深圳大鵬話核心詞性質的計量分析——兼論惠州話核心詞的性質 *

余鵬[1]　張梓杰[2]　曹用[3]
1. 香港中文大學（深圳）人文社科學院
2. 香港中文大學工程學院信息工程學系
3. 香港中文大學（深圳）經管學院
marcoyu@cuhk.edu.cn

提　要　深圳大鵬話的歸屬問題頗有爭議，有粵語說、客家話說、粵客混合方言說三種看法。本文選用斯瓦迪士二百詞表爲材料，對大鵬話與周邊方言作了相似度計算和聚類分析。結果表明，大鵬話的核心詞更接近粵語，尤其接近東莞，而與梅縣、深圳兩處客家話的相似度最低。這一研究以計量的方式否定了大鵬話核心詞的客家話性質，支持大鵬話核心詞更接近粵語的觀點，為核心詞性質提供了計量依據，從而爲進一步討論其方言歸屬提供了核心詞方面的計量證據。與前人的其他不同在於，我們搜集並公開了可對齊的詞彙材料，本文的計量結果是可以複現的。我們還發現，大鵬話與惠州話類似，音韻方面比較接近客家話，核心詞方面都比較接近粵語。不同的是，相比大鵬話，惠州話核心詞的客家話成分更多。暫無法排除惠州話核心詞的客、粵混合性質。

關鍵詞　大鵬話　惠州話　粵語　客家話　核心詞　相似度　聚類分析

壹　引言

關於深圳大鵬話的性質，有粵語說（中國社會科學院語言研究所等，2011: 圖 B1 17、117；劉鎮發，2010，記南澳口音；劉鎮發、袁方，2010；丘學強等，

* 本文受教育部人文社會科學研究青年項目“江淮官話黃孝片音韻比較研究”（22YJC740095）資助。本文曾在“第二十八屆國際粵方言研討會”（香港都會大學，2024 年 12 月 13-14 日）上報告。感謝與會專家、《粵語研究》匿名評審提出的寶貴意見。感謝李焱教授對詞彙相似度計算細節的指導。論文不足之處，由作者承擔責任。

2022，記鵬城口音）、客家話說（中國社會科學院等，1987，圖 B15，無文字討論；王莉寧，2024，密集布點）、客粵混合方言說（Chen，2016，記鵬城口音）三種看法。

劉鎮發兩篇文章都從音韻角度論證大鵬話的性質。王莉寧（2024）也主要從音韻角度論證，兼及個別詞彙。

前人研究中從詞彙討論大鵬話與周邊方言關係的，僅有陳利砼（Chen，2016: 90-93）、丘學強等（2022）。陳利砼考察了客家話和粵語的約 100 個特徵詞在大鵬話中的對應形式，發現這些詞中，大鵬話只有 10% 來自客家話，超過 80% 來自粵語。

丘學強等（2022:294-322）逐個考察了大鵬話的斯瓦迪士二百詞與周邊粵、客方言的異同後發現，不論高階詞（斯瓦迪士二百詞中的前 100 詞，更為基本）還是低階詞（後 100 詞），其中含有粵語成分的詞的數量都遠超過含有客家話成分的詞。在斯瓦迪士一百詞裡，與粵語同源的詞根語素共計 87 條，與客家話同源的詞根語素共計 73 條。其中明顯的粵語詞、客家話詞分別有 23 個、9 個。3 個詞條兼有粵、客說法（丘學強等 2022：316）。低階詞中，與粵語同源的詞根語素共 16 條，與客家話同源的詞根語素共 4 條，5 個詞條兼有粵客成分（丘學強等 2022：316）。

陳利砼（Chen，2016）的外部方言選用廣州粵語、梅州客家話共計 2 個點，僅展示小部分大鵬詞彙與粵、客的對齊語料。丘學強等（2022）用於比較的周邊方言很多，其中周邊方言語料多爲舉例式說明，一般不給出所有周邊方言的對齊語料。

惠州話的性質爭議也在客、粵之間，因此本文的討論也加入惠州話的語料。對前人研究的述評見項夢冰、曹暉（2005:164-165/2012:129-130）。侯小英（2017）認爲惠州話底層是客家話，受了粵方言較多影響。該文本質上也持客家話說。前人對惠州話的討論不涉及斯瓦迪士二百詞。

目前，尚無研究從計量角度討論大鵬話核心詞與周邊方言的異同及其聚類關係。目前學界對歸屬有爭議的其他方言的研究，多採用定性的分析，考察若干音韻、詞彙條目的相似情況。本文分析大鵬話核心詞與周邊方言的異同，計算詞彙的相似度，在此基礎上計算大鵬話與周邊方言的聚類關係，以確定大鵬話核心詞在多大程度上接近周邊的方言、最接近哪一種周邊的方言，並在此基礎上討論大鵬話的性質。

本文的附錄展示了本文所用語料，以顯示我們的研究是可驗證的。

大鵬話核心詞的性質這個題目，有助於深化我們對客、粵語關係的認識。在方法上，本文在前人方言間詞彙相似度的計算方法的基礎上，簡化了個別步驟的計算，進一步明確了計算方法的一些細節。此外，本文的計算方法也有助於以計量的方式討論漢語其他歸屬有爭議的方言核心詞的性質。

貳　材料

一、詞表選擇

我們採用斯瓦迪士一百詞（斯瓦迪士二百詞中的高階詞，即前 100 詞）來討論大鵬話、惠州話與周邊方言的關係。斯瓦迪士二百詞表爲斯瓦迪士（Swadesh）所創，最初用來統計詞彙變化速率和語言分化的年代。後來因爲其基本詞彙在語言演變中的穩定性質，這個詞表用來對民族語乃至東亞－太平洋語言作譜系分類（Wang，1995；趙志靖，2017），也用來判斷語言的接觸或同源關係（陳保亞，1996），在中國民族語學界影響較大。

這個詞表學界已用來討論漢語方言間的關係。Wang & Wang （2004/2008）用影響較大的三個詞表對漢語方言作了譜系分類，證明斯瓦迪士一百詞表更適合用來爲漢語方言分類。冉啟斌、丁俊（2023）討論漢語各大方言間的相似度和差異，所用材料為斯瓦迪士一百詞表。趙志靖（2024）用這個詞表研究吳語、徽語、江淮官話方言的相互關係。這些研究多討論漢語各大方言間的關係，較少涉及邊界方言的歸屬。據我們所知，僅江荻（2022）用斯瓦迪士一百詞表對 186 個漢語方言點作了自動聚類，其討論包括歸屬有爭議的若干邊界地區的漢語方言。

我們尚未見到前人以計量方式使用斯瓦迪士一百詞表討論漢語某個方言的歸屬問題。我們擬使用該詞表的材料，以計量的方式探索邊界地區歸屬有爭議的漢語方言的性質。爲印證前人的研究，並說明我們的計算方法的有效性，我們也考察了後 100 詞。

二、方言點選擇與材料來源

深圳大鵬話位於深圳。深圳境內有客家話、粵語兩種主要方言，二者都納入比較範圍。我們還考察了與深圳接壤的惠州話（指惠城區方言，2011 年出版的《中國語言地圖集》第二版將其劃入客家話粵台片梅惠小片）、東莞粵語、廣州粵語。東莞粵語納入比較，還因為劉鎮發（2010）引地方誌一類的材料提到“根據大鵬

居民記載，他們的祖先來自東莞一帶，在明代初年在大鵬半島建立‘衛所’”。廣州話是粵語的代表，在廣東省各方言中極爲强勢，當然也納入比較。客家方言的代表點梅州客家話也納入比較。

深圳大鵬話材料來自丘學強等（2022）。東莞（莞城）、廣州（荔灣）、深圳（寶安）三處粵語，深圳（坪山）、梅州兩處客家方言和惠州話，其語料皆來自林倫倫等（2022）。除惠州外，大鵬周邊這些地方的方言歸屬都無爭議。

爲表述簡明，後文深圳（寶安）粵語簡稱深圳粵語，深圳（坪山）客家話簡稱深圳客家話，廣州（荔灣）粵語簡稱廣州粵語。

叁　計算方法

一、詞彙相似度算法

漢語兩個方言的詞彙相似度有兩大類算法。第一大類，長期以來的研究計算的多是漢字（即漢字反映的語素或詞）的差異（鄭錦全，1988；王士元、沈鍾偉，1992；游汝杰、楊蓓 1998；Wang & Wang，2004/2008；李焱，2013）。這類計算立足於漢語方言大部分詞有漢字可寫的特點，相同漢字表示同源語素 。第二大類算法，即近年來出現計算音標差異的編輯距離算法（趙志靖，2017；江荻，2022；冉啟斌、丁俊，2023），適用於包括漢語在內的所有語言，其計算使用 Python 軟件，需要一定的編程語言能力。第一類算法得出了具有語言學意義的結果，例如 Wang & Wang （2004/2008）、李焱（2013）等都成功地對漢語十種方言作了譜系分類，得出的分類結果符合漢語方言學界的一般看法。可見，計算詞彙相似度時，以漢字作爲計算的對像是可行的，可以得出符合常識的結論。由於本文討論的語言都是漢語方言，因此本文採用前者的方法。

本文詞彙相似度的計算方法在李焱（2013）的基礎上有所改動，聚類方法則完全根據該文。根據李焱（2013），詞彙相似度的計算基於漢字，不考慮構詞方式（有兩個例外，見後文） 。我們根據原文的說明，並請教該文作者後，對一對多和多對多兩種情況的計算作出一些改動，說明如下。

（一）一對一

待比較項如果是一對一，例如大鵬“肚胈”、東莞“屎肚”，總字數是 3，則二者相似度

$$S_{\text{肚胈 - 屎肚}} = 1/3$$

用公式表示爲：

$$S_{一對一}=A/B$$

S 表示 similarity 的首字母，指方言對應詞的相似度，其值在 0 ～ 1 之間。A 表示方言對應詞共有字的個數。B 表示方言對應詞的總字數（不計重複）。這種情況我們的算法與李焱（2013）相同。

（二）一對多與多對多

待比較項的對應詞的數量如果不同，則數量少的一邊設置一個或多個空形式詞，使得兩邊對應詞的數量相同，兩邊形成一對一的幾個配對，總配對數等於“一對多”裡“多”的一邊的對應詞數量。例如，待比較項如果是一對三，則“一”的這邊設置兩個空形式詞，這樣形成三個一對一的配對。

1. 計算結果與原文結果相同

一對多或多對多時，優先將相似度大的對應詞配對，再按公式算出每個配對的相似度，最後除以總配對數。例如，大鵬“阿爸”、廣州“老豆、阿爸”，其中“阿爸－阿爸”完全相同，是相關配對，相似度為 100%，優先配對。“空形式詞－老豆”是無關配對，不優先配對。因此：

$$S_{阿爸－老豆、阿爸}=（2/2+0/2）/2=1/2$$

其中分母 2 為總配對數。此處我們與原文算法不同的是，原文將左邊的“阿爸”與右邊的兩個對應詞分別配對，所以兩個配對分別是：“阿爸－老豆”“阿爸－阿爸”，其中左邊的“阿爸”與右邊的兩個對應詞配對了兩次，而我們的算法左邊的“阿爸”只與右邊的“阿爸”配對一次，不再與右邊的“老豆”繼續配對。即，原書對於一對多的情況，在“一”的這邊不設置空形式詞。爲了簡化計算，我們簡化了配對規則，不允許一個對應詞同時與另一個方言的多個對應詞配對。只要一個詞目在兩個方言裡的對應詞數量不同，都設置空形式詞，使每個對應詞都只與另一個方言的一個對應詞配對。

待比較項如果是“多對多”，例如深圳粵語“抹、繳”、大鵬“擦、繳、捽”，則

$$S_{抹、繳－擦、繳、捽}=（1/1+0+0）/（1+1+1）=1/3$$

其中，左右兩邊的“繳”完全相同，相似度 100%，優先配對。左邊剩下“抹”，右邊剩下“擦”“捽”，左右都沒有相同的字。左邊比右邊少一個詞形，左邊因此設置一個空形式詞。剩下兩對無論怎樣配對，相似度都是 0。分母 3 表示 3 個配對（含左邊的空形式詞）。以上這個例子我們的算法的過程和結果與原文相同。

2. 計算結果與原文結果不同

我們與原文不同的是，原文對於深圳粵語“如果、假如”、廣州“如果、若果”的配對，由於一個對應詞可以與另一個方言的多個對應詞組成相關配對（兩邊一共 4 個形式，其中 3 個形式都含有“如”字，3 個形式都含有“果”字），原文會形成“如果 - 如果”“如果 - 若果”“假如 - 如果”三個相關配對，“假如 - 若果”一個無關配對，一共 3 個配對。那麼總相似度為

$$S_{\text{如果、假如－如果、若果}}=（1+1/3+1/3）/（1+1+1+1）=5/12$$

而本文只允許 2 個配對。在左右所有可能的配對中，左邊的“如果”與右邊的“如果”相似度最大，為 1，因此，左邊的“假如”與右邊的“若果”配對，相似度為 0。這樣

$$S_{\text{如果、假如－如果、若果}}=1/（1+1）=1/2$$

計算結果比原文的結果略大。

（三）詞彙內部結構

以上計算方法不考慮詞彙內部結構，有兩個例外。

其一，大鵬“灰塵”、深圳客家“塵灰”這樣詞序相反的配對，相似度要再乘以 0.5。

第二，重疊形式的每個疊字視為單獨的漢字，計入總字數。因為“星－星星”的配對，重疊與否有類型學含義，計算時重新分析為：星 $_1$—星 $_1$ 星 $_2$。即一個漢字 X 在一個詞中重出 N 次，或一個漢字代碼 X 重出 N 次，則兩個疊詞視為 X1、X2、XN……，以此類推。重疊形式不一定相鄰[1]，例如“碎 1 支碎 2 點 - 零碎 1”，相似度為 1/5。

兩個方言的重疊字不同的情況。“紅紅的－赤赤的”，標記為“紅 $_1$ 紅 $_2$ 的—赤 $_1$ 赤 $_2$ 的”，相似度為 1/5。其中分子表示交集“的”的數量爲 1。分母 5 表示幷集（紅 $_1$、紅 $_2$、的、赤 $_1$、赤 $_2$）的漢字數量。

（四）其他情況

調查不出來的詞，即缺乏語料的詞目，都視爲零形式。

無漢字可寫的單音節詞或語素，用大寫字母替代漢字，以便計算機程序批量計算。同源的單音節詞或單音節語素用同一個大寫字母表示，如果不同源則用不同大寫字母表示。

1　此類情況在本文的語料裡暫未遇到。這裡仍說明，以便將來處理更多的語料。

二、聚類算法

李焱（2013）對多種聚類算法進行了比較。他首先提到了最短系聯法，認爲“這種方法會縮小類別與類別之間的距離，效果並不好”。該文也使用了鄭錦全提供的相似度矩陣對比了鄰接法（neighbor-joining method）和非加權平均值系聯法，發現前者會導致許多無法解釋的混亂現象，後者的結果比較符合學界對納入計算的漢語方言的歸屬的一般看法。

該文又對非加權平均值系聯法中點對簇和簇對簇的相似度算法及聚類算法進行進一步完善，即本文所採用的聚類算法。

我們的聚類方法採用平均值系聯法，計算方法都根據李焱（2013）。該文改進了點對簇、簇對簇之間的相似度的計算方法。點對簇的相似度等於點與簇中每個點相似度的平均值。簇對簇的相似度為簇內點與對方簇成員之間相似度的平均值。

若相似度最高的是尚未成簇的方言之間的相似度，就將這兩種方言聚成一簇。

若相似度最高的是尚未成簇的方言和簇之間的相似度，就將這種方言和此簇聚成新的簇。

若相似度最高的是簇和簇之間的相似度，就將這兩簇聚成新的簇。

具體而言，聚類的步驟為：首先，計算尚未成簇的兩方言之間的相似度。接著，計算尚未成簇的方言和簇之間的相似度，即尚未成簇的方言搭配簇中各方言的平均相似度。最後，計算簇和簇之間的相似度，即甲簇中各方言分別搭配乙簇中各方言的平均相似度。

肆　計量分析的結果與討論

一、大鵬話與周邊方言核心詞的相似度

根據上文的算法，大鵬話與周邊方言斯瓦迪士二百詞、高階詞（前 100 詞）、低階詞（後 100 詞）的詞彙相似度分別如表 1、表 2、表 3 所示。

這些方言都是漢語，因此其高階詞的相似度都應該高於低階詞相似度。而三張表中，每兩個方言之間，高階詞的相似度都大於低階詞，部分說明瞭本文採用的詞彙相似度的計算方法的可靠性。

三張表顯示，不論是二百詞還是高階詞，與大鵬話相似度最高的方言前兩位由高到底都是東莞粵語、深圳粵語，與大鵬話相似度最低的方言前兩位由低到高

都是梅州客家話、深圳客家話。

大鵬話核心詞更接近粵語，這印證了丘學强等（2022）的結論。大鵬話核心詞與梅縣客家話相似度最小，這與音韻角度比較的結果（劉鎮發 2010；王莉寧 2024）不同。

表 13 大鵬話與周邊方言二百核心詞的相似度

	大鵬	深圳客	梅州客	惠州	廣州粵	東莞粵	深圳粵
大鵬		0.65281	0.54898	0.68445	0.70291	0.72967	0.70650
深圳客			0.67403	0.69766	0.61475	0.64824	0.60789
梅州客				0.58079	0.54558	0.56961	0.54268
惠州					0.68801	0.69634	0.66260
廣州粵						0.79624	0.76677
東莞粵							0.78526
深圳粵							

表 14 大鵬話與周邊方言高階詞（前 100 詞）的相似度

	大鵬	深圳客	梅州客	惠州	廣州粵	東莞粵	深圳粵
大鵬		0.69881	0.58274	0.71528	0.73214	0.75893	0.75099
深圳客			0.70089	0.71825	0.64980	0.67401	0.65377
梅州客				0.60615	0.58681	0.61210	0.59127
惠州					0.73115	0.73214	0.72619
廣州粵						0.82143	0.83433
東莞粵							0.86210
深圳粵							

表 15 大鵬話與周邊方言低階詞（後 100 詞）的相似度

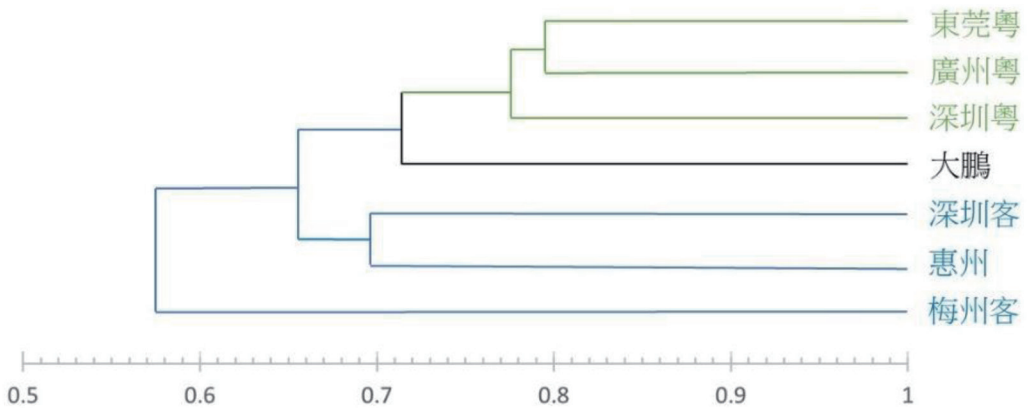
	大鵬	深圳客	梅州客	惠州	廣州粵	東莞粵	深圳粵
大鵬		0.60451	0.51354	0.65208	0.67222	0.69896	0.65979
深圳客			0.64583	0.67604	0.57795	0.62118	0.55972
梅州客				0.55417	0.50229	0.52500	0.49167
惠州					0.64271	0.65875	0.59583

廣州粵						0.76979	0.69583
東莞粵							0.70458
深圳粵							

二、大鵬話與周邊方言核心詞相似度的聚類分析

圖 1 是根據表 1 畫的二百核心詞的聚類圖。聚類方法見前文第三節。

圖 4　大鵬話與周邊方言二百核心詞相似度的聚類圖



這個聚類圖中，三種典型粵語方言聚成簇，接著大鵬話與三種典型粵語聚集成更大的簇。接下來，惠州話與深圳客家話聚成客家話簇。梅州客家話作爲客家話的代表，却不在這個客家話簇中，是因爲斯瓦迪士二百詞反映的譜系關係不如一百詞準確（Wang & Wang，2004/2008:351）。

大鵬話在聚類圖中的位置，介於典型粵語簇、典型客家話之間。

再看圖 2 高階詞聚類圖，根據表 2 畫出。東莞、深圳、廣州三處的粵語首先聚成一簇，接著大鵬與這個典型粵語簇聚成更大的簇。到這一步，斯瓦迪士一百詞的聚類圖與二百詞的聚類圖相同。接下來與圖 1 不同。惠州話與前述四點聚成的簇一起，聚成更大的簇。然後，深圳客家話與梅州客家話聚成一簇，這是典型客家話簇，符合方言學界對這兩處方言的認識。這印證了 Wang & Wang（2004/2008:351）提出的“高階詞適用於漢語方言譜系分類”。至此，所有的方言點都在方言簇中。

圖 2 的刻度表明，大鵬話與典型粵語簇的相似度略高於惠州話，明顯高於深圳客家話、梅州客家話。

圖 2　大鵬話與周邊方言一百核心詞相似度的聚類圖

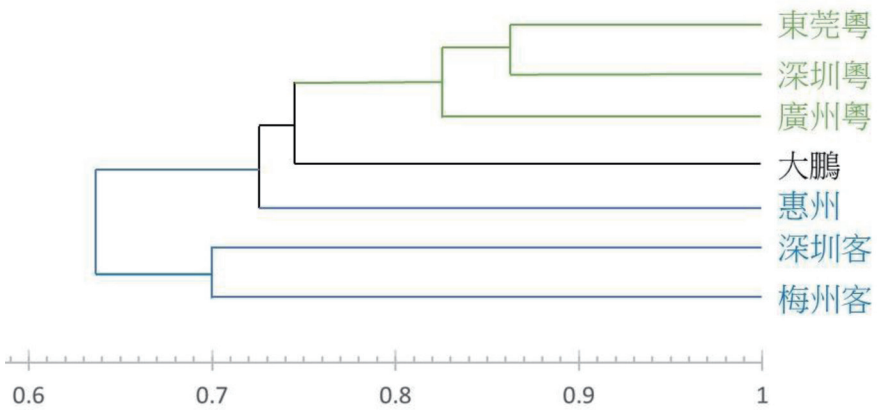


圖 2 的惠州話在大鵬話之後，先與典型粵語聚集成簇，而不先與典型客家話聚集成簇。中國社會科學院語言研究所等（2011: 圖 B1-17、117）將惠州話歸爲客家話，認爲“惠州市區方言是一種受粵語影響很深的客家話，與周圍客家話很不相同”。因此，惠州話在圖 2 中的位置是可以解釋的，不能說明我們的聚類方案和圖 2 的聚類結果無效。

三、混合方言的可能性：對計算結果的進一步分析

一百核心詞的聚類圖中，大鵬話、惠州話的位置介於典型粵語簇、典型客家話簇之間。因此，接下來，我們看二者有無可能解釋成客、粵混合方言。

爲便於理解，下文討論的大鵬話、惠州話與相關方言或方言簇的詞彙相似度如表 4 所示。大鵬話與典型粵語簇的相似度（0.74735）高於深圳客家話與梅州客家話的相似度（0.70089），大鵬話與典型客家話簇的相似度（0.64078）又低於深圳客家話與典型粵語簇的相似度（0.65919）。既然深圳客家話的歸屬沒有疑義，那麼大鵬話在一百核心詞層面是粵語，不然深圳客家話的歸屬也會受到質疑。

再看惠州話。雖然惠州話與典型粵語簇的相似度（0.72983）、與深圳粵語的相似度（0.72619）都高於深圳客家話與和梅州客家話的相似度（0.70089），但是惠州話與典型客家話簇的相似度（0.66220）不夠低，其數值高於深圳客家話與典型粵語簇的相似度（0.65919）。因此，難以直接認定惠州話的一百核心詞是粵語性質。此外，惠州話與典型粵語簇的相似度（0.72983）、惠州話與深圳粵語（即與惠州話相似度最低的粵語）的相似度（0.72619）都低於大鵬話與典型粵語簇的相似度（0.74735），也低於大鵬話與廣州粵語（即與大鵬話相似度最低的粵語）的相似度（0.73214）。而惠州話與典型客家話簇的相似度（0.66220）又高於大鵬話與典型客家話簇的相似度（0.64078），這意味著：從核心詞看，惠州話比大

鵬話更靠近客家話，且與粵語距離更遠。因此，暫不能排除惠州話具有客家話與粵語混合的性質。

表 16 關於大鵬話、惠州話混合方言可能性的數據

	大鵬話	惠州話		深圳客
與典型粵語簇	0.74735	0.72983	與梅州客	0.70089
- 其中最低者	0.73214 （與廣州粵）	0.72619 （與深圳粵）		/
與典型客家話簇	0.64078	0.66220	與典型粵語簇	0.65919

註：（1）“其中最低者”全稱是“與典型粵語簇內方言相似度最低者”。（2）表中只列出本節上文討論過的數字。

四、大鵬話、惠州話詞彙性質、音韻性質的不平行現象

丘學強等（2022:276-278）比較了大鵬話和深圳（南頭）粵語、客家話[2]、惠州話 30 個音韻條目的異同，發現大鵬話與深圳（南頭）粵語相同的條目為 16 條，與客家話相同的爲 11 條，與惠州話相同的為 17 條（數量最多）。

王莉寧（2024）對大鵬半島作了密集布點的音韻研究，認為：（1）大鵬話“在重要的音類分合關係上，與客家話粵台片更接近”。（2）大鵬話的底子是客家話，而其粵語特點是面子。（3）大鵬話應歸爲客家話。

至於惠州話，項夢冰、曹暉（2005:160-170/2012:126-138）比較了惠州話的音韻特點與粵語中心區的音韻特點、客家話中心區的音韻特點，認爲（項夢冰、曹輝 2005:164-165/2012:129-130）。

結論是顯而易見的：惠州話既不能劃歸粵語中心區，也不能劃歸客家話中心區，它是一種過渡方言。如果不設立過渡區，就需要採用 5.1.7.7 所提出的‘弱勢方言的逾越擴展原則’來確定惠州話的歸屬。

該書指出：（1）惠州話古全濁聲母今讀送氣塞音、塞擦音，不符合粵語中心區的特徵。（2）而廣東境內粵客接觸地帶的客家話是弱勢方言，其粵語特徵可以看做是粵語衝擊的結果。（3）如果惠州話屬於粵語，我們無法解釋作爲强勢方言

2 引者註：原文未說明該書所言客家話是何處的客家話。

的粵語的古全濁聲母今讀會是客家話類型[3]。

侯小英（2017）考察了惠州話的語音、詞彙，認為惠州話是“粵色客底的粵化客方言”，“具有更多的客方言語音成分”，詞彙上“客粵雜糅，粵語成分略佔優勢”。

以上研究說明大鵬話、惠州話的共同點：共時語音層面上客家話成分佔優勢[4]，共時詞彙層面上粵語成分佔優勢。

伍　結論

大鵬話與周邊共計 7 種方言中，所有方言間高階詞的相似度都明顯高於低階詞的相似度。三個典型粵語兩兩之間的相似度都最高，在聚類圖上都聚成一簇。歸屬無爭議的深圳客家話、梅州客家話聚成一簇，且與典型粵語簇相似度都明顯最低。這些都說明我們採用的詞彙相似度算法及聚類方法能反映方言的親疏關係。

丘學強等（2022：291、321、封底）提出：“大鵬話屬東莞、深圳一帶受客家話滲透較深的粵語”，“大鵬話本身是以東莞、深圳等地的粵語為底層方言的”，大鵬話“核心詞彙以粵語爲主，……。從歷時角度來看，大鵬話以粵語爲基礎方言，從共時角度來看，它呈現出了粵、客混合的語言面貌”。我們的計量研究支持了大鵬話核心詞以粵語爲主的觀點。在本文考察的方言中，不論是整個 200 詞還是高階詞，大鵬話都與東莞粵語最相似，與深圳粵語次相似。與大鵬話相似度最低的兩個方言從小到大依次是梅縣客家話、深圳客家話。這都說明大鵬話的核心詞的主體是粵語而非客家話的。

我們比較的方言還包括惠州話。惠州話與大鵬話在方言歸屬上都有爭議，共時層面，二者音韻上都是客家話成分佔優勢，詞彙上都是粵語成分佔優勢。與大

3　引者註：惠州話古全濁聲母的今讀不符合粵語中心區的特徵，其類型與粵語非中心區的粵語吳化片相同。粵語吳化片古全濁聲母字今清化後，一律讀送氣清音（參看中國社會科學院語言研究所，2012：130）。前人（劉鎮發 2004）將惠州話與粵語四邑片比較，而未與吳化片比較，大概是由於惠州話與粵語四邑片相似點很少。

4　有不同看法。劉鎮發（2010）、劉鎮發、袁方（2010）認爲，大鵬話的音系接近香港（圍頭）粵語，與客家話只是表層的接觸關係。劉鎮發（2004）分析了惠州話的音韻特點，認爲惠州話既非客家話，也非粵語。“……粵客贛均是宋代南方官話方言的後裔，將客贛方言重新歸併為一個方言，那麼惠州話群的大多數方言還是和大部分的客贛方言經歷過同樣的音變，例如濁上、濁去、濁入清化後送氣。”惠州方言群與嘉應方言群都是客贛方言的一個次方言，二者是平行的關係。（劉鎮發（2004：122））

鵬話不同的是，在典型粵語與典型客家話之間的連綫上，惠州話的核心詞比大鵬話略靠近典型客家話。我們暫無法排除惠州話的客、粵混合性質。

我們的計算結果表明，大鵬話的核心詞絕非梅州式的，甚至不同於深圳的客家話。相反，大鵬話的核心詞非常接近東莞、深圳、廣州等典型粵語。惠州話的核心詞也更靠近粵語，但其核心詞相比大鵬話有更多的客家話成分。

不論是否維持目前一級方言客、贛分立的方案，都必須解釋為什麼大鵬話、惠州話的音韻面貌相對接近客贛，而核心詞接近典型粵語。這是本課題下一步的研究方向。

參考文獻

陳保亞 1996《論語言接觸與語言聯盟：漢越（侗台）語源關係的解釋》，北京：語文出版社

侯小英 2017〈惠州話：粵色客底的粵化客方言〉，《學術研究》第 7 期，159-169 頁

李焱 2013〈從詞彙看漢語十大方言的親疏關係〉，董紹克等《漢語方言詞彙比較研究》，北京：商務印書館，254-283 頁

林倫倫、甘于恩、莊初升 2022《中國語言資源集 廣東 詞彙卷》，北京：中國社會科學出版社

江荻 2022〈漢語方言自動聚類與分區及相關計算方法〉，《暨南學報（哲學社會科學版）》第 3 期，10-23 頁

劉叔新 2007《東江中上游土語群研究——粵語惠河系探考》，中國社會出版社

劉鎮發 2004〈再論惠州話群的歸屬問題〉，《方言及方言史論集》，劉鎮發著，111-123，香港霞明出版社

劉鎮發 2010〈從深圳大鵬話看粵語和客家話的接觸關係〉，《南方語言學》第 2 輯，42-46 頁

劉鎮發 2010〈惠州話是粵語的一支嗎？——與劉叔新教授的商榷〉，《客語千秋：第八屆國際客方言學術研討會論文集》，羅肇錦、陳秀琪主編，文鶴出版有限公司，470-481 頁

劉鎮發、袁方 2010〈大鵬話語音系統概述〉，《粵語研究》第 6-7 期，13-17 頁

丘學強、溫育霖 2022《方言島 深圳大鵬話研究》，中國社會科學出版社

冉啟斌、丁俊 2023〈漢語方言的相似度與差異——基於 ASJP 模式語言距離計算的考察〉，《語文研究》第 2 期，17-26 頁

王莉寧 2024，《交界地帶方言的特徵描寫和性質鑒別》，“實驗語言學 +”雲上論壇直播報告，6 月 11 日

王士元、沈鍾偉 1992〈方言關係的計量表述〉，《中國語文》第 2 期，81-92 頁

項夢冰、曹暉 編著 2005/2012《漢語方言地理學》，北京：中國文史出版社 / 中國書籍出版社

游汝杰、楊蓓 1998〈廣州話、上海話和普通話詞彙接近率的計量研究〉，《漢語計量與計算機研究》，鄒嘉彥、黎邦洋、陳偉光、王士元主編 , 香港城市大學，57-78 頁

趙志靖 2017《東亞－太平洋區域語言相似關係計量研究》，南京：南京大學出版社

趙志靖 2024《吳語徽語江淮官話方言關係計量研究》，北京：中國社會科學出版社

鄭錦全 1988 漢語方言親疏關係的計量研究，《中國語文》第 2 期，87-102 頁

中國社會科學院、澳大利亞人文科學院 1987《中國語言地圖集》，香港：朗文出版社

中國社會科學院語言研究所、中國社會科學院民族學與人類學研究所、香港城市大學語言資訊科學研究中心 2012《中國語言地圖集》（第二版），北京：商務印書館

Chen, Litong. 2016. Dapeng Dialect [D]. Ohio State University.

Wang, Feng & Wang S-Y William. 2004/2008. Basic words and language evolution. *Language and Linguistics* 5.6: 43-62. 漢譯見汪鋒、王士元著，谷峰譯，汪鋒校 2008 基本詞彙與語言演變，《語言學論叢》第 33 輯，340-358 頁

Wang, William S-Y. 1995. A quantitative study of Zhuang-Dong Languages. In *In Memory of Mantaro J. Hashimoto*. Eds. A.O. Yue & M. Endo. Tokyo: Uchiyama.

附錄：大鵬話與周邊方言斯瓦迪士二百詞

200 詞序	普通話	英語	大鵬	東莞	深圳粵	廣州	惠州	深圳客	梅州
1	我	I	我	我	我	我	我	我	我
2	你	you	你	你	你	你	你	你	你
3	我們	we	我哋	我哋	我哋	我哋	我	我兜 | 我啲	我等人
4	這	this	呢	呢個	呢個	呢個 | 咿個	A 隻	迮隻	B 隻 | 個隻
5	那	that	嗰	嗰個	嗰個	嗰個	蓋隻	介隻	A 隻 | 個隻
6	誰	who	哪霞 | 哪人	乜誰	A 個	邊個	哪誰	哪人	B 人 |B
7	什麼	what	乜嘢	乜嘢	垢	乜嘢 | 咩嘢 | 咩	A 惹 | 一惹	脈個	脈 = 個
8	不	not	唔	唔	唔	唔	唔	唔	唔
9	全部	all	全部 | 通通	都	都	都 | 冚唪吟	都	都	都
10	多	many	多	多	多	多	多	多	多
11	一	one	一	一	一	一	一	一	一
12	二	two	二	二	二	二	二	二	二
13	大	big	大	大	大	大	大	大 | 大隻	大
14	長	long	長	長	長	長	長	長	長
15	小	small	細	細	細	細	細	細 | 細隻	細
16	女人	woman	女人	夫娘	夫娘	婆 | 女人	婦女	姑娘	婦人家
17	男人	man	男人	佬	男人 | 佬	佬 | 男人	男欸	老哥	男子人
18	人	person	人	人	人	人	人	人	人
19	魚	fish	魚	魚	魚	魚	A 欸	魚	魚子
20	鳥	bird	雀仔	雀 | 雀仔	雀仔	雀仔	雀欸	鳥仔	鳥子
21	狗	dog	狗	狗	狗	狗	狗	狗	狗子

22	蝨子	louse	狗虱 | 虱𧏛	虱𧏛	𧏛虱	虱𧏛	虱𧏛	虱嫲	虱嫲
23	樹	tree	樹	樹	樹	樹	樹	樹	樹
25	葉子	leaf	葉	葉	葉	葉	葉欸	葉	葉子
33	雞蛋	egg	雞春	雞春	雞春	雞蛋	雞春	雞春	雞卵
35	尾巴	tail	尾	尾	尾	尾	尾	尾	尾巴
37	頭髮	hair	頭毛	頭毛	頭毛	頭髮	頭毛	頭顱毛	頭顱毛
38	頭	head	頭殼	頭	頭	頭	頭顱	頭顱殼	頭顱
39	耳朵	ear	耳吉	耳吉	耳吉	耳仔 | 耳	耳朵	耳公	耳公
40	眼睛	eye	眼	眼	眼	眼	眼	眼	眼珠 | 目珠
41	鼻子	nose	鼻公	鼻哥	鼻哥	鼻哥	鼻拱	鼻公	鼻公
42	嘴	mouth	嘴	嘴	嘴	口	嘴	啜角	啜
43	牙齒	tooth	牙 | 牙齒	牙	牙	牙	牙	牙	牙齒
44	舌頭	tongue	腩	腩	腩	腩	腩	舌嫲	舌嫲
45	爪子	claw	爪	爪	腳	爪	爪	爪	爪子
46	腳	foot	腳	腳	腳	腳	腳	腳	腳
47	膝蓋	knee	膝頭	膝頭	蓋膝頭	膝頭 | 膝頭哥	膝頭	膝頭	膝頭
48	手	hand	手	手	手	手	手	手	手
49	肚子	belly	肚肶	屎肚	肚	肚	肚肶	肚肶	肚屎
50	脖子	neck	頸骨	頸	頸	頸	頸	頸	頸筋
51	乳房	breasts	晕仔	晕	媽仔	晕	晕 | 晕牯	晕牯	晕牯
54	喝	drink	飲	飲	飲	飲	飲	飲	食
55	吃	eat	食	食	喫	食	食	食	食
56	咬	bite	咬	咬	咬	咬	齧	齧	齧
57	看見	see	睇到	睇	睇	睇	睇	看	看
58	聽到	hear	聽到	聽	聽	聽	聽	聽	聽
59	知道	know	知到	知 | 知道	知	知	知	知到	知得
60	睡	sleep	睏覺	睏	睏	睏	睏	睡	睡目
61	死	die	死	死	死	死	死 |A	死	死
62	殺	kill	剉 | 殺	剉	剉	剉	剉	剉	治
65	走	walk	行	行	走 | 行	行	行	行	行
67	躺	lie	睏	睏	睏	睏	睏	睡	睡目
68	坐	sit	坐	坐	坐	坐	坐	坐	坐
69	站	stand	徛	徛	徛	徛	徛	徛	徛

70	給	give	畀	畀	畀	畀	畀	分	分
71	說	say	講	講 | 話	講	講	講	講	講 | 話
72	太陽	sun	熱頭	熱頭	熱頭	日頭 | 熱頭	日頭	日頭	日頭
73	月亮	moon	月光	月	月光	月光	月光	月光	月光
74	星星	star	星	星	星	星	星	星	星子
75	水	water	水	水	水	水	水	水	水
76	雨	rain	水	雨	雨	雨	水	水	雨
77	石頭	stone	石牯	石頭	石頭	石	石牯	石牯	石頭
78	沙子	sand	沙粒	沙	沙	沙	沙	沙	沙子
79	土地	earth	地	泥	泥	泥	泥	泥	泥
80	雲	cloud	雲	雲	雲	雲	雲	雲	雲
81	煙	smoke	煙	煙	煙	煙	煙	煙	煙
82	火	fire	火	火	火	火	火	火	火
83	灰	ash	灰	火灰	灰	灰	灰	火灰	灰
84	燒	burn	燒	燒	燒	燒	燒	燒	燒
85	路	path	路	路	路	路	路仔	路	路
86	山	mountain	嶺崗	山頭 | 山	山	山	山	嶺崗	山
87	紅	red	紅	紅	紅	紅	紅	紅	紅
88	綠	green	青 | 綠	綠	綠	綠	綠	綠	青
89	黃	yellow	黃	黃	黃	黃	黃	黃	黃
90	白	white	白	白	白	白	白	白	白
91	黑	black	烏	黑	黑	黑	烏	烏	烏
92	晚上	night	晚頭黑	晚頭夜	晚黑	晚黑	黑晚	暗晚晨	夜晡頭
93	熱	hot	熱	熱	熱	熱 | 燮	熱	熱	熱
94	冷	cold	冷 | 凍	凍	凍 | 冷	凍 | 冷	冷	冷	寒
96	新	new	新	新	新	新	新 | 新淨	新	新
97	好	good	好	好	好	好	好	好	好
98	圓	round	圓	圓	圓	圓	圓	圓	圓
99	乾	dry	燥	燥	燥	乾	糟	燥	燥
100	名字	name	名	名	名	名	名	名	名
101	和	and	同	同	同	同	同 | 湊 | 摎	摎	同
103	背	back	腰骨背	背脢	背	背脊	背脢	背梁	背梁
104	壞	bad	壞 | 衰	衰	衰	衰	衰	衰	壞

108	孩子	child	細佬仔	細蚊仔	細路哥	細路	細佬哥	阿僬	阿僬子 | 細 A 蟻子
110	砍	cut	斬	斬	斫	斬	斬	斬 | 斫	倒
113	髒	dirty	A	B 虱 | 鹹濕	癩 C	汙糟 | 邋遢	DE	囉嗦	F
114	呆、笨	dull	懵 | 傻	傻	傻	傻 | 戇居	懵 | 傻	傻	戇
115	塵土	dust	灰塵	塵	灰塵	塵	塵灰	塵灰	塵灰
116	掉	fall	跌	跌	冇	跌	跌	跌	跌
117	遠	far	遠	遠	遠	遠	遠	遠	遠
118	父親	father	阿爸	老子	老子	老豆 | 阿爸	阿爸	阿爸	爺子
119	怕	fear	驚 | 惶	惶	驚	驚	惶 | 驚	惶	驚
120	少	few	少	少	少	少	少	少	少
121	打架	fight	打交	打交	打交	打交	打交	打交	打交欸
122	五	five	五	五	五	五	五	五	五
125	花	flower	花	花	花	花	花	花	花
126	霧	fog	霧	霞霧	霧	霧	霧	霧水	雾紗
127	四	four	四	四	四	四	四	四	四
129	水果	fruit	水果	生果	生果	生果	生果	水果	青果
130	草	grass	草	草	草	草	草	草	草
132	他	he	渠	渠	渠	渠	渠	渠	渠
133	這裡	here	呢定	呢處	呢度	咿度 | 呢度 | 呢處	A 笪 |A 定	迂位	B 欸 | 個欸
134	打	hit	打	打	打	打	打	打	打
135	拿	hold-take	擰	搦	擰	擰 | 拎 | 搦	撿 | 攎	拿	拿
136	怎麼	how	點	點樣	點	點 | 點樣	樣邊	哪樣般	樣欸
138	丈夫	husband	老公	老公	老公	老公	老公	老公	老公
139	冰	ice	冰	冰	冰	冰	冰	冰	A
140	如果	if	假如	如果	如果 | 假如	如果 | 若果	係	係話	係話
141	在	in	在	起	在	喺 | 響	在	在	喺在
142	湖	lake	湖	湖	湖	湖	湖	湖	湖
144	左邊	leftside	左邊	左邊	左便	左便 | 左邊	左手析	左邊析	左析
145	腿	leg	腳	腳	大髀	腳	腳	腳髀	腳
147	母親	mother	阿奶 | 阿嬤	老母	奶	老母 | 阿媽	阿奶 | 阿媽	阿嬤	娭子
148	窄	narrow	狹	窄	窄	窄	窄 | 狹 | 窄狹	狹	狹

149	近	near	近	近	近	近	近	近	近
150	老	old	老	老	老	老	老	老	老
151	玩	play	嫽	嫽	嫽下	逛	嫽	嫽	嫽
153	推	push	A| 推	A	推 |A	推	A	攪	攪
154	右邊	rightside	右邊	右邊	右便	右便 | 右邊	右手析	右邊析	右析
155	對	correct	啱	啱	啱	啱	著 | 啱	著	著
156	江	river	河	海 | 江	江	江	大江	江	河 | 江
157	繩子	rope	繩	繩	繩	繩	繩欸 | 索欸	索嫲	索子
158	腐爛	rotten	爛	黵	黵爛	溫爛	糜 | 歿	歿	歿
159	擦	rub	擦 | 繳 | 捽	繳	抹 | 繳	抹	抹	捽	拭
160	鹽	salt	鹽	鹽	鹽	鹽	鹽	鹽	鹽
161	抓	scratch		捉	捉	捉	捉	捉	捉
164	尖	sharp	尖 | 利	尖	尖	尖	尖	尖	尖
165	短	short	短	短	短	短	短	短	短
166	唱	sing	唱	唱歌	唱歌	唱歌	唱歌	唱歌	唱歌
168	聞	smell	聞	聞	聞	聞	A	鼻	鼻
169	平	smooth	平	平	平	平	平	平	平
170	蛇	snake	蛇	蛇	蛇	蛇	蛇	蛇	蛇哥
171	雪	snow	雪	雪	雪	雪	雪	雪	雪
172	吐	spit	唾		吐		魯 | 吐	噴	退
173	撕裂	split	撕	撕	A	撕	撕	脈	扯
174	壓	squeeze		磧	壓	磧	磧	磧	磧
175	刺	stab	築 | 戳	刺	刺 |	刺	藜	藜	
176	棍子	stick	棍	棍	棍	棍	棍欸	棍	棍子
177	直	straight	直	直	直	直	直	直	直
178	吮	suck	啜	啜	嘿	啜	嘬	嘬	嘬 | 吮
179	腫	swell	腫	腫	腫	腫	腫 | 鳳	腫	腫
180	那兒	there	嗰定	嗰處	嗰頭	嗰度 | 嗰處	蓋笪 | 蓋定	介位	A 欸 | 個欸
181	他們	they	渠哋	渠哋	渠哋	渠哋	渠	渠家兜 | 渠家 | 渠家啲	渠等人
182	厚	thick	笨	笨	笨	厚	笨	笨	笨
183	薄	thin	薄	薄	薄	薄	薄	薄	薄
184	想	think	諗	諗	諗	諗	諗	諗	偭

185	三	three	三	三	三	三	三	三	三
186	扔	throw	劈			抌|掉	劈	拂	B
187	捆	tie	綯	綁	紮	紮|綁	綁|捆	綁	綁
188	轉	turn	轉	轉	轉	轉	捩	顛	顛
189	嘔吐	vomit	嘔	嘔	吐	嘔	嘔	嘔	翻
190	洗	wash	洗	洗	洗	洗	洗	洗	洗
191	濕	wet	濕	濕	濕	濕	濕	濕	濕
192	哪裡	where	哪定	始處	A 處	邊度|邊處	耐呢|耐定	哪位	哪欸
193	寬	wide	闊	闊	闊	闊	闊	闊	闊
194	妻子	wife	老婆	老婆	老婆	老婆	老婆	老婆	老婆
195	風	wind	風	風	風	風	風	風	風
196	翅膀	wing	翼	翼	翼	翼	翼	翼	翼胛
197	重	heavy	重	重	重	重	重	重	重
199	蟲	worm	蟲	蟲	蟲	蟲	蟲	蟲	蟲子

說明：（1）由於方言比較強調材料對齊，我們的材料裡，實際上高階詞（前 100 詞）只有 86 詞，低階詞（後 100 詞）只有 80 詞，共計 164 詞。斯瓦迪士二百詞的其他條目《中國語言資源集》詞表未收。（2）表中大寫字母表示無漢字可寫的單音節語素，材料來源原文以國際音標表示。

Computational Analysis of the Nature of the Dapeng Basic Words in Shenzhen – Also on the Nature of the Huizhou Bacis Words

YU Peng[1], ZHANG Zijie[2], CAO Yong[3]

1.School of Humanities and Social Sciences, The Chinese University of Hong Kong, Shenzhen

2.Department of Information Engineering, Faculty of Engineering, The Chinese University of Hong Kong

3.School of Management and Economics, The Chinese University of Hong Kong, Shenzhen

marcoyu@cuhk.edu.cn

Abstract

The attribution of Shenzhen's Dapeng dialect is rather controversial, with three viewpoints: a Yue dialect, a Hakka dialect, or a mixed dialect of Yue and Hakka. Based on the Swadesh 200-word list, this paper employs lexical similarity calculation and cluster analysis to explore the relationship between Dapeng and surrounding dialects. The results indicate that the basic words of Dapeng are closer to Yue dialects, especially those in Dongguan and Shenzhen, and they have the least similarity with Meixian and Shenzhen Hakka. This study quantitatively refutes the Hakka nature of the Dapeng basic words, supporting the view that the Dapeng dialect is close to the Yue dialects and providing new evidence for its dialect attribution. Unlike predecessors, we have collected and attached aligning lexical materials, ensuring the repeatability of our computational results. In addition to the relationship between the Dapeng, Yue, and Hakka dialects, we also discover that Dapeng and Huizhou share two linguistic features. Their phonological features are more similar to Hakka, while their basic lexicon is similar to Yue dialects. The difference is that compared with Dapeng, the Huizhou basic words have more Hakka dialect components. For the time being, there is no way to exclude the mixed nature of the Huizhou basic words.

Key words: Dapeng dialect, Huizhou dialect, Yue, Hakka, basic words, similarities, cluster Analysis