Computational Linguistics and Chinese Language Processing
Vol. 18, No. 2, June 2013, pp. 19-34

以中文十億詞語料庫爲基礎之兩岸詞彙對比研究

Cross-Strait Lexical Differences: A Comparative Study based on Chinese Gigaword Corpus

洪嘉馡*、黃居仁+

Jia-Fei Hong and Chu-Ren Huang

摘要

近幾年來，由於兩岸交流頻繁，兩岸使用的詞彙，也因此互相影響甚重，語言學界對於漢語詞彙的研究，不論在語音、語義或語用上的探討，發現兩岸對使用相同漢語時的詞彙差異有著微妙性的區別。而兩岸卻又的確是使用漢字體系的書寫系統，只有字形上有可預測的規律性對應。本文在以兩岸皆使用中文文字的原則上，在繁體中文與簡體中文的使用狀況來比對兩岸使用詞彙的特性與現象，以探究與語義對應與演變等相關的議題。

首先，在 Hong 和 Huang (2006) 的對應上，藉以英文 WordNet 爲比對標準，藉由比較北京大學的中文概念辭典(Chinese Concept Dictionary (CCD))與中央研究院語言所的中文詞網(Chinese Wordnet (CWN))兩個 WordNet 中文版所使用的詞彙，探討兩岸對於相同概念詞彙的使用狀況。本文進一步使用中文概念辭典與中文詞網所使用的詞彙，在 Gigaword Corpus 中繁體語料與簡體語料的相對使用率，探究兩岸對於使用相同詞彙，或使用不同詞彙的現象與分佈情形，並以 Google 網頁中所搜尋到的繁體資料與簡體資料進行比對、驗證。

關鍵詞：CCD, CWN, WordNet, Gigaword Corpus, Google, 兩岸詞彙, 詞義, 概念

* 國立臺灣師範大學 National Taiwan Normal University
E-mail: jiafeihong@gmail.com
+ 香港理工大學 The Hong Kong Polytechnic University
E-mail: churenhuang@gmail.com

Abstract

Studies of cross-strait lexical differences in the use of Mandarin Chinese reveal that a divergence has become increasingly evident. This divergence is apparent in phonological, semantic, and pragmatic analyses and has become an obstacle to knowledge-sharing and information exchange. Given the wide range of divergences, it seems that Chinese character forms offer the most reliable regular mapping between cross-strait usage contrasts. In this study, we take general cross-strait lexical wordforms to discovery of cross-strait lexical differences and explore their contrasts and variations.

Based on Hong and Huang (2006), we discuss the same conceptual words between cross-strait usages by WordNet, Chinese Concept Dictionary (CCD) and Chinese Wordnet (CWN). In this study, we take all words which appear in CCD and CWN to check their lexical contrasts of traditional Chinese character data and simplified Chinese character data in Gigaword Corpus, explore their appearances and distributions, and compare and demonstrate them via Google website.

Keywords: CCD, CWN, WordNet, Gigaword Corpus, Google, Cross-Strait Lexical Wordforms, Semantics, Concepts

1. 緒論

兩岸使用詞彙的差異問題，在目前兩岸人民的各種交流中，早就已經呈現出許多無法溝通、理解困難，或是張冠李戴，表達不合宜的錯誤窘境。探討兩岸使用詞彙的差異性時，不僅讓大量使用詞彙的記者們，感受到兩岸的差異 (如：華夏經緯網，2004；南京語言文字網，2004；廈門日報，2004)，甚至，近年來未了因應對岸人民來台旅行，台灣交通部觀光局也整理了「兩岸地區常用詞彙對照表」(中華民國交通部觀光局，2011)，這些皆成爲漢語詞彙學與詞彙語義學上研究的重要課題(如：王鐵昆與李行健，1996；姚榮松，1997)。

以往對於這個議題的研究，不論語言學學者或文字工作者注意到這個問題時，僅能就所觀察到特定詞彙的局部對應，來提出分析與解釋而缺乏全面系統性的研究。本文的研究方法，第一是延續 Hong 和 Huang (2006)、洪 等人(洪嘉馡與黃居仁，2008)的研究方法，先以 WordNet 做詞義概念的判準，比對中文概念辭典與中文詞網裡，概念相同、語義相同的詞彙使用狀況；第二、是以有大量兩岸對比語料的 Gigaword Corpus 作爲實證研究的基礎，驗證中文概念辭典與中文詞網對於相同概念語義的詞彙，使用上，確實有其差異性的存在。這是一個以實際語料、實際數據進行比對，且具有完整性、全面性、概括性的研究。

又 Miller 等人 (Miller, Beckwith, Fellbaum, Gross & Miller, 1993)認爲他們可透過使用同義詞集來表現詞彙概念和描述詞彙的語義內容，所以他們建立了 WordNet，近年來，

也有不少研究團隊在處理以 WordNet 爲出發點的不同語言翻譯。

值得一提的是，同屬於漢語詞彙系統的繁體中文系統與簡體中文系統，在中央研究院語言所與北京大學計算語言所的研究團隊裡，也針對此議題，做了不少相關的研究，因此，本文想要探討的是，相同概念的漢語詞彙語義，在繁體中文與簡體中文的使用狀況。

另外，對於繁體中文系統與簡體中文系統的對應，我們以 WordNet 當作研究語料的基礎，是爲了可以建立一套符合詞彙知識原則並能運用於英中對譯的系統，如此一來，即可比對出兩個中文系統，在語言使用上的差異性。

2. 研究動機與目的

自兩岸交流日趨頻繁之後，本屬於同文同種的漢語系統，確有不少知識與信息交流的障礙，造成這樣的原因，莫過於兩岸詞彙使用的差異。相同的詞形，卻代表不同的詞義；或相同的語義，卻有兩種不同的表達詞彙。這種問題，已經讓許多文字工作者費盡心思，試圖來解決這樣的窘境；而語言學者對於這種現象，也試圖從語音、語義、語用等方面著手，希望從各種與語言相關的角度，來探究兩岸詞彙的差異。

在研究議題上，光是觀察到兩岸選擇以不同的詞形來代表相同的語義，如下述例子(1)、(2)，這樣是不夠的。以 Gigaword Corpus 的語料呈現台灣/ 大陸使用的狀況及其在 Gigaword Corpus 中的詞頻，如下：

(1) 台灣的「煞 (155/ 65)」、大陸的「非典 (354/ 33504)」
（「Sars (SEVERE ACUTE RESPIRATORY SYNDROME)」「嚴重急性呼吸道綜合症」的翻譯）

(2) 台灣的「計程車 (22670/ 68)」、大陸的「出租車 (422/ 5935)」

在詞彙語義學研究上，我們必須進一步追究，這些對比的動機，語言的詞彙與詞義演變的動力是否相關，對比有無系統性的解釋等。Chinese GigaWord Corpus 包含了來自兩岸的大量語料，其中，有約 5 億字新華社資料(XIN)、約 8 億字中央社資料(CNA)，可以看出台灣和大陸對於同一概念而使用不同詞彙的實際狀況與分佈。

如要追究動機與解釋等理論架構問題，當然不能只靠少數觀察到的例子，而必須建立在數量較大的語料庫上，以便做全面深入的分析。以上兩個對比爲例，其實在大陸與台灣的語料中，都有相當多的變例出現。

3. WordNet和中文詞網(CWN)

WordNet 是一個電子詞彙庫的資料庫，是重要語料來源的其中一個語料資料庫，WordNet 的設計靈感源自於近代心理語言學和人類詞彙記憶的計算理論，提供研究者在計算語言

學，文本分析和許許多多相關的研究(Miller *et al*., 1993; Fellaum, 1998)。在 WordNet 中，名詞、動詞、形容詞、副詞，這四個不同的詞類，分別設計、組合成同義詞集(synsets)的格式，呈現出最基本的詞彙概念，在這當中，以不同的語義關係連結各種不同的同義詞集，串成了 WordNet 的整個架構，也呈現了 WordNet 整個全貌。

自從 Miller 等人 (1993)、Fellaum (1998) 發展 WordNet 以來，WordNet 就持續不斷地更新版本，目前最新的版本是 WordNet 3.0 版，這些版本間的差異，包括了同義詞集的量和他們的詞彙定義。然而，對於拿 WordNet 來做研究語料的學者，多數還是以 WordNet 1.6 版爲最多，因爲這個版本是目前最多計算語言學學者使用的。在 WordNet 1.6 版裡，有將近 100,000 的同義詞集。

我們知道，雙語領域分類，可以增加我們各種領域詞彙庫的發展，同樣的，在上一段的內容，我們也提到關於以 WordNet 爲基礎，發展出繁體中文系統(Chinese Wordnet, CWN)與簡體中文系統(Chinese Concept Dictionary, CCD)的對譯，我們使用雙語詞網，作爲詞彙知識資料庫來實現、支持我們在詞彙概念上的研究。

在中英雙語詞網中，每一個英文的同義詞集，我們都會給予三個最適合且對等中文翻譯，而這些翻譯，如果不屬於真正的同義詞，我們也會標註他們的語義關係(Huang, Tseng, Tsai & Murphy, 2003)，又這些雙語詞網，也在中研院語言所詞網小組團隊的發展，將每一個同義詞集都與 SUMO 概念節點連結，進而開發出 Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) (Huang, Chang & Li, 2010)。當我們無法直接取得中英相對應的詞彙，我們在雙語詞網的資料庫裡，可以利用這些語義關係，進而發展並預測領域分類。

4. WordNet和中文概念辭典(CCD)

CCD，中文概念辭典(Chinese Concept Dictionary)，是一個中英雙語的詞網，整個架構發展也是來自於 WordNet (于江生與俞士汶，2004；于江生、劉揚與俞士汶，2003；劉揚、俞士汶與于江生，2003)。在 CCD 的發展手冊裡記載，研究團隊描述這些詞義的首要條件，是不可以破壞原本 WordNet 對於同義詞集定義概念與其語義關係的架構。另一方面，CCD 的研究團隊考量到可以存在許多在中文與英文的不同描述架構，所以，他們不止表現對中文詞彙內涵的表達，也發展了中文詞彙語義與概念的關係性，以利於強調中文的特質。

CCD 的研究團隊專注在整個 CCD 的架構，提出同一概念的同義詞集的定義，其所呈現的概念、定義和概念網的上下位語義關係，每一個同義詞集都有其基本關係，彼此之間亦有語義關係的存在。至於 CCD 的邏輯推演原則在語義網上的呈現，是運用到數學的形式而來的，是可以幫助研究者在中文語義分析上的使用。

自從 2000/09 開始，北京大學計算語言學研究所就已經開始著手以 WordNet 爲基準，研究 CCD，並建立一個中英雙語的詞網，一個可以提供各種不同研究的詞網，如機器翻譯(MT)，訊息擷取(IE)…等等。

基於 WordNet 英文概念與 CCD 中文概念是屬於兩個不同知識背景，也因此 CCD 中，他們兩者間的相互關係與概念，是非常複雜、繁瑣的。CCD 包括了大量且繁雜的成對、成組的小網絡，大致上，差不多有 10^5 的概念節點和 10^6 的成組小網絡的概念關係，他們的關係，呈現如下圖：

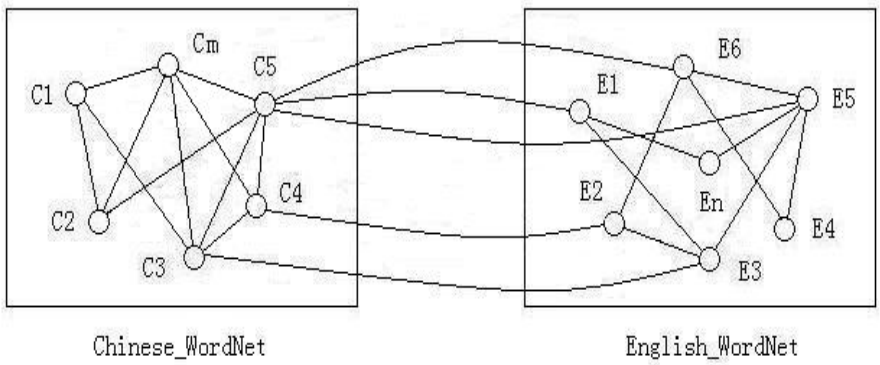

圖 1. WordNet 小網絡中複雜的關係結構

5. 文獻探討

對於兩岸詞彙對比的探討，過去的研究，多半著重在表面語言特徵的區別。如列舉語音方面、詞彙方面的對比(南京語言文字網，2004)；或以語音、詞彙、語法及表達方式等方面來分析語言差異的現象 (如：王鐵昆與李行健，1996；姚榮松，1997；許斐絢，1999；戴凱峰，1996)。

近年來，對於兩岸詞彙對比的研究，比較新的研究方法，是以 WordNet 爲基礎，取兩岸語料庫資料作比較，進而分析兩岸詞彙的對比(如：Hong & Huang, 2006)；或以 Chinese Gigaword Corpus (2005)爲基礎，探索兩岸對於漢語詞彙在使用上的差異現象，例如：相關共現詞彙(collocation)的差異、台灣或大陸獨用的差異、特定語境下的特殊用法的差異、語言使用習慣的差異等等(如：洪嘉馡與黃居仁，2008)。

6. 研究方法

本研究以英文的 WordNet、繁體中文系統的中文詞網(CWN)、以及簡體中文系統的中文概念辭典(CCD)等三大資料庫爲主，對於繁體中文系統的英中對譯與簡體中文系統的英中對譯，我們先進行比對，試圖在比對中，尋找出兩者之間的差別與使用分佈。

相同的概念，本歸屬於一個同義詞集，但因兩岸在詞彙使用上的差異，而有所不同，儘管如此，仍舊有一些兩岸使用相同的詞彙來表達相同的概念語義。本文中將從繁體中文系統與簡體中文系統的英中對譯資料裡，集中探究同一個同義詞集，在兩岸使用的詞彙是完全相同、完全不同的狀況。然後，再將這些完全相同、完全不同的詞彙，以 Gigaword Corpus 爲基礎，分析這些詞彙在這個語料庫裡，所呈現出兩岸使用的狀況。

接著，本文再以語料庫爲研究出發點，是以約十四億字的 Chinese Gigaword Corpus 爲主要語料來源，以中文詞彙速描爲搜尋語料工具 Chinese Gigaword Corpus (2005)、Chinese Word Sketch Engine、Kilgarriff *et al.* (2005)。Chinese Gigaword Corpus 包含了分別來自大陸、臺灣、新加坡的大量語料，包括約 5 億字新華社資料(XIN)、約 8 億字中央社資料(CNA)，及約 3 千萬字新加坡聯合早報資料(Zaobao)。本研究，僅就大陸新華社資料與臺灣中央社資料進行比對，因此，本文研究可以提供兩岸詞彙差異的大量詞彙證據。

最後，本文亦視 Google 爲一個擁有大量繁體中文、簡體中文的語料庫，試圖根據 Google 所搜尋到的繁體中文網頁與簡體中文網頁的資料，進行並驗證兩岸在詞彙使用差異上的實際使用證據。

爲了可以比較 Chinese Gigaword Corpus 的繁體中文與簡體中文，及兩者的使用差異性，我們採用中文詞彙速描系統(Chinese Word Sketch)進行檢驗。中文詞彙速描系統裡，有四大搜尋功能，分別爲：concordance、word sketch、Thesaurus、Sketch-Diff，其中「Sketch-Diff」這個功能就是比較詞彙差異的工具，可以看出兩岸對於同一概念而使用不同詞彙的實際狀況與分佈，也可以看出同一語義詞彙在兩岸的實際語料中，所呈現的相同點與差異性。我們主要利用中文詞彙速描中詞彙速描差異(word sketch difference)的功能。詞彙速描差異的實際操作介面，如圖 2：

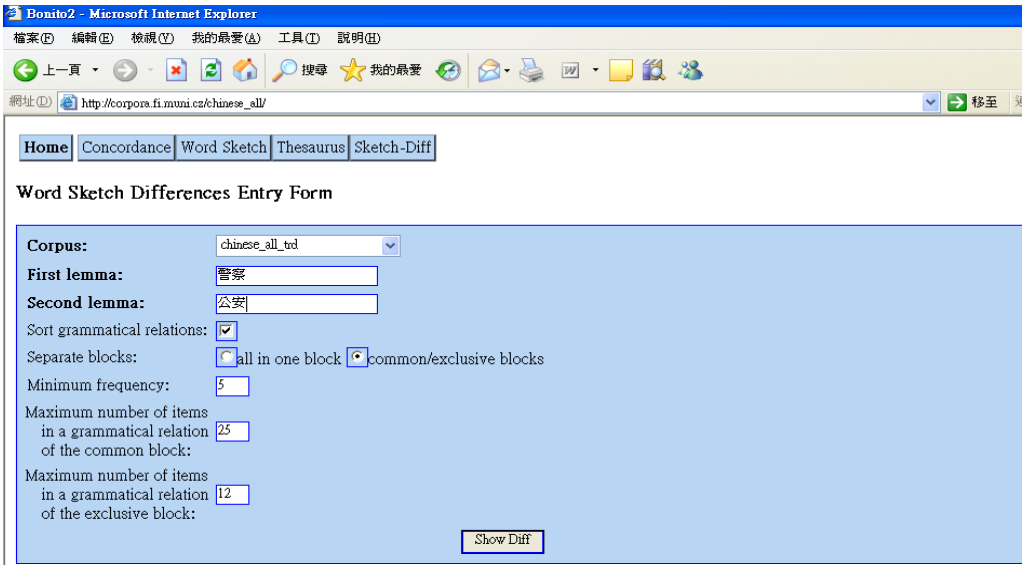

圖 2. 中文詞彙速描系統的詞彙描素對比

在此功能下，我們將已經比對過 CCD 與 CWN 對應不同對譯的詞彙，進一步探究兩詞彙的使用狀況與分佈。在本文中，主要是以比對兩岸詞彙詞頻爲主，倘若在 CCD 與 CWN 的對應中，確實是相同語義，卻在兩岸使用完全相同或完全不同的詞彙，那麼其各自使用的詞彙，在 Gigaword Corpus 裡繁體語料與簡體語料交叉比對後所得的詞頻，也應當會有近似的分佈現象，藉此數據，不但可以證明 CCD 和 CWN 在英中對譯上，繁體

中文系統與簡體中文系統，是有差別的，也可以證明，確實有兩岸使用不同詞彙來表達相同概念語義的用法，進而了解兩岸詞彙的實際現象，以進行本研究的分析。

7. CCD與CWN語料分析

繁體中文系統的英中對譯(CWN)與簡體中文系統的英中對譯(CCD)，依不同詞類，區分成：名詞、動詞、形容詞和副詞四大類來進行對比，以 WordNet 爲主，檢測在同一個同義詞集中(Synset)，繁體中文系統的對譯詞彙和簡體中文系統的對譯詞彙，然後再進行比對。

在四大詞類中，我們可以清楚得知，在同一個同義詞集中(Synset)，繁體中文系統，可能有多個相對應的對譯詞彙，同樣地，簡體中文系統也可能有個相對應的對譯詞彙。在這些對譯詞彙裡，又有可能是兩邊使用的對譯詞彙完全一樣，稱之「完全相同」；如果，兩邊使用的對譯詞彙，沒有一個相同的，稱之「完全不同」，也就是「真正不同」；或者，只有使用其中一個或一個以上對譯詞彙，這個狀況，稱之「部份相同」，而在「部份相同」的對譯詞彙，如果兩邊的對譯詞彙使用的詞首相同，稱之「詞首相同」，如果只是使用到相同的字，則稱之「部份字元相同」，如：

表 1. CCD 和 CWN 對譯的各種分佈狀況

Synset	CCD 對譯詞彙	CWN 對譯詞彙	
bookshelf	書架、書櫃、書櫥	書架、書櫃、書櫥	完全相同
lay off	下崗	解雇	完全不同
immediately	立即	立刻	詞首相同
according	據報	根據	部分字元相同

對於 CWN 與 CCD 的對比，總共有 70744 個 Synset 是對譯相同的，分屬於形容詞、副詞、名詞和動詞這四個詞類當中，其中，以名詞在 CWN 與 CCD 的完全相同對譯中，所佔比例最高，有 66.79%；反之，動詞所佔比例最低，僅有 4.05%，其詳細的分佈情況，如下圖顯示：

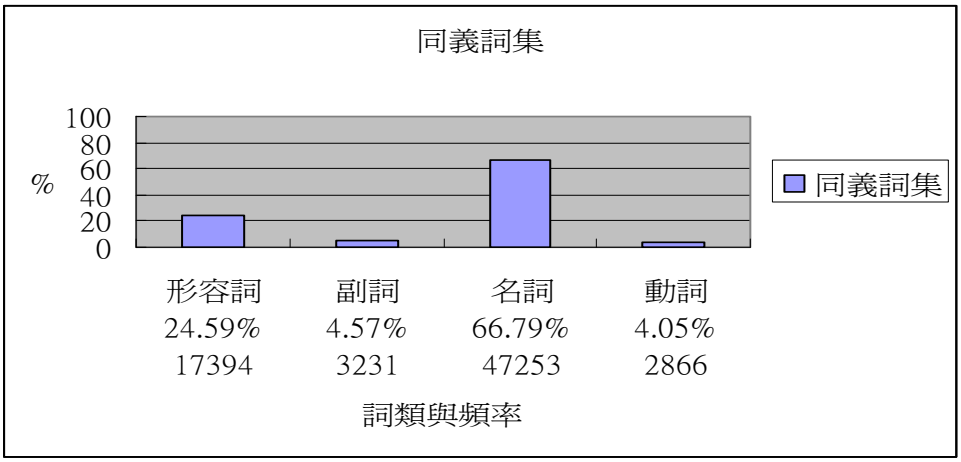

圖 3. CCD 和 CWN 依不同詞類呈現對譯相同的分佈

以 WordNet 爲基礎，CWN 所對譯的繁體中文與 CCD 所對譯的簡體中文，兩者使用完全不相同的情況，依各詞類的分佈情形，如下圖顯示：

表 2. CCD 和 CWN 對譯不同的分佈狀況

	形容詞	副詞	名詞	動詞	**總數**
同義詞集數量	17915	3575	66025	12127	**99642**
不同對譯數量	521	344	18772	9261	**28898**
	2.91%	9.62%	28.43%	**76.37%**	**29.99%**
	最少			**最多**	

值得一提的是，在 CCD 和 CWN 翻譯不同的分佈狀況裡，很清楚得看到，「動詞」在兩岸的使用狀況，有極大的差異性，不過，由於在我們實際使用漢語時，常會以同類近義詞或語義相近相關詞來取代原本的詞彙，所以，我們又更進一步，更仔細地分析，希望將每一個詞類中，有這樣的使用情形分類出來，以得到真正兩岸使用不同詞彙的現象。

我們以詞彙的「詞首相同」、「部份字元相同」和「真正不同」這三大類爲主，分析 CCD 和 CWN 在形容詞、副詞、名詞和動詞這四個詞類當中，翻譯不同的分佈狀況，如下表顯示：

表 3. CCD 和 CWN 在各詞類中，對譯不同的分佈狀況

形容詞的分佈				
類別	**詞首相同**	**部分字元相同**	**真正不同**	**總數**
同義詞集	169	175	177	521
	344			
百分比	34.44%	33.59%	33.97%	100%
	66.03%			

副詞的分佈				
類別	**詞首相同**	**部分字元相同**	**真正不同**	**總數**
同義詞集	77	114	153	344
	191			
百分比	22.38%	33.14%	44.48%	100%
	55.52%			

名詞的分佈				
類別	**詞首相同**	**部分字元相同**	**真正不同**	**總數**
同義詞集	7113	7843	3816	18772
	14956			
百分比	37.89%	41.78%	20.33%	100%
	79.67%			

動詞的分佈				
類別	**詞首相同**	**部分字元相同**	**真正不同**	**總數**
同義詞集	3269	3316	2676	9261
	6586			
百分比	35.30%	35.80%	28.90%	100%
	71.10%			

從表 1 到表 3，我們可以清楚知道對於各詞類，CCD 和 CWN 在對譯不同的詞彙裡，仍然有些算是語義相近的相關詞彙，扣除這些相關詞彙後，兩岸詞彙在使用上的真正不同，就可清楚呈現。至於，上文中，所提及關於「動詞」是兩岸詞彙中，使用最多不同的狀況，我們從表 3 的分析得知，在動詞的使用上，因爲較常出現同類近義詞或語義相近相關詞來取代原本的詞彙的狀況，所以「詞首相同」和「部分字元相同」這兩類佔了

很大的因素，在 9261 個詞彙裡，就有 6586 個詞彙，大約是 71.10%，其真正兩岸對於動詞的不同使用，則有 2676 個詞彙，大約是 28.90%。

我們將以圖 3 和表 3 中，四種詞類裡，使用完全相同的詞彙與真正不同的詞彙，藉由 Gigaword Corpus 來分析兩岸人民對於詞彙使用的實際狀況。

8. 實驗設計與詞彙差異分析

以 WordNet 爲中心所對譯出 CCD 的簡體中文和 CWN 的繁體中文，比較兩者的對譯，有完全相同、完全不同與部份相同等三大類，在此，本研究僅就前兩類的資料，再以 Gigaword Corpus 爲依據，檢測實際語料中所呈現的狀況，同時，也以目前在網路上搜尋功能相當強大的 Google 作爲驗證的對象，比對利用在 Google 所搜尋的資料來驗證兩岸詞彙的對比。

8.1 Gigaword Corpus

首先，我們先取兩岸使用詞彙「完全相同」的資料，檢測這些資料在 Gigaword Corpus 中，分屬在繁體中文與簡體中文的使用頻率，再計算每個詞彙的頻率在繁體中文資料與簡體中文資料裡所佔的比例，如此，即可知道每一個詞彙，在繁體中文與簡體中文裡，出現和使用的情形。理想的想法，如果一個詞彙在兩岸使用的情況是非常接近的，其兩者詞頻比例的差距，應該是非常小的。我們試著將同一詞彙在兩岸使用的詞頻比例相減，以便檢測這些使用上完全相同的詞彙，又因其差距的數值過小，所以我們以放大 100000 倍後的數值來呈現，其分佈情形，如下圖所示：

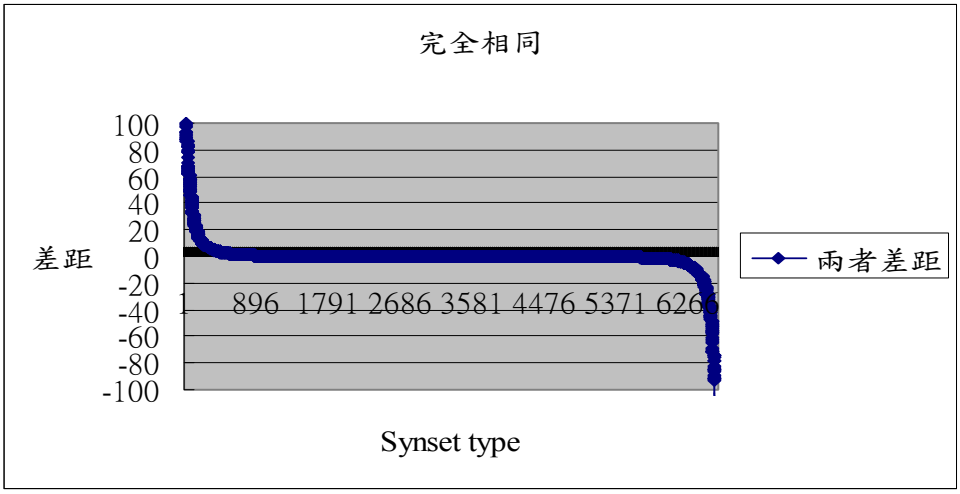

圖 4. 兩岸使用完全相同詞彙的分佈情況

從圖 4 來看，曲線彎曲的前後兩端，代表兩者的差距較大，靠左邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠右邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。在使用完全相同詞彙中，在分析數據呈現上仍有些使用差異的現象，這是值得我們深入

探討的議題。在 CCD 的簡體中文和 CWN 的繁體中文裡，有 6637 筆兩岸使用完全相同詞彙，我們以 Gigaword Corpus 的語料進行檢測，發現中央社/新華社語料所使用分佈差異的平均值爲 0.0143%。Gigaword Corpus 對於兩岸詞彙使用差異分佈在這個平均值內的詞彙，共計有 5880 筆。換句話說，兩岸使用完全相同的詞彙裡，在 Gigaword Corpus 使用狀況較爲相近的有 5880 筆，使用狀況較爲不相同的仍有 757 筆。在中央社與新華社語料中分別有 354 筆和 403 筆。這 757 筆資料是「同中有異」的詞語，值得我們將來進一步分析。

在 6637 個兩岸使用完全相同詞彙中，圖 4 雖然顯示其頻率差距幾乎是零。但是，如果我們由差距最小的第 3076 個詞，依前後各取 30%的(就是第 2153 個詞彙取到第 4144 個詞彙），將差距再放大呈現如圖 5。從圖 5 的差距數值顯示，是非常非常小的。一方面證明兩岸使用這些詞彙的情形，是非常非常接近的。另一方面，當間距放大後，我們看到差異分佈呈平滑的 S 字型，這也與預期中自然語料分佈的狀況相符。

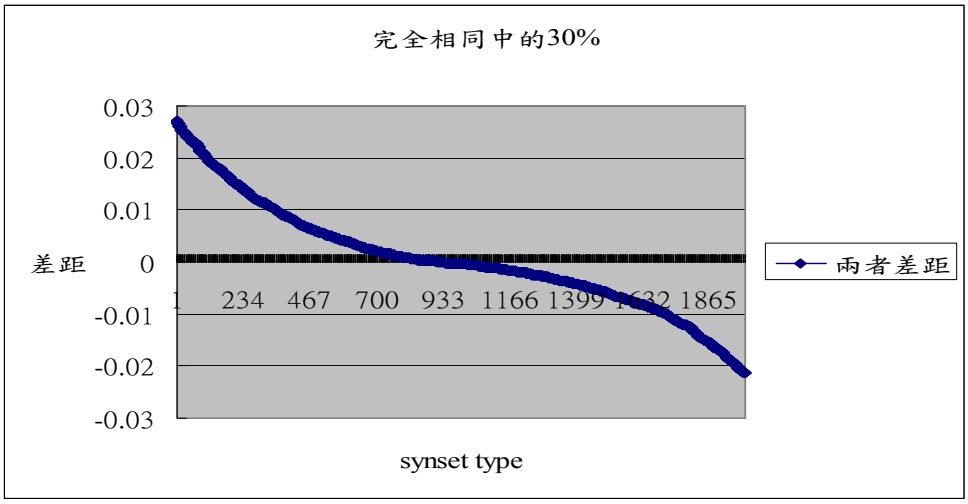

圖 5.兩岸使用完全相同詞彙中，差距最小的 30%的分佈情況

下面表 4，說明兩岸對於相同詞彙，在 Gigaword Corpus 中 CAN 的繁體中文與 XIN 的簡體中文，兩者使用狀況是非常接近的。

表 4. 兩岸使用完全相同詞彙的分佈狀況示例

詞彙	詞頻		附註
	CNA (繁體中文)	XIN (簡體中文)	
酒桶	32 (0.157μ)	20 (0.155μ)	使用狀況非常接近
絲瓜	1380 (6.78μ)	96 (0.748μ)	使用狀況有差異
柳葉刀 雙刃小刀	2 (0.00982μ)	273 (2.13μ)	使用狀況有顯著差異

接著，我們以相同的實驗方法與步驟來檢測兩岸使用完全不同的詞彙，檢測這些資料在 Gigaword Corpus 中呈現的分佈，在此，本文僅取數量較大的名詞和動詞來做比對，並且擷取語料的原則是出現在 CCD 所使用的詞彙，是 XIN 的詞頻大於 CNA 的詞頻；出現在 CWN 所使用的詞彙，是 CNA 的詞頻大於 XIN 的詞頻，其分佈情形，如下圖所示：

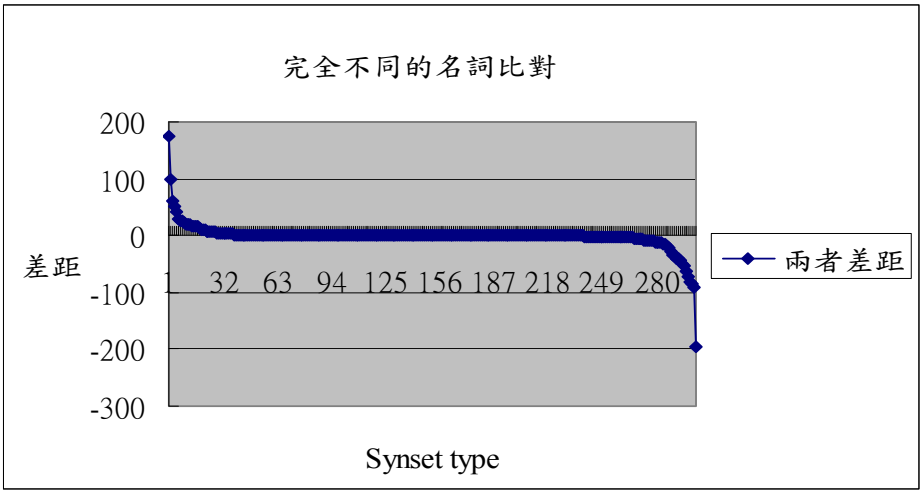

圖 6. 兩岸使用完全不同的名詞詞彙的分佈情況

在兩岸使用完全不同的名詞詞彙裡，共計有 302 筆資料，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們一樣採取兩者差距最小的 30%來檢測，其計有 91 筆資料，從圖 7 的差距數值來看，可以證明，這些不同的詞彙，在所屬的語言系統裡，其使用狀況的獨特性，換言之，同一個詞彙，在繁體中文系統裡，使用的頻率較高，在簡體中文系統裡，使用的頻率較低，反之亦然，而呈現相對之分佈狀態，這樣的情形，在圖 7 的差距數值和表 5 例子中得到驗證。

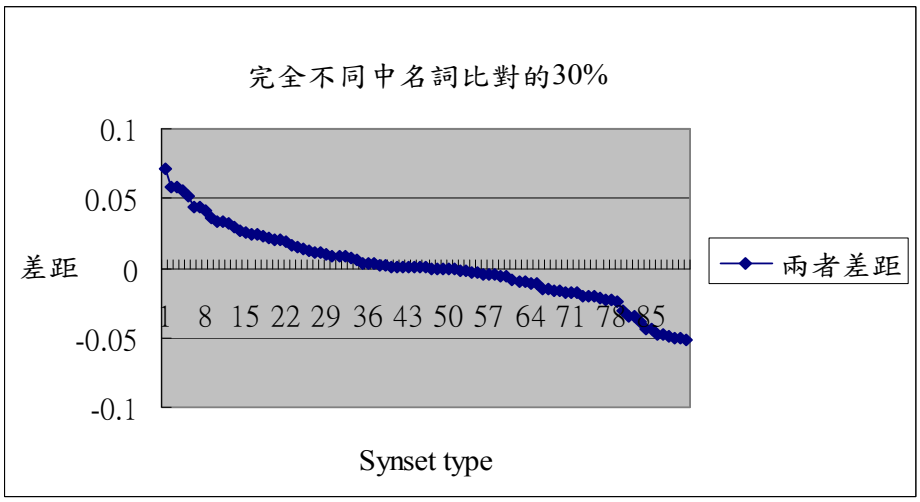

圖 7. 兩岸使用完全不同的名詞詞彙中，差距最小的 30%的分佈情況

下面表 5，說明兩岸對於完全不同的名詞詞彙，在 Gigaword Corpus 中 CAN 的繁體中文與 XIN 的簡體中文，兩者使用的分佈狀況。

表 5. 兩岸使用完全不同的名詞詞彙的分佈狀況示例

詞彙		詞頻				附註
CCD	CWN	CCD		CWN		
		XIN	CNA	CNA	XIN	
風帽	頭罩	10 (0.0779μ)	2 (0.0098μ)	101 (0.4963μ)	37 (0.2882μ)	使用狀況對比明確
雙休日	週末	1383 (10.7736μ)	25 (0.1228μ)	17194 (84.4908μ)	6105 (47.558μ)	使用對比較不明確
屏幕 CRT 屏幕	映像管	3086 (24.04μ)	118 (0.5798μ)	427 (2.0983μ)	1 (0.0078μ)	使用對比較不明確

至於在兩岸使用完全不同的動詞詞彙裡，共計有 461 筆資料(如圖 8 所示)，靠右邊的彎曲曲線部份，是台灣呈現強勢詞彙的現象，靠左邊的彎曲曲線部份，則是大陸呈現強勢詞彙的現象。我們採取一樣的方式來進行檢測，其 30%的資料，共計有 140 筆，從圖 8、圖 9 的差距數值來看，確實可以證明這些使用不同的動詞詞彙，在繁體中文系統與簡體中文系統，有其使用狀況的對比性。

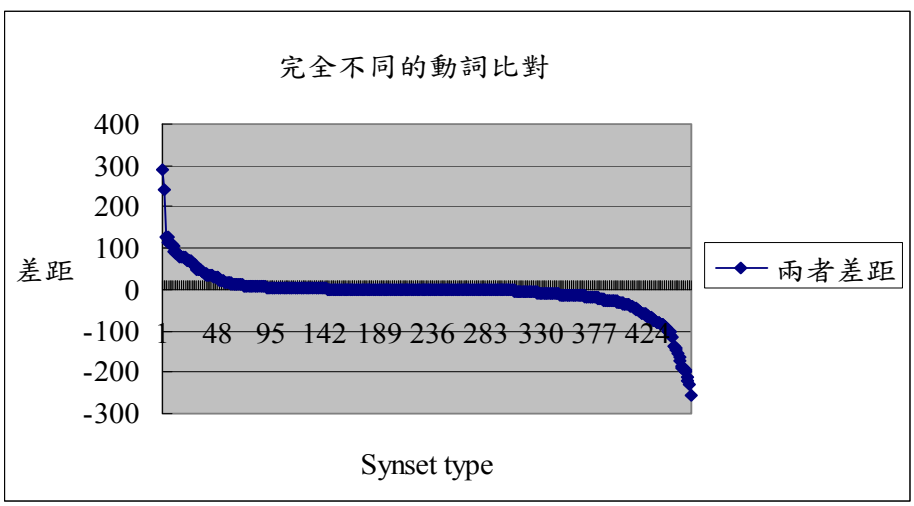

圖 8. 兩岸使用完全不同的動詞詞彙的分佈情況

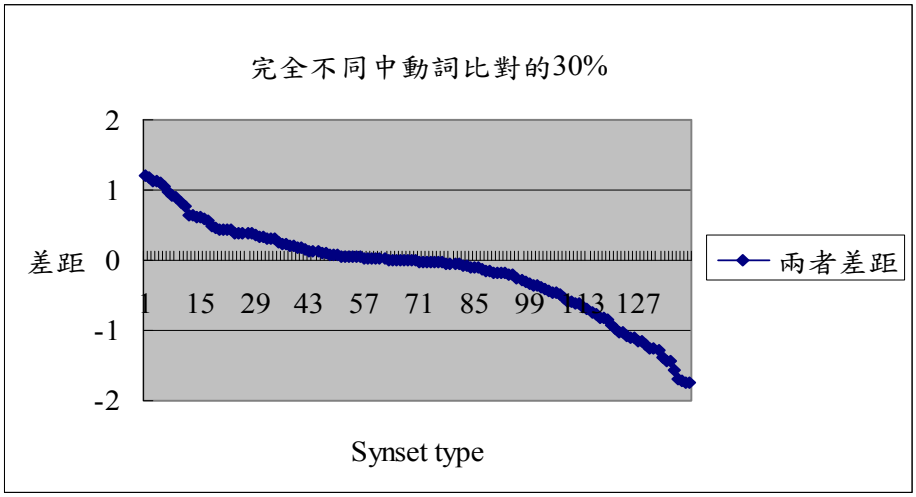

圖 9. 兩岸使用完全不同的動詞詞彙中，差距最小的 30%的分佈情況

兩岸使用完全相同詞彙的平均值是 0.0143%，那麼，理論上，兩岸使用不同詞彙的比例，應該大於這個平均值，倘若小於這個平均值，則有可能是兩岸使用相同概念詞彙時，產生混用的現象。在使用不同的名詞詞彙中，以大陸獨有詞的比例來排序，發現有 174 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 168 筆資料小於這個平均值；在使用不同的動詞詞彙中，以大陸獨有詞的比例來排序，發現有 320 筆資料小於這個平均值；以台灣獨有詞的比例來排序，則有 348 筆資料小於這個平均值。這個數據證實了一個直覺的觀察，就是說兩岸詞彙互相影響滲透的現象日益顯著。以目前的數據看來，台灣的用法影響大陸略強於於大陸的用法影響台灣。

8.2 Google搜尋引擎

對於兩岸詞彙對比研究而言，除了根據具有學術性質的語料庫的資料來進行對比之外，我們也利用一般民眾每天都會使用的網路資料來進行對比，試圖了解民眾在日常生活中對於兩岸詞彙的使用狀況。因此，我們選定以 Google 搜尋引擎所找到的資料做爲對於兩岸詞彙對比研究的對象，進行搜尋後所得到的結果，即可觀察到「所有中文網頁」與「繁體中文網頁」的訊息，雖然沒有直接顯示「簡體中文網頁」的資訊，但在「所有中文網頁」的筆數與結果，可以看到包含「繁體中文網頁」與「簡體中文網頁」的訊息，換句話說，「簡體中文網頁」的筆數與結果，就是「所有中文網頁」的筆數扣掉「繁體中文網頁」的筆數，例如：

(3) 出租車

所有中文網頁：約有 141,000,000 項結果
繁體中文網頁：約有 4,330,000 項結果
簡體中文網頁：約有 136,670,000 項結果

由(3)的查詢結果顯示，「出租車」在簡體中文網頁的使用頻率多於繁體中文網頁的使用，表示「出租車」一詞，一般民眾在大陸地區是比較常使用的；相反地，在台灣地區則是比較少使用的。

有些是關於音譯的詞彙，在兩岸的使用上也有所不同，例如：美國總統 Obama，台灣的音譯名是「歐巴馬」，大陸的音譯名是「奧巴馬」，從 Google 搜尋的網頁資料，如(4)所示，台灣使用「歐巴馬」的筆數多於大陸；反之，大陸使用「奧巴馬」的筆數多於台灣，如此一來，顯示音譯詞彙方面在台灣與大陸皆有獨特使用的對應詞彙，也可以看出兩岸對於音譯詞的差異性及使用的頻率。

(4) 台灣的「歐巴馬 (4,670,000/ 1,230,000)」、
大陸的「奧巴馬 (10,100,000/ 115,900,000)」

再者，在兩岸人民的生活中，也有因爲一些制度、環境、日常生活、習慣而產生出的特殊用語，例如：台灣的「學測」、「免洗筷子」與大陸的「維穩」、「一次性筷子」…等，皆可從 Google 搜尋網頁的資料顯示出兩岸對於某些詞彙的獨用，或者對於相同的概念卻以不同的詞彙來呈現。

儘管 Google 搜尋網頁的資料可以顯示繁體中文網頁的偏用或簡體中文網頁的偏用，以呈現台灣、大陸的兩岸詞彙使用差異性，然而，兩岸人民在各方面的交流、接觸日益頻繁狀況下，彼此使用對方詞彙的狀況也日趨頻繁，以致於漸漸失去所謂台灣繁體中文系統獨用的詞彙或大陸簡體中文系統獨用的詞彙，例如：警察與公安。根據洪 等(洪嘉馡與黃居仁，2008)的研究結果，「警察」一詞應屬於較常被使用在台灣繁體中文系統，但是，在 Google 搜尋網頁的資料卻發現，「警察」一詞亦已在大陸地區廣泛被使用了。

(5) 警察
所有中文網頁：約有 335,000,000 項結果
繁體中文網頁：約有 24,500,000 項結果
簡體中文網頁：約有 310,500,000 項結果

藉由 Google 搜尋網頁的資料，雖然可以呈現出台灣、大陸的兩岸詞彙對比使用狀況，但是，畢竟網路上的資源是比較多元化、也比較具有複雜性，而且，我們無法從網頁訊息得知兩岸網頁總數各若干，所以，按常理推斷，大陸的網頁應該比台灣多很多。此外，兩岸目前交流較頻繁，常有互相引用，無法排除，目前，即可看出大陸用「警察」用法是愈來愈多。這也說明，Google 所搜尋到的資料，僅可以當作目前台灣與大陸一般民眾對於某些詞彙的使用狀況，而無法真正提供兩岸詞彙或世界華語對比的研究。

9. 結論

兩岸詞彙在使用上的相同、不同或些許的差異，甚或混雜使用，在交流頻繁的情形下，已經日趨明顯，如何區分並釐清兩岸詞彙的個別語義架構，又能在其架構下，增加我們對於漢語詞彙語義系統性演變脈絡的理解，是我們從事語言研究者不容忽視的議題。本文藉由 WordNet 所發展出的繁體中文系統 CWN 與簡體中文系統 CCD，進行兩岸詞彙的比對，再將比對過後的詞彙，以收集實際大量語料的 Gigaword Corpus 爲基礎，檢測兩岸在詞彙上使用的現象與分佈狀況；亦可由 Gigaword Corpus 所呈現的狀況，證明繁體中文系統 CWN 與簡體中文系統 CCD 在比對上的正確度與可靠性；也證實了 CCD 和 CWN 將兩岸詞彙對比的使用狀況質化呈現，而 Gigaword Corpus 則是以實際語料來驗證兩岸詞彙對比的使用狀況量化呈現。我們更進一步發現了兩岸共用詞彙有「同中有異」的現象，而對比詞彙也產生了互相滲透影響的現象。值得更深入探討研究。同時，應用具有大量繁體中文、簡體中文的 Google 搜尋網頁的資料進行兩岸人民使用詞彙的對比與差異分析，在此，發現具有學術性質的語料庫，如本文所使用的 Gigaword Corpus 在作爲兩岸詞彙對比研究或世界華語對比研究時，其研究成果與學術價值是比 Google 所提供的資料高很多的。

參考文獻

Chinese Word Sketch Engine: http://wordsketch.ling.sinica.edu.tw/.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Hong, J.-F., & Huang, C.-R. (2006). WordNet Based Comparison of Language Variation - A study based on CCD and CWN. Presented at *Global WordNet (GWC-06)*. 61-68. January 22-26. Jeju Island, Korea.

Huang, C.-R., Chang, R.-Y., & Li, S.-b. (2010). *Sinica BOW: A bilingual ontological wordnet*. In: Chu-Ren Huang *et al*. Eds. Ontology and the Lexicon. Cambridge Studies in Natural Language Processing. Cambridge: Cambridge University Press.

Huang, C.-R., Tseng, E. I. J., Tsai, D. B. S., & Murphy, B. (2003). Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Languages and Linguistics*. *4*(3), 509-532.

Kilgarriff, A., Huang, C.-R., Rychly, P., Smith, S., & Tugwell, D. (2005). *Chinese Word Sketches*. ASIALEX 2005: Words in Asian Cultural Context. June 1-3. Singapore.

Lexical Data Consortium. 2005. Chinese Gigaword Corpus 2.5.: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993). Introduction to WordNet: An On-line Lexical Database. In *Proceedings of the fifteenth International Joint Conference on Artificial Intelligence.*

于江生、俞士汶(2004)。中文概念詞典的結構。*中文信息學報(Journal of Chinese Information Processing)*, *16*(4), 12-21。

于江生、劉揚、俞士汶(2003)。中文概念詞典規格說明。*Journal of Chinese language and Computing*, *13*(2), 177-194。

王鐵昆、李行健(1996)。兩岸詞彙比較研究管見。*華文世界*，第 81 期，台北。

中華民國交通部觀光局 (2011)。*兩岸地區常用詞彙對照表*，http://taiwan.net.tw/m1.aspx?sNo=0016891.

姚榮松(1997)。論兩岸詞彙差異中的反向拉力。*第五屆世界華語文教學研討會*，世界華語文教育協進會主辦，1997 年 12 月 27-30 日，台北劍潭。

南京語言文字網 (2004)。*兩岸普通話大同中有小異*，http://njyw.njenet.net.cn/news/shownews.asp?newsid=367.

洪嘉馡、黃居仁(2008)。語料庫爲本的兩岸對應詞彙發掘(A Corpus-Based Approach to the Discovery of Cross-Strait Lexical Contrasts). *Language and Linguistics*, *9*(2), 221-238, 2008. Taipei, Nankang: Institute of Linguistics, Academia Sinica。

許斐絢(1999)。*台灣當代國語新詞探微*。國立台灣師範大學華與文教學研究所碩士論文，台北。

華夏經緯網(2004)。*趣談海峽兩岸詞彙差異*，http://www.huaxia.com/wh/zsc/00162895.html.

廈門日報(2004)。*趣談兩岸詞彙差異*，http://www.csnn.com.cn/csnn0401/ca213433.htm.

劉揚、俞士汶、于江生(2003)。CCD 語義知識庫的構造研究。*中國計算機大會(CNCC'2003)*。

戴凱峰(1996)。*從語言學的觀點探討台灣與北京國語間之差異*[A Linguistic Study of Taiwan and Beijing Mandarin]。政治作戰學校外國與文學系碩士論文，台北。