[7][8]， 翻譯的 語言對 是 英文法文。 當 給定 一法語句 F ， 以及 其可能的 英語翻譯句 E， 我們 給于 一機率 $Pr(E|F)$， 統計式 翻譯 是 在 所有可能的 E 中， 選擇 條件機率 最大者：

$$\hat{E} = \arg\max_{E} \Pr(E|F).$$

由於 $\Pr(E|F) = \Pr(F|E)\Pr(E) / \Pr(F)$， 而且 分母部份 與 E 無關， 因此

$$\hat{E} = \arg\max_{E} \Pr(F|E)\Pr(E).$$

跟隨[7]， 後面項 稱爲 語言模型(language model)， 前面項 稱爲 翻譯模型(translation model)。

關於 語言模型 $\Pr(E) = \Pr(E_1 \cdots E_n)$， 可以使用 n-gram 模型 逼近：

$$\Pr(E) \doteq \prod_{i=1}^{n} \Pr(E_i|E_{i-n+1} \cdots E_{i-1})$$

其中 $E_{i-n+1} \cdots E_{i-1} \stackrel{\text{def}}{=} E_1 \cdots E_{i-1}$ 如果 $i - n + 1 < 0$， $\Pr(E_1|E_{1-n+1} \cdots E_{1-1}) \stackrel{\text{def}}{=} \Pr(E_1)$。稍後的實驗 使用 bigram 語言模型， 或者 $n = 2$。

關於 翻譯模型， Brown 等 [8] 提出 一系列 逐漸複雜的 模型 1 到 模型 5， 並且 詳細討論 牽涉到$\Pr(F|E)$的 各項參數的 估計方法， 是 近來 以詞爲底的 統計式自動翻譯的 基礎， 而且 吸引了 很多 後續研究。

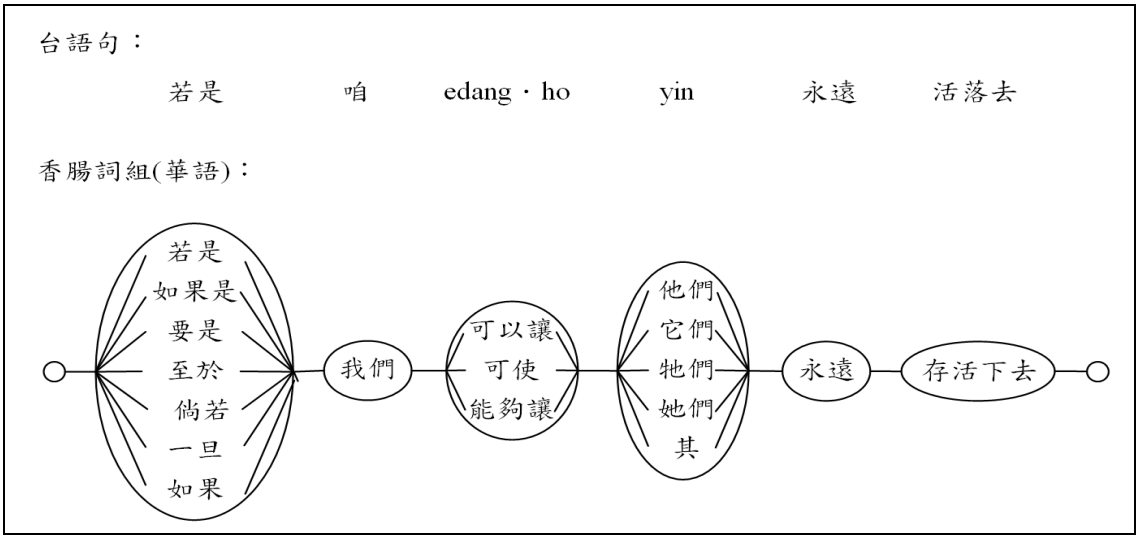
由於 我們的語料 不是 以詞爲底， 而是 以詞組爲底， 詞組數 遠大於 詞數； 而且 語料量 也不足， 此地 我們沒有 應用 Brown 等人的 方法， 而 另外提出 一個簡化的 香腸詞組的 翻譯法， 並且 初步報告 結果。

我們 舉例說明 香腸詞組。 圖八 是

台語句： “若是 咱 endang・ho yin 永遠 活落去”

(華語句： “如果 我們 可以讓 他們 永遠 活下去”)

所 對應 華語 香腸詞組串。 注意到 香腸詞組串 是 由 香腸詞組 組成， 每個香腸詞組 只是 每個 台語詞組的 可能翻譯的 包裹而已。



圖八、 香腸詞組 舉例。

如果 我們 限制 在 香腸詞組串 當中的 可能路徑 去尋找 機率最大者，那麼 最佳解的 搜尋範圍 可以 有效控制， 而且 翻譯模型的 機率估計 也可以簡化： 設 台語句 $D=(d_1,\cdots,d_I)$ 有 I 個詞組， S_{d_i} 表示 d_i 的 (華語)香腸詞組， $S_D=\{(m_1,\cdots,m_I):m_i\in S_i\}$ 表示 香腸詞組串的 所有可能的 路徑， 符號 m_i 表示 華語詞組， 那麼

$$(\hat{m}_1,\cdots,\hat{m}_I)=\underset{(m_1,\cdots,m_I)\in S_D}{\arg\max}\prod_{i=1}^{I}\Pr(m_i|d_i)$$ 不使用語言模型

$$(\hat{m}_1,\cdots,\hat{m}_I)=\underset{(m_1,\cdots,m_I)\in S_D}{\arg\max}\prod_{i=1}^{I}\Pr(m_i|d_i)\Pr(m_i|m_{i-1})$$ 使用語言模型

似乎可以 當做 最佳翻譯的 二種準則： 前者 完全不管 語言模型， 單純選擇 香腸詞組中 機率 最大者； 後者 再加上 語言模型 篩選。

對台華語 語言對 而言， 因爲 詞序接近， 共同詞 眾多， 使用 語言模型的 香腸詞組 翻譯模型 做爲 初步的 翻譯模型， 應 不算過分。 香腸詞組 有其優點： 實作 與 概念 簡單； 計算量 較少； 當然， 此地的 香腸詞組 也有限制： 譬如， 此地 沒有考慮 詞序不同的 情況； 詞組 若是 跳位對齊 則 無法處理 等等。 對本文而言， 主要是 初步探討， 我們 沒有追求 最一般的模型。 而且 實作上 有些 較小、較煩瑣的 細節， 如 句首、句尾的 機率、 多對多 詞組對應的 問題(使用 Brown 等人的 fertility 機率)， 我們就 不在此 詳述。

使用 訓練語料， 我們 估計各項 機率 如下：

$$Pr(w_1)\leftarrow \#(w_1)/N$$

$$Pr(w_i\mid w_{i-1})\leftarrow \frac{\#(w_{i-1},w_i)/N}{\#(w_{i-1})/N}=\frac{\#(w_{i-1},w_i)}{\#(w_{i-1})},\ i=2,\ldots,n$$

$$Pr(d \mid m) \leftarrow \frac{\#(d,m)}{\#(m)}$$

以下 報告 二種情況 之下的 台華互譯的 結果：(1) 使用 語言模型 (2)不使用 語言模型。 如表五，我們 從 讓格 09 平行語料庫 的 18363 組 對應句子 中，從中 隨機選取 90 %， 16527 個句子 來做 訓練語料， 其餘 1836 個句子 做爲 外部測試； 再從 訓練語料 中， 隨機選取 約 5 % 826 個句子 做爲 內部測試。

表五、 各資料的總句數

資料	句子數	測試有效句數*
讓格 09 平行語料庫	18363	
訓練語料	16527	
內部測試	826	821
外部測試	1836	338

*有效句數 請見內文。

在 測試句 中， 我們 可以預期 以下的 情況：

(1) 香腸詞組 組合數 過多。 雖然 香腸詞組 已經減少 可能組合， 有時 測試語句 組合數 仍然 過於龐大。 因此 我們 爲 此組合數 設下 一門檻值—— 50 萬，若 組合數 大於 50 萬句， 則 我們 就 不翻譯 此句。

(2) 空的 香腸詞組。 受限於 有限 平行語料， 我們 不能永遠 找得到 對應詞組， 致使產生 空的 香腸詞組。這 在 outside test 極容易發生。 如果 有出現 空的 香腸詞組， 那麼 我們 就 不翻譯 此句。

(3) 有效測試句。 扣掉 組合數過多，以及 有 空的 香腸詞組 的 測試句，剩餘的 稱爲 有效測試句。

(4) 「也可」正確句。 翻譯的結果 可能 與 語料庫的答案 完全一樣，也可能 雖與 答案不同， 但是 意思相同， 也可以 當做 翻譯正確。 Brown 等人 分別稱呼爲「Exact」、「Alternate」。 圖九 是 一個例。 此種 正確形式， 主要依靠 人工檢查， 算是麻煩。 實際的 正確率 應是 此兩種情形的 加總。

Exact

台語句 ： Mgor 最後 番仔火 勝利 a ，

標準答案： 不過 最後 火柴棒 勝利了 ，

翻譯句 ： 不過 最後 火柴棒 勝利了 ，

Alternate

台語句 ： 其他 e 人 ma 攏 看 gah qang 去 。

標準答案： 其他的人 也都 看到 發愣 。

翻譯句 ： 其他的人 也都 看得 愣住了 。

圖九、 「exact」與「Alternate」正確句

在 以上的設定 之下，我們 使用 Python 3.0.1 [11] 實作 台華語互譯，結果如 表六 和表七。 其中 F-量度 是 正確率 以及 召回率 的 調和平均。

表六、 翻譯結果：比較 使用 語言模型的 效果 (台翻華)

語言模型	資料	有效句數	正確率 (Exact)	正確率 (Alternate)	正確率	召回率	F-量度
不使用	inside (826)	821	47.87% (393/821)	20.46% (168/821)	68.33% (561/821)	67.92% (561/826)	67.53
	outside (1836)	338	47.63% (161/338)	24.56% (83/338)	72.19% (244/338)	13.29% (244/1836)	22.45
使用	inside (826)	821	98.29% (807/821)	1.34% (11/821)	99.63% (818/821)	99.03% (818/826)	99.33
	outside (1836)	338	68.05% (230/338)	19.82% (67/338)	87.87% (297/338)	16.18% (297/1836)	27.33
註：其中的正確率 是以 有效句數 做分母， 非 句子總數。因為 我們的語料庫 過少， 而導致 台語詞組 無法找到 其對應的 華語詞組。							

表七、 翻譯結果-比較使用 語言模型的 效果(台翻華)

語言模型	資料	有效句數	正確率 (Exact)	正確率 (Alternate)	正確率	召回率	F-量度
不使用	inside (826)	826	31.60% (261/826)	23.24% (192/826)	54.84% (453/826)	54.84% (453/826)	54.84
	outside (1836)	343	37.03% (127/343)	28.86% (99/343)	65.89% (226/343)	12.31% (226/1836)	20.74
使用	inside (826)	826	97.94% (809/826)	1.57% (13/826)	99.52% (822/826)	99.52% (822/826)	99.52
	outside (1836)	343	61.52% (211/343)	29.74% (102/343)	91.25% (313/343)	17.05% (313/1836)	28.73
註：其中的正確率 是以 有效句數 做分母， 非 句子總數。因為 我們的語料庫 過少， 而導致 台語詞組 無法找到 其對應的 華語詞組。							

從 表六 及 表七， 我們 至少 可以結論：

(1) 沒有 對應詞組 當然 不能翻譯， 因此， 建立 對應詞組， 應當可以 提高 翻譯的效率。 這 不能單靠 詞組爲底的 平行語料， 應該可以 由詞爲底的 模型入手。 讓格書寫的 基礎單位 是 簡短詞組，小過 傳統翻譯的 大詞組(句)，也許會 較容易些。

(2) 由於採用 較大單位的 簡單詞組，詞組爲底的 語言模型 的 參數估計 更加容易 發生 轉移機率 估計爲 0 的 問題。除了 傳統 語言模型的 平滑技巧，我們 值得研究 詞組 轉移機率 使用 詞轉移機率的 平滑方法。

六、 結論

讓格書寫 提供了 一個 新舞台。讓格書寫 實質上 讓 弱勢語言 如 台語客語 站在較好的 基礎上。 至少， 我們 可以不用 再受限於 斷詞， 可以 有效的 進行 進一步的 研究， 如翻譯等。 事實上， 讓格書寫 也同樣利益 強勢的華語， 使用 讓格書寫的 華文， 也可以 比較快速 使用 新想法， 得到 較佳的結果。 試想， 如果 英語現在規定 去掉 空白字元， 那麼 英語的 語言技術， 應該是 嚴重退步。 我們 沒有有必要 故意阻礙 台華客文的 語言技術。

從 技術的角度， 讓格的 分簡短詞組 接近分詞， 因此 比起 傳統的 分句書寫，讓格書寫的 分詞問題 較容易處理(如果 需要的話)； 另外一面， 比如 語言剖析的 需要， 讓格書寫的 簡短詞組 已經結合 前後詞， 實質上 語句的 單位數目 較少， 可以減低 剖析的模糊度， 因此 統計式的方法 也許 可以得到 較佳的結果。

過去 我們的 台語 自然語言 處理嘗試， 一直受限於 語料不足， 採用 讓格書寫之後， 在短時間中， 我們 也製作了 一套 句對齊、 詞組對齊 的 台華語 平行語料庫： 讓格 09・平行語料庫， 並且 用來探討 台華語 對應詞組問題， 構思出 香腸詞組， 初探 台華語 翻譯問題。

有趣的是， 同樣的方法 也利益 強勢華語。

致謝

讓格 09・平行語料庫 是 由多人完成， 除了 本文作者 參與 以外， 也受益 以下 諸位先生， 在此致謝： 呂菁菁、 游政榮、 吳德祥、 陳俊良。

參考文獻

[1] 江永進、張春凰、呂菁菁(2009). “讓格書寫： 意義、理由 gah 簡則”， *台灣風物*59 卷 1 期，2009。

[2] 張春凰(1994). “母語寫作經驗”， *青春ｅ路途*， 台北：台笠。

[3] Saegner, Paul(1997). *Space Between Words: The Origin of Silent Reading*. Stanford University Press, Stanford, California, USA.

[4] 自由時報(2008). “中英對照 讀新聞”， 每日一則， 全年。

[5] 林慧婷、陳則伊、謝旻男(2009). *發明的故事*(台譯本)， 將出版(時行台語文會)，2009。

[6] 游政榮(2006)，*青鳥*(台譯本). 時行台語文會出版，2006。

[7] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin (1990). “A Statistical

Approach to Machine Translation," *Computational Linguistics Volume 16, Number 2, June 1990.*

[8] Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, (1993). "A Statistical Approach to Machine Translation," *Association for Computational Linguistics, 1993*.

[9] 林淑卿(2009). "從 台華平行 語料庫 擷取 對應詞組典"， 國立 清華大學 統計所 碩士論文。

[10] 楊佩琪(2009). "讓格書寫下 統計式 台華翻譯 初探"， 國立 清華大學 統計所 碩士論文。

[11] Python 3.0.1 (2009). http://www.python.org .