行政院國家科學委員會專題研究計畫 成果報告

台語文語法結構樹建置 (3/3)
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 97-2221-E-122-004-
執 行 期 間 ：97 年 08 月 01 日至 98 年 07 月 31 日
執 行 單 位 ：大漢技術學院資訊工程系

計 畫 主 持 人：楊允言
共 同 主 持 人：張學謙
計畫參與人員：碩士班研究生-兼任助理人員：袁崇祐
講師級-兼任助理人員：梁淑慧
其他-兼任助理人員：陳德樺

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 98 年 10 月 30 日

行政院國家科學委員會補助專題研究計畫 ■ 成 果 報 告
□期中進度報告

台語文語法結構樹建置

計畫類別：■ 個別型計畫　　□ 整合型計畫
計畫編號：NSC 97-2221-E-122 -004
執行期間：2008 年 8 月 1 日至 2009 年 7 月 31 日

計畫主持人：大漢技術學院資訊工程系助理教授 楊允言
共同主持人：國立台東大學華語文學系副教授 張學謙
協同主持人：中央研究院資訊科學研究所研究員 陳克健
計畫參與人員：梁淑慧、陳德樺、廖淑鳳、賴淑玲、袁崇祐、林俊育、謝佑明、林素朱、詹金來、李承泰

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：大漢技術學院資訊工程系

中　華　民　國　9　8　年　1　0　月　3　1　日

研究摘要：

本計畫名稱為「台語文語法結構樹建置」，預計利用三年時間，將包括全羅馬字書寫及漢羅合用的台語文語料庫做進一步處理，製作台語加工語料庫。第一年對語料做語法標記，主要採用統計方法，利用台語華語辭典及中研院詞庫小組開發的工具。第二年的計畫，主要做台語漢字/羅馬字(聲音標記)互轉系統、變調標記，並根據語法標記的結果提供帶詞類的語詞搭配資料。現在提第三年計畫，要建立台語語法結構樹(tree bank)，以做為進一步台語文計算語言學相關研究的基礎。

因為資源很有限，本計畫以依存語法為基礎，並以中研院詞庫小組中文句結構樹的做法為主要學習對象，實作出台語文語法結構樹。若有餘力，則嘗試用其它語法理論來做做看，例如連結語法、詞彙語法等。

所有的成果，也將開發成 web 介面供大眾使用。

關鍵詞：語料庫，台語文，語法結構樹，剖析器

This 3-years-long project entitled "The Construction of Taiwanese Tree Bank". We plan to refine the written Taiwanese raw corpus that includes both Han-Romanization and full Romanization scripts. The task of the first year is to annotate the syntactic marker mainly using the statistical-based method via Taiwanese-Mandarin dictionary and tools made by CKIP group. The second year project is to implement an online conversion system between the Taiwanese Han-Romanization and full Romanization scripts, an online Taiwanese tone sandhi marker system, and an online Taiwanese word collocation system with the POS tagging information. New we propose the following work, use these bases to construct the Taiwanese tree bank which can be the most important basis of the written Taiwanese computational linguistics related research.

Due to lack of resource, this project bases on dependency grammar, follows the steps of Sinica treebank, and want to implement ths Taiwanese treebank. We also intend to try other grammar theory such as link grammar or word grammar et. al.

We also want to offer the web interface tools made by this project to all users.

Keywords: corpus, written Taiwanese, treebank, parser

一、　計畫內容簡介 [Kè-ōe lōe-iông kán-kài]

本計畫預計以台語文語料為基礎，利用三年的時間，逐步建立台語文句的語法結構樹，此為第二年的計畫，主要目標是做全羅、漢羅文本互轉及台語文語詞搭配(collocation)。

建立語法結構樹，需要投入大量的研究人力，然而所核准的研究人力費，只能聘請兼任助理來執行此計畫。無論如何，我們只能盡力朝目標前進。

因為台語文的特殊性，包括書寫尚未標準化、台語漢字一字多音情形比華語還嚴重等等，所以我們認為，如果要建立台語文語料庫，語音標記是有必要的，而全羅馬字書寫是台語文書寫的傳統之一，可直接以全羅馬字來標記發音。

第一年的計畫成果，整理了一些漢羅及全羅台語文對齊的語料，數量雖然不算多，不過可以提供基礎。計畫成果置於網站上 http://iug.csie.dahan.edu.tw/TGB/tagging/tagging.asp 。

第二年的計畫成果為台語文漢羅文本和羅馬字文本互轉，背後利用語料的統計。其中，漢羅轉羅馬字，平均可達 95%的正確率，效果不錯；但是羅馬字轉漢羅效果較差，正確率不到 90%。計畫的成果也置於網站上 http://iug.csie.dahan.edu.tw/TGB/CLHLMI/clhlmi.asp 。

這一年的計畫，以第一年台語 tagging 的成果為主要基礎，想要建立句法結構樹。然而在實作上，遭遇到很多問題，以致於實際的產出比原先設定的進度少了許多。雖然日後暫時沒有研究經費的支持，我們仍將設法繼續未完的工作。

二、　執行步驟 [Chi̍p-hêng pō͘-sò͘]

要建立台語文剖析器，需要較多的人力，而本計畫的經費連一個專任助理都無法聘請，只能用較簡便的方式來執行。

首先，我們請求中央研究院資訊科學所詞庫小組的協助，在中研院資訊所辦了一場小型研習，由詞庫人員告訴本計畫的工作團隊關於分詞、詞性標記及句法剖析的相關技術與中文現有的資源。

因為限於資源，本計畫的人力主要用於整理語句，工作流程如下：

1. 選定資料，我們挑選 1923 年出版的台語教科書《台語大成》(劉克明編)及《台灣教會公報》1920 年代的文本，各挑 500 句，打字；
2. 《台語大成》為漢字文本，利用第二年的計畫成果，將文本轉成羅馬字文本，並由人工校對結果;《台灣教會公報》為羅馬字文本，同樣利用第二年的計畫成果，將文本轉成漢羅文本，並由人工校對結果
3. 將前述的漢羅、羅馬字對齊文本，以第一年的計畫成果得到詞性標記的結果，並以人工校對；
4. 詞庫小組的中文詞性標記系統，得到結果後，還有一個工具可以輔助合詞、分詞、改詞類等工作，但是因為文本的格式不同，不能套用在台語。因此，人工修改詞性標記結果，變得很繁複；另外，經由手動更改的結果往往影響到格式的挪動，例如全型、半型的問題；多一個空格或少一個空格的問題等，導致下一步驟 parser 程式無法判讀，來來回回處理費時費力；
5. 接著，我們利用中研院資訊所詞庫小組的 parsing 工具處理，這個 parser 以 PCFG (Probabilistic Context-Free Grammar)為基礎，利用中文的訓練語料，進行中文語句剖析，正確率約為 75%；問題是，拿來做台語，正確率又下降許多，機器在「語意角色」及「句法層次」上經常誤判，必須加以校正。幾乎所有的句子都需要人工再校正，而且操作頗費功夫，節點的挪移很花時間；

利用中文 parsing 工具處理台語，以下是我們遇到的一些問題:

- 機器在處理台語由代名詞代表的所有格形式時，常有誤判的情形。如：阮（我的、我們的）、恁（你的、你們的）、in（他的、他們的）常誤判為動詞 VH 或形容詞 A；
- “在(tī)”由機器分詞出來一律為介詞 P，但當句子沒有主要動詞時，“在(tī)”是擔任句中的主要動詞，表示「位於」之意。如 1.1 句：恁厝在(tī)何位(toh-ūi)？
- 有時機器分析出來為動詞組 VP，但實際上應為句子 S(如 2.1、10.1

句)。

- 有時機器分析出來為名詞組 NP，但實際上應為句子 S(如 8.1 句)。
- 有時機器分析出來為名詞組 NP，但實際上應為動詞組 VP。
- 有時機器分析出來為句子 S，但實際上應為動詞組 VP。
- 有時機器分析出來少一個節點，必須增加方為完整(如 4.1 句)。句中的"此個[Chit-ê]"是代表「這個人」之意，省略定量詞組後的名詞組，而實際即代表一個完整的「定、量、名」結構，擔任句中的 theme 角色。
- 因為台語語料中的數量詞組沒有連字符，而剖析出來的圖表因為位置的錯置而不清楚，必須將位置加以調整(如 5.1 句)。定量詞組也有這種情形。
- 因為這個 parser 的設計是以華語語料為主，因此有些句子在台語是合語法的，但用 parser 分析卻無法成為合法的句子（機器不接受），只好稍做調整(如 9.1 句)。
- "著 tio̍h"是表示「義務」deontics；而不是「評價」evaluation。
- 當羅馬拼音較長時，無法完全顯示出來，尾巴會被切掉。

三、 實驗結果及分析 [Si̍t-giām kiat-kó kap hun-sek]

在此列舉其中 10 句，我們將陸續整理出來，放在計畫成果網站上。

1. (#1.1) 恁[Lín](Nh) 厝[chhù](Na) 在[tī](VC1) 何位 [toh-ūi](Ncd) ?(QUESTIONCATEGORY)
2. (#2.1) 汝[Lí](Nh) 是[sī](SHI) in[in](Nh) 甚人 [siáⁿ-lâng](Nh) ?(QUESTIONCATEGORY)
3. (#4.1) 此個[Chit-ê](Neqa) 是[sī](SHI) 恁[Lín](Nh) 後生[hāu-siⁿ](Na) 不[m̄](T) ?(QUESTIONCATEGORY)
4. (#5.1) 何[Toh](Nep) 一[chi̍t](Neu) 位[ūi](Nf) 是[sī](SHI) 恁[Lín](Nh) 令郎[lēng-lông](Na) ?(QUESTIONCATEGORY)
5. (#6.1) 阮[Gún](Nh) 查晡人[cha-po͘-lâng](Na) 無[bô](D) 在得

[tī-teh](VH) .(PERIODCATEGORY)

6. (#7.1) 返來[tńg-lâi](VA) 即[chiah](Da) 共[kā](P) 伊[i](Nh) 講[kóng](VE) .(PERIODCATEGORY)
7. (#8.1) 阮[Gún](Nh) 查某人[cha-bó͘-lâng](Na) 昨昏暗[cha-hng-àm](Nd) 拾[khioh](VC) 囝仔[gín-á](Na) .(PERIODCATEGORY)
8. (#9.1) 恁[Lín](Nh) 的[ê](DE) 犬仔[káu-á](Na) 走來[cháu lâi](VA) 在[tī](P) 阮[gún](Nh) 兜[tau](Nc) .(PERIODCATEGORY)
9. (#10.1) 此等[Chiah-ê](Neqa) 官舍[koaⁿ-sià](Nc) 攏[lóng](D) 是[sī](SHI) 支廳[chi-thiaⁿ](Nc) 的[ê](DE) .(PERIODCATEGORY)
10. (#12.1) 此等[Chiah-ê](Neqa) 是[sī](SHI) 常常[siông-siông](D) 有[ū](V_2) 的[ê](DE) 事情[tāi-chì](Na) .(PERIODCATEGORY)

(以上句子是經過人工校正後的結果)

1. #1:1.[0] S(theme:NP(possessor:Nh:恁[Lín]|Head:Na:厝[chhù])|Head:VC1:在[tī]|Head:Ncd:何位[toh-ūi])#?(QUESTIONCATEGORY)
2. #2:1.[0] S(theme:NP(Head:Nh:汝[Lí])|Head:SHI:是[sī]|range:NP(possessor:Nh:[in]|Head:Nh:甚人[siáⁿ-lâng]))#?(QUESTIONCATEGORY)
3. #4:1.[0] S(theme:NP(quantity:Neqa:此個[Chit-ê])|Head:SHI:是[sī]|range:NP(possessor:Nh:恁[Lín]|Head:Na:後生[hāu-siⁿ])|particle:T:不[m̄])#?(QUESTIONCATEGORY)
4. #5:1.[0] S(theme:NP(quantifier:Nep:何[Toh]|Head:DM:一位[chi̍t ūi])|Head:SHI:是[sī]|range:NP(possessor:Nh:恁[Lín]|Head:Na:令郎[lēng-lông]))#?(QUESTIONCATEGORY)
5. #6:1.[0] S(theme:NP(possessor:Nh:阮[Gún]|Head:Na:查晡人[cha-po͘-lâng])|negation:D:無[bô]|Head:VH:在得[tī-teh])#.(PERIODCATEGORY)
6. #7:1.[0] VP(Head:VA:返來[tńg-lâi]|complement:VP(quantity:Da:即[chiah]|target:PP(Head:P:共[kā]|DUMMY:NP(Head:Nh:伊[i]))|Head:VE:講

[kóng]))#.(PERIODCATEGORY)

7. #8:1.[0] S(agent:NP(possessor:Nh:阮[Gún]|Head:Na:查某人 [cha-bó͘-lâng])|time:Nd:昨昏暗[cha-hng-àm]|Head:VC:拾 [khioh]|theme:NP(Head:Na:囝仔[gín-á]))#.(PERIODCATEGORY)
8. #9:1.[0] S(theme:NP(possessor:N·的(head:Nh:恁[Lín]|Head:DE:的 [ê])|Head:Na:犬仔[káu-á])|Head:VA:走來[cháu lâi]|location:PP(Head:P:在 [tī]|DUMMY:NP(possessor:Nh:阮[gún]|Head:Nc:兜 [tau])))#.(PERIODCATEGORY)
9. #10:1.[0] S(theme:NP(quantifier:Neqa:此等[Chiah-ê]|Head:Nc:官舍 [koaⁿ-sià])|evaluation:D:攏[lóng]|Head:SHI:是[sī]|range:NP(property:N·的 (head:Nc:支廳[chi-thiaⁿ]|Head:DE:的[ê])))#.(PERIODCATEGORY)
10. #12:1.[0] VP(theme:NP(quantity:Neqa:此等[Chiah-ê])|Head:SHI:是 [sī]|range:NP(predication:VP·的(head:VP(time:D:常常 [siông-siông]|Head:V_2:有[ū])|Head:DE:的[ê])|Head:Na:事情 [tāi-chì]))#.(PERIODCATEGOR)

(以上是機器 parsing 後，人工校正後的結果)

四、 產出資料 [Àn-sǹg sán-chhut chu-liāu]

本計畫原本希望能產出 1000 句台語句法結構樹，卻因為遭遇到種種的問題，並沒有達成此目標。我們會陸續整理，並將成果置於 http://iug.csie.dahan.edu.tw/TGB/parsing 。

五、 參考文獻 [Chham-khó bûn-hiàn]

Iunn, Un-Gian, Lau, Kiat-Gak, Tan-Tenn, Hong-Giau, Lee, Sheng-An, and Kao, Cheng-Yan, 2007, “Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods”, International Journal of Computational Linguistics and Chinese Language Processing Vol. 12, No. 4, pp. 349-370

Yu-Ming Hsieh, Duen-Chi Yang and Keh-Jiann Chen, 2007, “Improve Parsing Performance by Self-Learning”, Computational Linguistics and Chinese

Language Processing, vol. 12, No. 2, June 2007, pp.195-216
楊允言、戴嘉宏、劉杰岳、陳克健、高成炎, 2008, “利用統計方法及中文訓練資料處理台語文詞性標記”, 第二十屆自然語言與語音處理研討會論文集 pp166-179,台北：台師大資訊系，2008/9/4-5

六、 計畫成果自評 [Kè-ėk sêng-kó chū phêng]

研究內容與原計畫相符,但是限於經費與資源,我們並沒有達成預期目標(實作出台語剖析器),僅只整理出一些台語句法結構樹。研究成果之學術價值不足，但是應用價值頗高。