語料庫導向之方位短句於固定框架的共現概念統計分析

A Corpus-driven Pattern Analysis in Locative Phrases: A Statistical Comparison of Co-appearing Concepts in Fixed Frames

趙逢毅　August F.Y. Chao
國立政治大學資訊管理學系
Department of Management Information Science
National Chengchi University
fychao.tw@gmail.com

鍾曉芳　Siaw-Fong Chung
國立政治大學英國語言學系
Department of English
National Chengchi University
sfchung@nccu.edu.tw

摘要

中文的方位詞組主要可以前飾詞(以、之)與後綴詞(邊、面、頭)，結合明確的方向指引(如：前後、上下、左右、裡外等)組合而成。這樣的組成在實際使用上，卻會有避免使用或不存在的組合邏輯，同時這樣的現象亦發生在方位短語構成上。本研究試使用計算統計方法，分析在 Sketch Engine 中取得的方位名詞組的概念合成模式。在詞彙概念方面，我們使用具知識層級架構的中文同義詞詞林[1]進行將詞彙的概念探索，並計算方位短句裡所包含的知識概念組成模式，最後試從統計方法上尋得詮釋概念與方位詞組組合模式的實證資訊。在本研究之中，我們使用了資訊度量方法中的互斥資訊(Point-wise Mutual Information, PMI)進行統計分析兩個詞組概念間的相關性，並使用多變數互斥資訊 (Multivariate Mutual Information, MMI)[2]進行三個概念間的相關分析。本研究的統計結果除了解所選用的語料庫中使用方位名詞的情況外，亦從單一及成對出現的語境概念內容 (描述人、物、時空…等在同義詞林中第一階層的名詞)，分析各種方位短語使用的前飾/後綴語的搭配方式，以冀期精萃出來的結果，能對方位詞彙的分析上能提供參考的模式。

Abstract

This paper analyzes synonym groups appearing in fixed frames containing Chinese locative phrases such as [*zái* noun phrase (*yǐ*/*zhī*) *shàng*/*xià*/etc. *biān*/*miàn*/etc.] by using statistical methods. We collected locative phrases from Sketch Engine using 11 monosyllabic locative words and 5 locative compound-formation patterns, and we aligned these compounds with Chinese Synonym Forest [1] before clustering. Different noun phrases were mapped to their collocating synonym groups to as to enable mutual information comparisons between different combinations. When analyzing concept combinations, we used point-wise mutual information to compare two synonym groups, and adopt multivariate mutual information

(MMI)[2] to examine three groups. The results showed that behaviors of using suffixes and prefixes to forming locative nouns in different context (combination of 1 or 2 top level synonym groups), and the statistic results can be used in further analyzing locative nouns in different fields.

關鍵詞：中文方位詞，同義詞詞林，互斥資訊，多變數互斥資訊

Keywords: Chinese Locative Nouns, Chinese Synonym Forest, PMI, MMI.

一、緒論

方位名詞表達了從某個參考物件或事項而產生的方向資訊。在中文裡， Li 與 Thompson[3]指出方位名詞主要是以下列的方式出現：

在 **名詞片語** ~ (方位名詞單元)

在這個結構之中，方位名詞可以是單音字或是雙音字的組合。單音字如**上/下**、**前/後**、**左/右**、**裡/外**、**東/西**、**南/北**及**內/中**等；雙音字組合則是前述的單音字搭配***以***與***之***作爲前飾詞，或是***邊***、***面***與***頭***做爲後綴詞。然而，並非所有的組合在表達方向時都會被使用到。依據盛玉麒在《現代漢語網絡課程》[4]中的方位詞分析(如下表一)，在 14 個方位詞與五種前飾/後綴詞的組合並非經常被用到(或不會被用到)。

表 1 中文方位多音詞組合表

	後綴詞			前飾詞	
	~邊	**~面**	**~頭**	**以 ~**	**之 ~**
上	上邊	上面	上頭	以上	之上
下	下邊	下面	下頭	以下	之下
前	前邊	前面	前頭	以前	之前
後	後邊	後面	後頭	以後	之後
左	左邊	左面	N/A	N/A	N/A
右	右邊	右面	N/A	N/A	N/A
裡	裡邊	裡面	裡頭	N/A	N/A
外	外邊	外面	外頭	以外	之外
東	東邊	東面	東頭	以東	之東
西	西邊	西面	西頭	以西	之西
南	南邊	南面	南頭	以南	之南
北	北邊	北面	北頭	以北	之北
內	N/A	N/A	N/A	以內	之內
中	N/A	N/A	N/A	N/A	之中

許多研究也從不同的觀點對中文方位進行討論。如從參照框架(Frames of Reference)的概念進分析「上」[5]與「前」[6]方位詞的特性，及依意象圖式(Image Schema)來探討《詩

經》中的方位詞「下」[7]與足部動作詞的空間隱喻[8]。上述的研究都只局限在單一方位名詞的探討，並不能將前述表一之中各項方位詞組合進行綜合比較，因此無法較全面了解各方位詞的組合之間有何差異。

本研究中，我們接續先前的研究[9]從 Sketch Engine 裡收集在中文十億字語料庫(Chinese Giga-Word Corpus[1])[10]中包括出現在表一裡的各種組合的短語(在此，我們稱爲方位短語)，並且將所收集得到的短語切段(segmentation)爲詞組後，再透過同義詞詞林[1]轉成其知識架構中的同義詞組代號。在先前的研究之中，我們希望透過視覺化的查詢工具，以呈現較爲明顯的詞群，其中包括了較高出現率(High Frequency)與分群鑒別率(Cluster Discrimination)。在此我們採取不同於先前研究的分析策略，旨在了解詞組所轉換詞群關係間的相依關係。首先依詞組多寡使用不同*互斥*資訊(Mutual Information)的計算原則，以分析在語料庫中每一方位短語裡所存在的知識概念組合模式。在計算*互斥*資訊時，因多變數(由三個詞組所構成的三個概念間)計算不能直接使用兩變數(兩個概念)的*互斥*資訊計算原則，從而我們使用多變數*互斥*資訊[2]來計算。研究結果除了可以提供在同義詞詞林中，知識層級較高的同義詞組在不同的方位短語裡的常見出現模式，亦可擷取常見的中間層級同義詞組在方位短語的使用情況。爲了避免混淆，在本文之中所使用的字句單元大小關係，我們定義爲如下：「方位詞」(如上下、左右…等)，是組成「方位詞組」的重要單元；「方位(名)詞組」可以是單音詞的「方位詞」，或是與前飾/後綴詞組成的多音詞；「方位詞短語」則是符合 Li 與 Thompson[3]所指結構的短語；最後「方位短句」則是由 Sketch Engine 所取得的符合搜尋結果，其句中雖然包括「方位詞短語」，但因受系統限制無法取得完整句子。

本篇論文的架構如下：在第二節，我們先回顧*互斥*資訊計算的相關原則與方法，並說明同義詞詞林的知識概念架構；接著我們報導整個研究過程，其中包括資料收集、處理、同義詞概念轉換與*互斥*資訊相關計算；在第四節中，我們將研究結果則以高階層知識概念進行報告與討論；最後是結論與討論。

二、文獻探討

在這節中，我們說明*互斥*資訊計算的相關原則與同義詞詞林的內容。*互斥*資訊計算是本研究中用來評估概念之間的相*互*關係計算原則，而同義詞詞林則是參考其具系統架構知識分類，以協助我們了解在方位詞短語之中的概念組成原則。

（一）*互斥*資訊 Mutual Information

（1）PMI, Point-wise Mutual Information

從訊息理論(Information Theory)所延用而來的*互斥*資訊計算原則，是指兩發生**事件**之間的相關性參考指標。在此**事件**則是指某單一詞或是單一知識概念(於同義詞詞林之中的同義詞代號)出現在句子之中的情況。例如從 Chinese Giga-word Corpus 中取得的例子"*在/P21 國家_Na 的_DE 邊界/Ncb 之外/Ng*"，我們則稱在句子中可以「找到邊界一詞出現在句子中」的事件發生。接著我們將所有在語料庫中，包括"*之外*"方位詞組的 41612 條短句進行統計(此數字爲 Sketch Engine 回傳的符合搜尋條件的資料總數)，且逐一計次後了解"*邊界*"出現的次數共計有 25 次，我們便可使用條件機率概念表示此**事件**-Cb14A01(Cb14A01, 爲同義詞詞林裡的同義詞群代號，於下 2.2 節中說明。)發生在包

[1] 中文十億字語料庫包括了 2466840 篇台灣中央社(CNA)與大陸新華社(XIN)新聞文本。

括“*之外*”所有句子總數裡的機率爲 $P(\text{"邊界 "}) = \frac{25}{41612}$ ；同理我們亦可以同樣的方式，計算在所有“*之外*”的短句子裡，同時也出現“*的*”的詞組，其結果計算機率的結果爲 $P(\text{"的 "}) = \frac{15632}{2466840}$ 。

而計算兩詞組或兩知識概念之間的 PMI 計算公式如下：

$$PMI\ (x; y) = \log_2 \frac{p(x \cap y)}{p(x)\,p(y)} \quad \ldots (1)$$

所以爲了計算“*的*”與“*邊界*”兩詞組的 PMI 值，我們亦需要尋找同時“*的*”與“*邊界*”出現在短句中次數，即 $P(\text{"的"} \cap \text{"邊界 "}) = \frac{6}{2466840}$ ，則兩者間的 PMI 值爲：

$$PMI\ (\text{"的";"邊界 "}) = \log_2 \frac{\frac{6}{2466840}}{(\frac{15632}{2466840}) \times (\frac{25}{2466840})} = 5.243$$

若我們將“*的*”-“*邊界*”同時比較其它三個詞組“*耳語*”, “*決議*”與“*人口*”：$PMI\ (\text{"的";"耳語 "}) = 7.30$ 、 $PMI\ (\text{"的";"決議 "}) = 5.53$ 、 $PMI\ (\text{"的";"人口 "}) = 3.33$ 時，我們可以知道“*耳語*”與“*的*”之間的相關性高過其它的詞組，也就是“*耳語*”與“*的*”相較於“*決議*”與“*的*”、 “*邊界*”與“*的*”、與“*人口*”與“*的*”在“*之外*”的短句之中出現的。以上的計算原則可以讓我們了解在特定的語料庫中，任兩詞組之間共同出現的相依關係。而此特定語料庫亦可使用有限制的方位詞替代之，以了解在方位詞限制之下兩特定詞組的共同出關係情況。

（2）Multivariate Mutual Information

在使用 PMI 計算相關性時有一限制是，僅能計算兩兩概念或詞組之間的相關性。當面臨三個(或三個以上)事件的相關性比較時，則是透過條件*互斥*資訊(Conditional Mutual Information)值來進行擴展。我們以三個事件的*互斥*資訊爲例，它的數值範圍如下[11]：

$$-\min\{I(X;Y \mid Z), I(Y;Z \mid X), I(X;Z \mid Y)\} \le I(X;Y;Z) \le \min\{I(X;Y), I(Y;Z), I(X;Z)\}$$

其中， $I(X;Y \mid Z)$ 、 $I(Y;Z \mid X)$ 、 $I(X;Z \mid Y)$ 則是各別在 *Z, X, Y* 條件下，計算 *PMI(X;Y), PMI(Y;Z), PMI(X;Z)*的數值之後再取最小值，並與在總體樣本下再計算一次 *PMI(X;Y), PMI(Y;Z), PMI(X;Z)*。這樣的計算要經過 2^n-1 次，十分複雜。從而我們參考[2]的多變數資訊*互斥*計算方法中的具體交*互*資訊(*SI*$_1$, Specific Interaction Information)，做爲三個事件相關性的比較原則。在[2]的計算中，將上式中求多變數*互斥*資訊值 $I(X;Y;Z)$ 化簡爲交*互*資訊(*SI*$_1$)的一般式爲：

$$SI_1(X;Y;Z) = \log \frac{p(x,y)\,p(y,z)\,p(x,z)}{p(x)\,p(y)\,p(z)\,p(x,y,z)}$$

從而此 *SI*$_1$ 即可用於三個(或三個以上)的事件相關分析之中。在此要特別說明，在[2]中

亦定義 SI_2 定爲具體相關性，但在此不使用的原因是：SI_2 是用將多個事件視爲一個整體對特定情境的相關性，一般是使用在資訊挖掘(Information Retrieval)領域中屬性選擇(feature selection)方法上。而本研究裡的取得的方向短句都是具體使用到特定的方向名詞，即所有短句裡的詞組在語料庫中都是我們要研究的對象，並沒有詞組選擇上的問題。

（二）同義詞詞林

同義詞詞林(梅家駒等,1983)收錄來自詞素、詞組、成語、方言詞與古語等詞等共五萬三千多詞彙數，並且依照同義詞分類涵義有系統地區分爲人(A)、物(B)、時間/空間(C)、抽象事物(D)、特徵(E)、動作(F)、心理活動(G)、活動(H)、現象與狀態(I)、關聯(J)、語助(K)、敬語(L)等十二組大類以及若干中類與小類。同類型詞語依照「相對、比較」的排序原則每行依同義/近義程度在同類型中由左自右排列，詞語所屬類別與列舉位置則隱含有作者們的巧思。而電子化的同義詞林擴展版是由哈爾濱工業大學信息檢索研究室(HIT IR Lab)所提供，除了整理、除了刪除舊詞與罕用詞外，並依新聞語料加入常用新詞。此外再對原始的分類也擴展到五層，其中加入「相等、同義」(=)、「不等、同類」(#)及「自我封閉、獨立」(@)等相關涵義。

表 2 同義詞詞林擴展版 例

Cb01A01= 方向 方位 方面 方 向
Cb02A01= 東南西北 四方
Cb03A01= 上 上面 上邊 上頭 上端 頂端 頭 上方
Dm01A01= 政府 內閣 閣 當局 朝
Dm01A05= 朝廷 宮廷 廟堂 王室 朝 廷 皇朝 清廷
Aa01B03# 良民 順民
Bg03A01@ 火

在表 2 中可看得出，同義詞詞林擴展版都保留了分類類別、字彙及同義詞彙，且沒有針對該類別給予明確的類別涵義定義，亦沒有對類別中的詞彙給予明確定義。在分類之中，每行最前面的英文與數字符號代表其同義詞組的編號，以 Cb01A01, Cb02A01, Cb03A01 三組看來，我們可以大體上從語義了解 Cb 一類是方向性的詞彙；同理，Dm01A01 與 Dm01A05 兩組詞組爲不同時代的政府機構名詞。另一個同義詞詞林所存在的問題是，一字/詞多義會同時被歸入不同的詞組之中。例如在 Dm01A01 與 Dm01A05 裡，我們可以看到"*朝*"字被同時列在兩同義詞組裡。最後，同義詞詞林的分類代碼上可以看出，第一碼爲高階層知識概念層級，即前述的十二組大類。而 Cb(方向)即爲時間/空間(C)的中階層子類別，Dm(機構)則爲抽象事物(D)的知識概念中階層子類別。

三、研究方法與結果

（一）資料搜集、處理及分析方法

我們將本研究進行的概念流程圖呈現在次頁的圖 1，詳細說明如下：首先我們先依照方位詞的前飾/後綴組合表，建立合適的方位詞組合搜尋名詞組合，接著到 Sketch Engine

之中的十億字中文語料庫尋找有出現待尋找的方位名詞組。待擷取完成所有的資料之後，我們先進行計算不同方位名詞組合的敘述統計結果，以驗證漢語文法上的前飾/後綴組合文法情況。然後我們便進行短句的過濾，將方位名詞短句切割出來，最後透過同義詞詞林進行同義詞組代碼轉換，完成資料清理的動作。

在建立待尋找的方位名詞組時，我們比較表 1 中文方位多音詞組合表中的組合內容。研究過程中，我們除掉左、右、內、中的方位名詞，因爲從表 1 中可知此 4 個方位單音名詞的組合出現較少(即左/右唯有後綴用法、內/中僅有前飾用法)。在進行擷取的過程裡，我們使用 Sketch Engine 中的 Collocate 功能，並將結果與詞組詞性內容(part of speech)都儲存下來。而過濾方位短句時，本研究以 Sketch Engine 傳回詞組單位(compound segment)爲基準，並以所搜尋的方位名詞開始往前，若三個詞組內有"*在*"出現，則收錄"*在*"至方位名詞之間的所有詞組單位；若三個詞組內沒有"*在*"，則僅收錄最多三個詞組單位，做爲方位名詞短句。在這樣過濾原則下，我們可以確定所收錄到的詞組單位是小於等於三，以利後續分析。

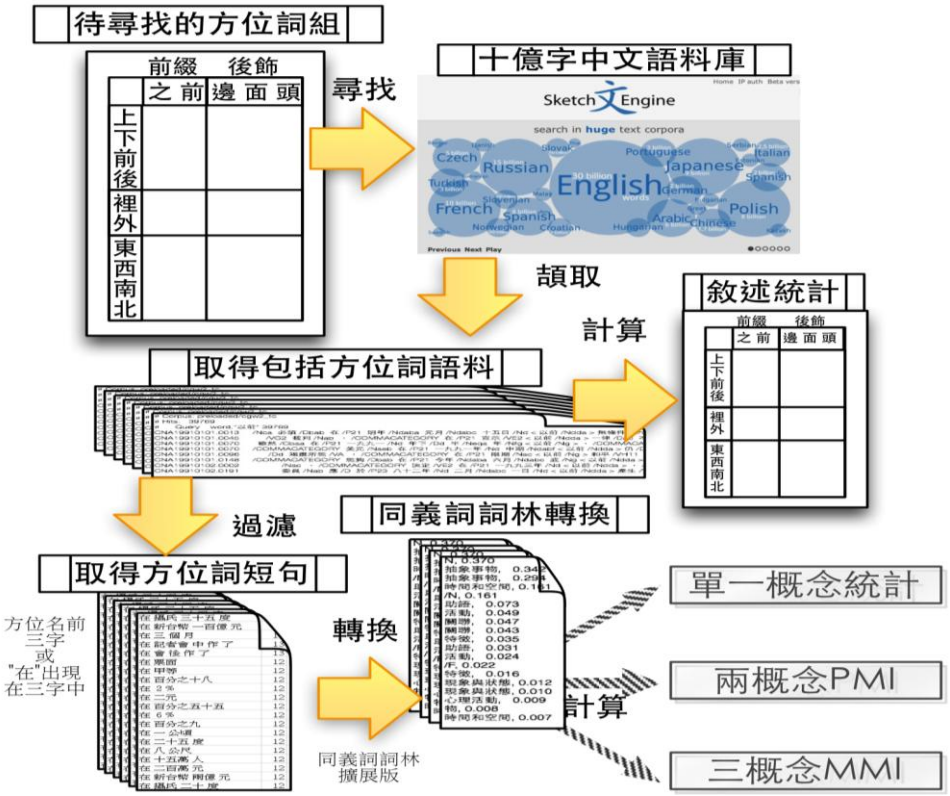


圖 1 研究流程圖

在進行同義詞代碼轉換過程之前，因爲同義詞詞林爲簡體字碼編寫，所以我們使用了維基百科的繁簡分歧詞表進行繁簡轉換。維基百科的繁簡分歧詞表包括了大陸、台灣、香港與新加坡各地的漢語編碼與詞彙互換原則，例如 hardware 一詞在大陸稱作"*硬件*"、台灣則稱做"*硬體*"，使同義詞詞林更切合台灣用語。在進行轉換過程中，我們以 Sketch Engine 傳回詞組單位爲基準，在同義詞詞林中尋找完全符合的同義詞代碼。在先前已提及，同義詞詞林會有多義字同時並分列於不同同義詞組之中，而造成一字有多個同義詞

組代碼。在這裡，我們爲求能精確地找到概念間的組合，所以我們則以排列組合的方式將所有可能的組合情況都羅列在內。最後計算概念關聯時，則會依代碼組合的數目逐一計算互斥資訊計算公式。在計算概念關聯性時，以高階層知識概念爲基準，即將取得的詞組代碼以第一碼(即 A~L)進行計算，以得到一般性概念在方位短句中的組成模式。

（二）研究結果

在前述的計算過程，我們先對取得的語料進行初步的敘述統計討論後，再對不同的知識概念關聯進行探討。初步的敘述統計主要是想了解從十億中文語料庫中所取得的實際語料統計結果，是否能與漢語方位名詞的組成方式有相同。而在後續的概念關聯分析，主要是想了解概念間的組成關係是否會因方位詞組成不同而有所影響。相關的內容分述如下列。而各詞組的概念是以同義詞詞林之中最高階層的同義詞類爲代表，除了我們可以依循同義詞類代碼尋找所屬的高階層代號之外，亦可以避免過多中階層知識概念的交錯影響，而失去焦點。

（1） 方位詞的敘述統計

我們使用了 10 個方位詞(上/下、前/後、裡/外、東/西、南/北)，及 5 個不同的組合方式(前飾詞：以、之；後綴詞：邊、面、頭)，到十億中文語料庫中擷取方位名詞所存在的句子，並透過前述的過濾原則(由方位詞爲基準，向前計算，遇"*在*"即停，最多三組)，進行敘述統計分析，其結果如下表：

表 3 從十億中文語料庫中擷取的方位短句分佈

	後綴詞			前飾詞		計次	佔總比率
	~邊	~面	~頭	以 ~	之 ~		
上	**11**	788	51	1557	15559	17966	15%
下	**6**	169	3	8273	7547	15998	13%
前	**3**	1085	154	31618	12596	45456	38%
後	**9**	1028	215	22051	3751	27054	23%
裡	**9**	1086	97	**0**	**0**	1192	1%
外	**28**	1254	154	4370	1918	7724	7%
東	139	**33**	**0**	**0**	424	596	1%
西	147	**66**	**3**	**0**	874	1090	1%
南	118	**20**	**0**	**0**	390	528	0%
北	199	**78**	**0**	**0**	731	1008	1%
	669	5607	677	67869	43790	118612	
	1%	5%	1%	57%	37%		

在表 3 我們將數字較少的區塊特別以粗線條框出，並比較表 1 漢語文法中所指出的方位詞組合原則，我們得到下列的結果：(a) 佔有比率分析：因爲我們所選用的語料是用來

報導時事的文本資料，所以統計結果在方位詞的使用上偏重於上/下、前/後 4 個方位名詞，此外前飾詞以/之的用法相對於後綴詞較爲頻繁。而在新聞語料之中，地理的方位名詞(東、南、西、北)的使用情況則相對於其它地方用法則少了很多(約在都不到 2%)。此外使用"*外*"的情況相對於"*裡*"的用法上較頻繁。(b) 上/下、前/後分析：這四個方位詞在組合成爲多音方位詞時，主要是以前飾字(*之*/以)爲主，而後綴詞的使用上主要是以"*面*"爲主。(c) 裡/外的分析：我們所得到的結果在"*裡*"的沒有前飾用法上是與表 1 裡的預期是相同的。而以"*外*"字來說"*以外*"相對於"*之外*"較爲白話、通俗，這與我們所選用的文本特性亦有相關。(d) 東/南/西/北的分析：這四個方位詞在新聞語料的使用上，不常出現。而在使用的時候，則會以前飾詞"*之*"及後綴詞"*邊*"進行組合。

綜合來看，當新聞語料在進行報導的時候，作者在構思方位詞上/下、前/後、裡/外時，會直接使用"*以*"而不會使用"*之*"，因爲在描述事件報導的是以通俗考量。但在說明東/南/西/北時，此時構詞的行爲則會與前述的結果相反，而使用前飾詞"*之*"，在此我們推論是因爲"*之南*"相對於"*以南*"所指的隱喻地理範圍較小(求謹慎)。

（2）方位詞短語組成模式—高階層知識概念

接著我們將所取得的語料依照同義詞詞林進行同義詞群組轉換後，並以其較高層級的同義詞類代碼(即代碼中的第一碼)進行概念相關性的計算。概念相關性的計算會依照詞組單位的多寡而決定該使用那一種計算方式：單一詞組爲計次、兩詞組使用 PMI 計算、三詞組則使用 MMI 進行統計比較。此外因爲我們所選用的語料庫特性，所以本研究只著重在上/下、前/後、裡/外等六個方位詞，其它的方位詞因爲樣本比例太少在本研究不討論。

2.1 單一知識概念分析

首先我們對在方位短句中包括有單一同義詞代碼的資料進行統計分析，並列出其同義詞代表意義與該代碼佔所有此資料類別(僅出現單一同義詞代碼)的比例如下表 4。

同義詞林中的*助語*詞類包括了*疏狀*、*中介*、*聯接*、*輔助*、*呼嘆*及*擬聲*六類，而"*的*"則是被歸在*輔助*類中。所以從從表 4 中可發現*助語*類常常在與方位名詞連用，如"*的後面*"。而"*之*"與"*的*"同時都有修飾方位詞的關係，所以從表 4 裡亦可發現這兩個詞不會重覆使用在方位短語之中。接著我們從表 4 的縱向來分析不同的方位詞組成方式，"*邊*"、"*面*"與"*頭*"的後綴詞組成方式裡物出現有次數比較在前飾詞"*之*"與"*以*"裡面多出許多；同理，在前飾詞"*之*"與"*以*"裡面*抽象事物*、*活動*、*特徵*也相較於"*邊*"、"*面*"與"*頭*"的後綴詞組成計次中要多出許多。從這點亦可以看出，"*邊*"、"*面*"與"*頭*"的後綴詞組成方式會依照參考點爲實體存在的名詞組成方位短句；而*活動*、*特徵*詞組等不是實體存在的名詞則比較傾向與前飾詞"*之*"與"*以*"組成。

而從橫向討論來看，雖然上/下都可以用來述描*物*與*抽象事物*詞類，但*特徵*類只會出現在"*之上*"與"*上邊*"的組合；而"*下*"則是會出現*時間和空間*與*活動*的詞類。而在橫向的前/後中，我們發現針對*時間和空間*(例如："*學期*")的方位短句組成模式，有"*之前*"、"*之後*"與"*後邊*"三種情況，但沒有"*前邊*"的使用方式。同樣的情況如*活動*(例如："*革命*")在裡/外的組合之中，僅有"*之外*"與"*裡邊*"兩種情況。從這樣的分析，我們得到(a)方位名詞的組合不一定都是對稱的。也就是如前面所指出的*時間和空間*(例如："*學期*")的方位短句組成不會有"*前邊*"的情況，就算方位名詞組成文法是的正確的。(b)特定詞類會有習慣上的使用原則。如前述*活動*在裡/外的組合之中僅有"*之外*"與"*裡邊*"組成情況與*特徵*類

只會出現在"*之上*"與"*上邊*"的組合。

表 4 方位名詞單一概念之共現比例(取前 3)

	~邊		~面		~頭		以 ~		之 ~	
上	特徵	0.5	助語	0.6	助語	0.5	抽象事物	0.4	抽象事物	0.9
	助語	0.4	物	0.2	物	0.3	助語	0.2	物	0
	物	0.1	抽象事物	0.1	抽象事物	0.2	關聯	0.2	特徵	0
下	助語	0.5	助語	0.5	物	1	抽象事物	1	抽象事物	0.5
	時間和空間	0.2	物	0.4			關聯	0	活動	0.3
	關聯	0.1	抽象事物	0.1			活動	0	關聯	0.1
前	助語	1	助語	0.4	助語	0.6	助語	0.5	特徵	0.4
			抽象事物	0.3	抽象事物	0.2	時間和空間	0.4	活動	0.3
			物	0.2	物	0.1	抽象事物	0	抽象事物	0.1
後	物	0.9	助語	0.6	助語	0.6	時間和空間	0.6	活動	0.4
	時間和空間	0.1	物	0.2	抽象事物	0.2	助語	0.2	抽象事物	0.3
			抽象事物	0.1	現象與狀態	0.1	活動	0.1	現象與狀態	0.1
裡	抽象事物	0.4	抽象事物	0.4	抽象事物	0.4				
	物	0.2	物	0.3	物	0.2				
	活動	0.2	助語	0.2	助語	0.2				
外	抽象事物	0.4	助語	0.5	抽象事物	0.5	抽象事物	0.8	抽象事物	0.7
	助語	0.3	抽象事物	0.3	助語	0.3	時間和空間	0.1	活動	0.2
	物	0.2	物	0.1	物	0.1	物	0.1	助語	0

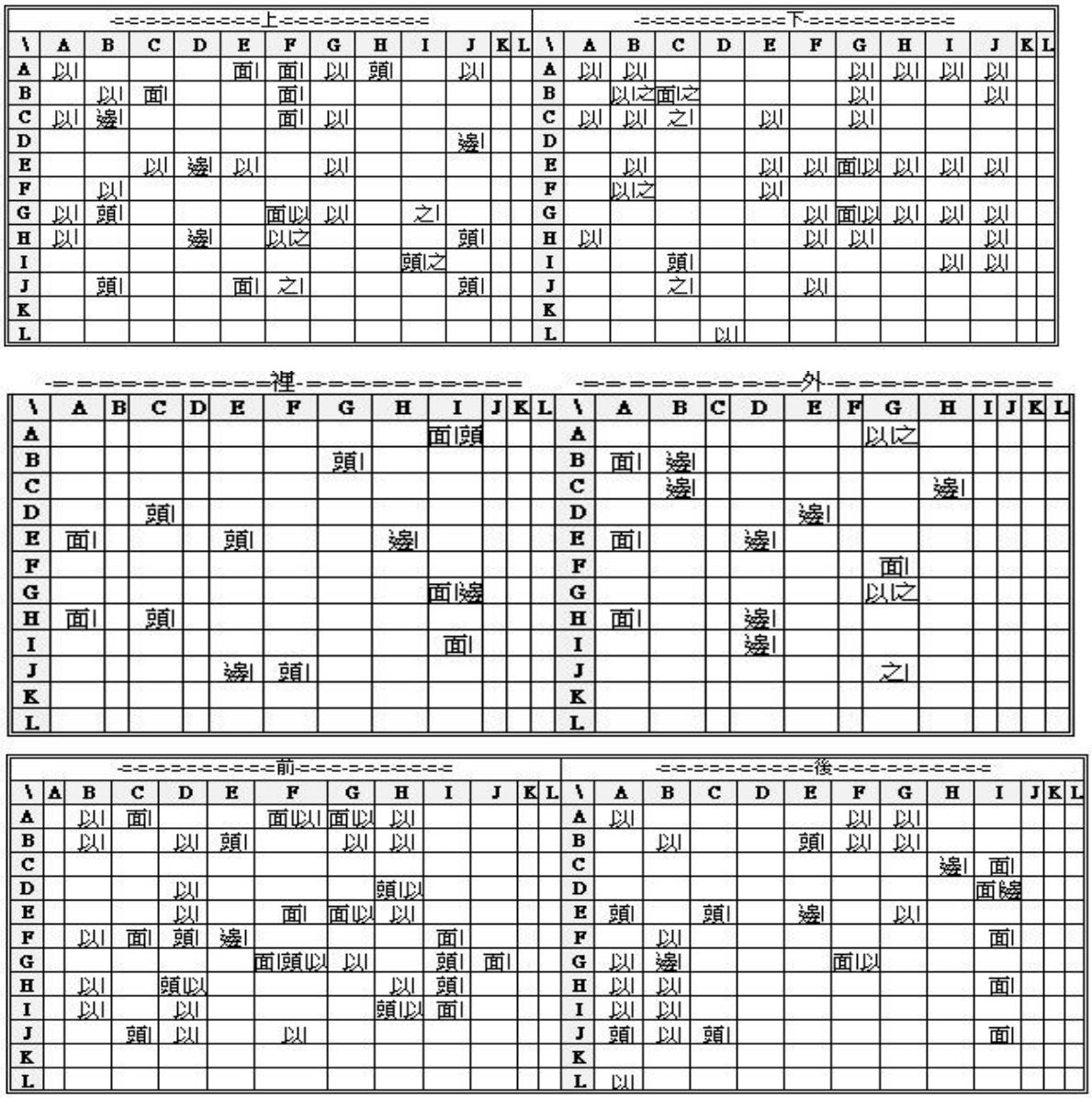
*表中的數值爲出現頻率

2.2 兩知識概念分析

接著，我們將包括兩個同義詞代號的方位短語進行相關分析。配合前飾與後綴詞的組成原則，我們分析結果依上/下、前/後與裡外分列如下表 5。

表 5 中的 A~L 是同義詞類代號中的第一碼，分別爲人(A)、物(B)、時間/空間(C)、抽象事物(D)、特徵(E)、動作(F)、心理活動(G)、活動(H)、現象與狀態(I)、關聯(J)、語助(K)、敬語(L)等十二組大類。在每張子表的左方軸是距離方位名詞二個位置的詞組(即 window size 爲-2，往前數第二個詞組)，而上面軸是距離方位名詞一個位置的詞組(即 window size 爲-1，往前數第一個詞組)。所以我們用"*上*"子表爲例，左方爲 A 上方爲 A 的交集出現"*以*"的情況，是表示方位短語組合必需是第一組(方位名詞前二位置)爲人詞類之下的詞組與第二組(方位名詞前一位置)亦爲人詞類之下的詞組，最後方位名詞的爲"*以上*"的情況，例如："*助理(A-人) 教授(A-人) 以上*"、 "*主任(A-人) 檢察官(A-人) 以上*"的方位短句。

表 5 包括兩概念的方位詞組成相關表



-=-=-=-=-=-=-=-=-=上-=-=-=-=-=-=-=-=-												
\	A	B	C	D	E	F	G	H	I	J	K	L
A	以I				面I	面I	以I	頭I		以I		
B		以I	面I			面I						
C	以I	邊I				面I	以I					
D										邊I		
E			以I	邊I	以I		以I					
F		以I										
G	以I	頭I				面I以I	以I		之I			
H	以I			邊I		以I之				頭I		
I								頭I之				
J		頭I			面I	之I				頭I		
K												
L												

-=-=-=-=-=-=-=-=-=下-=-=-=-=-=-=-=-=-												
\	A	B	C	D	E	F	G	H	I	J	K	L
A	以I	以I					以I	以I	以I	以I		
B		以I之	面I之				以I			以I		
C	以I	以I	之I		以I		以I					
D												
E		以I			以I	以I	面I以I	以I	以I	以I		
F		以I之			以I							
G						以I	面I以I	以I	以I	以I		
H	以I					以I	以I			以I		
I			頭I						以I	以I		
J			之I			以I						
K												
L				以I								

-=-=-=-=-=-=-=-=-=裡-=-=-=-=-=-=-=-=-=												
\	A	B	C	D	E	F	G	H	I	J	K	L
A									面I頭			
B							頭I					
C												
D			頭I									
E	面I				頭I			邊I				
F												
G									面I邊I			
H	面I		頭I									
I									面I			
J					邊I	頭I						
K												
L												

-=-=-=-=-=-=-=-=-=外-=-=-=-=-=-=-=-=-												
\	A	B	C	D	E	F	G	H	I	J	K	L
A							以I之					
B	面I	邊I										
C		邊I						邊I				
D					邊I							
E	面I			邊I								
F							面I					
G							以I之					
H	面I			邊I								
I				邊I								
J							之I					
K												
L												

-=-=-=-=-=-=-=-=-=前-=-=-=-=-=-=-=-=-													
\	A	B	C	D	E	F	G	H	I	J	K	L
A		以I	面I			面I以I	面I以I	以I				
B		以I		以I	頭I		以I	以I				
C												
D				以I				頭I以I				
E				以I		面I	面I以I	以I				
F		以I	面I	頭I	邊I				面I			
G						面I頭I以I	以I		頭I	面I		
H		以I		頭I以I				以I	頭I			
I		以I		以I				頭I以I	面I			
J			頭I	以I		以I						
K												
L												

-=-=-=-=-=-=-=-=-=後-=-=-=-=-=-=-=-=-													
\	A	B	C	D	E	F	G	H	I	J	K	L
A	以I					以I	以I					
B		以I			頭I	以I	以I					
C								邊I	面I			
D									面I邊I			
E	頭I		頭I		邊I		以I					
F		以I							面I			
G	以I	邊I				面I以I						
H	以I	以I							面I			
I	以I	以I										
J	頭I	以I	頭I						面I			
K												
L	以I											

直覺上可以看到前飾詞“*以*”充滿了表 5 之中各子表內容，而漢語文法上不存在的“*以*”與“*裡*”的組成情況，亦可在表 5 之中觀察到。而針對“*外*”的組合上，也僅只有前飾詞與“*之*”、“*外*”有出現在十億語料庫之中。這樣的例子如：“*我們(A-人) 期望(G-心理活動) 以外*”、“*感到(G-心理活動) 滿意(G-心理活動) 以外*”、“*超出(J-關聯) 控制(G-心理活動) 之外*”等。在此需要說明的事，“*控制*”詞組同時存在多義涵且分在同義詞詞林中的三個類別分別爲(Gb-心理活動)，(Hc-行政管理)與(Je-影響)之中。但因爲我們無法了解完整句子之中“*控制*”詞組的明確語義，所以我們將原本的方位短句擴展成爲三組“*控制*”語義，並與“*超出*”(Jb-異同)進行 1*3 次的兩詞組間 PMI 計算。

而在縱向的分析中，*A-人*詞類在上/下、前後等情況都是使用前飾詞“*以*”來組成，而在裡/外的方位詞中，則是以“*面*”在組成方位詞短句，如：“*安全(E-特徵) 考慮(G-心理活動) 以外*”、“*培育 (H-活動) 人才(A-人) 之外*”。同樣特例如當方位名詞的前一組詞爲 *G-*

*心理活動*與六個方位名詞的組合上，多數是使用"*以*"做爲前飾詞，除了"*外面*"、"*下面*"、"*前面*"與"*裡頭*"。期中與"*下*"的組合方式，因"*以*"與"*下*"沒有此用法所以僅能使用"*下面*"組成。而 *G-心理活動，與*"*外面*"的組合只有兩種情況發生在語料庫中，分別爲"*嚷嚷(F-動作) 想到(G-心理活動) 外面*"與"*探頭(F-動作) 看(G-心理活動) 外面*"(此處的看是同義詞詞林表中的"*Gb02B01= 認爲 以爲 覺得 道 看 當 覺著*"。其它"*前面*"與"*裡頭*"例子如下："*她(A-人) 聽到(F-動作) 前面*"、"*她(A-人) 認爲(G-心理活動) 裡頭*"等。

2.3 三知識概念分析

在三組知識概念的組合中，我們使用多變數*互斥*資訊計算裡的交*互*資訊(SI_1)做爲評量標準，以避免繁雜的多變數條件機率下*互斥*資訊的比較計算。最後，因爲三種知識概念的排列組合結果非常多，所以僅保留在不同的方位名詞所組成的短句中，所有組合計算結果的平均交*互*資訊(SI_1)高於 0.8 的結果進行討論，如下表 6。

表 6 三組知識概念下交*互*資訊(SI_1)>0.8 之結果

方向詞	往前第三位詞組	往前第二位詞組	往前第一位詞組	方向名詞	(SI_1)數值
上	抽象事物	活動	活動	以上	2.1
		特徵			1.9
	活動	助語			0.9
下	關聯	動作	物	下面	0.9
	抽象事物	抽象事物	**時間和空間**	以下	1.7
	物	活動	助語		1.6
	物	**時間和空間**	助語		0.9
	關聯	特徵	**時間和空間**	下邊	1.9
	時間和空間	抽象事物	助語		1.2
前	抽象事物	助語	**時間和空間**	之前	2.6
	現象與狀態	特徵			1.7
	抽象事物	現象與狀態			1.4
	特徵	抽象事物			1.3
	活動	特徵			1.3
	助語	**時間和空間**	助語		1.2
	特徵	活動	**時間和空間**		1.1
	時間和空間	**時間和空間**			1.1
	助語				1
	人	助語			0.9
	現象與狀態	**時間和空間**			0.9
	抽象事物		助語		0.9
	特徵	助語	**時間和空間**		0.9
	抽象事物	助語	關聯	以前	0.9
後	時間和空間	時間和空間	現象與狀態	後頭	1.5
裡	助語	現象與狀態	抽象事物	裡邊	1.4
外	抽象事物	助語	物	以外	0.8
	時間和空間	助語	關聯	外邊	1.3

與 2.2 相同，我們僅討論上/下、前/後、裡/外這六個方位名詞的知識概念組成模式，因爲東/南/西/北在我們所使用的語料庫中，使用情況較少。表 5 中方位詞組排列較高交*互*資訊(SI_1)的結果，並依知識概念在方位短句中出現順序：方位詞前第三、第二、第一位

詞組，進行排列。在方位詞"*上*"之中三知識概念組合模式較高 SI_1 結果，可以看到是"*(D-抽象事物) (H-活動)或(E-特徵)或(K-助語) (H-活動) 以上*"，其中"*(H-活動)或(E-特徵)或(K-助語) (H-活動)*"組合模式是*(H-活動)*詞組的明確性(specific)說明，如"*組織(D-抽象事物) 負責(H-活動) 接受(H-活動) 以上*"、"*簡報(D-抽象事物) 後(E-特徵) 作(H-活動) 以上*"、"*目標(D-抽象事物) 時(K-助語) 作(H-活動) 以上*"。所以我們可以知主要"*以上*"都是用在描述*活動*的詞組方位性。且在"*上*"組合模式之中，僅"*以上*"的"*(H-活動)*"使用情況較其它前飾後綴的方位用語組合來說較爲固定。方位詞"*下*"可分成"*下面*"與"*以下*"/"*下邊*"兩種類別，其中"*下面*"的情況較固定的組合模式是"*(J-關聯) (F-動作) (B-物) 下面*"，如"就是*(J-關聯) 對(F-動作) 海床(B-物) 下面*"。此外"*以下*"/"*下邊*"的組合模式就都會包括 *C-時間和空間*或 *H-活動*等，如"*大陸(B-物) 面臨(H-活動) 了(K-助詞) 以下*"、"*一(J-關聯) 個(E-特徵) 村莊(C-時間和空間) 下邊*"等。方位名詞"*之前*"則是很明確的出現*(C-時間和空間)*詞組在不同的組合上，如"*在(K-助詞) 一月(C-時間和空間) 十五日(C-時間和空間) 之前*"等。其它的方位名詞組合也者有知識概念專屬的使用情況，在此不一一綴述。

五、結論與討論

本研究試以統計語料庫的觀點，討論方位詞在不同的組合情況下，方位短語的知識概念組合情況。而知識概念，在本研究之中是以同義詞詞林的同義詞組架構爲主，主要是因爲同義詞詞林包括的詞組與其知識架構是較爲完備的參考基準。接者以十億中文語料庫中，我們擷取的其包括了不同組合的方位詞，除了透過敘述統計與漢語文法中方位詞組成原則比較外，亦對方位詞短語中知識概念組合原則透過相關性計算後，分析方位名詞的組合方式與短語中知識概念之間的關係。在相關性的計算上，因爲 PMI 計算無法直接計算多概念間相關，所以我們引用多變數的交互資訊(SI_1)做爲評量標準。在單一、兩組、三組知識概念的分析結果之中，我們透過語料庫中新聞文本的統計資料佐證，更清楚地了解方位名詞在使用前飾詞"*之*"、"*以*"與後綴詞"*邊*"、"*面*"、"*頭*"在新聞文本的使用習慣上的差異。

透過相關性統計資料僅能提供在許多知識概念裡的選出特徵值較高的組合，並沒辦法完全透過統計資料完整解釋方位短語中所有知識概念在語義上的組合情況。此外同義詞詞林的簡體編撰、分類架構與多義詞在其架構的定位上，亦會造成本研究結果的偏差。再者，本研究所使用的是搜集新聞語料的，所以這些新聞語料的內容亦會讓使用上的用法與習慣存有偏差。最後，使用多變數的交互資訊(SI_1)做爲評量標準缺少更多的實驗結果的驗證資訊，這亦是本研究的問題所在。然而在華語教學在方位語的需求，與協助方位名詞在訓詁的領域上，本研究則提供分析方向供研究者參考。

Acknowledgements

This research is supported by National Science Council grant 101-2410-H-004-176-MY2 directed by Siaw-Fong Chung.

參考文獻

[1] 梅家駒, 竺一鳴, 高蘊琦, 殷鴻翔. 同義詞詞林. 香港: 商務印書館, 1984.

[2] Cruys, T. Van de, “Two Multivariate Generalizations of Pointwise Mutual Information”, Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo’2011), pp. 16-20, 2011.

[3] Li, C. N. and Thompson, S. A., “Mandarin Chinese: A functional reference grammar”, University of California Press, 1989.

[4] 盛玉麒, Modern Chinese online course 現代漢語網絡課程, [Online]. Available: http://www.yyxx.sdu.edu.cn/chinese/, visited on 2013/06/01.

[5] 許雅臻與戴浩一，“空間方位詞「上」在三個參照框架中的分析”, 國立中正大學語言學研究所未發表論文, 2001.

[6] 梁闞元與王松木, “從參照框架分析現代漢語前之意象圖式”, 國立成功大學華語文教學研究所未發表論文, 2010.

[7] 黃翠芬, “從意象圖式探測詞義發展— 以《詩經》方位詞「下」爲例”, 朝陽人文社會學刊, vol. 9, no. 1, pp.235-266, 2011.

[8] 邱湘雲, “漢語足部動作詞的空間隱喻”, 彰化師範大學文學院學報, vol. 6, pp. 225-242, 2012.

[9] Chao, A. and Chung, S. F., “A Lexico-Semantic Analysis of Chinese Locality Phrases - A Topic Clustering Approach”, Forthcoming in Generative Lexicon and Distributional Semantics 6th International Conference, Italy.

[10] Ma, W. Y. and Huang, C. R., “Uniform and effective tagging of a heterogeneous giga-word corpus”, In 5th International Conference on Language Resources and Evaluation (LREC2006), pp. 24-28, 2006.

[11] Sunil, S., “A review on multivariate mutual information”, Univ. of Notre Dame, Notre Dame, Indiana, 2008.