# 行政院國家科學委員會補助專題研究計畫成果報告
# [ Hêng-chèng-īⁿ Kok-ka Kho-hák Úi-ôan-hōe Pó-chō Choan-tôe Gián-kiù Kè-ōe Sêng-kó Pò-kò ]
# 台語文語料庫蒐集及語料庫為本台語書面語音節詞頻統計
# [ Tâi-gú-bûn Gú-liāu-khò Sơ-chi̍p kap Gú-liāu-khò ûi Pún Tâi-gú Su-bīn-gú Im-chiat Sû-pîn Thóng-kè ]
# Taiwanese Corpus Collection and Corpus Based Syllable / Word Frequency Counts for Written Taiwanese

計畫類別：個別型計畫
[Kè-ōe lūi-pia̍t : Kò-pia̍t-hêng kè-ōe ]

計畫編號：NSC 93-2213-E-122-001-
[Kè-ōe pian-hō : NSC 93-2213-E-122-001- ]

執行期間：2004 年 8 月 1 日至 2005 年 7 月 31 日
[Chi̍p-hêng kî-kan : 2004 nî poeh-gė̍h chhe-it kàu 2005 nî chhit-gė̍h saⁿ-cha̍p-it hō ]

計畫主持人：大漢技術學院資訊工程系助理教授 楊允言
[Kè-ōe chú-chhî-jîn : Tāi-hàn Ki-sut Ha̍k-īⁿ Chu-sìn Kang-têng-hē chō-lí kàu-siū Iûⁿ Ún-giân ]

共同主持人：台東大學語文教育學系副教授 張學謙
[Kiōng-tông chú-chhî-jîn : Tâi-tang Tāi-ha̍k Gú-bûn Kàu-io̍k-hē hù kàu-siū Tiuⁿ Ha̍k-khiam ]

執行單位：大漢技術學院 資訊工程系
[ Chi̍p-hêng tan-ūi : Tāi-hàn Ki-sut Ha̍k-īⁿ Chu-sìn Kang-têng-hē ]

## 一、計畫內容簡介 [ Kè-ōe lōe-iông kán-kài ]

本計畫預計要建立至少三百萬音節的台語文語料庫，文類涵蓋學術論文、報導性文章（新聞報導、訪談）、各類創作（小說、劇本、散文、新詩、笑話、寓言故事、童謠）、民間文學、書信、...等各類型。 [ Pún kè-ōe àn-sǹg beh kiàn-li̍p chì-chió saⁿ-pah-bān im-chiat ê Tâi-gú-bûn gú-liāu-khò, bûn-lūi pau-koah ha̍k-sut lūn-bûn, pò-tō-sèng bûn-chiuⁿ ( sin-bûn pò-tō, hóng-tâm ), kok lūi chhòng-chok ( siáu-soat, ke̍k-pún, sàn-bûn, sin-si, chhiò-khe, gū-giân kò-sū, gín-á koa-si ) , bîn-kan bûn-ha̍k, phoe-sìn, ... téng-téng kok lūi-hêng. ]

經過一年的努力，我們很高興超過預期的目標許多，收集到的台語文語料超過九百萬音節。 [ Keng-kè chi̍t-tang ê phah-piàⁿ, gún chin hoaⁿ-hí chhiau-kè àn-sǹg ê bo̍k-phiau chin chōe, siu-chi̍p tio̍h ê Tâi-gú-bûn gú-liāu chhiau-kè

káu-pah-bān im-chiat. ]

以目前的語料規模，尚不足以建置一平衡語料庫。不過，我們先以收集到的語料做基礎，計算最基本的音節/語詞的頻率/互訊息(Mutual Information) /相關度(Correlation)等資料，這些統計結果，除了可以對目前國中小鄉土語言教學的教材編寫方向，提出具體的建議，也希望能對相關研究有所貢獻。 [ Í bo̍k-chiân ê kui-bô, iah-koh bô châi-tiāu kiàn-li̍p chi̍t-ê pêng-hêng gú-liāu-khò. M̄-koh, gún seng iōng siu-chi̍p ti̍oh ê gú-liāu chòe ki-chhó, kè-sìng siōng ki-pún ê im-chiat / gú-sû ê pîn-lu̍t / hō-sìn-sit / siong-koan-tō téng ê chu-liāu, chiâ ê thóng-kè kiat-kó, tû-liáu ē-tàng tùi bo̍k-chiân kok-bîn tiong-sió-ha̍k hiong-thó gú-giân kàu-ha̍k ê kàu-châi pian-siá hong-hiòng, thê-chhut kū-thé ê kiàn-gī, mā hi-bōng ē-tàng tùi siong-koan gián-kiù ū kòng-hiàn. ]

語料庫的建置，並不是短時間的工作，必須投注許多的人力物力。以華文的語料庫來說，包括台灣中央研究院的現代漢語平衡語料庫，中國的漢語詞頻統計語料庫、現代漢語研究語料庫及漢語加工語料庫，香港的中文五地區共時語料庫，都有大批的研究人力及經費投入，當然相對地，也展現出豐碩的成果。[ Gú-liāu-khò ê kiàn-li̍p, m̄-sī té sî-kan ē-tàng ôan-sêng ê khang-khè, pit-su ài tâu-ji̍p chiok chōe ê jîn-le̍k kah chu-gôan. Í Hôa-bûn ê gú-liāu-khò lâi kóng, pau-koah Tâi-ôan Tiong-iong-gián-kiù-ī$^{n}$ ê hiān-tāi Hàn-gú pêng-hêng gú-liāu-khò, Tiong-kok ê Hàn-gú sû-pîn thóng-kè gú-liāu-khò, hiān-tāi Hàn-gú gián-kiù gú-liāu-khò kah Hàn-gú ka-kang gú-liāu-khò, Hiong-káng ê Tiong-bûn ngó-tōe-khu kiōng-sî gú-liāu-khò, lóng ū tōa-liōng ê jîn-le̍k kah keng-hùi tâu-ji̍p, siong-tùi lâi kóng, mā ū chin phong-phài ê sêng-kó. ]

反觀台語文，雖然台灣社會已經逐漸重視台語，但是目前尚未建置完成一個可公開共用的台語文語料庫；此外，目前台語文界還爲了音標系統及用字問題爭論不休，又使建置語料庫這項重要工作的難度加大。[ Tn̂g-lâi khòa$^{n}$ Tâi-gú-bûn, sui-jiân chit-má Tâi-ôan siā-hōe í-keng chiām-chiām khai-sí tih tiōng-sī Tâi-gú, m̄-koh kàu ta$^{n}$ ûi-chí, iah-bô chi̍t-ê kong-khai ê Tâi-gú-bûn gú-liāu-khò kiàn-li̍p hó-sè; lēng-gōa, bo̍k-chiân Tâi-gú-bûn-kài iah-koh ūi-ti̍oh im-phiau hē-thóng kah iōng-jī būn-tôe tih sio-chì$^{n}$, hō kiàn-li̍p gú-liāu-khò chit-kiā$^{n}$ tiōng-iàu ê khang-khè koh jú phái chìn-hêng. ]

不過，經過十多年來相關台語文電子檔案的累積，以及相關台語文自然語言處理技術的發展，已經讓台語文語料庫的建置有了最起碼的基礎。

本計畫的最終目的在於建立一個往後可公開的台語文語料庫，以期建立台語文自然語言研究的堅實基礎。 [ M̄-koh, keng-kè chap̍-gōa-tang í-lâi, Tâi-gú-bûn lúi-chek chiâⁿ chōe tiān-chú tóng-àn, koh ka-siōng siong-koan Tâi-gú-bûn chū-jiân gú-giân chhú-lí ki-sut̍ ê hoat-tián, í-keng thòe Tâi-gú-bûn gú-liāu-khò ê kiàn-li̍p phah chit̍-ê chiâⁿ hó ê ki-chhó. Pún kè-ōe chòe-āu ê bo̍k-tek sī beh kiàn-li̍p chit̍-ê kong-khai ê Tâi-gú-bûn gú-liāu-khò, chiâⁿ-chòe Tâi-gú-bûn chū-jiân gú-giân chhú-lí siōng chāi ê ki-chhó. ]

## 二、本計畫參與人員 [ Pún kè-ōe chham-ú jîn-ôan ]

本計畫需要資訊科學、語言學及台語文等相關背景的人才共同參與。除了直接參與此計畫的人員之外，間接參與本計畫的台語文語料提供者，也扮演了非常重要的角色。 [ Pún kè-ōe su-iàu chu-sìn kho-ha̍k, gú-giân-ha̍k kah Tâi-gú-bûn siong-koan pōe-kéng ê jîn-châi kiōng-tông chham-ú. Tû-liáu tit̍-chiap chham-ú chit-ê kè-ōe ê jîn-ôan í-gōa, kàn-chiap chham-ú pún kè-ōe ê Tâi-gú-bûn gú-liāu thê-kiong-chiá, mā pān-ián hui-siông tiōng-iàu ê kak-sek. ]

直接參與計畫人員，包括：[ Tit̍-chiap chham-ú kè-ōe jîn-ôan pau-koah : ]

1. 計畫主持人：楊允言（大漢技術學院資訊工程系助理教授）[ Kè-ōe chú-chhî-jîn : Iûⁿ Ún-giân (Tāi-hàn Ki-sut̍ Ha̍k-īⁿ Chu-sìn Kang-têng-hē chō-lí kàu-siū ) ]
2. 共同計畫主持人：張學謙（台東大學語文教育學系副教授）[ Kiōng-tông kè-ōe chú-chhî-jîn : Tiuⁿ Ha̍k-khiam ( Tâi-tang Tāi-ha̍k Gú-bûn Kàu-io̍k-hē hù-kàu-siū )]
3. 計畫助理： [ Kè-ōe chō-lí : ]
   (i) 劉杰岳（台大資訊工程研究所畢業，具資訊科學及台語文專長，主要負責程式設計）； [ Lâu Kia̍t-gák (Tâi-tāi Chu-sìn kang-têng gián-kiù-só pit-gia̍p, ū chu-sìn kho-ha̍k kah Tâi-gú-bûn choan-tióng, chú-iàu hū-chek thêng-sek siat-kè ) ; ]
   (ii) 廖麗雪（具台語文專長，曾擔任台語文書刊編輯及台語文相關研究計畫助理，主要負責語料整理）； [Liāu Lē-soat ( ū Tâi-gú-bûn choan-tióng, pat chòe kè Tâi-gú-bûn su-khan pian-chi̍p kah Tâi-gú-bûn siong-koan gián-kiù ê kè-ōe chō-lí, chú-iàu hū-chek gú-liāu chéng-lí ) ; ]
   (iii) 陳德樺（淡江大學畢業，具台語文專長，曾擔任台語文刊物主編，

主要負責語料整理） [Tân Tek-hôa ( Tām-kang tāi-hȧk pit-giȧp, ū Tâi-gú-bûn choan-tióng, pat chòe kè Tâi-gú-bûn khan-bu̍t chú-pian, chú-iàu hū-chek gú-liāu chéng-lí ) ]

間接參與計畫人員，主要爲提供語料，包括：賴伯年（工程師，旅居加拿大，曾任《台文通訊》編輯）、林俊育（旅居美國，海外台灣公論報《蕃薯園》台文專刊總編輯）、李勤岸（語言學博士，台師大台文所助理教授）、呂興昌（成大台文所退休教授）、蕭平治（彰化田中國小退休教師）、王寶漣（《TGB 通訊》總編輯）、陳廷宣（前《滇根母語文》總編輯）、盧永芳（目前旅居越南，台語文師資）、鄭詩宗（醫師，《台灣字》總編輯）、張復聚（醫師，南社母語組組頭）、張裕宏（語言學博士，台大語言學研究所退休教授）、鄭良偉（語言學博士，交大電信所客座教授）、莊惠平（具台語文專長）、李自敬（具台語文專長）、...等人。[ Kàn-chaip chham-ú kè-ōe jîn-ôan, chú-iàu sī thê-kiong gú-liāu, pau-koah : Lōa Pek-nî ( kang-têng-su, tòa tī Ka-ná-tà, pat chòe kè “Tâi-bûn thong-sìn” pian-chi̍p ), Lîn Chùn-io̍k ( tòa Bí-kok, hái-gōa Tâi-ôan kong-lūn-pò “Hân-chû-hn̂g” Tâi-bûn choan-khan chóng pian-chi̍p, Lí Khîn-hōaⁿ ( gú-giân-hȧk phok-sū, Tâi-su-tāi Tâi-bûn-só chō-lí kàu-siū ) , Lī Heng-chhiong ( Sêng-tāi Tâi-bûn-só thòe-hiu kàu-siū ) , Siau Pêng-tī ( Chiong-hòa Tiân-tiong kok-hāu thòe-hiu kàu-su ) , Ông Pó-liân ( “TGB thong-sìn” chóng-pian-chi̍p ) , Tân Têng-soan ( chêng “Thòaⁿ-kun Bó-gú-bûn” chóng- pian-chi̍p) , Lô Éng-hong ( bo̍k-chiân tòa tī Ò̍at-lâm, Tâi-gú-bûn su-chu ), Tēⁿ Si-chong ( i-su, “Tâi-ôan-jī” chóng-pian-chi̍p ) , Tiuⁿ Ho̍k-chû ( i-su, Lâm-siā Bó-gú-chơ chơ-thâu ) , Tiuⁿ Jū-hông ( gú-giân-hȧk phok-sū, Tâi-tāi gú-gián-só thòe-hiu kàu-siū) , Tēⁿ Liông-úi ( gú-giân-hȧk phok-sū, Kau-tāi Tiān-sìn-só kheh-chō kàu-siū ) , Chng Hūi-pêng ( ū Tâi-gú-bûn choan-tióng) , Lí Chū-kèng ( ū Tâi-gú-bûn choan-tióng) … téng lâng .]

## 三、 產出資料 [ Sán-chhut chu-liāu ]

經過過去的準備工作以及計畫執行期間的努力，本計畫蒐集到超過九百萬音節的台語文語料，包括漢羅台語文 556 萬多音節及全羅台語文 346 萬多音節。礙於上述語料並未全部徵得原作者同意授權，這些語料暫時無法公開給社會大眾使用，不過，本計畫提供台語文語詞檢索系統讓使用者查詢，並完成相關統計資料。[ Keng-kè kè-khì ê chún-pī khang-khè kah kè-ōe chi̍p-hêng kî-kan ê phah-piàⁿ, pún kè-ōe siu-chi̍p tio̍h chhiau-kè káu-pah-bān

im-chiat ê Tâi-gú-bûn gú-liāu, pau-koah Hàn-lô Tâi-gú-bûn 556-bān-gōa im-chiat kah chôan-lô Tâi-gú-bûn 346-bān-gōa im-chiat . In-ūi chiâ ê gú-liāu pēng m̄-sī lóng tit-tio̍h gôan-chok-chiá tông-ì siū-kôan, chiâ ê gú-liāu chiām-sî iah bô-hoat-tō͘ kong-khai hō͘ siā-hōe tāi-chiòng sú-iōng, m̄-koh, pún kè-ōe thê-kiong Tâi-gú-bûn gú-sû kiám-sek hē-thóng hō͘ sú-iōng-chiá chhâ-sûn, mā ôan-sêng siong-koan ê thóng-kè chu-liāu. ]

1. 台語文語詞檢索查詢系統 [ Tâi-gú-bûn gú-sû kiám-sek chhâ-sûn hē-thóng ]

   分全羅及漢羅語料兩部分，由使用者輸入欲查詢的語詞，系統會從語料中找出此語詞並顯示出前後文供使用者閱覽，透過真實語料學習此語詞的用法。網址在：[ Hun chôan-lô kah Hàn-lô gú-liāu nn̄g pō-hūn, sú-iōng-chiá su-ji̍p beh chhâ-sûn ê gú-sû, hē-thóng ē ùi gú-liāu lāi-to͘e chhē chhut chit-ê gú-sû kah gú-sû ê chêng-āu-bûn hō͘ sú-iōng-chiá khòaⁿ, thàu-kè chin-si̍t gú-liāu lâi o̍h chit-ê gú-sû ê iōng-hoat. Bāng-chí tī : ]
   http://iug.csie.dahan.edu.tw/TG/Concordance/form.asp

2. 統計表 [ Thóng-kè-pió ]

   (a) 漢羅台語文部分 [ Hàn-lô Tâi-gú-bûn pō-hūn ]

      i. 音節頻率統計表 [ Im-chiat pîn-lu̍t thóng-kè-pió ]

      共有 8,527 個音節，我們根據音節頻率高到低做排序，將覆蓋率達 90%的前 1,183 個音節列出供查詢。網址在：[ Lóng-chóng ū 8,527-ê im-chiat, gún kun-kù im-chiat pîn-lu̍t ùi kôan kàu kē pâi-sū, chiong jia-khàm-lu̍t 90% ê chêng 1,183-ê im-chiat liā̍t chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
      http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/shf.asp

      ii. 音節互訊息(Mutual Information)統計表[ Im-chiat hō-sìn-si̍t thóng-kè-pió ]

      共有 570,903 筆音節互訊息的資料，我們取互訊息的值>=10.0 且此雙連音節串(bigram)頻率>=10 的 2,309 筆音節互訊息排序資料列出供查詢。網址在：[ Lóng-chóng ū 570,903-pit im-chiat hō-sìn-si̍t ê chu-liāu, gún kéng hō-sìn-si̍t >= 10.0 ji̍-chhiáⁿ chit-ê siong-liân im-chiat-chhòan pîn-lu̍t >= 10 ê 2,309-pit hō-sìn-si̍t pâi-sū chu-liāu, liā̍t chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
      http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rsl

t/shm.asp

iii. 音節相關度(Correlation)統計表 [ Im-chiat siong-koan-tō thóng-kè-pió ]

共有 570,903 筆音節相關度的資料，我們取相關度的值>=20,000.0 且此雙連音節串(bigram)頻率>=40 的 2,492 筆音節相關度排序資料列出供查詢。網址在： [Lóng-chóng ū 570,903-pit im-chiat siong-koan-tō ê chu-liāu, gún kéng siong-koan-tō >= 20,000.0 jî-chhiáⁿ chit-ê siong-liân im-chiat-chhòan pîn-lu̍t >=40 ê 2,492-pit siong-koan-tō pâi-sū chu-liāu, li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/shc.asp

iv. 語詞頻率統計表[ Gú-sû pîn-lu̍t thóng-kè-pió ]

共有 47,130 個語詞，我們根據語詞頻率高到低做排序，將覆蓋率達 80%的前 2,255 個語詞列出供查詢。網址在：[ Lóng-chóng ū 47,130-ê gú-sû, gún kun-kù gú-sû pîn-lu̍t ù kôan kàu kē pâi-sū, chiong jia-khàm-lu̍t 80% ê chêng 2,255-ê gú-sû li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/whf.asp

v. 語詞互訊息統計表 [ Gú-sû hō-sìn-sit thóng-kè-pió ]

共有 962,470 筆語詞互訊息的資料，我們取互訊息的值>=10.0 且此雙連語詞串(bigram)頻率>=10 的 2,348 筆語詞互訊息排序資料列出供查詢。網址在：[ Lóng-chóng ū 962,470-pit gú-sû hō-sìn-sit ê chu-liāu, gún kéng hō-sìn-sit >= 10.0 jî-chhiáⁿ chit-ê siong-liân gú-sû-chhòan pîn-lu̍t >= 10 ê 2,348-pit gú-sû hō-sìn-sit pâi-sū chu-liāu, li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/whm.asp

vi. 語詞相關度統計表[Gú-sû siong-koan-tō thóng-kè-pió ]

共有 962,470 筆語詞相關度的資料，我們取相關度的值

>=90,000.0 且此雙連語詞串(bigram)頻率>=18 的 2,457 筆語詞相關度排序資料列出供查詢。網址在： [ Lóng-chóng ū 962,470-pit gú-sû siong-koan-tō ê chu-liāu, gún kéng siong-koan-tō >= 90,000.0 jî-chhiáⁿ chit-ê siong-liân gú-sû-chhòan pîn-lu̍t >= 18 ê 2,457-pit gú-sû siong-koan-tō pâi-sū chu-liāu, liát chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ] http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/whc.asp

(b) 全羅台語文部分[ Chôan-lô Tâi-gú-bûn pō-hūn ]

i. 音節頻率統計表[Im-chiat pîn-lu̍t thóng-kè-pió ]

共有 3,525 個音節，我們根據音節頻率高到低做排序，將覆蓋率達 90%的前 623 個音節列出供查詢。網址在：[ Lóng-chóng ū 3,525-ê im-chiat, gún kun-kù im-chiat pîn-lu̍t ùi kôan kàu kē pâi-sū, chiong jia-khàm-lu̍t 90% ê chêng 623-ê im-chiat liát chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ] http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/shf.asp

ii. 音節互訊息統計表 [ Im-chiat hō-sìn-sit thóng-kè-pió ]

共有 297,426 筆音節互訊息的資料，我們取互訊息的值>=7.0 且此雙連音節串(bigram)頻率>=7 的 2,819 筆音節互訊息排序資料列出供查詢。網址在： [ Lóng-chóng ū 297,426-pit im-chiat hō-sìn-sit ê chu-liāu, gún kéng hō-sìn-sit >= 7.0 jî-chhiáⁿ chit-ê siong-liân im-chiat-chhòan pîn-lu̍t >= 7 ê 2,819-pit hō-sìn-sit pâi-sū chu-liāu, liát chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ] http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/spm.asp

iii. 音節相關度統計表 [ Im-chiat siong-koan-tō thóng-kè-pió ]

共有 297,426 筆音節相關度的資料，我們取相關度的值>=6,000.0 且此雙連音節串(bigram)頻率>=12 的 2,484 筆音節相關度排序資料列出供查詢。網址在：[Lóng-chóng ū 297,426-pit im-chiat siong-koan-tō ê chu-liāu, gún kéng siong-koan-tō >= 6,000.0 jî-chhiáⁿ chit-ê siong-liân im-chiat-chhòan

pîn-lu̍t >=12 ê 2,484-pit siong-koan-tō pâi-sū chu-liāu, li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/spc.asp

iv. 語詞頻率統計表 [ Gú-sû pîn-lu̍t thóng-kè-pió ]

共有 73,258 個語詞，我們根據語詞頻率高到低做排序，將覆蓋率達 80%的前 1,463 個語詞列出供查詢。網址在：[Lóng-chóng ū 73,258-ê gú-sû, gún kun-kù gú-sû pîn-lu̍t ùi kôan kàu kē pâi-sū, chiong jia-khàm-lu̍t 80% ê chêng 1,463-ê gú-sû li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/wpf.asp

v. 語詞互訊息統計表 [Gú-sû hō-sìn-sit thóng-kè-pió ]

共有 563,875 筆語詞互訊息的資料，我們取互訊息的值>=7.0 且此雙連語詞串(bigram)頻率>=7 的 2,349 筆語詞互訊息排序資料列出供查詢。網址在： [Lóng-chóng ū 563,875-pit gú-sû hō-sìn-sit ê chu-liāu, gún kéng hō-sìn-sit >= 7.0 ji̍-chhia$^{n}$ chit-ê siong-liân gú-sû-chhòan pîn-lu̍t >= 7 ê 2,349-pit gú-sû hō-sìn-sit pâi-sū chu-liāu, li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/wpm.asp

vi. 語詞相關度統計表[Gú-sû siong-koan-tō thóng-kè-pió ]

共有 563,875 筆語詞相關度的資料，我們取相關度的值>=3,000.0 且此雙連語詞串(bigram)頻率>=6 的 2,411 筆語詞相關度排序資料列出供查詢。網址在：[Lóng-chóng ū 563,875-pit gú-sû siong-koan-tō ê chu-liāu, gún kéng siong-koan-tō >= 3,000.0 ji̍-chhia$^{n}$ chit-ê siong-liân gú-sû-chhòan pîn-lu̍t >= 6 ê 2,411-pit gú-sû siong-koan-tō pâi-sū chu-liāu, li̍at chhut-lâi thê-kiong chhâ-sûn. Bāng-chí tī : ]
http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/rslt/wpc.asp

3. 發表論文：[ Hoat-piáu lūn-bûn ]

我們利用蒐集到的語料的一部分：台語新約聖經，因爲有兩個版本，包括 1916 年巴克禮聖經及 1974 年紅皮聖經，以此材料來計算台語歷時的語言流失情形。計算的結果令人驚訝，若以 1916 年的台語爲基礎，從詞型的觀點看，六十年後，台語語詞流失了 43%。 [Gún lī-iōng siu-chi̍p ti̍oh ê gú-liāu ê chi̍t-pō-hūn : Tâi-gú Sin-iok Sèng-keng, in-ūi ū nn̄g-ê pán-pún, pau-koah 1916 nî Pa-khek-lé Sèng-keng kah 1974 nî Âng-phê Sèng-keng, í chit nn̄g-hūn chhâi-liāu lâi kè-sìng Tâi-gú le̍k-sî ê gú-gián liû-sit chêng-hêng. Kè-sìng ê kiat-kó chin kia$^{n}$-lâng, nā-sī iōng 1916 nî ê Tâi-gú chòe ki-chhó, ùi sû-hêng ê koan-tiám lâi khòa$^{n}$, 60 tang āu, Tâi-gú gú-sû liû-sit 43% . ]

## 四、 技術資料 [ Ki-sù̍t chu-liāu ]

1. 語料儲存格式：[ Khìng gú-liāu ê keh-sek ]

本計畫所蒐集的語料，羅馬字的部分，主要包括 HOTSYS、HOTSYS 2000 及 Taiwanese Package 1.50 三種軟體的編碼。我們利用劉杰岳開發的 Taiwanese Package 做轉碼，將各種格式資料轉成 plain text。其中，調符以數字表示，如"lé --> le2"、"kòng --> kong3"、"hôa --> hoa5"、"bēng --> beng7"、"tit --> tit8"...等，另外，後元音"ɔ"以"ou"表示，鼻音"$^{n}$"以"N"表示。[ Pún kè-ōe sɔ-chi̍p ti̍oh ê gú-liāu, Lô-má-jī ê pō-hūn, chú-iàu pau-koah HOTSYS, HOTSYS 2000 kah Taiwanese Package 1.50 sa$^{n}$-chióng nn̄g-thé ê pian-bé. Gún lī-iōng Lâu Kia̍t-gák khai-hoat ê Taiwanese Package chòe chóan-bé, chiong kok-chióng keh-sek ê chu-liāu chóan chòe Plain Text. Kî-tiong, tiāu-hû iōng sò-jī piáu-sī, chhin-chhiū$^{n}$ "lé --> le2", "kòng --> kong3", "hôa --> hoa5", "bēng --> beng7", "tit --> tit8" … téng, lēng-gōa, āu-gôan-im "ɔ" iōng "ou" piáu-sī, phī$^{n}$- im "$^{n}$" iōng "N" piáu-sī. ]

2. 斷詞程式 [ Tn̄g-sû thêng-sek ]

我們以台文華文線上辭典的六萬多詞爲依據，用 backward maximal matching 演算法對漢羅台文斷詞，使用 Java 程式語言。[ Gún iōng Tâi-bûn Hôa-bûn sòa$^{n}$-téng sû-tián ê la̍k-bān-gōa sû chòe kun-kù, iōng Backward Maximal Matching ián-sòan-hoat tùi Hàn-lô Tâi-gú-bûn chòe tn̄g-sû, sú-iōng Java thêng-sek gú-giân. ]

因爲書寫並不一致，加上目前斷詞程式並未處理定量詞、專有名詞等，所以斷詞結果比實際的詞次(word tokens)多，也比實際的詞型(word

types)少。(例如有一個雙音節詞 AB，也有人寫成 AB'，而辭典只收錄 AB 沒有 AB'，詞 AB'可能會被拆成 A 和 B'兩個語詞，導致詞次增加、詞型減少) [ In-ūi su-siá hong-sek pēng bô it-tì, koh ka-siōng bo̍k-chiân ê tn̄g-sû thêng-sek pēng bô chhú-lí tēng-niû-sû, choan-iú bêng-sû téng-téng, só-í tn̄g-sû ê kiat-kó pí sit-chè ê sû-chhù khah chōe, mā pí sit-chè ê sû-hêng khah chió. ( Khó-pí kóng, ū chit-ê siang-im-chiat sû AB, mā ū-lâng siá chòe AB', m̄-koh sû-tián kan-na siu AB bô siu AB', an-nē, AB' chit-e sû khó-lêng ē hông thiah-chòe A kah B' nn̄g-ê gú-sû, tì-sú sû-chhù cheng-ka, sû-hêng kiám-chió ) ]

3. 語詞檢索程式 [ Gú-sû kiám-sek thêng-sek ]

資料部分，每篇文章是一個文字檔，我們將檔名及其目錄等訊息放在資料庫中，根據資料庫的內容，逐一開啓檔案並與輸入的語詞做比對，比對出來後，列出左右固定長度的前後文。這是 web 介面程式，使用 ASP，作業系統爲 Microsoft Windows 2000 Server。 [ Chu-liāu pō-hūn, múi chit-phi$^{n}$ bûn-chiu$^{n}$ sī chit-ê bûn-jī tóng-àn, gún chiong tóng-miâ kah i ê bo̍k-lio̍k téng-téng ê sìn-sit hē tī chu-liāu-khò lāi-tóe, kun-kù chu-liāu-khò ê lōe-iông, kā khui chit-ê chit-ê ê tóng-àn phah-khui, pēng-chhiá$^{n}$ kah su-ji̍p ê gú-sû lâi pí-tùi, pí-tùi tio̍h liáu-āu, liat-chhut kò-tēng tn̂g-tō̍ ê chêng-āu-bûn. Che-sī Web kài-bīn ê thêng-sek, sú-iōng ASP, chok-giap hē-thóng sī Microsoft Windows 2000 Server . ]

我們還將使用者所查詢的語詞、每日查詢次數等訊息記錄下來，提供日後改進系統的參考。 [ Gún koh ū chiong sú-iōng-chiá só chhâ-sûn ê gú-sû, ta̍k-kang chhâ-sûn kúi pái téng-téng ê sìn-sit kì-lo̍k lo̍h-lâi, thang chiâ$^{n}$-chòe āu-ji̍t kái-chìn hē-thóng ê chham-khó . ]

4. 頻率/互訊息/相關度統計程式[ Pîn-lu̍t / hō-sìn-sit / siong-koan-tō̍ thóng-kè thêng-sek ]

使用 Java 程式語言。[ Sú-iōng Java thêng-sek gú-giân ]

互訊息的公式爲，對語料庫中相鄰的兩個音節(或語詞) A,B，[ Hō-sìn-sit ê kong-sek sī, tùi gú-liāu-khò lāi-tóe sio-óa ê nn̄g-ê im-chiat ( iah-sī gú-sû ) A kah B , ]

$$MI(AB) = - \log \frac{P(AB)}{P(A)\,P(B)}$$

相關度的公式爲，對語料庫中相鄰的兩個音節(或語詞) A,B，

[ Siong-koan-tō ê kong-sek sī, tùi gú-liāu-khò lāi-tóe sio-óa ê nn̄g-ê im-chiat ( iah-sī gú-sû ) A kah B , ]

$$rel(AB) = \frac{n\,(n_{11} \times n_{22} - n_{12} \times n_{21})^2}{n_{1*} \times n_{2*} \times n_{*1} \times n_{*2}}$$

其中，[ Kî-tiong ]

| | $B$ | $\sim B$ | $\Sigma$ |
|---|---|---|---|
| $A$ | $n_{11}$ | $n_{12}$ | $n_{1*}$ |
| $\sim A$ | $n_{21}$ | $n_{22}$ | $n_{2*}$ |
| $\Sigma$ | $n_{*1}$ | $n_{*2}$ | $n$ |

5. 頻率/互訊息/相關度查詢程式[Pîn-lu̍t / hō-sìn-sit / siong-koan-tō chhâ-sûn thêng-sek ]

我們將統計結果放入資料庫，開發 web 介面程式做查詢，並有計數器紀錄這些統計表被查詢的次數。使用 ASP，作業系統爲 Microsoft Windows 2000 Server。[ Gún chiong thóng-kè ê kiat-kó hē-jip chu-liāu-khò, khai-hoat Web kài-bīn thêng-sek lâi chhâ-sûn, pēng-chhiáⁿ kì-lo̍k chiâ ê thóng-kè-pió hông chhâ kúi pái. Sú-iōng ASP, chok-giap hē-thóng sī Microsoft Windows 2000 Server . ]

## 五、 資料說明 [ Chu-liāu soat-bêng ]

本計畫的語料來源，主要來自以下幾個部分：[ Pún kè-ōe ê gú-liāu lâi-gôan chú-iàu ū í-hā kúi-ê pō-hūn : ]

1. 台文刊物：包括《台文通訊》(1991 年創刊)、《台文罔報》(1996 年創刊)、《TGB 通訊》(1999 年創刊)、《蓮蕉花》(1999 年創刊)、《台灣字》(2000 年創刊，全羅馬字)、《淡根》母語文雜誌(2002 年創刊，現已停刊)、《台灣公論報》蕃薯園台文專刊(2003 年創刊)、...等。 [ Tâi-bûn khan-bu̍t : pau-koah “Tâi-bûn Thong-sìn” (1991 nî chhòng-khan), “Tâi-bûn Bóng-pò” ( 1996 nî chhòng-khan ), “TGB Thong-sìn” ( 1999 nî chhòng-khan ) , “Liân-chiau-hoe” ( 1999 nî chhòng-khan ), “Tâi-ôan-jī” ( 2000 nî chhòng-khan ) , “Thòaⁿ-kun bó-gú-bûn” chha̍p-chì ( 2002 nî chhòng-khan, chit-má í-keng thêng-khan ), “Tâi-ôan kong-lūn-pò” Han-chû-hn̂g Tâi-bûn choan-khan ( 2003 nî chhòng-khan ), … téng-téng .]
2. 專書或論文：主要由作者或編者提供。[ Choan-su iah-sī lūn-bûn : chú-iàu sī chok-chiá iah-sī pian-chiá thê-kiong . ]

3. 研究計畫成果：主要爲國家台灣文學館籌備處委託成功大學台灣文學系執行的「台灣白話字文學資料蒐集整理計畫」中已經數位化的電子檔。計畫主持人爲呂興昌教授。 [ Gián-kiù kè-ōe sêng-kó : chú-iàu sī kok-ka Tâi-ôan Bûn-hák-kóan Tiû-pī-chhù úi-thok Sêng-kong Tāi-hák Tâi-ôan-bûn-hák-hē chi̍p-hêng ê “Tâi-ôan Pe̍h-ōe-jī bûn-hák chu-liāu sơ-chi̍p chéng-lí kè-ōe” lāi-tóe í-keng sờ-ūi-hòa ê tiān-chú tóng-àn. Kè-ōe chú-chhî-jîn sī Lī Heng-chhiong kàu-siū. ]

本計畫所蒐集到的語料，數量爲：台語羅馬字部分有 3,462,367 音節 (tokens / 3,525 types)，漢羅合用台語文有 5,568,057 音節 (tokens / 8,527 types )，共計 9,030,424 音節，爲原先預期的三倍。計算音節數時，我們將不合法的台語羅馬字去除（有可能是打字錯誤，或者是夾雜在語料中的英文）。若以語詞來計算，台語羅馬字部分有 2,436,599 詞次(73,258 詞型)，漢羅合用台語文有 4,051,195 詞次(47,130 詞型)。理論上，漢羅的詞型應該不少於台語羅馬字，其原因應該包括下列因素： [ Pún kè-ōe siu-chi̍p tio̍h ê gú-liāu ê sờ-liōng : Tâi-gú Lô-má-jī pō-hūn ū 3,462,367-ê im-chiat ( Tokens / 3,525 Types ), Hàn-lô hā̍p-iōng Tâi-gú-bûn ū 5,568,057-ê im-chiat ( Tokens / 8,527 Types ) , lóng-chóng 9,030,424-ê im-chiat, sī thâu-khí-seng àn-sǹg ê saⁿ-pē. Kè-sǹg im-chiat-sờ ê sî-chūn, gún chiong bô hā̍p-hoat ê Tâi-gú Lô-má-jī kā thé̍h-tiāu ( ū khó-lêng sī phah-jī chhò-gō͘, iah-sī lām tī gú-liāu lāi-tóe ê Eng-bûn ). Nā-sī iōng gú-sû lâi sǹg, Tâi-gú Lô-má-jī pō-hūn ū 2,436,599-ê sû-chhù ( 73,258-ê sû-hêng ), Hàn-lô hā̍p-iōng Tâi-gú-bûn ū 4,051,195-ê sû-chhù ( 47,130-ê sû-hêng ). Lí-lūn siōng, Hàn-lô ê sû-hêng eng-kai khah chōe kè Tâi-gú Lô-má-jī, chit-ê gôan-in, èng-kai pau-koah ē-bīn kúi-ê in-sờ :]

1. 台語羅馬字連字符書寫不一致，導致詞型增加；[ Tâi-gú Lô-má-jī liân-jī-hû su-siá bô it-tì, tì-sú sû-hêng cheng-ka ; ]
2. 因爲腔調或文白問題，導致台語羅馬字的詞型增加（例如"phê-ôe" 和 "phôe-ê"漢字書寫都是「皮鞋」，屬於腔調問題；"gêng-chiap" 和 "ngiâ-chih"漢字書寫都是「迎接」，屬於文白問題）； [ In-ūi khiuⁿ-kháu iah-sī bûn-pe̍h ê būn-tôe, tì-sú Tâi-gú Lô-má-jī ê sû-hêng cheng-ka ( khó-pí "phê-ôe" kah "phôe-ê" Hàn-jī su-siá lóng-sī "皮鞋" , che-sī khiuⁿ-kháu ê būn-tôe; "gêng-chiap" kah "ngiâ-chih" Hàn-jī su-siá lóng-sī "迎接", che-sī bûn-pe̍h būn-tôe) ; ]

3. 斷詞程式沒有處理定量詞，導致漢羅詞型減少；[ Tīng-sû thêng-sek bô chhú-lí tēng-niû-sû, tì-sú Hàn-lô sû-hêng kiám-chió ; ]
4. 斷詞程式沒有處理專有名詞，導致漢羅詞型減少；[ Tīng-sû thêng-sek bô chhú-lí choan-iú bêng-sû, tì-sú Hàn-lô sû-hêng kiám-chió ; ]
5. 漢羅書寫不一致，本來應該會增加詞型，但是辭典可能只有其中一種寫法，導致漢羅詞型沒有增加（但是詞次增加）。 [ Hàn-lô su-siá bô it-tì, pún-chiâⁿ èng-kai ē cheng-ka sû-hêng, m̄-koh sû-tián khó-lêng kan-na ū kî-tiong chit-khóan siá-hoat, tì-sú Hàn-lô sû-hêng bô cheng-ka ( m̄-koh sû-chhù cheng-ka ) .]

語料各文類的分佈情形如下表，我們直接以檔案大小來計算各文類的比例。 [ Gú-liāu kok bûn-lūi hun-pò ê chêng-hêng liàt tī ē-bīn ê pió, gún tit-chiap iōng tóng-àn tōa-sòe lâi kè-sìng kok bûn-lūi ê pí-lē. ]

| 文類[Bûn-lūi] | 漢羅[Hàn-lô] | 全羅[Chôan-lô] |
|---|---|---|
| 學術[Hák-sùt] | 7.48% | 2.01% |
| 報導[Pò-tō] | 4.23% | 2.54% |
| 訪談[Hóng-tâm] | 1.42% | 0.00% |
| 傳記[Tōan-kì] | 2.90% | 5.03% |
| 評論[Phêng-lūn] | 4.87% | 4.39% |
| 其它[Kî-tha] | 1.20% | 0.34% |
| 小說[Siáu-soat] | 29.31% | 59.08% |
| 散文[Sàn-bûn] | 35.78% | 17.16% |
| 新詩[Sin-si] | 5.30% | 3.42% |
| 劇本[Ke̍k-pún] | 3.43% | 3.42% |
| 兒童[Gín-á] | 0.41% | 0.97% |
| 笑話[Chhiò-khe] | 0.27% | 0.24% |
| 寓言[Gū-giân] | 0.24% | 0.12% |
| 對話[Tùi-ōe] | 0.38% | 0.04% |
| 書信[Phoe-sìn] | 1.04% | 0.58% |
| 民間文學[Bîn-kan bûn-hák] | 0.72% | 0.11% |
| 演講[Káng-ián] | 1.02% | 0.54% |

## 六、 未來工作 [ Bī-lâi khang-khè ]

台語文語料庫的蒐集整理只是一個起步，除了繼續蒐集整理，我們也

很希望能繼續延伸出相關應用。包括：[Tâi-gú-bûn gú-liāu-khò ê siu-chip chéng-lí kan-na sī chit-ê khí-pō, tû-liáu kè-siok siu-chip chéng-lí, gún mā chiâⁿ hi-bōng ē-tàng kè-siok iân-sin siong-koan ê èng-iōng. Pau-koah : ]

- 繼續整理語料，除了這段時間台語文刊物及專書、論文繼續在出版外，網路上如台語網及台語文的 Wikipedia（線上免費百科全書）等，語料也持續在增加；[ Kè-siok chéng-lí gú-liāu, tû-liáu chit tōaⁿ sî-kan Tâi-gú-bûn khan-but kah choan-su, lūn-bûn kè-siok ū tih chhut-pán í-gōa, bāng-lō téng-bīn ê chu-liāu, chhin-chhiūⁿ Tâi-gú-bāng kah Tâi-gú-bûn ê Wikipedia ( sòaⁿ-téng bián-hùi pek-kho-chôan-su ) téng-téng, gú-liāu mā kè-siok tih cheng-ka ; ]
- 改進斷詞程式；[ Kái-chìn tn̄g-sû thêng-sek ; ]
- 語料加工：標注詞性，或是句法樹的建立等；[ Gú-liāu ka-kang : phiau-chù sû-sèng, iah-sī kiàn-lip kù-hoat-chhiū téng-téng ; ]
- 利用台語文語料來建立台語常用句型，這對台語學習應該有極大的幫助；[ Lī-iōng Tâi-gú-bûn gú-liāu lâi kiàn-lip Tâi-gú siông-iōng kù-hêng, che tùi Tâi-gú hák-sip èng-kai ū chiâⁿ tōa ê pang-chān ; ]
- 利用台語文語料庫來編纂台語辭典； [ Lī-iōng Tâi-gú-bûn gú-liāu-khò lâi pian-chip Tâi-gú sû-tián ; ]
- 建立平衡語料庫；[ Kiàn-lip pêng-hêng gú-liāu-khò ; ]
- 取得作者授權，未來將語料庫公開，促成相關基礎及應用研究的進行；[ Tit-tioh chok-chiá siū-kôan, bī-lâi chiong gú-liāu-khò kong-khai, chhiok-sêng siong-koan ki-chhó kah èng-iōng gián-kiù ê chìn-hêng ; ]
- ...

## 七、補充說明 [ Pó-chhiong soat-bêng ]

本報告採用華文及台語羅馬字書寫，一方面配合慣例，另一方面則希望本報告完成後，又能增加台語文語料。[ Pún pò-kò chhái-iōng Hôa-bûn kah Tâi-gú Lô-má-jī su-siá, chit-hong-bīn phòe-hap̍ kòan-lē, lēng-gōa chit-hong-bīn mā hi-bōng pún pò-kò ôan-sêng liáu-āu, koh ē-tàng cheng-ka Tâi-gú-bûn gú-liāu. ]

## 八、參考資料 [ Chham-khó chu-liāu ]

Taiwanese Package http://www.phahng.idv.tw/

中研院現代漢語平衡語料庫 [ Tiong-gián-īⁿ hiān-tāi Hàn-gú pêng-hêng gú-liāu-khò]

http://www.sinica.edu.tw/SinicaCorpus/

台文華文線頂辭典 [ Tâi-bûn Hôa-bûn sòaⁿ-téng sû-tián ]

http://iug.csie.dahan.edu.tw/TG/sutian/

台語文語詞檢索系統[ Tâi-gú-bûn gú-sû kiám-sek hē-thóng ]

http://iug.csie.dahan.edu.tw/TG/concordance/form.asp

台語文線上百科全書 [ Tâi-gú-bûn sòaⁿ-téng pek-kho-chôan-su ]

http://zh-min-nan.wikipedia.org/

白話字&萬國碼[ Pe̍h-ōe-jī kah Bān-kok-bé ] http://iug.csie.dahan.edu.tw/TG/Unicode/

本計畫統計資料 [ Pún kè-ōe thóng-kè chu-liāu ]

http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp

楊允言，2005，〈台語文書面語語料庫簡介及典藏相關建議〉，語言政策的多元文化思考系列研討會之三：台灣閩南語(Holo話)的活力與傳承，中央研究院語言所 [ Iûⁿ Ún-giân, 2005, 'Tâi-gú-bûn su-bīn-gú gú-liāu-khò kán-kài kah tián-chông siong-koan kiàn-gī', Gú-giân chèng-chhek ê to-gôan bûn-hòa su-khó hē-liát gián-thó-hōe chi saⁿ : Tâi-ôan Bān-lâm-gú (Holo ōe) ê o̍ah-le̍k kah thôan-sêng, Tiong-iong gián-kiù-īⁿ Gú-giân-só͘ ]

http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/iankang/guliaukhou/TGBguliaukhou.asp

楊允言等，2004，〈台語流失Kah變化e探討—以台語新約聖經做例〉，語言人權與語言復振學術研討會論文集p237-249，台東：台東大學語教系 [ Iûⁿ Ún-giân téng, 2004, 'Tâi-gú liû-sit kah piàn-hòa ê thàm-thó – í Tâi-gú sin-iok Sèng-keng chòe lē', Gú-giân jîn-kôan kah gú-giân ho̍k-chìn ha̍k-sut gián-thó-hōe lūn-bûn-chi̍p p237-249, Tâi-tang : Tâi-tang tāi-ha̍k Gú-kàu-hē ]

http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/gusupianhoa/nttu-gsph.asp

劉杰岳、楊允言，2002，〈白話字電腦文書處理ê研究〉，第四屆台灣語言及其教學國際學術研討會論文集p341-349，高雄，中山大學中文系 [ Lâu Kia̍t-ga̍k, Iûⁿ Ún-giân, 2002, 'Pe̍h-ōe-jī tiān-náu bûn-su chhú-lí ê gián-kiù', tē-sì-kài Tâi-ôan gú-giân kah kàu-ha̍k kok-chè ha̍k-sut gián-thó-hōe lūn-bûn-chi̍p p341-349, Ko-hiông, Tiong-san tāi-ha̍k Tiong-bûn-hē ]

http://iug.csie.dahan.edu.tw/iug/Ungian/Chokphin/Lunbun/POJtiannau/POJtiannau-TIongsan.htm

附錄[ Hù-liȯk ]

## 查詢[見笑] ê結果—台語文Concordance 程式

Beh chhē ê詞：見笑

| | | | |
|---|---|---|---|
| 1 | 一方面爲倒陽ê tāi-chì | 見笑 | tńg siū$^{n}$氣，一方面無歡喜 |
| 2 | chiah 知i mā 是寫詩--ê， | 見笑 | ，我soah m̄ bat--i。小 |
| 3 | khioh來做，i m̄ 知thang | 見笑 | ，koh講gah嘴角全波，台á kha |
| 4 | 慶--a hō A 猴罵一下起 | 見笑 | ，無意無意，跛--le 跛--le |
| 5 | bih、koh 鑽bih 無路ê | 見笑 | tāi——i 有影做錯--a，又koh |
| 6 | tī 電視頂hiān-sì ， | 見笑 | tāi--o͘h，我大人大種--a |
| 7 | 堪tit倒--來，我有做歹事連累父母 | 見笑 | 。」伊ê 老母無罵--伊，抱伊ê |
| 8 | 伊ê 頭殼疼痛伊講：「落難tú著 | 見笑 | 倒轉--來，母子受安慰kiám m̄ |
| 9 | 平安，咱兩個著求上帝贊伊好膽干證主無 | 見笑 | 。」(今印明白) |
| 10 | 也敢thí 嘴講beh去，真bē | 見笑 | 咧！」Hō in嬸佮兩個小妹責備 |
| 11 | 備伊講：「你這個若鬼也teh bē | 見笑 | 。」煞無意無意koh kiu入塗炭間 |
| 12 | 嬸佮寶珠、寶玉攏罵--伊：「你bē | 見笑 | 也敢出去穿。」欽差聽見講：「M̄-thang |
| 13 | 差你來，kám m̄是好？」那想那 | 見笑 | ，因爲從前伊kap伊ê 父母是拜上 |
| 14 | beh 展伊奇巧ê計策，後--來抵著 | 見笑 | ，反轉見出伊ê慇慢，án-ni號 |
| 15 | 大官攏無teh 想伊ê婬女怎樣ê | 見笑 | ，就伊ê子婿攏m̄ 知內中ê緣故 |

圖一 台語文語詞檢索查詢結果畫面

[ Tô it Tâi-gú-bûn gú-sû kiám-sek chhâ-sûn kiat-kó ōe-bīn]

## Hàn-lô Tâi-gú-bûn sû-pîn thóng-kè chu-liāu 漢羅台語文詞頻統計資料 2236

第 1 2 [3] 4 5 6 7 8 9 10 11 12 頁 後壁10頁，總共113頁, 2255筆資料

| 編號 | 語詞 | 頻率 | 比例 | 頻率合總 |
|---|---|---|---|---|
| 41 | 若 | 10,722 | 0.2647% | 29.9607% |
| 42 | 這 | 10,374 | 0.2561% | 30.2167% |
| 43 | thang | 9,876 | 0.2438% | 30.4605% |
| 44 | mā | 9,275 | 0.2289% | 30.6895% |
| 45 | 所 | 9,244 | 0.2282% | 30.9176% |
| 46 | 年 | 8,493 | 0.2096% | 31.1273% |
| 47 | 對 | 8,273 | 0.2042% | 31.3315% |
| 48 | 阮 | 8,059 | 0.1989% | 31.5304% |
| 49 | 好 | 8,000 | 0.1975% | 31.7279% |
| 50 | 因爲 | 7,886 | 0.1947% | 31.9226% |
| 51 | 想 | 7,883 | 0.1946% | 32.1171% |
| 52 | 就是 | 7,232 | 0.1785% | 32.2957% |
| 53 | tùi | 7,166 | 0.1769% | 32.4725% |
| 54 | ài | 7,061 | 0.1743% | 32.6468% |
| 55 | á | 6,804 | 0.1680% | 32.8148% |
| 56 | bē | 6,780 | 0.1674% | 32.9821% |
| 57 | 這個 | 6,713 | 0.1657% | 33.1478% |
| 58 | chiah | 6,632 | 0.1637% | 33.3116% |
| 59 | leh | 6,503 | 0.1605% | 33.4721% |
| 60 | lóng | 6,406 | 0.1581% | 33.6302% |

轉去首頁 上頭前 頭前10頁 頂一頁 下一頁 後壁10頁 上尾頁

圖二 詞頻統計查詢畫面 [ Tô jī Sû-pîn thóng-kè chhâ-sûn ōe-bīn]