台灣南島語新詞表達推薦研究初探
A Preliminary Study on New Word Suggestion in Formosan Languages

林川傑 Chuan-Jie Lin†
國立臺灣海洋大學資訊工程學系
Department of Computer Science and Engineering
National Taiwan Ocean University
cjlin@email.ntou.edu.tw

宋麗梅 Li-May Sung†
國立臺灣大學語言學研究所
Graduate Institute of Linguistics
National Taiwan University
limay@ntu.edu.tw

劉沛綸、張肇豐、蔡孟睿、王鈞宇、吳金民
國立臺灣海洋大學資訊工程學系
Department of Computer Science and Engineering
National Taiwan Ocean University
{11157029, 01257020, 01257026, 01257131, 11057056}@email.ntou.edu.tw

摘要

臺灣南島語這類低資源語言因詞彙量較少，在進行機器翻譯研究時常會碰到新觀念不知如何翻譯的問題。本論文試著從音譯、近義單詞、詞組表意等各種方式來推薦新詞的族語寫法。推薦音譯做法是建立中文漢語拼音與各族語聲韻母之間的對應關係，就可以將各漢語拼音音節用各族語拼寫出來。推薦單詞的做法是估算新詞語意與辭典裡各詞條語意的相似度，推薦同義詞或最高分的近義詞。推薦詞組的做法則是詢問了大型語言模型，我們提出一個三步驟的做法請大型語言模型挑選 2~5 個已知詞來拼組出新詞的語意。最後參考原住民族語言研究發展基金會的族語新詞創制計畫成果來觀察這些推薦方法的可行性，進而探討未來還需針對哪些族語的自然語言處理技術，以增進新詞推薦可行性。

Abstract

Being low-resource languages, Formosan languages often have small size vocabulary which often lead to OOV problem in machine translation, especially for those new concepts in Chinese. This paper proposes methods to suggest new words in Formosan languages by transliteration, word similarity, and phrase construction. To do transliteration, mapping tables of vowels and consonants between Chinese and Formosan languages are constructed. Transliterated new words are generated according to these mapping tables. To make a single-word suggestion, the word in the Formosan dictionary who has the highest sense similarity with the target Chinese new word is suggested. To suggest a phrase to represent the new word, we design a three-step procedure to ask LLM to select 2~5 words in the Formosan dictionary where the combination can represent the target new word. These methods are examined based on the results of New Word Creation Projects held by Indigenous Languages Research and Development Foundation. More important future works are also discussed.

關鍵字：臺灣南島語、新詞、新詞推薦、借詞、大型語言模型

Keywords: Formosan languages, neologism, new word suggestion, loan word, large language model

1 緒論

在進行機器翻譯研究時，未知詞的翻譯常常是個難題。然而當翻譯的目標語言屬於低資源語言 (low-resource language) 的時候，如果該語言的詞彙數量比來源語言少很多，可想見有許多來源語言的詞彙無法在目標語言的辭典中被找到，也是一種未知詞問題。

† 通訊作者 corresponding authors

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 236–245
November 6-7, 2026 ©2026 Association for Computational Linguistics

以台灣南島語為例，在原住民族語言研究發展基金會 (簡稱原語會) 的各族語線上辭典[1]能查詢到的詞彙數量平均是 6000 多 (以網站全文下載版本為準)，遠少於中文詞彙量 (教育部國語辭典簡編本約 45000 詞、重編修訂本約 167000 詞)，由中文翻譯至台灣南島語時一定會碰到這類問題。過往對未知詞翻譯的策略也無法套用到低資源語言的情形。

在思考此問題的可能解法時，我們發現語發中心的新詞創制計畫提供了另一方向切入的研究成果。這個計畫主要在探討如何用族語表達現代社會新概念的詞彙。我們會以這項成果出發，研究人工智慧系統可以怎樣提供新詞創制的協助，也可以提供未知詞機器翻譯的方向。

2 族語新詞計畫簡介

新詞創制計畫是由原住民族語言研究發展中心 (簡稱語發中心) 執行的一系列計畫，定義及動機如下 (張裕龍, 2016; 戴佳豪, 2019)：

> 新詞 (neologism) 指的是語言中用來描述新知、概念的詞或詞組。…由於科技不斷進步、生活型態的轉變以及外來語言的流入刺激之下，帶來了新的觀念…產生詞彙的變化以及產生許多新語彙。…為發展原住民新語言，能與社會語言充分的表達與交流，本研究持續進行新創詞研發，來擴充族語詞彙，填補詞彙空缺以活絡族語為目的。

計畫從民國 103 年起，各執行年度會挑選一些屬於不同領域的中文詞，各族各自討論出這些中文詞對應的族語詞。截至 108 年止，各族已至少有 560 個新詞創制的成果，並由原語會編撰為新詞辭典發布在網站上。

2.1 新詞創制規則類型

本研究最需要的資訊是各種可能的新詞創制規則，可惜沒有論文直接討論這點。不過在各年度的「制定族語新創詞」研究報告中，曾針對族語老師們所擬出的族語新創詞做了類型及構詞的分析，極具參考價值。其中 105 年度報告 (張裕龍, 2016) 將之歸納為 6 大類型，在此簡介如下：

(1) 借詞 (音譯)：以族語的發音方式拼出其他語言的發音。像「海報」的阿美詞 haypao 是中文音譯，而「粉筆」的阿美詞 hakofoko 則是日文音譯 (白墨、はくぼく)。

(2) 複合結構：直接使用中文詞構詞組成的對應族語詞。例如泰雅語「單行道」tqutux wwahan = 一 tqutux + 通道 wwahan、「通緝令」bbiru' pkzyap = 文件 bbiru' + 抓捕 pkzyap 都是屬於這種類型。

(3) 詞組意譯：採語意解釋的方式來表達中文詞的語意。例如泰雅語「紅綠燈」psyax sinwalan wahan = 發亮 psyax + 允許 sinwalan + 來 wahan、「列印機」qeqaya' ppslung m'mu' = 器具 qeqaya' + 模仿 ppslung + 壓住 m'mu'

(4) 衍生：在構詞學指的是在詞幹上加上詞綴來產生新意義的詞 (Bauer, 1983)，105 年度報告中明確舉出的例子只有太魯閣新創詞的「徒步區」ksaan = ksa-an (走-處所焦點，即「走路的地方」) 和「作業」ptasun = patas-un (讀書-受事焦點，即「被學習的東西」)。

(5) 借代：指的是將原本已存在的詞的詞意擴張比喻替代另一種意義。105 年度報告中沒有明確舉出的例子，理解上應該是以近義詞來推薦。

(6) 縮略：指的是將較長的詞或詞組簡稱的做法，像是中文用「北車」來稱呼「台北車站」這樣。105 年度報告中只有撒奇萊雅語出現這種類型，由於範例太少無明確規則，本論文暫不考慮這種型式的推薦法。

這些類別中，借詞是比較容易判斷出來的類別，其餘不論衍生、近義詞、複合結構或詞組意譯都需要具有語言學背景的人才能判斷。比較單純的分群依據是：衍生、近義詞所推薦的會是單一一個族語詞，複合結構及詞組意譯則會是多詞的詞組。本論文也改以這三種情形分別提出推薦方法。

[1] https://e-dictionary.ilrdf.org.tw/

2.2 族語新詞資料來源

本研究所使用的族語新詞資料來自兩處。一是語發中心新詞創制計畫的報告，範例請見表 1，可看到各年度提供資訊不同，有的是新詞的組合意義，有的是例句。但借詞資訊標得較清楚。

另一個來源是「原住民族語言線上辭典」網站[2]，在各族語頁面的「下載專區」→「全文下載」可以找到「新詞辭典」，收錄了 105~108 年新詞創制計畫提出的族語新詞。內容如圖 1 所示，但僅有族語、中文對照，沒有更多資訊可用 (借詞會標為「音譯詞」或「直譯」)。

合併這兩者資料時，若同一新詞出現不同建議，會以新詞辭典的版本為主。此外，各族語有各不同方言的新詞建議時，因為我們希望查詢線上辭典來得知中文語意，因此只挑選了線上辭典所採用的方言，包括賽考利克泰雅語、北排灣語、郡群布農語、南王卑南語、德固達雅賽德克語。阿美語及魯凱語新詞辭典並未分出方言。

表 2 統計了各族蒐集到的新詞數量，已移除重複出現在兩種來源的詞條。可以看到借詞數量比例較少，而各族使用單詞或詞組來表達新詞的比例各各不同。

詞彙	**泰雅語 (賽考利克)**
村長	mrhuw qalang 部落首領
自助餐	sinlungan niqan 隨心吃東西的地方
古早味	ssbkisan 酷似古老的樣子、味道等

(a) 103、104 年度新創詞 (摘錄).

詞彙	「校長室」(泰雅語各方言)
族語	kakiyan kucyu' (澤) kocyo sicu (賽) patawawan na kucyu' (汶) tman mrhuw pwasan biru (宜) ptiyawan kocyu (四) pumahan kocyo (萬)
例句/翻譯	kakiyan patawayan kucyu. 校長辦公的處所(賽) sicu kocyo ga pcywagan na kocyo. 校長室是校長辦公處所.

(b) 108 年度新創詞 (摘錄).

表 1. 新創詞計畫推薦結果列表.

— Aa —

ali nSeediq/ jiyaxq qrasun seejiq/ diyax sngayan sediq 原住民族日

ali sskuwan 保存期限

ayu huda 冰川 (河)

ayus cceka knlangan (Tg) / ayus kska knglangan (Tr) / ayus cka knlangan (To) 赤道

ayus tdaya knlangan (Tg, To) / ayust tgdaya knglangan (Tr) 北回歸線

ayus thunac knlangan (Tg, To) /ayus thunac knglangan (Tr) 南回歸線

圖 1. 賽德克語新詞辭典 (節錄).

3 借詞推薦做法

借詞型新詞指的是直接在族語中使用別的語言的詞彙來表達新詞，常以音譯型式、也就是以族語的發音拼出該語言的發音。

在我們拿到資料中，有明確標出「中借」、「日語借詞」、「音譯」等標籤就歸類為借詞，共有 519 個。來源語言包括中文、日文、閩南

族語	**借詞**	**單詞**	**詞組**	**總計**
阿美	65	263	260	588
泰雅	28	119	460	607
排灣	75	190	318	583
布農	2	378	211	591
卑南	80	144	376	600
魯凱	43	320	226	589
鄒	16	138	427	581
賽夏	10	186	382	578
達悟(雅美)	7	95	459	561
邵	25	87	461	573
噶瑪蘭	24	159	403	586
太魯閣	2	54	535	591
撒奇萊雅	16	543	45	604
賽德克	51	59	487	597
拉阿魯哇	7	129	442	578
卡那卡那富	68	158	345	571
總計	519 5.5%	3022 32.3%	5837 62.2%	9378

表 2. 原語會新詞辭典詞彙數量.

[2] https://e-dictionary.ilrdf.org.tw/

語、客語等等，數量如表 3 所示。「音譯」及「直譯」沒有標明語言因此獨立統計，目測大多是中文借詞。本研究只處理中借的情形，因為我們認為會用日語借詞等情形應該都是早已在族語中存在的借詞詞彙，未來若要發明新的音譯詞理論上會以中文詞為主。

進行中文借詞音譯的第一步，是製作各族語聲韻母與中文聲韻母的對應關係。中文採用漢語拼音系統，漢字與拼音的關係來自 cdict[3] 漢英字典，還會寫程式將聲、韻母切開，方便接下來的對應統計。

族語拼音資料則是蒐集自線上辭典裡的族語詞條，一樣地寫程式將族語音節及聲韻母切開，也先統計各聲、韻母的使用頻率。

在製作音譯所需的語言聲、韻母對應表時，訓練資料就是新詞計畫資料裡的中文借詞詞彙。依上述方法將中文新詞轉成漢語拼音、切開聲韻母，也將族語音譯的音節及聲韻母切開，統計它們之間的對應關係。

例如「健保」漢語拼音 jian-bao，拆成[(j,ian) (b,ao)]兩個音節。對應的阿美語新詞 cinpaw 則可拆成[(c,in) (p,aw)]兩個音節，因此得到漢語聲母 j 對應到阿美語 c、漢語 b 對應到阿美語 p 的關係，韻母亦同。

音譯常常會碰到音節數不一致的情形，像三字詞「平交道」ping-jiao-dao 的阿美語新詞 piciyawtaw 是 4 個音節，也就是 jiao 要對應到 ci-yaw 兩個音節。

我們的做法是：若族語音節的聲母是半元音 y 或 w，但中文不是，則可將此族語音節與前一音節合併去做對齊。上例中，漢語[(p,ing)]對到阿美語[(p,i)]、漢語[(j,iao)]對到阿美語[(c,i) (y,aw)]、漢語[(d,ao)]則對到阿美語[(t,aw)]。雙音節在統計聲韻母對應時，只有第一音節子音對應中文聲母，其餘都對應到中文韻母。

原本標為「中借」、「音譯」等可視為中借的詞有 428 個。對齊後發現有 12 個族語音節數量比中文少，須排除不使用。這些有的是英文縮寫 (像「光碟」CD、「公斤」kg)，有的是簡稱，像「流行性感冒」卑南只說感冒 kanmaw、「高速鐵路」魯凱只說高鐵 kauti。排除後共有 416 個借詞用來進行本實驗。

各族語各自統計中文聲韻母對應，取對應次數最多的當做本研究給該族語音譯的系統選擇，聲母對應請參考表 4。如果某個族語不曾出現某個漢語聲韻母的對應，就採用該族

族語	中借	日借	其他	音譯	總計
阿美	10	6	1	48	65
泰雅	3	20	1	4	28
排灣	6	8	0	61	75
布農	0	1	0	1	2
卑南	15	1	0	64	80
魯凱	3	7	0	33	43
鄒	2	8	0	6	16
賽夏	1	2	1	6	10
達悟(雅美)	0	0	0	7	7
邵	0	0	0	25	25
噶瑪蘭	5	4	0	15	24
太魯閣	0	0	0	2	2
撒奇萊雅	3	3	2	8	16
賽德克	1	15	2	33	51
拉阿魯哇	0	5	0	2	7
卡那卡那富	6	4	0	58	68

表 3. 原語會新詞借詞數量統計.

漢語拼音	b	ch	d	f	g	h	j	q	s	sh	t	x	z	zh
阿美	p	c	t	f	k	h	c	c	s	s	t	s	c	c
泰雅	p	c	t	f	k	h	c	c	s	s	t	s	c	c
排灣	p	c	t	v	k	h	c	c	s	s	t	s	c	c
布農	p	c	t	f	k	h	c	c	s	s	t	s	c	c
卑南	p	c	t	v	k	h	c	c	s	s	t	s	c	c
魯凱	p	c	t	f	k	h	c	c	s	s	t	s	c	c
鄒	p	c	t	f	k	h	c	c	s	s	t	s	c	c
賽夏	p	c	t	f	k	h	c	c	s	s	t	s	c	c
達悟(雅美)	p	c	t	f	k	h	c	c	s	s	t	s	c	c
邵	p	s	t	f	q	h	k	c	s	s	t	s	c	ts
噶瑪蘭	p	c	t	f	k	h	c	c	s	s	t	s	c	c
太魯閣	p	c	t	f	k	h	c	c	s	s	t	s	c	c
撒奇萊雅	p	c	t	f	k	h	c	c	s	s	t	s	c	c
賽德克	p	c	t	h	k	h	k	c	s	s	t	s	c	c
拉阿魯哇	p	c	t	f	k	h	c	c	s	s	t	s	c	c
卡那卡那富	p	c	t	f	k	h	c	c	s	s	t	s	c	c

表 4. 族語聲母與漢語拼音之對應表 (節錄).

[3] https://cdict.net/

語合法聲母中、其他族語最常採用的聲母對應，在表 4 中以灰色底色標示。

值得注意的是，有些字母雖然在漢語和族語拼音系統都存在，卻因發音不同而不是直接對應。像漢語 q 發音ㄑ就更接近族語的 c，族語 q 則是接近漢語的 k 或 g。

至於評估時，為了避免內部測試的情形，我們採用留一交叉驗證 (leave-one-out cross-validation)，每個新詞進行系統音譯時，這個詞就不放進訓練資料來提供聲韻母對齊。

借詞型新詞推薦評估結果請見表 5。雖然完全正確 (距=0) 的比例只有 39.4%，但大部份系統音譯都與正解高度相似、編輯距離 (edit distance) 多只在 3 之內。實際觀察拼音上的差異，多數是 o 和 u 的選擇、是否加上 y 等等。可見借詞推薦系統效能不錯。

有趣的是，這時檢視系統推薦結果與正解的編輯距離較大者，發現大多不是中文借詞。像「聖誕節」魯凱語 kulrisemase、排灣和卑南 kurisemas 都是譯自英文 Christmas，「便利商店」賽德克語 sebunrebun、卡那卡那富 sevun'irevun 都是譯自 7-11 的英文發音。未來也可以利用這點來快速過濾非中借詞彙。

4 單詞推薦做法

第二種推薦新詞的做法，是以一個族語詞彙延伸語意來表達新語意，直覺上是推薦一個辭典中已收錄、語意和新詞相近的族語詞。

為了確認此一前題是否成立，我們針對族語老師們推薦的單詞型新詞進行觀察，數量列在表 6 中。令人吃驚的是，有高達 85.6%左右的單詞型新詞並未收錄在辭典中。這些有可能是已知詞根的衍生詞，也有可能是真正的未知詞，當然還有少數是未標明的借詞。

雖然第 2.1 節談到能推薦單詞的規則包括衍生和借代兩種，可惜衍生需要族語的詞綴組合與語意變化之間的關係，目前尚無完整的研究成果可供參考。因此本論文主要探討借代新詞、利用語意相似度推薦新詞的做法。

因為無法直接估算一個族語詞與中文新詞的語意相似度，我們改用族語詞的中文語意去和中文新詞算相似度，因此只能推薦已收錄在族語線上辭典裡面的族語詞，實驗評估及觀察對象都只有表 6 裡那 435 個已知詞。辭典裡中文語意和新詞相似度最高者即為系統推薦的族語新詞。

族語	中借	距= 0	距≤ 1	距≤ 2	距≤ 3
阿美	58	24	43	49	52
泰雅	6	2	3	4	4
排灣	65	26	48	56	56
布農	1	1	1	1	1
卑南	77	44	56	66	69
魯凱	35	13	21	26	31
鄒	8	4	7	7	7
賽夏	7	1	1	2	5
達悟(雅美)	6	3	4	6	6
邵	25	5	13	18	22
噶瑪蘭	19	2	5	6	9
太魯閣	2	2	2	2	2
撒奇萊雅	9	1	5	5	7
賽德克	33	17	22	26	27
拉阿魯哇	1	1	1	1	1
卡那卡那富	64	18	39	52	57
總計	416	164	271	327	356
佔比(%)		39.4	65.1	78.6	85.6

表 5. 族語中文借詞推薦正確率.

族語	單詞數	未知詞	已知詞
阿美	263	248	15
泰雅	119	93	26
排灣	190	169	21
布農	378	336	42
卑南	144	118	26
魯凱	320	287	33
鄒	138	108	30
賽夏	186	132	54
達悟(雅美)	95	71	24
邵	87	71	16
噶瑪蘭	159	127	32
太魯閣	54	26	28
撒奇萊雅	543	504	39
賽德克	59	47	12
拉阿魯哇	129	102	27
卡那卡那富	158	148	10
總計	3022	2587 (85.6%)	435

表 6. 新詞計畫中單詞型新詞統計數量.

我們試用了三種語言模型的嵌入向量的夾角餘弦值來估算語意相似度，如表 7 所列，包括本實驗室自己的模型 (We)、fastText[4] (FF)、以及 sentenceTransformer[5] (ST)。本實驗室的詞嵌入模型是自行對中文維基百科斷詞後、用 Gensim fastText 工具訓練所得。

相似度最高分的推薦正確率 (和族語老師推薦詞相同的比例) 請見表 7。可看到正確率都不高，最好的是最適合估算語意相似度的 sentenceTransformer，正確率 22.3%。若加上優先選擇同義詞詞林裡收錄的同義詞 (系統代碼 ST2)、也就是將屬於詞林裡同一個同義詞詞群的詞語意相似度設為 1 的話，正確率可上升至 24.4%。

為了觀察為何語意相似度無法成功地挑出族語老師們所推薦的參考正解，我們另外統計了參考正解與中文新詞的 sentence Transformer 嵌入向量相似度分佈情形，詳見表 8。可看到相似度大於等於 0.8 的僅有 34.0% (仍高於正確率 24.4%)，甚至有近 35%相似度不到 0.6。這表示有兩大錯誤類型。

第一種錯誤類型是辭典中出現了比參考正解更相似的詞彙，例如「豪雨」在噶瑪蘭語辭典相似度最高的是 0.909 的 uzan「雨」，而參考正解是相似度 0.876 的 muzan「下雨」。

第二種錯誤類型更難預測，就是族語老師推薦了語意相似度不高的參考正解。例如「(家庭) 作業」，拉阿魯哇語參考正解是 asuhlatu「記下來」(相似度 0.100)，噶瑪蘭語參考正解是 sasudadan「要寫的」(相似度 0.154)，而卑南語參考正解是 tatililran「作業簿」(相似度 0.449)

再以「勘災」為例，阿美語參考正解是 minengneng「正在看」(相似度 0.285)，撒奇萊雅語參考正解是 miazih「在看」(相似度 0.319)，而拉阿魯哇語參考正解是 takuari「巡視」(相似度 0.459)。

未來工作有三個方向。一是詢問族語老師，語意相似度最高的詞彙是否也能被接受成為族語新詞。二是蒐集更多單詞型新詞的例子，觀察分析更多可用來設計系統的資訊。三是深入詞綴組合與語意的對應關係研究，表達新詞時可決定加上適當的詞綴組合變化。

族語	**總數**	**We**	**FF**	**ST**	**ST2**
阿美	15	2	2	3	4
泰雅	26	5	5	7	8
排灣	21	7	4	8	9
布農	42	8	7	8	8
卑南	26	4	4	8	8
魯凱	33	8	8	11	11
鄒	30	3	2	3	5
賽夏	54	7	7	14	15
達悟(雅美)	24	5	4	6	6
邵	16	1	1	3	3
噶瑪蘭	32	3	4	5	7
太魯閣	28	0	0	4	4
撒奇萊雅	39	2	2	3	4
賽德克	12	2	2	4	4
拉阿魯哇	27	1	2	6	6
卡那卡那富	10	5	3	4	4
總計	435	58	57	97	106
比例 (%)		13.3	13.1	22.3	24.4

表 7. 單詞型新詞推薦正確率.

相似度 ≥	1.0	0.8	0.6	0.4	0.2
阿美	3	5	8	13	15
泰雅	3	11	19	23	26
排灣	4	9	16	20	21
布農	6	15	26	37	42
卑南	4	9	14	25	26
魯凱	8	16	27	31	33
鄒	2	9	17	25	30
賽夏	7	19	36	51	54
達悟(雅美)	4	10	19	23	24
邵	1	3	11	15	16
噶瑪蘭	4	12	19	30	32
太魯閣	1	7	14	25	28
撒奇萊雅	2	7	20	37	38
賽德克	3	6	10	11	12
拉阿魯哇	2	6	15	23	27
卡那卡那富	3	4	7	8	10
總計	57	148	278	397	434
比例 (%)	13.1	34.0	63.9	91.3	99.8

表 8. 單詞型新詞與參考正解中文相似度分佈.

[4] https://fasttext.cc/

[5] https://sbert.net/

5 詞組推薦做法

第三種推薦新詞的做法，是以一個詞組 (兩個詞以上的組合) 來表達新語意。數量如表 2 所示，佔新詞計畫裡 62.2%，算是最大宗的情形。

想要了解詞組是如何將詞意拼組得到新詞所代表的語意，我們需要知道族語老師們推薦的詞組裡每個詞的意思，觀察數據請見表 9。可惜的是，與預期相同，5837 個詞組型新詞中僅有 19.8%的詞組裡所有詞都能在辭典中查到意思。其他 80.2%的詞組都至少有一個未知詞，使得語意拼組的研究不易進行。

直覺上第一個可能的詞組推薦做法，就是做機器翻譯。當然目前還沒有品質夠好的機器翻譯系統在線上 (又得能使用未知詞來翻譯)，我們會以開發中的翻譯系統試看看。

目前我們還能想到的其他方法，仍是以辭典中收錄的已知詞的語意，試著拼組出新詞的語意。這時會借用大型語言模型對於語意理解的能力，提出的做法請見後文。實驗資料只採用表 9 中 1155 個全是已知詞的詞組型新詞。

5.1 機器翻譯推薦新詞

我們首先嘗試的，是用本實驗室開發中的機器翻譯系統 (廖恒慶, 2023)，將中文新詞翻譯成族語。因為目前只開發了泰雅語、太魯閣語、賽德克語的機器翻譯，只有這三種族語有這種新詞推薦的實驗結果。

評估時為了避免全對才算對太過嚴苛，改用兩種指標：Dice 係數 (參考正解與翻譯結果的重疊程度) 和 F-score (依翻譯結果、參考正解、猜對詞數計算 precision、recall 後再算調和平均數)，以新詞為單位來算巨觀平均值。

實驗結果請見表 10 (a)，不論是 Dice 還是 F-score 都只在 0.08～0.22 之間，表示目前的機器翻譯確實很難提供與參考正解相同的詞組。

由於台灣南島語有豐富的焦點系統，一個詞根常常可以加上各種組合的詞綴形成不同的衍生詞，語意又高度相關。例如「代理人」太魯閣語的新詞推薦如下：

參考正解 (原始詞彙)：seejiq pririh
參考正解 (詞根形式)：seejiq ririh
機器翻譯 (原始詞彙)：rrihan seejiq
機器翻譯 (詞根形式)：ririh seejiq

原本比對原始詞彙，只會對中一個詞。改比對詞根形式後，兩個詞都能比對到。實驗結果請見表 10 (b) ，分數仍在 0.10～0.25 之間。

5.2 大型語言模型推薦新詞

我們利用大型語言模型 (LLM) 來推薦詞組型新詞的步驟如下：先請 LLM 依照自身語言能力，用幾個中文詞彙來組合出中文新詞的語意；接著至族語辭典搜尋與 LLM 建議詞語意相近的數個詞；最後請 LLM 再從這些近義詞

族語	詞組數	全已知	比例
阿美	260	25	9.6%
泰雅	460	83	18.0%
排灣	318	21	6.6%
布農	211	36	17.1%
卑南	376	87	23.1%
魯凱	226	25	11.1%
鄒	427	20	4.7%
賽夏	382	102	26.7%
達悟(雅美)	459	64	13.9%
邵	461	118	25.6%
噶瑪蘭	403	134	33.3%
太魯閣	535	205	38.3%
撒奇萊雅	45	4	8.9%
賽德克	487	120	24.6%
拉阿魯哇	442	63	14.3%
卡那卡那富	345	48	13.9%
總計	5837	1155	19.8%

表 10. 新詞計畫中詞組型新詞統計數量.

族語	Dice	Precision	Recall	F-score
泰雅	0.0860	0.0944	0.1305	0.1002
太魯閣	0.2143	0.2495	0.2203	0.2222
賽德克	0.1303	0.1456	0.1680	0.1499

(a) 原始詞彙比對.

族語	Dice	Precision	Recall	F-score
泰雅	0.1029	0.1080	0.1556	0.1175
太魯閣	0.2447	0.2870	0.2513	0.2540
賽德克	0.1623	0.1777	0.2125	0.1856

(b) 詞根形式比對.

表 9. 詞組型新詞機器翻譯推薦結果評估.

中挑選詞彙組合來表達中文新詞，它們對應的族語詞就是新詞推薦。

以實際實驗結果來舉例，請 LLM 用 2 到 5 個詞彙表達「獸醫」語意，會得到「動物、醫生、照顧」的建議。這三個詞在拉阿魯哇語辭典中都有收錄，因此系統的新詞建議就是 kuli'i maisasavusavuanu maiiatuhluhlu，這和族語老師建議的 maisasavusavuanu hliupaupangua kuli'i (醫生、養、動物) 對中了兩個詞。至於詞的順序不對，因為牽涉到語法，也會是未來研究工作之一。

再舉一個例子。請 LLM 表達「母親節」語意，會得到「媽媽、節日、紀念」的建議。這三個詞在泰雅語辭典沒有全收錄，挑選辭典中語意相近的詞：

媽媽 ≈ {母親、舅媽、...}
節日 ≈ {假日、日子、祭拜...}
紀念 ≈ {紀念、慶祝、...}

請 LLM 再在從這些近義詞其中挑選 2 到 5 個詞彙來組合出「母親節」的語意，得到「母親、日子、慶祝」的建議，因此系統的新詞建議就是 yaya' ryax smqqas，這和族語老師建議的 ryax sqqas yaya' (日子、慶祝、母親) 對中了兩個詞，而且沒對中的 smqqas 和 sqqas 應該是同一個詞根的衍生詞。

族語	Dice	Precision	Recall	F-score
阿美	0.168	0.182	0.174	0.168
泰雅	0.159	0.151	0.176	0.159
排灣	0.263	0.282	0.290	0.263
布農	0.238	0.214	0.280	0.238
卑南	0.223	0.248	0.227	0.223
魯凱	0.113	0.130	0.101	0.113
鄒	0.213	0.210	0.217	0.213
賽夏	0.151	0.171	0.145	0.151
達悟(雅美)	0.185	0.222	0.170	0.185
邵	0.181	0.205	0.170	0.181
噶瑪蘭	0.175	0.195	0.165	0.175
太魯閣	0.158	0.162	0.162	0.158
撒奇萊雅	0.125	0.125	0.125	0.125
賽德克	0.153	0.143	0.168	0.153
拉阿魯哇	0.299	0.286	0.325	0.299
卡那卡那富	0.346	0.312	0.399	0.346

表 11. 詞組型新詞大型語言模型推薦結果評估.

我們選擇的 LLM 是 GPT-OSS-20b，它可以在單機環境執行。第一步給 LLM 的完整提示詞請見附錄一。第二步比對語意相似度時仍是採用 sentenceTransformer 轉換的嵌入向量，取 cos 相似度 0.6 以上的前 5 名做為候選詞。第三步再請 GPT-OSS-20b 從候選詞集合中挑選詞組推薦，完整提示詞請見附錄二。

實驗結果請見表 11，各語言的 Dice 和 F-score 分數大多在 0.11～0.23 之間，排灣、拉阿魯哇、卡那卡那富三族分數最好，卡那卡那富語甚至有 0.346 分。但因為在泰雅、太魯閣、賽德克語系統表現沒有特別好，無法評判此方法與機器翻譯何者較好。

未來希望能有更多詞型組新詞的範例可供研究語法構成或是語意組成的規則。

6 結論

本論文針對臺灣南島語新詞創制的 AI 自動生成可行性進行了初步的討論。針對借詞、單詞、詞組三種型式都提出族語推薦的做法，並以族語新詞創制計畫裡族語老師們建議的當做參考正解，評估各種做法的正確率。

針對借詞型新詞推薦，本研究依據新詞計畫的參考正解製作了各族語聲韻母與中文聲韻母對照表，並用來產生中文借詞的族語拼法。正確率雖只有 39.4%，但有高達 85.6%的系統推薦與正解的編輯距離在 3 以內。

單詞型新詞推薦方法則是採用詞嵌入向量夾角餘弦值來模擬語意相似度。先查出中文新詞的詞嵌入向量，再去和字典中各詞中文語的詞嵌入向量計算相似度，以分數最高者做為推薦詞。詞嵌入向量選擇 sentence Transformer 效果最好，搭配同義詞詞林資訊後，正確率有 24.4%，仍有改善空間。

詞組型新詞推薦方法有二種，一個是使用本實驗室開發中的機器翻譯系統，測試泛泰雅語資料時 Dice 和 F-score 都只在 0.08 ~ 0.22 之間。

另一個詞組型新詞推薦方法是以三步驟詢問大型語言模型，讓它能從辭典中挑選詞彙來拼組出新詞的語意。各語言的 Dice 和 F-score 分數大多在 0.11～0.23 之間，可見本研究頗具困難度。

研究過程中發現未來若能有更針對臺灣南島語言現象的研究成果，會對新詞推薦有相

當大的幫助。這裡整理了一些重要的未來研究方向。

首先是未知詞的處理，不論單詞型或是詞組型新詞都充滿大量未知詞，需要能認識或解析這些未知詞才行。目前我們曾試著提出過解決方案 (Lin *et al.*, 2025)，但不知是否能解決未來會碰到的所有未知詞。

再來是構詞結構、也就是詞綴組合與單詞語意或是多詞詞組語意之間的關係，在我們能推薦出語意正確的詞根時，如何選擇、套用正確的構詞結構就是下個重要工作。

接下來是詞組的語法結構，可能是學習得到語法規則，也可以是增強的機器翻譯技術。由於新詞通常是名詞或動詞，未來可加強針對族語名詞或動詞詞組的語法結構研究。

當然我們希望未來能有更多數量的新詞資料可做為研究參考，希望在數量夠多時能採用機器學習方法學出更多推薦做法。

致謝

本論文之研究承蒙行政院國科會研究計畫 (NSTC 113-2221-E-019-037 -) 之經費支持，謹此致謝。

參考書目

Laurie Bauer. 1983. *English-Formation*, Cambridge Univ. Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information In *Transactions of the Association for Computational Linguistics*, Volume 5, pages 135-146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. 2020. Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, Volume 34, Number 05, pages 8236-8244.

Chuan-Jie Lin, Li-May Sung, Jing-Sheng You, Wei Wang, Cheng-Hsun Lee, and Zih-Cyuan Liao. 2020. Analyzing the morphological structures in Seediq words. In *International Journal of Computational Linguistics and Chinese Language Processing*, Vol. 25, No.2, pages 1-20.

Chuan-Jie Lin, Li-May Sung, Chun-Kai Yang, and Bo-Yi Jin. 2025. Formosan language morphological analyzer. In *Proceedings of the 2025 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, *IEEE International Workshop on Empirical Methods for Recognizing Inference in Text (EMRITE)*, pages 246-249.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3980–3990.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525.

宋麗梅 (2018) 賽德克語語法概論, 臺灣南島語言叢書 5, 原住民族委員會, 2nd Edition, Taiwan.

宋麗梅 (2022) 賽德克語詞類及其教學, 臺灣南島語言叢書 III, 財團法人原住民族語言研究發展基金會, Taiwan.

廖恒慶 (2023) 泛泰雅語群與中文之多語言機器翻譯研究, 國立臺灣海洋大學, 碩士論文.

張裕龍 (2016)「制定族語新創詞」報告, 原住民族語言研究發展中心.

戴佳豪 (2019) 族語新創詞報告, 原住民族語言研究發展中心.

李佩容 (2022) 太魯閣語詞類及其教學, 臺灣南島語言叢書 III, 財團法人原住民族語言研究發展基金會, Taiwan.

李佩容 and 許韋晟 (2018) 太魯閣語語法概論, 臺灣南島語言叢書 10, 原住民族委員會, 2nd Edition, Taiwan.

黃美金 (2022) 泰雅語詞類及其教學, 臺灣南島語言叢書 III, 財團法人原住民族語言研究發展基金會, Taiwan.

黃美金 and 吳新生 (2018) 泰雅語語法概論, 臺灣南島語言叢書 2, 原住民族委員會, 2nd Edition, Taiwan.

附錄一

詞組型新詞大型語言模型推薦法第一步，請大型語言模型用 2~5 個簡易中文詞來組合出新詞語意時，給的提示詞是：

```
你是族語新詞組合研究助理。請只輸出嚴格 JSON。
任務：為中文新詞選出 2 到 5 個「語意比較簡單」的中文詞彙。這些詞彙合起來要能解釋該新詞的意思，目的是拿去查族語字典的中文釋義。
規則：
不要機械拆字或照字面切詞；如果新詞是兩個字，不可以直接輸出這兩個字。
優先用基礎、常見、具體的語意詞，例如人、食物、工作、快、房間、機器。
可以改寫成更接近語意的詞；不要只複製新詞中的字。
禁止輸出 詞 1、詞 2、概念詞 1、族語 1 這類佔位字。
只輸出 JSON，不要解釋。

壞例子：主菜 → {"concepts":["主","菜"]}
好例子：主菜→ {"concepts":["主要","食物"]}
好例子：高速鐵路 → {"concepts":["快速","火車","路"]}

新詞：{新詞中文}
輸出 JSON，鍵名固定為 concepts。
```

附錄二

詞組型新詞大型語言模型推薦法第三步，請大型語言模型從給定的候詞集合中、挑選 2~5 個中文詞來組合出新詞語意時給的提示詞是：

```
你是族語新詞組合研究助理。請只輸出嚴格 JSON。
你要從候選族語字典詞中，挑 2 到 5 個語意比較簡單的項目。這些項目的中文釋義組合起來，要最符合該新詞的整體語意。只能選候選清單中的 candidate_id，不要自創族語詞。避免選到反義詞或只相關但不符合語意的詞。若同一中文釋義有多個族語詞，請選最適合的一個族語詞。selected_fm_words 必須完全來自候選清單的族語詞；proposed_fm_phrase 必須用空白串接 selected_fm_words。禁止輸出 詞 1、詞 2、族語 1、族語 2 或中文詞當作族語答案。只輸出 JSON，鍵名固定為 selected_ids、selected_fm_words、proposed_fm_phrase、reason。

新詞：{新詞中文}
GPT-OSS 拆出的概念詞：{概念詞 1、概念詞 2、概念詞 3...}

候選清單：
{candidate_id}.中文釋義={字典中文釋義}｜族語詞={族語詞 1；族語詞 2...}｜cosine={cosine 分數}｜查詢={是哪個概念詞查到}
{candidate_id}.中文釋義={字典中文釋義}｜族語詞={族語詞 1；族語詞 2...}｜cosine={cosine 分數}｜查詢={是哪個概念詞查到}
…
```