附加詞類訊息的台語語詞搭配佇教學上的應用

Hù-ka sû-luī sìn-sit ê Tâi-gí gí-sû tah-phuè tī kàu-ha̍k siōng ê ìng-iōng

楊允言， 大漢技術學院、資訊工程系， 助理教授。

Iûnn Ún-giân， Tuā-hàn Ki-su̍t Ha̍k-īnn, Tsu-sìn Kang-tîng Hē， tsōo-lí kàu-siū。

劉杰岳， 拋荒台文工作室， 負責人。

Lâu Kia̍t-ga̍k， Pha-hng Tâi Bûn Kang-tsok-sik， hū-tsik-jîn。

陳鄭弘堯， 台語文工作者。

Tân-Tīnn Hông-giâu， tâi gí-bûn kang-tsok-tsiá。

陳柏中， 國立清華大學、物理系， 助理教授。

Tân Peh-tiong， Kok-li̍p Tshing-huâ Tāi-ha̍k, Bu̍t-lí Hē， tsōo-lí kàu-siū。

摘要

Tiah-iàu

語詞搭配(collocation)是學習語詞按怎使用真好的工具，

Gí-sû tah-phuè(collocation) sī ha̍k-si̍p gí-sû án-nuá sú-iōng tsin hó ê kang-khū,

這會當透過 對語料的互訊息(mutual information) 佮 相關度(correlation) 的統計來得著， 毋過嘛會產生袂少無必要的資料(noise)。

tse ē-tàng thàu-kuè tuì gí-liāu ê Hōo-sìn-sit (mutual information) kah Siong-kuan-tōo (correlation) ê thóng-kè lâi tit--tio̍h, m̄-koh mā ē sán-sing bē-tsió bô pit-iàu ê tsu-liāu (noise).

本文提出利用詞類的訊息來進行篩選 ê 方法， 自動共無必要的資料提掉， 提升自動台語語詞搭配的品質， 幫贊台語語詞的學習。

Pún-bûn thê-tshut lī-iōng sû-luī sìn-sit lâi tsìn-hîng thai-suán ê hong-huat, tsū-tōng kā bô pit-iàu ê tsu-liāu the̍h-tiāu, thê-sing tsū-tōng Tâi-gí gí-sû tah-phuè ê phín-tsit, pang-tsān Tâi-gí gí-sû ê ha̍k-si̍p.

實驗結果顯示， 利用詞類 ê 篩選， 動詞—名詞(V—N)詞組 比 形容詞—名詞(A—N) 詞組 會當得著較好的結果；

Si̍t-giām kiat-kó hián-sī, lī-iōng sû-luī ê thai-suán, tōng-sû — bîng-sû (V—N) sû-tsoo pí hîng-iông-sû — bîng-sû (A—N) sû-tsoo ē-tàng tit-tio̍h khah-hó ê kiat-kó;

查詢一个動詞正爿的名詞 抑是一个名詞倒爿的動詞的語詞搭配，所揣出來的資料嘛真有學習參考的價值。

tshâ-sûn tsi̍t-ê tōng-sû tsiànn-pîng ê bîng-sû iah-sī tsi̍t-ê bîng-sû tò-pîng ê tōng-sû ê gí-sû tah-phuè, sóo tshuē--tshut-lâi ê tsu-liāu mā tsin ū ha̍k-si̍p tsham-khó ê kè-ta̍t.

毋過若欲進一步提升自動語詞搭配的品質， 咱猶閣有誠濟空課愛做。

M̄-koh nā beh tsìn-tsi̍t-pōo thê-sing tsū-tōng gí-sû tah-phuè ê phín-tsit, lán iah-koh ū tsiânn-tsē khang-khuè ài tsò.

關鍵詞：台語文、語詞搭配、互訊息、相關度、語文教學

Kuan-kiān-sû: Tâi-gí-bûn, Gí-sû Tah-phuè, Hōo-sìn-sit, Siong-kuan-tōo, Gí-bûn Kàu-ha̍k

1. 踏話頭：語料庫、語詞搭配

It. Ta̍h uē-thâu: Gí-liāu-khòo, Gí-sû Tah-phuè

「搭配」這个詞， 台語閣會使講「參佮」、「照佮」、「sann 佮」、「配搭」、...等等， 毋過為著佮華語詞對應， 所以採用「搭配」。

"Tah-phuè" tsit-ê sû, Tâi-gí koh ē-sái kóng "tsham-kah", "tsiàu-kah", "sann-kah", "phuè-tah", ... tíng-tíng, m̄-koh uī-tio̍h kah Huâ-gí-sû tuì-ìng, sóo-í tshái-iōng "tah-phuè".

語料庫就是囥語言材料的倉庫， 伊會使講是人類所講、所寫的語言文字的一个樣本，

Gí-liāu-khòo tō-sī khn̂g gí-giân tsâi-liāu ê tshng-khòo, i ē-sái kóng sī jîn-luī sóo kóng, sóo siá ê gí-giân bûn-jī ê tsi̍t-ê iūnn-pún,

透過這个樣本， 咱通透過語料庫了解這个語言、文字的使用情形。

Thàu-kuè tsit ê iūnn-pún, lán thang thàu-kuè gí-liāu-khòo liáu-kái tsit-ê gí-giân, bûn-jī ê sú-iōng tsîng-hîng.

二十世紀 60 年代， 英文的語料庫開始建立。

Jī-tsa̍p sè-kí lio̍k-khòng nî-tāi, Ing-bûn ê gí-liāu-khòo khai-sí kiàn-li̍p.

90 年代， 華文的語料庫嘛開始佇台灣、中國等所在建立。

kiú-khòng nî-tāi, Huâ-bûn ê gí-liāu-khòo mā khai-sí tī Tâi-uân, Tiong-kok tíng sóo-tsāi kiàn-li̍p.

台文的起步較晏， 一直到這世紀(21 世紀)才開始。

Tâi-bûn ê khí-pōo khah uànn, it-ti̍t kàu tsit sè-kí (jī-tsa̍p-it sè-kí) tsiah khai-sí.

對強勢語言來講， 語料庫的建立代表這个語言的重要性， 因為建立了後會帶動這的語言的相關研究， 嘛宣示伊佇學術界的地位；

Tuì kiông-sè gí-giân lâi kóng, gí-liāu-khòo ê kiàn-li̍p tāi-piáu tsit-ê gí-giân ê tiōng-iàu-sìng, in-uī kiàn-li̍p liáu-āu ē tài-tōng tsit-ê gí-giân ê siong-kuan gián-kiù, mā suan-sī i tī ha̍k-su̍t-kài ê tē-uī;

對弱勢語言來講， 語料庫的建立是這个語言欲復興的一个重要基礎，因為利用著主導世界行向的電腦科技， 而且有機會予這个語言佮其它的強勢語言做連結。

tuì jio̍k-sè gí-giân lâi kóng, gí-liāu-khòo ê kiàn-li̍p sī tsit-ê gí-giân beh ho̍k-hing ê tsi̍t-ê tiōng-iàu ki-tshóo, in-uī lī-iōng-tio̍h tsú-tō sè-kài kiânn-hiòng ê tiān-nóo kho-ki, jî-tshiánn ū ki-huē hōo tsit-ê gí-giân kah kî-tha ê kiông-sè gí-giân tsò liân-kiat.

語料庫的應用研究有誠濟， 佇遮，阮針對語文教學的語詞搭配做較詳細的探討。

Gí-liāu-khòo ê ìng-iōng gián-kiù ū tsiânn tsē, tī tsia, gún tsiam-tuì gí-bûn kàu-ha̍k ê gí-sû tah-phuè tsò khah siông-suè ê thàm-thó.

語詞搭配(collocation) 就是「語詞用會當預測的方式組合做伙的方式」，嘛有人講是「仝一个文本內底，佇真近的所在共同出現的兩个抑是兩个以上的語詞」。

Gí-sû tah-phuè(collocation) tō-sī "Gí-sû iōng ē-tàng ū-tshik ê hong-sik tsoo-hp̍ tsò-hué ê hong-sik", mā ū-lâng kóng sī "Kāng tsi̍t-ê bûn-pún lāi-tué, ī tsin kūn ê sóo-tsāi kiōng-tông tshut-hiān ê nn̄g ê iah-sī nn̄g ê í-siōng ê gí-sû".

語詞搭配會使學習著一个語詞 ê 慣用方法，

Gí-sû tah-phuè ē-sái ha̍k-si̍p-tio̍h tsi̍t-ê gí-sû ê kuàn-iōng hong-huat,

比一个例，

pí tsi̍t-ê lē,

咱訂火車單， 閣去車頭付錢提著車單， 咱會使講「車單拍好矣」，

lán tīng hé-tshia-tuann, koh khì tshia-thâu hù-tsînn the̍h-tio̍h tshia-tuann, lán ē-sái kóng "tshia-tuann phah hó--ah",

「拍」佮「摃」的意思相仝， 毋過咱袂使講「車單摃好矣」，

"phah" kah "kòng" ê ì-sù sio-kâng, m̄-koh lán bē-sái kóng "tshia-tuann kòng hó--ah",

所以， 「拍」kah「車單」就有搭配 ê 關係， 其中一个詞若是用同義詞共換掉， 這个搭配關係無一定會繼續存在。

sóo-í, "phah" kah "tshia-tuann" tō ū tah-phuè ê kuan-hē, kî-tiong tsi̍t-ê sû nā-sī iōng tông-gī-sû kā uānn-tiāu, tsit-ê tah-phuè kuan-hē bô-it-tīng ē kè-si̍o̍k tsûn-tsāi.

兩个語詞欲按怎鬥陣使用， 一般 ê 辭典並無特別 kā 註明， 這對語言 ê 學習加減有阻礙。

Nn̄g-ê gí-sû beh án-nuá tàu-tīn sú-iōng, it-puann ê sû-tián pīng-bô ti̍k-pia̍t kā tsù-bîng, tse tuì gí-giân ê ha̍k-si̍p ke-kiám ū tsóo-gāi.

台灣人學英文， 硬記英文語詞，

Tâi-uân-lâng o̍h Ing-bûn, ngē-kì Ing-bûn gí-sû,

欲用英文佮人溝通 ê 時陣， 利用遮的語詞鬥出一句咱掠做會通 ê 英文句， 毋過講英語 ê 人可能捎無貓仔門，

beh iōng Ing-bûn kah lâng koo-thong ê sî-tsūn, lī-iōng tsia ê gí-sû tàu tshut tsi̍t-kù lán li̍ah-tsò ē thong ê Ing-bûn-kù, m̄-koh kóng Ing-gí ê lâng khó-lîng sa bô niau-á-mn̂g,

因為咱 ê 語詞搭配錯誤， 錯誤可能包括語法 kah 慣用語。

in-uī lán ê gí-sû tah-phuè tshò-gōo, tshò-gōo khó-lîng pau-kuah gí-huat kah kuàn-iōng-gí.

所以語詞搭配成做語言學習真重要 ê 一部分。

Sóo-í gí-sû tah-phuè tsiânn-tsò gí-giân ha̍k-si̍p tsin tiōng-iàu ê tsi̍t-pōo-hūn.

英文 ê 部分， 出版袂少語詞搭配辭典， 嘛是為著濟濟英語 ê 學習者。

Ing-bûn ê pōo-hūn, tshut-pán bē-tsió gí-sû tah-phuè sû-tián, mā-sī uī-tio̍h tsuē-tsuē Ing-gí ê ha̍k-si̍p-tsiá.

台語 ê 學習敢有需要注意語詞搭配？

Tâi-gí ê ha̍k-si̍p kám ū su-iàu tsù-ì gí-sû tah-phuè?

當然有需要， 包括對想 beh 學台語 ê 外國人， 嘛包括咱這代 kah 下一代。

Tong-jiân ū su-iàu, pau-kuah tuì siūnn-beh o̍h Tâi-gí ê guā-kok-lâng, mā pau-kuah lán tsit-tāi kah ē-tsi̍t-tāi.

受著強勢 ê 華語影響， 誠濟自認為會曉講台語 ê 人， 台語講起來 mā 是 li-li-lak-lak，

Siū-tio̍h kiông-sè ê Huâ-gí íng-hióng, tsiânn tsē tsū-jīm-uî ē-hiáu kóng Tâi-gí ê lâng, Tâi-gí kóng--khí-lâi li-li-lak-lak,

嘛因為用華語思考， 有 ê 人「一尾魚」煞變做「一條魚」。

mā in-uī iōng Huâ-gí su-khó, ū-ê-lâng "tsi̍t-bué hî" suah pìnn-tsò "tsi̍t-tiâu hî".

建立台語 ê 語詞搭配辭典是台語教學真重要 ê 一部份，

Kiàn-lip̍ Tâi-gí ê gí-sû tah-phuè sû-tián sī Tâi-gí kàu-ha̍k tsin tiōng-iàu ê tsi̍t-pōo-hūn,

可惜目前猶無， 這是語言學界 ê 責任。

khó-sioh bo̍k-tsiân iah bô, tse-sī gí-giân-ha̍k-kài ê tsik-jīm.

本文是徛佇資訊科學 ê 角度， 提出現時另外一个通好運用 ê 方式： 利用台語文語料庫這个資源來建立台語語詞搭配 ê 資料。

Pún-bûn sī khiā tī tsu-sìn kho-ha̍k ê kak-tōo, thê-tshut hiān-sî līng-guā tsi̍t-ê thang-hó ūn-iōng ê hong-sik: Lī-iōng Tâi-gí-bûn gí-liāu-khòo tsit-ê tsu-guân lâi kiàn-lip̍ Tâi-gí gí-sû tah-phuè ê tsu-liāu.

2. 自動語詞搭配做法

Jī. Tsū-tōng Gí-sû Tah-phuè Tsò-huat

根據語料庫， 欲共共同出現而且互相關係密切 ê 語詞揣出來， 是利用統計 ê 方法。

Kin-kù gí-liāu-khòo , beh kā kiōng-tông tshut-hiān jî-tshiánn hōo-siong kuan-hē bi̍t-tshiat ê gí-sû tshuē--tshut-lâi, sī lī-iōng thóng-kè ê hong-huat.

有兩个公式會使做，

Ū nn̄g-ê kong-sik ē-sái tsò,

第一个是互訊息(mutual information，以下簡稱 MI)， 設使 A、B 是語詞， 這兩个語詞 ê MI ê 公式是寫做按呢。

tē-it-ê sī hōo-sìn-sit (mutual information, í-hā kán-tshing MI), siat-sú A, B sī gí-sû, tsit nn̄g-ê gí-sû ê MI ê kong-sik sī siá-tsò án-ne.

其中， 「P of A」是語詞 A ê 機率，

Kî-tiong, "P of A" sī gí-sû A ê ki-lu̍t,

設使 A 攏總出現「Frequence of A」擺，所有 ê 語料攏總有 N 个詞 (word tokens)，「P of A」等於 N 分之「Frequence of A」

siat-sú A lóng-tsóng tshut-hiān "Frequence of A" pái, sóo-ū ê gí-liāu lóng-tsóng ū N-ê sû (word tokens), "P of A" tíng-î N hun tsi "Frequence of A".

仝款，「P of B」是語詞 B ê 機率，「P of AB」是詞組 AB ê 機率。

Kāng-khuán, "P of B" sī gí-sû B ê ki-lu̍t, "P of AB" sī sû-tsoo AB ê ki-lu̍t.

咱分三種情形來討論：

Lán hun sann-tsióng tsîng-hîng lâi thó-lūn:

(a) 設使 A 後壁干焦會出現 B，B 頭前干焦會出現 A，這兩个詞一定成對出現，

(a) Siat-sú A āu-piah kan-na ē tshut-hiān B, B thâu-tsîng kan-na ē tshut-hiān A, tsit nn̄g-ê sû it-tīng sîng-tuì tshut-hiān,

這款情形下，「P of A」近倚「P of B」近倚「P of AB」，「MI of AB」近倚「負 log P of AB」，是一个較大 ê 正數；

tsit-khuán tsîng-hîng hā, "P of A" kīn-uá "P of B" kīn-uá "P of AB", "MI of AB" kīn-uá "hū log P of AB", sī tsi̍t-ê khah tuā ê tsiànn-sòo;

若是 AB 定定成對出現，毋過 A 後壁會接 B 以外其它 ê 詞，B 頭前嘛可能會接 A 以外其它 ê 詞，

nā-sī AB tiānn-tiānn sîng-tuì tshut-hiān, m̄-koh A āu-piah ē tsiap B í-guā ê sû, B thâu-tsîng mā khó-lîng ē tsiap A í-guā ê sû,

這个時陣，「MI of AB」可能嘛是正數，毋過會較細；

tsit-ê sî-tsūn, "MI of AB" khó-lîng mā-sī tsiànn-sòo, m̄-koh ē khah suè;

(b) 設使 A 佮 B 無關係（獨立事件），

(b) Siat-sú A kah B bô kuan-hē (to̍k-li̍p sū-kiānn),

「P of AB」近倚"P of A"乘"P of B"， 所以「MI of AB」近倚 0；

"P of AB" kīn-uá "P of A" sîng "P of B" , sóo-í "MI of AB" kīn-uá 0;

(c) 設使語詞 A 出現致使 B 較袂出現，

(c) Siat-sú gí-sû A tshut-hiān tì-sú B khah buē tshut-hiān,

「P of AB」 遠小於 "P of A" 乘 "P of B" , 所以「MI of AB」是負數。

"P of AB" uán sió î "P of A" sîng "P of B" , sóo-í "MI of AB" sī hū-sòo .

基本上， 語料愈大， 統計 ê 結果愈有代表性。

Ki-pún siōng, gí-liāu jú tuā, thóng-kè ê kiat-kó jú ū tāi-piáu-sìng.

第二个是相關度(correlation，以下簡稱 CR)， 這嘛是一个統計 ê 公式：

Tē-jī-ê sī siong-kuan-tōo (correlation, í-hā kán-tshing CR), tse mā-sī tsi̍t-ê thóng-kè ê kong-sik:

簡單講， 欲算「CR of AB」，就共語料內底 ê 詞組分做四个部分， 透過頂面 ê 公式來計算。

Kán-tan kóng, beh sǹg "CR of AB", tō kā gí-liāu lāi-té ê sû-tsoo hun-tsò sì ê pōo-hūn, thàu-kè tíng-bīn ê kong-sik lai kè-sǹg.

算出來 ê 數字攏是正數，

Sǹg--tshut-lâi ê sòo-jī lóng-sī tsiànn-sòo,

AB 兩个語詞若定定做陣出現， 「CR of AB」可能超過 10,000， 甚至超過 100,000。

AB nn̄g-ê gí-sû nā tiānn-tiānn tsò-tīn tshut-hiān, "CR of AB" khó-lîng tshiau-kuè tsi̍t-bān, sīm-tsì tshiau-kuè tsa̍p-bān.

愛注意 ê 是， 有時仔若拍字錯誤， 因為錯字 ê 詞頻真低， 會致使 MI 佮 CR 的分數變足懸。

Ài tsù-ì ê sī, ū-sî-á nā phah-jī tshò-gōo, in-uī tshò-jī ê sû-pîn tsin kē, ē tì-sú MI kah CR ê hun-sòo piàn tsiok kuân.

為著避免這个情形， 會使設詞頻 ê 限制，

Uī-tio̍h phiah-bián tsit-ê tsîng-hîng, koh ài siat sû-pîn ê hān-tsè,

就是講，欲計算 MI(AB)， 會使限制詞組 AB 的頻率超過一个數目（可比出現 10 擺以上）， 才來計算， 若無就共提掉。

tō-sī kóng, beh kè-sǹg MI(AB), ē-sái hān-tsè sû-tsoo AB ê pîn-lu̍t tshiau-kè tsi̍t-ê sòo-bo̍k (khó-pí tshut-hiān tsa̍p-pái í-siōng), tsiah lâi kè-sǹg, nā-bô tō kā thé̍h-tiāu.

3. 利用詞類訊息提升品質

Sann. Lī-iōng sû-luī sìn-sit thê-sing phín-tsit

用統計 ê 方式來揣語詞搭配， 佮語言專家利用伊專業智識揣出來 ê， 當然加減會無仝款。

Iōng thóng-kè ê hong-sik lâi tshuē gí-sû tah-phuè, kah gí-giân tsuan-ka lī-iōng i tsuan-gia̍p tì-sik tshuē--tshut-lâi--ê, tong-jiân ke-kiám ē bô kāng-khuán.

用統計方式， 咱有簡化， 限定佇相倚 ê 兩个語詞，

Iōng thóng-kè hong-sik, lán ū kán-huà, hān-tiānn tī sio-uá ê nn̄g-ê gí-sû,

實際上有 ê 互相搭配 ê 語詞並無相倚， 「拍車單」有可能講做「拍兩張車單」，

si̍t-tsè siōng ū-ê hōo-siong tah-phuè ê gí-sû pīng-bô sio-uá, "phah tshia-tuann" ū khó-lîng kóng tsò "phah nn̄g tiunn tshia-tuann",

嘛有可能是三个語詞互相搭配， 可比「除了……以外，猶閣 ……」。

mā ū khó-lîng sī sann-ê gí-sû hōo-siong tah-phuè, khó-pí " tû-liáu ... í-guā, iah-koh ...".

當然咱會使修正統計方式， 共無相倚毋過離無遠 ê 語詞嘛考慮入來， 抑是修改公式來計算三个語詞 ê 互相關係。

Tong-jiân lán ē-sái siu-tsìng thóng-kè hong-sik, kā bô sio-uá m̄-koh lī bô hn̄g ê gí-sû mā khó-lū--ji̍p-lâi, iah-sī siu-kái kong-sik lâi kè-sǹg sann-ê gí-sû ê hōo-siong kuan-hē.

總是按呢做加足厚工， 增加誠濟計算量， 改進 ê 成果嘛真有限。

Tsóng--sī an-ne tsò ke tsiok kāu-kang, tsing-ka tsiânn tsē kè-sǹg-liōng, kái-tsìn ê sîng-kó mā tsin iú-hān.

用統計方式做， 並毋是欲取代專家，

Iōng thóng-kè ê hong-sik tsò, pīng m̄-sī beh tshú-tāi tsuan-ka,

毋過， 咱利用電腦會當處理大量資料 ê 特性， 提供電腦計算出來 ê 結果， 對學習者佮專家攏有參考作用，

m̄-koh, lán lī-iōng tiān-nóo ē-tàng tshú-lí tuā-liōng tsu-liāu ê ti̍k-sìng, thê-kiong tiān-nóo kè-sǹg--tshut-lâi ê kiat-kó, tuì ha̍k-si̍p-tsiá kah tsuan-ka lóng ū tsham-khó tsok-iōng,

專家所列出來 ê 語詞搭配， 有 ê 因為實際上較少使用， 致使統計結果並無表現出來；

tsuan-ka sóo li̍at--tshut-lâi ê gí-sû tah-phuè, ū-ê in-uī si̍t-tsè siōng khah tsió sú-iōng, tì-sú thóng-kè kiat-kó pīng-bô piáu-hiān--tshut-lâi;

統計結果有， 毋過專家無列出來 ê， 有可能並無什麼意義，

thóng-kè kiat-kó ū, m̄-koh tsuan-ka bô li̍at--tshut-lâi ê, ū khó-lîng pīng bô siánn-mé ì-gī,

總是伊佇實際使用上有較懸的牽連， 嘛提供專家參考佮研究 ê 方向。

tsóng--sī i tī si̍t-tsè sú-iōng siōng ū khah kuân ê khan-liân, mā thê-kiong tsuan-ka tsham-khó kah gián-kiù ê hong-hiòng.

另外， 根據英文 ê 語料庫做出來 ê 語詞搭配， 介詞(in, at,...)、指示代名詞(that)、冠詞(a, the, ...) 因為頻率真懸， 遮 ê 語詞 ê 語詞搭配("of the", "that the", "to be", "in a", ...) 嘛攏排佇誠頭前。

Līng-guā, kun-kù Ing-bûn ê gí-liāu-khòo tsò--tshut-lâi ê gí-sû tah-phuè, kài-sû (in, at,...), tsí-sī tāi-bîng-sû (that), kuàn-sû (a, the,...) in-uī pîn-lu̍t tsin kuân, tsia ê gí-sû ê gí-sû tah-phuè ("of the", "that the", "to be", "in a", ...) mā lóng pâi tī tsiânn thâu-tsîng.

對遮咱發現，詞類會使提供咱一个真好 ê 線索，

Uì tsia lán huat-hiān, sû-luī ē-sái thê-kiong lán ts̍it-ê tsiânn hó ê suànn-soh,

假使咱會當利用詞類 ê 訊息， 對統計出來 ê 語詞搭配結果做分類整理，對學習者來講應該閣較有幫贊。

ká-sú lán ē-tàng lī-iōng sû-pîn ê sìn-sit, tuì thóng-kè--tshut-lâi ê gí-sû tah-phuè kiat-kó tsò hun-luī tsíng-lí, tuì ha̍k-si̍p-tsiá lâi kóng ìng-kai koh-khah ū pang-tsān.

有詞類訊息 ê 台語辭典， 包括日本天理大學 ê《現代閩南語辭典》(村上嘉英編，1981 年)、 台北中華語文研習所 ê《台英辭典》(Embree 編，1984)、...等， 可惜並無電子檔案通好運用。

Ū sû-luī sìn-sit ê Tâi-gí sû-tián, pau-kuah Ji̍t-pún Thian-lí tāi-ha̍k ê "Hiān-tāi Bân-lâm-gí sû-tián"(Murakami Yoshihide pian, 1981 nî), Tâi-pak Tiong-huâ gí-bûn gián-si̍p-sóo ê "Tâi-ing sû-tián" (Embree pian, 1984), ... tíng, khó-sioh pīng-bô tiān-tsú tóng-àn thang-hó ūn-iōng.

所以阮只好踅一輾， 以線頂台文華文辭典做基礎， 將台文 ê 詞條透過華文對譯， 對到中研院詞庫小組整理 ê 八萬目華文詞條， 對華文詞條得著詞類訊息。

Sóo-í gún tsí-hó sio̍h ts̍it-lìn, í suànn-tíng Tâi-bûn Huâ-bûn sû-tián tsò ki-tshóo, tsiong Tâi-bûn ê sû-tiâu thàu-kuè Huâ-bûn tuì-i̍k, tuì kàu Tiong-gián-īnn Sû-khòo sió-tsoo tsíng-lí ê peh-bān-bo̍k Huâ-bûn sû-tiâu, uì Huâ-bûn sû-tiâu tit-tio̍h sû-luī sìn-sit.

詞庫小組 ê 詞類分較幼， 有 46 个詞類， 包括動詞、名詞等攏繼續有分細類落去，

Sû-khòo sió-tsoo ê sû-luī pun khah iù, ū 46-ê sû-luī, pau-kuah tōng-sû, bîng-sû tíng lóng kè-sio̍k ū pun sè-luī lo̍h--khì,

毋過計算語詞搭配，咱無需要分遐爾幼， 用大類來分就好， 包括名詞、動詞、形容詞、 副詞、連接詞、介詞、語氣詞、感嘆詞等等。

m̄-koh kè-sn̂g gí-sû tah-phuè, lán bô su-iàu pun hiah-nī iù, iōng tuā-luī lâi pun tō hó, pau-kuah bîng-sû, tōng-sû, hîng-iông-sû, hù-sû, liân-tsiap-sû, kài-sû, gí-khì-sû, kám-thàn-sû tíng-tíng.

4. 實驗步數

Si̍t-giām pōo-sòo

下面是本實驗 ê 做法：

Ē-bīn sī pún si̍t-giām ê tsò-huat:

(a) 資料包括台語文語料、 台文華文線頂辭典（以下簡稱台華辭典）佮中研院詞庫小組八萬目詞(華文)。

(a) Tsu-liāu pau-kuah Tâi-gí-bûn gí-liāu, Tâi-bûn Huâ-bûn suànn-tíng sû-tián (í-hā kán-tshing Tâi-huâ sû-tián) kah Tiong-gián-īnn Sû-khòo-sió-tsoo pueh-bān-ba̍k sû (Huâ-bûn).

其中，台語文語料庫 ê 來源是台語文界 ê 朋友，楊允言負責整理，

Kî-tiong, Tâi-gí-bûn gí-liāu-khòo ê lâi-guân sī Tâi-gí-bûn-kài ê pîng-iú, Iûnn Ún-giân hū-tsik tsíng-lí,

包括全羅 3,462,367 个音節 佮漢羅 5,568,057 个音節；

pau-kuah tsuân-lô sann-pah sì-tsa̍p-la̍k-bān nn̄g-tshing sann-pah la̍k-tsa̍p-tshit ê im-tsiat kah Hàn-lô gōo-pah gōo-tsa̍p-la̍k-bān peh-tshing khòng gōo-tsa̍p-tshit ê im-tsiat;

台華辭典主要貢獻者是鄭良偉，閣有台語文界朋友鬥補充詞條， 目前有 6 萬 2 千外个詞條，每一个詞條包括漢羅、全羅、華文對譯三个欄位。

Tâi-huâ sû-tián tsú-iàu kòng-hiàn-tsiá sī Tēnn Liông-uí, koh ū Tâi-gí-bûn-kài pîng-iú tàu póo-tshiong sû-tiâu, bo̍k-tsiân ū la̍k-bān nn̄g-tshing-guā-ê sû-tiâu, muí si̍t-ê sû-tiâu pau-kuah Hàn-lô, tsuân-lô, Huâ-bûn tuì-i̍k sann-ê nuâ-uī.

(b) 全羅 ê 台語文語料無需要經過斷詞，毋過漢羅 ê 需要，

(b) Tsuân-lô ê Tâi-gí-bûn gí-liāu bô su-iàu king-kè tn̂g-sû, m̄-koh Hàn-lô--ê su-iàu,

斷詞 ê 做法是根據台華辭典，採用「倒頭上大比對」(backward maximal matching, BMM)演算法。

tn̂g-sû ê tsò-huat sī kun-kù Tâi-huâ sû-tián, tshái-iōng "tò-thâu siōng-tuā pí-tuì" (backward maximal matching, BMM) ián-suàn-huat.

語詞 ê 數量， 全羅有 2,436,599 个， 漢羅有 4,051,195 个。

Gí-sû ê sòo-liōng, tsuân-lô u nn̄g-pah sì-tsa̍p-sann-bān la̍k-tshing gōo-pah káu-tsa̍p-káu ê, Hàn-lô ū sì-pah khòng gōo-bān tsi̍t-tshing tsi̍t-pah káu-tsa̍p-gōo ê.

(c) 每一个語詞，透過台華辭典 ê 華文對譯， 對到中研院詞庫小組八萬目詞 ê 詞條， 揣出詞類， 詞類可能有幾 lō 个， 暫時無做處理。

(c) Muí tsi̍t-ê gí-sû, thàu-kè Tâi-huâ sû-tián ê Huâ-bûn tuì-i̍k, tuì kàu Tiong-gián-īnn sû-khòo sió-tsoo peh-bān-ba̍k sû ê sû-tiâu, tshuē-tshut sû-luī, sû-luī khó-lîng ū kuí-lō ê, tsiām-sî bô tsò tshú-lí.

因為無處理含糊性問題， 所以會影響著實驗結果 ê 品質。

In-uī bô tshú-lí hâm-hôo-sìng ê būn-tuê, ē íng-hióng-tio̍h si̍t-giām kiat-kó ê phín-tsit.

另外，本實驗採用簡化 ê 詞類， 干焦分大類：動詞、名詞、形容詞、...等等，

Līng-guā, pún si̍t-giām tshái-iōng kán-huà ê sû-luī, kan-na pun tuā-luī: tōng-sû, bîng-sû, hîng-iông-sû, ... tíng-tíng,

其中，動詞 VH 小類是「狀態不及物述詞」， 親像「浪漫」、「特別」、「辛苦」、「豐富」、「心酸」、「感動」， 阮共改做形容詞。

kî-tiong, tōng-sû VH sió-luī sī "Tsōng-thài put-ki̍p-bu̍t su̍t-sû", tshin-tshiunn "lōng-bān", "ti̍k-pia̍t", "sin-khóo", "hong-hù", "sim-sng", "kám-tōng", gún kā kái-tsò hîng-iông-sû.

按呢修改，有 ê 無問題，部分可能有問題。

An-ne siu-kái, ū-ê bô būn-tuê, pōo-hūn khó-lîng ū būn-tuê.

(d) 計算兩个相倚 ê 語詞的 MI 佮 CR，閣經過詞類 ê 篩選。

(d) Kè-sǹg nn̄g-ê sio-uá ê gí-sû ê MI kah CR, koh king-kè sû-luī ê thai-suán.

因為資料量誠大， 這擺實驗干焦揣 A-N（形容詞——名詞）、V-N（動詞——名詞）詞組。

In-uī tsu-liāu-liōng tsiânn tuā, tsit-pái si̍t-giām kan-na tshuē A-N (hîng-iông-sû —— bîng-sû), V-N (tōng-sû —— bîng-sû) sû-tsoo.

5. 實驗結果

Gōo. Si̍t-giām kiat-kó

本實驗 ê 程式用 Java 寫，共結果輸出到資料庫， 資料量誠大。

Pún si̍t-giām ê thîng-sik iōng Java siá, kā kiat-kó su-tshut kàu tsu-liāu-khòo, tsu-liāu-liōng tsiânn tuā.

本節利用表格佮文字說明來展示實驗成果。

Pún-tsiat lī-iōng pió-keh kah bûn-jī suat-bîng lâi tián-sī si̍t-giām sîng-kó.

5.1 MI 佮 CR 共同詞組數量

Gōo tiám it MI kah CR kiōng-tông sû-tsoo sòo-liōng

表一 共 MI 佮 CR 頭前 30 / 100 / 500 / 1,000 / 3,000 / 5,000 个共同詞組 ê 數量列出來， 漢羅佮全羅分開算， 待討論：(這个詞組，佇 MI 佮 CR 兩个統計表攏有出現)

Pió-it kā MI kah CR thâu-tsîng 30 / 100 / 500 / 1,000 / 3,000 / 5,000 ê kiōng-tông sû-tsoo ê sòo-liōng li̍at--tshut-lâi, Hàn-lô kah tsuân-lô hun-khui sǹg, Thāi thó-lūn：(tsit-ê sû-tsoo , tī MI kah CR nn̄g-ê thóng-kè-pió lóng-ū tshut-hiān)

其中，漢羅 ê 部分，詞組至少出現 10 攏， 全羅 ê 部分，詞組至少出現 5 攏:

kî-tiong, Hàn-lô ê pōo-hūn, sû-tsoo tsì-tsió tshut-hiān 10-pái, tsuân-lô ê pōo-hūn, sû-tsoo tsì-tsió tshut-hiān 5-pái:

表一 互訊息 kah 相關度 共同詞組數量比例

Piáu tsi̍t hōo-sìn-sit kah siang-kuan-tōo kiōng-tông sû-tsoo sòo-liōng pí-lē

對表一看起來， 漢羅 ê 部分，MI 佮 CR 算出來 ê 結果較一致，

Uì pió-it khuànn--khí-lâi, Hàn-lô ê pōo-hūn, MI kah CR sǹg--tshut-lâi ê kiat-kó khah it-tì,

這應該是因為， 漢羅 ê 語料量較濟，統計出來 ê 結果較有可信度。

tse ìng-kai sī in-uī, Hàn-lô ê gí-liāu-liōng khah tsē, thóng-kè--tshut-lâi ê kiat-kó khah ū khó-sìn-tōo.

5.2 A-N V-N 詞組

Gōo-tiám jī A-N V-N sû-tsoo

本實驗用詞類是 A-N 佮 V-N 做篩選，

Pún si̍t-giām iōng sû-luī sī A-N kah V-N tsò thai-suán,

只要這个詞組有可能是 A-N 抑是 V-N 就共掠出來，

tsí-iàu tsit-ê sû-tsoo ū khó-lîng sī A-N iah-sī V-N tō kā li̍ah--tshut-lâi,

毋過因為雜訊袂少， 表 2 佮 表 3 是對漢羅佮全羅， MI ê 統計表對懸到低分別用人工揣出是 A-N 抑是 V-N 的詞組：

m̄-koh in-uī tsa̍p-sìn bē tsió, pió-jī kah pió-sann sī uì hàn-lô kah tsuân-lô, MI ê thóng-kè-pió uì kuân kàu kē hun-piàt iōng lâng-kang tshuē-tshut sī A-N iah-sī V-N ê sû-tsoo:

A-N 詞組較僫揣， 揀出來 ê 有 ê 看起來誠勉強。

A-N sû-tsoo khah oh tshuē, kíng--tshut-lâi--ê ū-ê khuànn--khí-lâi tsiânn bián-kióng.

親像漢羅部分 ê「旺梨」、「ām 瓜」， 自本應該是一个詞，

Tshin-tshiūnn Hàn-lô pōo-hūn ê " ōng lâi", "ām kue", tsū-pún ìng-kai sī tsi̍t-ê sû,

毋過台華辭典內底有「王梨」無「旺梨」， 有「醃瓜」無「ām 瓜」，致使這幾个詞予斷詞系統切做兩个詞，

m̄-koh Tâi-huâ sû-tián lāi-tué ū "ông-lâi" bô "ōng-lâi", ū "am-kue" bô "ām-kue", tì-sú tsit kuí-ê sû hōo tn̄g-sû hē-thóng tshiat-tsò nn̄g-ê sû,

煞閣因為切開 ê「這兩个詞」ê 結合性懸， 佇 A-N 詞組表現出來。

suah koh in-uī tshiat-khui ê " tsit nn̄g-ê sû" ê kiat-ha̍p-sìng kuân, tī A-N sû-tsoo piáu-hiān--tshut-lâi .

對這个角度看， 自動語詞搭配統計表， 對辭典詞條 ê 收錄，會當提供一寡建議。

Uì tsit-ê kak-tōo khuànn, tsū-tōng gí-sû tah-phuè thóng-kè-pió, tuì sû-tián sû-tiâu ê siu-lo̍k, ē-tàng thê-kiong tsi̍t-kuá kiàn-gī.

V-N 詞組 ê 部分，全羅 ê 部分真好， 對 MI 統計表頭前 64 个詞組內底就通揀出真正符合 V-N ê 詞組， 毋過漢羅 ê 部分就誠 bái。

V-N sû-tsoo ê pōo-hūn, tsuân-lô ê pōo-hūn tsin hó, uì MI thóng-kè-pió thâu-tsîng 64-ê sû-tsoo lāi-té tō thang kíng-tshut tsin-tsiànn hù-ha̍p V-N ê sû-tsoo, m̄-koh Hàn-lô ê pōo-hūn tō tsiânn bái.

主要 ê 原因是斷詞， 因為語料內底 ê 寫法佮辭典無一致。

Tsú-iàu ê guân-in sī tn̄g-sû, in-uī gí-liāu lāi-té ê siá-huat kah sû-tián bô it-tì.

5.3 名詞 ê 動詞搭配佮動詞 ê 名詞搭配

Gōo-tiám sann Bêng-sû ê tōng-sû tah-phuè kah tōng-sû ê bîng-sû tah-phuè

表四 查詢「V-舌」

Pió-4 Tshâ-sûn "V-tsi̍h"

「MI 小於零」ê 部分無列出來

"MI sió-î ling5" ê pōo-hūn bô li̍at--tshut-lâi

表五 查詢「損-N」

Pió-5 Tshâ-sûn "kòng-N"

假使咱欲知影一个名詞頭前會使用 siánn-mé 動詞（倒爿搭配）， 抑是一个動詞後壁會用使用 siánn-mé 名詞（正爿搭配）， 用 MI 抑是 CR 來查詢，查出來 ê 結果真有參考價值，

Ká-sú lán beh tsai-iánn tsi̍t-ê bîng-sû thâu-tsîng ē-sái iōng siánn-mé tōng-sû (tò-pîng tah-phuè), iah-sī tsi̍t-ê tōng-sû āu-piah ē-sái iōng siánn-mé bîng-sû (tsiànn-pîng tah-phuè), iōng MI iah-sī CR lâi tshâ-sûn, tshâ--tshut-lâi ê kiat-kó tsin ū tsham-khó kè-ta̍t,

舉一个例， 表四用「V-舌」查詢，表五用「kòng-N」查詢:

kú tsi̍t ê lē, pió-sì iōng " V-tsi̍h" tshâ-sûn, pió-gōo iōng "kòng-N" tshâ-sûn:

這个查詢方式， 對語言學習應該有真大 ê 幫贊。

Tsit-ê tshâ-sûn hong-sik, tuì gí-giân ha̍k-si̍p ìng-kai ū tsin tuā ê pang-tsān.

5.4 主要問題

5.4 Tsú-iàu būn-tuê

實驗 ê 結果並無原來按算 ê 遐爾好， 下面分析問題出佇佗位：

Si̍t-giām ê kiat-kó pīng bô guân-lâi àn-sǹg ê hiah-nī hó, ē-bīn hun-sik būn-tê tshut tī tó-uī:

(1) 漢羅用字無一致， 致使斷詞效果無好：

(1) Hàn-lô iōng-jī bô it-tì, tì-sú tn̄g-sû hāu-kó bô hó:

V-N ê 部分，漢羅 ê 結果真 bái，

V-N ê pōo-hūn, Hàn-lô ê kiat-kó tsin bái,

人工 tih 看遮的詞組 ê 時， 發現有誠濟詞組是斷詞錯誤 ê，

lâng-kang tih khuànn tsia-ê sû-tsoo ê sî, huat-hiān ū tsiânn-tsē sû-tsoo sī tn̄g-sû tshò-gōo--ê,

有 ê 是台華辭典內無這个詞，

ū-ê sī Tâi-huâ sû-tián lāi bô tsit-ê sû,

有 ê 是語料 ê 漢羅寫法 kah 台華辭典內 ê 無仝， 實際上， 語料內底 ê 寫法就無一致，

ū-ê sī gí-liāu ê Hàn-lô siá-huat kah Tâi-huâ sû-tián lāi ê bô-kâng, si̍t-tsè siōng, gí-liāu lāi-té ê siá-huat tō bô it-tì,

辭典無 ê 詞， 若 beh 改進， 包括加詞、專有名詞辨識、定量詞處理等等， 寫法無一致 ê 問題 tō 真僫解決， kā 所有 ê 寫法 lóng kā lok tī 辭典內底並 m̄是好 ê 解決方式；

sû-tián bô ê sû, nā beh kái-tsìn, pau-kuah ka sû, tsuan-iú-bîng-sû piān-sik, tīng-niû-sû tshú-lí tíng-tíng, siá-huat bô it-tì ê būn-tuê tō tsin oh kái-kuat, kā sóo-ū ê siá-huat lóng kā lok tī sû-tián lāi-tué pīng m̄-sī hó ê kái-kuat hong-sik;

(2) 詞類問題：

(2) Sû-luī būn-tuê:

語言翻譯是多對多， 咱透過華語對譯詞掠詞類， 這个過程就有可能會 tiòng 一寡無必要 ê 詞類出來；

Gí-giân huan-i̍k sī to tuì to, lán thàu-kè Huâ-gí tuì-i̍k-sû lia̍h sû-luī, tsit-ê kè-thîng tō ū khó-lîng ē tiòng tsi̍t-kuá bô pit-iàu ê sû-luī tshut-lâi;

另外， 一个詞有至少兩个詞類 mā 是真普遍 ê 情形， 因為無處理， 造成雜訊誠濟；

līng-guā, tsi̍t-ê sû ū tsì-tsió nn̄g-ê sû-luī mā-sī tsin phóo-phiàn ê tsîng-hîng, in-uī bô tshú-lí, tsō-sîng tsa̍p-sìn tsiânn tsē;

建立台語詞 ê 詞類，佮確認正確詞類， 這兩件代誌攏是大工程， 毋過早慢愛做，咱才有法度閣進一步。

kiàn-li̍p Tâi-gí-sû ê sû-luī, kah khak-jīm tsìng-khak sû-luī, tsit nn̄g-kiānn tāi-tsì lóng-sī tuā-kang-thîng, m̄-koh tsá-bān ài tsò, lán tsiah ū-huat-tōo koh tsìn-tsi̍t-pōo .

6. 未來方向

Li̍ok Bī-lâi hong-hiòng

本實驗雖然有一寡成果， 總是，值得拍拚 ê 所在猶閣誠濟，包括：

Pún si̍t-giām sui-jiân ū tsi̍t-kuá sîng-kó, tsóng--sī, ta̍t-tit phah-piànn ê sóo-tsāi iah-koh tsiânn tsē, pau-kuah:

(a) 開發線頂台語語詞搭配查詢系統， 幫贊台語相關研究 ê 進行；

(a) Khai-huat suànn-tíng Tâi-gí gí-sû tah-phuè tshâ-sûn hē-thóng, pang-tsān Tâi-gí siong-kuan gián-kiù ê tsìn-hîng;

(b) 除了 A-N V-N 詞組以外， 閣整理其它詞組， 親像 P-N(介詞—名詞)、 N-V(名詞—動詞)、 D-A(副詞—形容詞)、 D-V(副詞—動詞)、 V-R(動詞—代名詞)、 ……等；

(b) Tû-liáu A-N V-N sû-tsoo í-guā, koh tsíng-lí kî-tha sû-tsoo, tsin-tshiūnn P-N (kài-sû — bîng-sû), N-V (bîng-sû — tōng-sû), D-A (hù-sû — hîng-iông-sû), D-V (hù-sû — tōng-sû), V-R (tōng-sû — tāi-bîng-sû), … tíng;

(c) 維護台華辭典， 共 MI 抑是 CR 分數較懸 ê 兩个單音節詞 ê 語詞搭配 掠出來做人工檢查， 考慮將遮 ê 語詞加入辭典；

(c) Uî-hōo Tâi-huâ sû-tián, kā MI iah-sī CR hun-sòo khah kuân ê nn̄g-ê tan-im-tsiat-sû ê gí-sû tah-phuè lia̍h tshut-lâi tsò jîn-kang kiám-tsa, khó-lū tsiong tsia ê gí-sû ka-ji̍p sû-tián;

(d) 改進斷詞系統， 利用 BMM 演算法結果做訓練資料重做， 對有至少兩種斷法 ê 文句， 用統計方法提懸正確率， 猶閣有專有名詞、 定量詞處理；

(d) Kái-tsìn tn̄g-sû hē-thóng, lī-iōng BMM ián-suàn-huat kiat-kó tsò hùn-liān tsu-liāu tîng tsò, tuì ū tsì-tsió nn̄g-tsióng tn̄g-huat ê bûn-kù, iōng thóng-kè hong-huat thê-kuân tsìng-khak-lu̍t, iah-koh ū tsuan-iú bîng-sû, tīng-niû-sû tshú-lí;

(e) 整理台語詞類， 總是這是大工程， 嘛需要詳細 ê 規劃；

(e) Tsíng-lí Tâi-gí sû-luī, tsóng--sī tse-sī tuā kang-thîng, mā su-iàu siông-sè ê kui-uē;

(f) 利用統計方法自動標記詞類， 解決含糊性問題， 嘛進一步處理較幼 ê 詞類；

(f) Lī-iōng thóng-kè hong-huat tsū-tōng piau-kì sû-luī, kái-kuat hâm-hôo-sìng būn-tê, mā tsìn-ts̍it-pōo tshú-lí khah iù ê sû-luī;

(g) 做台語文文句 ê 分拆(parsing)， 利用語法樹(parsing tree)改進語詞搭配 ê 正確率。

(g) Tsò Tâi-gí-bûn bûn-kù ê hun-thiah(parsing), lī-iōng gí-huat-tshiū(parsing tree) kái-tsìn gí-sû tah-phuè ê tsìng-khak-lu̍t.