中 華 心 理 學 刊
民98，51卷，4期，415-435

Chinese Journal of Psychology
2009, Vol. 51, No. 4, 415-435

中文語意空間建置及心理效度驗證：以潛在語意分析技術爲基礎

陳明蕾[1]　王學誠[2]　柯華葳[1]

[1]國立中央大學學習與教學研究所
[2]麻州州立大學波士頓校區資訊科學學系

論文編號：08A17；初稿收件：2008 年 11 月 19 日；完成修正：2009 年 9 月 25 日；正式接受：2009 年 10 月 30 日
通訊作者：柯華葳　320 中壢市五權里 2 鄰中大路 300 號　國立中央大學學習與教學研究所（hwawei@ncu.edu.tw）

近年來，運用大型語料為基礎，進行比對及描繪詞彙間的語意關係，是心理語言學中新興的研究取向。本文採用潛在語意分析技術，建立一個能表徵中文詞彙間語意關聯性的語意空間。此一語意空間的特色是可以將隱藏在詞彙背後的語意關連性呈現出來，而且也能以向量的方式，進行詞彙與句子、詞彙與文件之間語意關聯性的比對。實驗結果顯示，以潛在語意分析技術所建置的中文語意空間，能反應中文讀者內在心理詞彙表徵間之語意關聯性。

關鍵詞：語意空間、語意關聯性、潛在語意分析

緒論

理解生活環境中的語彙意義，是人類相當重要的認知行爲之一。人對語彙的理解，不會只來自單一的字詞，而是從一連串的詞彙而來。例如，在對話的情境中，說話的人若只說「月亮」一詞，聽話的人通常無法判斷在此對話情境下「月亮」的意思是什麼。如果說話的人是說「今晚的月亮眞美」，聽者即可以理解說者此時所欲表達的是「月亮」在今晚的狀態。句意脈絡或是文章脈絡之所以可以幫助聽者或讀者有效理解單一字詞的意義，乃是因爲單一字詞往往有不同的意義，唯有在該字詞與其它字詞一起出現時，人們才能更正確的理解該字詞的意義。例如，「月亮影響了地球的潮汐變化」與「今晚的月亮眞美」，這兩句話裏的「月亮」對於讀者而言，其內在表徵所觸發的意義並不完全相同。

掌握衆多詞彙所隱含錯綜複雜的語意關係，一直是人類學習語言歷程中的重要一環。Landauer（2006）認爲人捕捉詞彙間語意關係的歷程，很類似理解地圖上任一城市在此地圖上的意義，是藉由了解城市間彼此對應的位置。例如，地圖上台中在台北的南邊，基隆在台北的北邊，讀者即可以判斷出基隆也在台中的北邊。同樣的，藉由比對詞彙在句意脈絡下的對應位置，我們也能慢慢捕捉到不同句子裏兩兩詞彙的語意關聯性，例如下面的三句話：(1)「學校裏面有很多認眞教學的老師」、(2)「老師正一筆一畫的教學生學習新字的筆順」以及(3)「這星期眼科醫師來學校幫忙做視力檢查」。雖然句二並沒有出現「學校」這個詞彙，但藉由老師這個詞彙的重疊，讀

誌謝

感謝國科會千里馬計劃（計劃編號：NSC095-2917-I-008-001）與教育部「發展國際一流大學及頂尖研究中心計畫-中文詞彙學習與讀寫歷程探究之系列研究」（五年五百億專案計劃）對本研究的經費支持。此外，本研究的完成也要感謝Professors Walter Kintsch and Eilleen Kintsch所給予的指導與協助。

者所建置出的內在語意表徵中，學生和學校會有著一定程度的語意關聯性。同樣的，句三雖然沒有出現老師和學生，藉由「學校」這個詞彙的聯繫，仍然能讓讀者將「醫師」、「學生」及「老師」的語意關係建置起來。此一語意建置的歷程，就像人將日常生活中所接收的語料，放進一個多維度的向量空間，然後，以向量的方式表徵字詞在該空間中的位置，之後，即再藉由兩兩向量的比對，來呈現兩兩字詞間語意相同及相異之處。因此，Landauer 和他的同事採用一種可以用來分析大量語料資訊的擷取技術－潛在語意分析（Latent Semantic Analysis, LSA）描繪出一個能反映人心理表徵的語意空間（Landauer, 2002; Landauer, Foltz, & Laham, 1998）。

潛在語意分析和語意空間

潛在語意分析一開始是由Deerwester等人（Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990）所提出，其基本概念是先以二維的矩陣空間表徵文字和原始文件間的關係，再利用奇異值分解（Singular Value Decomposition, SVD）的方式，找出字詞對應文件的語意結構，此一技術的特色是將許多原本字面看不到的資訊呈現出來，因此能大幅提昇資訊擷取的有效性，而這也是Deerwester將此分析方式稱之爲「潛在」語意分析的原因。Landauer、Foltz及Laham（1998）則將此技術引進心理語言分析的研究領域。若有一個大型的語料來源能適當的反映人所擁有的語彙知識，即可以藉用LSA的技術建置出一個能反應這些語彙知識背後語意關係的語意空間。Landauer與他的同事以葛羅里學術百科全書（Grolier Encyclopedia）做爲他們建置語意空間的語料來源。首先，他們建立一個有60,768個詞（term），以及30,473個段落的共生矩陣（term-to-document co-occurrence matrix）。此一共生矩陣，呈現的是百科全書中每個關鍵詞在每個段落中的出現次數，完全沒有涉及關鍵詞間的語意關係。接著利用SVD的數學演算方式以及維度化約（Dimension Reduction）方式，將此一大型矩陣簡化成一個約有300個維度的語意空間後，就可以將隱藏在關鍵詞背後的語意關連性計算出來（Landauer et al., 1998）。

LSA所建置的語意空間，不論是單獨的詞、整段的段落、或是整篇的文章，都是以向量的方式呈現該詞彙、段落或是文章在語意空間裡的相對位置。兩個文件間的相似度可以用兩向量角度的餘弦值（cosine value）來表示。餘弦值愈大，表示兩個向量的夾角愈小。兩個向量的夾角愈小，表示這兩個向量愈接近。在語意空間中，所顯示的意義是這兩個向量的語意愈相似。因此，不論是字詞與字詞、字詞與段落、以及段落與段落間語意相似性的評估，在LSA所建置的語意空間中都能從這兩個向量間的向量餘弦值中推算出來（Landauer et al., 1998）。Landauer等人的嘗試，不僅打開了一個語意知識表徵的新方法，也讓心理語言學家能用一種較精確又自動化的方式分析詞與詞、句子與句子、或是文章與文章間的語意關聯程度。

以Landauer爲首的LSA研究團隊一方面致力於語意空間心理效果的檢測 （例如同義詞及多義詞的測試，Landauer et al., 1998），同時也致力發展語意空間的應用（例如文章連貫性計算及學生摘要的評分，Foltz, Kintsch, & Landauer, 1998）。爲使LSA的應用更爲廣泛，科羅拉多大學的LSA研究團隊，除了採用不同語料來源建置出不同的語意空間，同時也建置一個網站（http://lsa.colorado.edu/），供其它研究人員使用。該網站目前提供的功能包括有計算鄰近詞、句子間的語意相關性、以及篇章間的語意相關性（Dennis, 2006）。此外，由於LSA具有不需要使用文法或事先定義語彙的特性，使得LSA的技術可以不受語言環境的限制。也就是說，即是使用者所使用的是非英語的語料庫，只要使用者建立好關鍵詞在文件中出現次數的矩陣，就能利用LSA的技術建置出該語系的語意空間。近年來，科羅拉多大學LSA研究團隊也分別與德國、法國學者合作，成功的建置了法文與德文語料的語意空間，這兩種語言的語意空間在其語言環境中的心理效度都已得到支持（Quesada, 2006）。

在中文環境，也有研究者曾使用LSA技術進行研究，例如張國恩與宋曜廷 （2005）即利用LSA技術建兩個主題（族群與群落、端午節）的語意空間，其詞彙量分別是1557個詞及2921個詞。再由這兩個語意空間爲基礎，設計一個可以自動評量小六學生閱讀摘要寫作的系統，結果發現以向量餘弦值爲計算標準的評分方式，和教師所評閱的分數之間有不錯的相關。此外，葉鎮源（2002）也曾經以新新聞週刊中的100份新聞文件爲基礎，建置了一個約有1600個關鍵詞大小的語意空間，該研究發現以向量餘弦值爲基礎所進行的文件摘要工作的確有不錯的效果。

上述兩個研究顯示以LSA所建置的中文語意空間的確能捕捉到中文讀者知識語彙間的語意關係。若能進一步採用更大型的語料建置一個中文語意空間，應該能更完整的表徵中文讀者語彙知識間的語意關係。此一語意空間的完成，不僅可以讓中文研究者用一種完整精確又自動化的方式分析句子與句子間或是文章與文章之間的語意關聯程度，也可以擴展後續應用的可能性，例如文件擷取的功能或是學生摘要寫作的評

量。有鑑於此，本研究的主要目的乃在建構一個有大型語料庫爲基礎的中文語意空間供研究使用。

所謂的大型的語料來源，以LSA研究團隊而言，他們採用的TASA（Touchstone Applied Science Associates, Inc.）語料有9百萬個詞，而法文的語料庫有320萬個詞，德文的語料庫有500萬個詞（Dennis, 2006）。中研院所建置的平衡語料庫（3.0版）約有500萬個詞，是目前繁體中文語料庫中最爲豐富的一個，相當適合做爲本研究之語料來源。以下首先說明中文語意空間建置之過程；其次，依序說明中文現有之語意空間網頁中的各項功能；最後，並就此語意空間的效度進行驗證。

中文語意空間之建置

基本上，採用LSA技術建置語意空間，需要三個步驟，分別是：(1) 建立一個以段落爲列以詞爲欄的共生矩陣；(2) 運用SVD轉換矩陣；(3) 以維度約化的方式重新建置矩陣（Quesada, 2006）。以下即就這三個步驟依序說明本研究之中文語意間之建置細節。

字詞－文件矩陣之建立

本研究所使用之現代漢語平衡語料庫，乃是中央研究院2006年發售之語料版本，共計9,277份文件，五百萬詞。首先我們從文件裏尋找詞做爲矩陣所需要之關鍵詞，本研究所定義的關鍵詞，是在文件中曾出現兩次以上的詞。每個關鍵詞（此即所謂的type）只要出現兩次以上，即會在矩陣中標出該詞在該文件所出現的次數（此即所謂的token）。整體而言，分析（parsing）每份文件中每個關鍵詞出現次數的步驟，在語料來源不大的情形下，大多可以採用人工的方式，但是當語料量甚大的情形，即需要有一個能自動分析文件且能自動計算關鍵詞出現次數的的程式方能進行文件－字詞矩陣的建置。

目前常用的文件分析軟體中，General Text Parser（*GTP*）（Giles, Wo, & Berry, 2003）不只可以自動分析文件中的關鍵詞，同時也因爲該軟體也能執行SVD的計算，所以LSA研究團隊在建置英文的文件－字詞矩陣時，即採用該軟體來分析其語料的關鍵詞。GTP在進行關鍵詞搜尋與登錄時，文件中原的有詞界限（Word Boundary）可以有效的幫助程式識別關鍵詞，同時也能藉此去掉文件中不必要的符號（例如去掉數字、特殊符號等）。雖然中文文件的書寫系統並沒有給每個詞一個明顯的詞界限，但是，中研院的平衡語料庫所使用的語料都已標記詞界限，因此，本研究一開始輸入給GTP程式的語料，即是中研院已完成斷詞的語料，如此GTP程式即可以自動進行中文關鍵詞的搜尋與次數的計算。

由GTP程式所建立的文件－字詞矩陣，還不能直接進行SVD的運算。要進行SVD的計算之前，需要將一些出現次數過高的功能詞排除（以英文而言，像of, an, the等即是，以中文而言，像的、這、那等詞），一方面是因爲這些功能詞並未帶有太多的意義，同時也是因爲它的出現次數高而干擾了語意比對的效果。除了出現頻繁的功能詞需要篩選出來，只出現一次的詞也需要篩選出來。因爲，只出現一次的詞在共生矩陣中缺乏比對的對象，也會干擾LSA語意比對的效果（Quesada, 2006）。爲能有效的進行關鍵詞篩選的工作，LSA研究團隊是以中止名單（Stop list）的方式，讓文件分析的程式在登錄文件時會自動忽略這些關鍵詞。目前本實驗室關鍵詞矩陣是分兩階段進行，第一階段將平衡語庫中只要曾經出現一次的詞即登錄爲一欄，此一階段程式所定義的關鍵詞共有55,303個詞；第二個階段即是設定詞篩選的標準，分別是：(1)出現兩次以上的詞才列爲本研究之關鍵詞；(2) 刪除出現頻繁的功能詞，本語意空間之刪除標準乃是以中研院詞頻計算在前百分之一的高頻詞爲主。此一階段完成後，自動登錄程式所建置完成的是一個有49,021關鍵詞與9,277篇文章的矩陣。以自動篩選的方式，將被列入中止名單中的詞刪去。

平衡語料庫中原有的9,277篇文章，每篇文章的字數長短差距甚大（M = 964 字，SD = 2335 字。LSA研究團隊在建立其文件－字詞共生矩陣時，每一份文件的字數是接近的，他們的做法不以文章爲文件的單位，而是改以段落爲其文件單位，因爲，每一個段落所包含的關鍵詞數量不會有太大的差別（Landauer & Dumais, 1997）。因此，我們進一步將文件進行分段，每一段落大約有200個中文字。原有的9,277篇文章，進行分段後，形成了40,463個文件，字數長短差異相對較小（M = 219字，SD = 66字），最後，再以此文件數量爲列，重新登錄建置一個49,021（詞）乘40,463（段落）的文件－字詞矩陣。

執行SVD轉換矩陣

由GTP所建立的文件－字詞矩陣，基本上所表徵的是每一個關鍵詞在文件所出現的次數，此一矩陣並未表現關鍵詞彼此間的語意關係。但是藉由SVD的運算過程，可以算出每個關鍵詞在對角矩陣中的特徵值，一般來說特徵值愈大的向量，表示具有較大的訊息量，反之則只有微小的訊息量。經過SVD轉換後的

矩陣，關鍵詞和文件的關係，就不是原本出現次數的關係了，取而代之的是表徵關鍵詞在文件中的語意關係。

目前有許多可以免費使用的SVD套裝軟體，一般來說，若要轉換的矩陣不大，則大多數的SVD套裝軟體都能處理。但是建立一個語意空間所需要的文件－字詞矩陣都相當龐大，例如TASA的矩陣就是由上百萬個關鍵詞所形成，因此，建置語意空間時，研究者最常面臨的問題是，如何在有限的電腦硬體設備中讓程式有效的執行如此龐大矩陣的運算。在不增加電腦設備的前提下，本研究採用Tedlab（http://tedlab.mit.edu/~dr/ SVDLIBC/）所發展的SVDlibc套裝程式，最主要的原因是SVDlibc設計兩種矩陣輸入的格式，分別是Sparse Matrix與Dense Matrix，二者主要的差異在於前者只列出在矩陣中不爲零的值，後者則是將所有的值都列出。SVDlibc所設計的Sparse Matrix輸入格式因爲只列出不爲零的值，因此有效的節省了執行SVD時所需的硬體空間。以我們目前的經驗，中研院平衡語料庫所建立的文件－字詞矩陣（即49,021【詞】乘40,463【段落】）的矩陣，我們在PIII 2.8G，1G Ram的電腦上執行SVD的矩陣轉換工作，約 1 個小時即可以完成矩陣轉換的工作。

維度約化

經由SVD轉換後的矩陣，應該要留下多少個奇異值，也是語意空間建置過程中，相當重要的一個議題。在共生矩陣中，每個關鍵詞和文件都是以向量的方向呈現，本研究所完成的共生矩陣，每個詞的向量都會有40,463個維度。在維度約化的過程中，如果只用兩到三個維度來表徵每個關鍵詞的向量，則每個關鍵詞之間的相似性會過高；反之，如果保留所有的維度來表徵每個關鍵詞的向量，則每個關鍵詞的相似性就又會幾近於零。到底每個關鍵詞應該要保留多少個維度，才能達到最理想的語意關聯性評做的效率，是LSA能否建置一個合理語意空間的重要議題。Landauer與Dumais（1997）的研究發現，在維持100、200以及300個維度的情形下，對同義詞的測試都能有不錯的效果。有鑑於此，本研究使用Tedlab軟體進行SVD的矩陣轉換時，即同時建置了含有100、200以及300個維度的中文語意空間。

字單位及詞單位的語意空間

雖然，中文詞在中文閱讀的理解歷程扮演重要角色，但是中文字仍然是中文書寫系統的基本單位，而且中文詞多數由單字或雙字所組成，偶也有三字或四字詞（詳細論述請參閱：高千惠、胡志偉、曹昱翔、羅明，2009）。那麼是該以字單位或是以詞爲單位來建立中文語意空間，才能最逼近中文讀者內在語意知識的表徵，是一個尚待驗證的議題。在未有明確定論之前，本研究先以同樣的語料，分別以字及詞爲單位建立兩個語意空間。關於字單位語意空間的使用或是詞單位語意空間的使用，文後會有進一步的說明。在此，僅著墨於兩個語意空間的可能限制。

在以詞爲單位的語意空間中，最大的限制是當有一個詞彙，不曾出現在中研院的平衡語料庫時，此一以詞爲單位的語意空間就無法針對這一個詞進行詞對詞，或是詞對文件的語意比對。例如，本研究所使用的語料來源中並沒有「奈米」一詞，則以詞爲單位的語意空間即無法計算「奈米」與「毫米」這兩個關鍵詞的語意關係或是和「物理計量單位」這句話間的語意關係。若是將奈米放入一句話中，例如「奈米是物理學的計量單位」，則不論是兩兩文件比對，或是文件對關鍵詞的比對，此時語意空間所形成的向量表徵，並沒有包含「奈米」一詞，而是由文件中其它的關鍵詞組成此一向量。此時，雖然可以得到向量餘弦值，但使用者不能忽略原本的向量裡就少了一個關鍵詞的限制。

以字爲單位的語意空間中，若是有生僻少見的中文字（例如「杢」）未曾出現在中研院的平衡語料中，此時，也會無法針對這個少見字進行語意比對。不過這種情形出現的比率會比以詞爲單位的語意空間來得低。大體而言，使用者幾乎可以針對所有常用中文字進行語意的比對。但是因爲LSA的語意計算過程中會忽略關鍵詞在句中出現的順序，此一現象使得以字爲單位的中文語意空間，遇到所謂的顛倒詞，例如「領帶」與「帶領」，或是「球拍」與「拍球」，即無法有效的評估這兩個詞在該語意空間中的語意關聯性。因爲這些顛到詞在以字爲單位的語意空間中，會被視爲是兩個完全相同的向量，兩個完全相同的向量進行語意的比對，其向量餘弦值爲1。不過，若是將這些顛倒詞放入句子的脈絡中，例如「選手買球拍」與「選手拍球」，則以字爲單位的語意空間，仍然可以就這兩個含有顛倒詞的句子進行兩句話之間的語意比對。

中文語意空間網頁主要功能介紹

中文語意空間的建置，爲的是要讓研究者能有一種較精確又自動化的方式分析中文句子與句子間或是文章與文章間的語意關聯程度。爲能提供研究社群一

個更簡單便利的介面，同時考量非中文研究者使用的便利性，我們初步先建置一個以中英文並列描述的中文語意空間網頁（http://www.lsa.url.tw）。

本網頁目前區分出兩大區塊，首先是語意關聯性的分析（圖1左方上之LSA應用），在此功能區中有兩個主要的功能，分別是「成對比對（pairwise comparison）」，在此功能中使用者可以查詢兩兩關鍵詞、兩兩段落、兩兩文章、或關鍵詞與段落之語意關聯性，以及「最接近詞排列（Nearest Neighbors）」，使用者也可以查詢每個關鍵詞在本語意空間中，語意最爲相近的關鍵詞。其次，是語彙功能區，在此功能區中我們提供自動斷詞與詞頻計算的功能。以下，即簡略說明每一功能區的使用方式：

（一）成對比對（Pairwise Comparison）

在成對比對的功能區中，我們分別提供「以詞爲單位」與「以字爲單位」的語意空間。使用者可依其研究需要，選擇所欲使用的語意空間。圖2所呈現的是本網頁所提供的兩兩向量比較的功能區。

使用者首先需依其需要選擇所要使用的語意空間（ASBC中研院平衡語料庫，是以詞爲單位所建置的語意空間；另一個有以括號標出字單位的選項，是以字爲單位所建置的語意空間）。之後，依向量的內

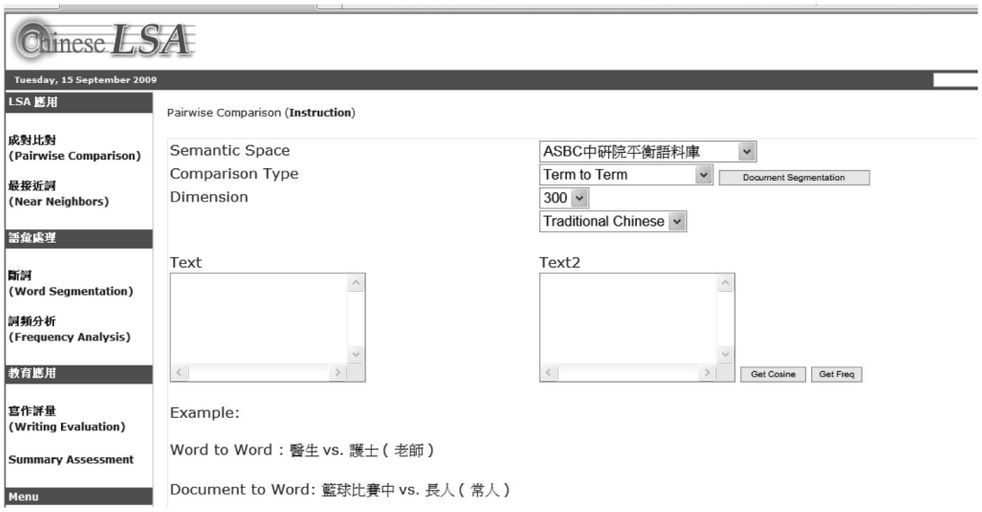

圖1：中文LSA網頁首頁及其三類主要功能

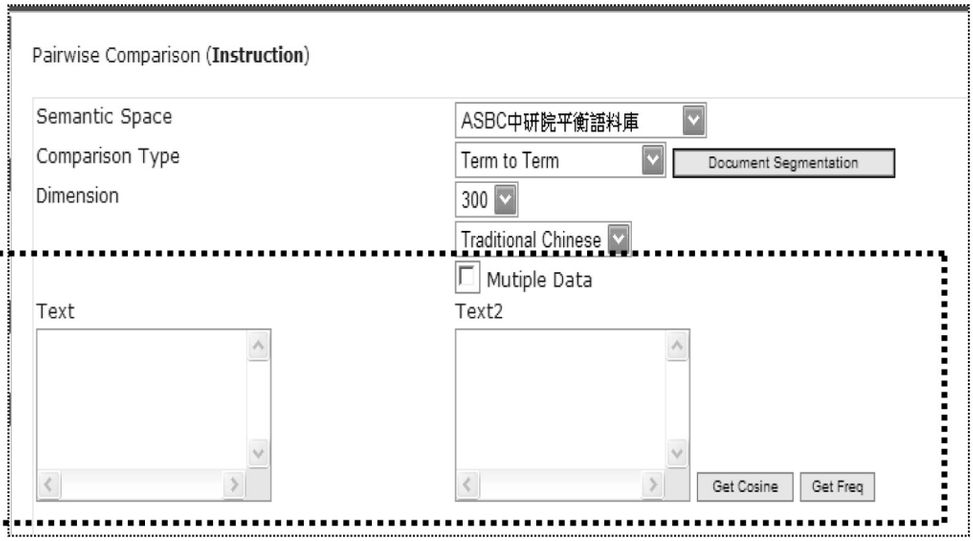

圖2：兩兩向量成對比對功能區

容選擇比對的類型，目前提共有四種類型的比對，分別是：「term-to-term」，所提供的是兩兩關鍵詞的比對；「Document-to-term」與「term-to-document」的比對，則是一邊爲文件，另一邊爲關鍵詞的比對；「document-to-document」的比對，則是兩兩文件的比對使用以詞爲單位的語意空間。之後，使用者可以依其輸入的中文字選擇繁體中文或是簡體中文（系統預設值是繁體中文）；最後，使用者在虛線方框內的文字輸入施區分別輸入所欲比對的文字，然後按下虛線方框內的「Get cosine」功能鍵，即可以得到這兩個向量的語意關聯值，當兩個向量的cosine值愈高，即表示這兩個向量的語意關聯性愈接近。表1所呈現的即是以字爲單位與以詞爲單位的語意空間中，四類功能的比對結果。

由表1也可以看出，在以字爲單位的語意空間進行語意關聯性的比對，只有輸入超過一個字以上，系統就會將此視爲是文件的類型，使用者不需另外在每個字之間輸入空白。但是，若使用者選擇以詞爲單位的語意空間進行兩兩向量語意關聯性的比對，只要所欲比對的文字超過一個詞以上，在本系統中就稱之爲文件的比對（除了term-to-term之外的三類功能都包含有文件的比對）。因此，在文字區輸入文字時，可以事先以詞爲單位，在每個詞之間加入空格（例如，學生 快樂 的 嬉戲），或是以沒有詞空格的方式輸入文字，再以此功能區所提供的自動斷功能（即畫面上顯示的”Document Segmentation”）進行斷詞的工作，系統即會自動將讀者所輸入的文字，以空白鍵區分出詞界限。

當使用者所欲比對的資料不只一筆，例如，想要知道一篇文章中，每一句話之間的語意關聯性（document-to-docuement）；或是使用者想要知道某一個關鍵詞，在篇章中每句話的語意關聯性（term-to-document），則使用者可以在兩個文字輸入區中，同時輸入多筆資料對並勾選網頁中的「multiple Data」(註一)，然後再按下「Get cosine」，此時，畫面下方會出現「Download cosine result」，使用者點選該連結後，即可以看到多筆資料的比對結果。

（二）詞或文件的最接近詞（Nearest Neighbors）的尋找及排列功能

每個詞或文件在語意空間中，都有其語意上最爲相近的詞，我們稱之爲最接近詞。在此功能區中，使用者可以查詢與某一個關鍵詞或是某一句話，在語意上最爲相近的關鍵字或詞有哪些。圖3顯示的畫面是本網頁中的Nearest Neighbors功能區。

要使用此一功能區，使用者仍然需要先選擇是要是以字爲單位或是以詞爲單位所建立的語意空間。之後，使用者可依其需求設定所欲搜尋關鍵詞的詞頻範圍；同時；使用者也可以依其研究需要設定要多

表1 中文語意空間兩兩向量比對範例

	以字爲單位的語意空間		
	Text1	Text2	Cosine
term-to-term	學	生	0.74
term-to-document	學	校園充滿讀書聲音	0.36
docutment-to-term	學生快樂的嬉戲	校	0.36
docutment-to-document	學生快樂的嬉戲	校園充滿讀書聲音	0.43
	以詞爲單位的語意空間		
	Text1	Text2	Cosine
term-to-term	學生	校園	0.53
term-to-document	學生	校園 充滿 讀書 聲音	0.26
docutment-to-term	學生 快樂 的 嬉戲	校園	0.37
docutment-to-document	學生 快樂 的 嬉戲	校園 充滿 讀書 聲音	0.23

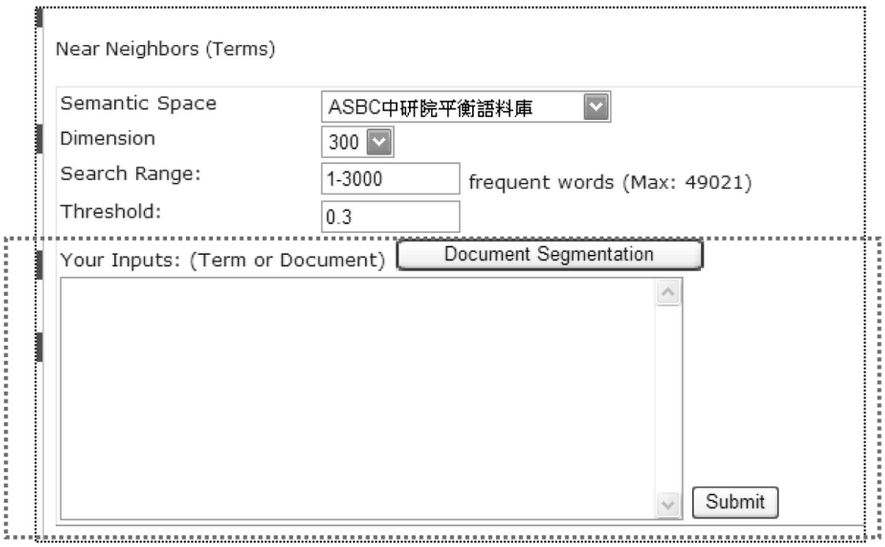

圖3：語意最接近詞尋找排列功能區

表2　中文語意空間語意最接近詞比對範例

月亮		月亮 從 東方 升起	
Term	Cosine	Term	Cosine
月亮	1.00	從	0.74
星星	0.80	月亮	0.51
太陽	0.73	東方	0.43
天空	0.71	星	0.405
亮	0.68	星星	0.402
星	0.66	角度	0.403
光芒	0.65	天空	0.391
明亮	0.64	太陽	0.382
繞	0.59	光芒	0.355
望遠鏡	0.59	明亮	0.346

少語意相關以上的詞才納入搜尋範圍，即畫面中的Threshold選項，在這邊研究者可填入從0到1之間的任一個數值。舉例來說，如果使用者填入0.3這個數值，此時，系統僅會針對與關鍵詞有0.3以上的語意相關性的詞進行排列。最後，使用者將所欲查詢的關鍵詞或是一份文件放進文字輸入區（即圖3虛線方框處），再按下送出（submit）功能鍵，即可得到所欲查詢的結果。表2左邊所呈現的結果是將詞頻設爲1-6000詞的情形下，將「月亮」這個關鍵詞放進文字輸入區後，系統依照cosine值高低所排序出來的前10個語意相近詞。

當使用者選擇使用以字爲單位的語意空間，進行最接近詞的比對時，一旦使用者所輸入的文字超過一個中文字，則使用者需要在字與字之間加入空格，這樣系統才能將這些文字視爲文件處理。此外，由於是最接詞的尋接及排列的功能，所以不論使用者一

開始輸入的是單一個詞，或是由數個中文詞所組成的文件，系統比對給使用者的結果都是與該詞或該文件在語意上最接近的詞彙。右邊則是將「月亮從東方升起」這個句子放入文字輸入區後，系統依cosine值高低所排列出來的前10個語意相近詞。

（三）語彙功能區：斷詞與字、詞頻之自動計算

關於斷詞與詞頻計算功能區的存在，主要是爲了提供使用者查詢文件中所使用語詞的頻次高低。例如，使用者如果想要了解某一篇文章所使用的常用字或詞的比率高低，即可透過此一功能進行自動化的處理。使用者首先將所欲分析的文章貼入圖4畫面中的虛線方框處，在沒有斷詞的情形下，使用者按下「Get CharFre」可以得到該篇文章使用字的字頻分配情形。

若要得到文章中使用詞的分配情形，則需要先將文章進行斷詞，網頁中的「Segmentation」提供自動斷詞的功能，斷詞之後的材料，按下「Get WordFre」即可以得到該篇文章使用詞的分配情形。以康軒出版社小二的國語課本中的課文「小鎮的柿餅節」這一課爲例，該課文總共有198個字（含標點符號）。表三所呈現的即是該篇文章的字及詞的頻次及分配情形。不論是字頻或詞頻的計算，標點符號以及語料中未曾納入的語詞，因爲沒有詞頻的資料，系統就會將這些標點符號和詞彙歸入「OOV」這一類別中，同時也會顯示其數量多少及在整篇文章中所佔的比率。本系統除了會顯示文章用字或用詞的比率分配情形，也會依照文字順序將每個字或詞的頻次詳列給使用者，以方便使用者知道每個字或詞的字詞頻高低。

中文語意空間之效度

目前不論是英文、德文、或法文環境所建立的LSA語意空間，都是以詞爲「字詞－文件」矩陣建置的基本單位。本研究雖然同時建置有以字及以詞爲單位的中文語意空間，然而考量中文詞在中文閱讀理解歷程中扮演著基本的意義單位（鄭昭明，1981；彭瑞元、陳振宇，2004），以及與其它拚音文字的語意空間有相同建置單位的前提下，本研究將針對以詞爲單位所建置的中文語意空間，能否有效的反應中文讀者內在知識語意表徵進行驗證。以下分別從詞與詞的語意關聯性，以及句與句之語意關聯性這兩個層面來驗證。

實驗一：詞彙間語意關聯性之驗證

在中文語意空間中兩兩詞彙間cosine值可以用來表徵詞彙間的語意關聯性，當兩個關鍵詞的cosine值愈高，表示這兩詞的語意關聯性就愈高例如，例如月亮和地球的cosine值是.55，但是月亮和沙漠的cosine值是.16。這兩個數據看似符合大多數人的直覺，但仍需有更有系統的驗證本研究所建置之中文語意空間能否捕捉中文讀者內在心理詞彙間之語意關聯表徵。

胡志偉、陳貽照、張世華及宋永麒（1996）曾經以600個多義詞完成中文多義詞自由聯想常模，該次實驗共有300名大學生參加。該研究所整理出來的常模和相關的聯想詞，相當程度能反應多義詞在中文成熟讀者內在心理語意表徵的關性。若中文語意空間能合理的表徵中文讀者內在心理詞彙之語意關係，則本語意空間所計算的多義詞聯想常模，應能與胡志偉等人所研究之實徵資料相呼應。因此，檢視中文語意空間能否合理反應中文讀者內在心理表徵中之多義詞表徵，應是驗證此一語意空間是否有實在效度的合理途徑之一。

方法

1. 研究材料

本研究之材料之要採用胡志偉等人（1996）所建

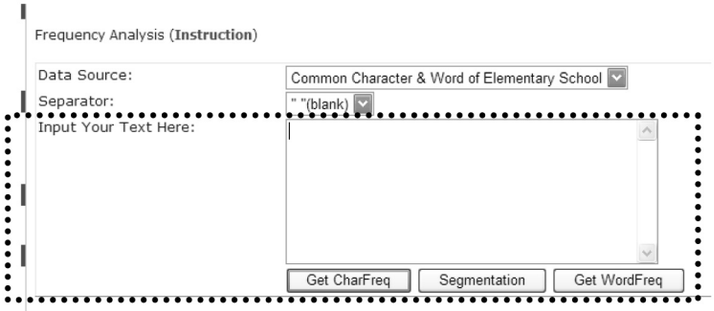

圖4：斷詞及詞頻計算功能區

表3 「小鎮的柿餅節」字頻及詞頻比率分配情形

Range	Number	Percentage (%)	items(s) (Repeated items are not shown)
	字頻比率		
1-500	137	69.19	天,來,了,節,的,活,動,開,始,,,新,小,也,熱,,起,,當,車,子,,進,,時,一,,,風,過,,許,多,做,,人,家,前,後,都,滿,,,在,陽,光,下,,片,金,黃,看,,,很,可,愛,爸,,說,每,年,九,月,到,十,二,,這,裡,,,和,,,會,把,,,個,,,變,成,
501-1000	21	10.61	秋,靜,,鎮,鬧,,陣,,香,迎,吹,屋,,排,
1501-2000	7	3.54	餅,,
2501-3000	2	1.01	埔,烘,
3001-3500	8	4.04	柿,
OOV	23	11.62	，,,
	詞頻比率		
1-3000	75	54.74	秋天,來,了,節,的,活動,開始,,靜靜,,也,熱鬧,,起來,當,車子,時,一陣,,,香,許多,做,,人家,屋,前,,後,都,,在,陽光,下,片,很,可愛,爸爸,說,每,年,到,,和,,會,把,個,,變
3001-6000	4	2.92	外地,心想事成,事事,如意,
6001-9000	3	2.19	迎風,都會,親友,
9001-12000	3	2.19	柿子,金黃,
12001-15000	3	2.19	小鎮,,
30001-33000	1	0.73	到站,
33001-36000	1	0.73	秋風,
OOV	47	34.31	，,柿餅,,新埔,,。,開進,,,,吹過來,,,,排滿,,一,

置之多義詞聯想常模中的兩筆數據，分別是：讀者對主、次意義的總反應次數，以及讀者針對主、次意義進行聯想所反應詞彙的個別總次數。茲分別明於下。

首先，多義詞詞義擷取的辨識歷程中深受詞彙本身相對頻率（relative frequency）的影響，依主、次意義相對頻率相差懸殊的情形，可大分爲強勢語義（Biased）的多義詞，以及均勢語義（Balanced）的多義詞。在詞彙聯想作業（word association task）中將讀者所反應的詞彙依其語意進行歸類，即可依其懸殊性的差異情形區分出強勢語義與均勢語義。以胡志偉等人（1996）的研究爲例，「一刀」這個詞可以是指把東西、動物或人砍一刀的意思，也可以是指稿紙的計算單位。在該研究中，有85位讀者聯想的詞彙和主要語義有關；只有1位讀者聯想到稿紙這個次要語義。此時，「一刀」這個詞即是所謂的強勢語義的多義詞。再以「比畫」一詞爲例，該詞是指兩人動武的意思，但也是以手勢作模擬的意思，在該研究中有45位讀者聯想出和「兩人動武」有關的詞彙，另外則有38位讀者聯想出和「以手勢作模擬」有關的詞彙。此時，「比畫」一詞即是所謂的均勢語義多義詞。表4是本研究依胡志偉等人所歸類的反應次數所計算出來的強勢語義多義詞和均勢語義多義詞的平均反應次數。

其次，者針對主、次意義進行聯想所反應詞彙的個別總次數，是指在該次研究中，參與實驗的讀者根據主、次要意義所反應的個別詞彙的總次數。以「打氣」一詞爲例，讀者共聯想出22個詞彙。在主要意義中，「加油」一詞的總次數是56次；在次要意義中「輪胎」一詞的總次數是9次（詳見表5）。

2. 資料分析

首先，將600個多義詞視爲「term」，再將胡

表4　胡志偉等人（1996）將讀者依主、次要意義反應次數之平均數與標準差

	主要意義		次要意義		
	M	*SD*	*M*	*SD*	*N*
強勢語義	81.00	9.13	7.15	6.26	289
均勢語義	52.29	13.42	27.46	10.48	311

表5　「打氣」一詞所產生聯想詞之反應次數

主要意義	總反應次數	聯想詞	個別總次數	次要意義	總反應次數	聯想詞	個別總次數
鼓勵	74	加油	56	把空氣注入輪胎或皮球		輪胎	9
		鼓勵	5			籃球	2
		比賽	4			腳踏車	2
		鼓舞	2			氣球	2
		拍肩	1			打氣筒	1
		努力	1			筒	1
		啦啦隊	1			輪子	1
		加油棒	1			球	1
		精神支柱	1			車子	1
		振作	1			吹氣球	1
		人	1			充氣	1

志偉等人（1996）在該研究中針對主、次意義的描述設定爲「documents」，接著使用中文語意空間中「term-to-documents」的功能，我們預期強勢語義多義詞的主要意義和該多義詞的語意關聯性會高於次要意義，而均勢語義的多義詞，其主、次要意義的語意關聯性會相似。其次，一樣將原本使用的600個多義詞視爲「term」，再將個別聯想詞，也視爲一個「term」，接著使用「term-to-term」 的功能，針對多義詞和聯想詞的語意關係進行比對，我們預期聯想詞的個別總次數愈高的詞，和該多義詞的語意關聯性也會愈高；反之，個別聯想詞的總次數愈低，則該詞和多義詞的語意關聯性就會愈低。

結果

首先，在表六呈現的是415個多義詞與主、次要語意間cosine值的平均數與標準差。之所以只有415個多義詞能計算出「term-to-documents」的cosine值，是因爲有些多義詞在中文語意空間中並不是以詞（term）的方式而是以文件的方式呈現，例如「一場春夢」這一個多義詞，在中文語意空間被視爲是由三個詞所組成的文件（documents）。

資料顯示「主、次要意義」與「多義詞的類型」有交互作用存在，$F(1, 413) = 15.81$，$p < .001$，$\eta^2 = .037$。進一步分析發現，當多義詞爲強勢語義的多義詞，與主要意義的cosine值會顯著高於該多義詞與次要意義的cosine值，$F(1, 206) = 29.57$，$p < .001$，$\eta^2 = .129$。當多義詞爲均勢語義的多義詞，與主、次要意義的cosine值的差異情形則沒有達到統計上的顯著水準，$F(1, 207) = .013$，$p > .05$。

在多義詞與聯想詞的語意關聯性上，依其個別聯想詞的總反應次數，可以分爲低、中、及高反應次數三組。圖六呈現的是低、中、及高反應次數的cosine值。結果顯示，反應次數有主要效果，$F(1, 434) = 7.95$，$p < .001$，$\eta^2 = .036$。其中反應數次數低組，其多義詞與聯想詞的cosine平均值最低，而反應次數高組，其多義詞與聯想詞的cosine平均值最高，三組間都的差異情形亦都達統計上的顯著水準。

上述「term-to-documents」與「term-to-term」結果，顯示中文語意空間不僅能適切的反應多義詞與主

表6　多義詞與主、次要意義cosine之平均數與標準差（term-to-documents）

	主要意義		次要意義		
	M	*SD*	*M*	*SD*	*N*
強勢語義	0.137	0.145	0.081	0.113	207
均勢語義	0.098	0.112	0.097	0.113	208

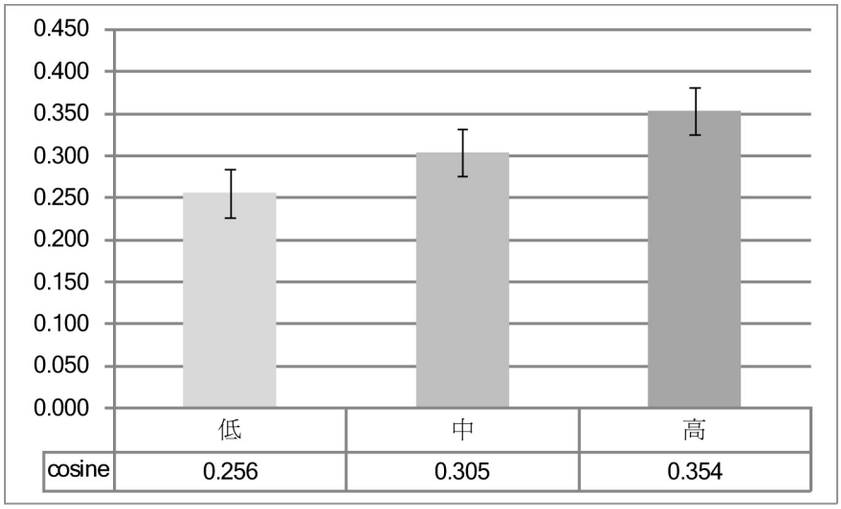

圖5：低、中、高反應次數組之cosine值

次意義間的語意關聯性；而且中文語意間所捕捉到的兩兩詞的語意關聯性，相當程度能反應中文讀者內在知識表徵之兩兩詞的語意關聯程度。

實驗二：句間語意關聯性之驗證

實驗二旨在驗證中文語意空間所計算的句子與句子間之語意關聯性，能否反應中文讀者內在之心理知識表徵。

方法

1. 研究對象

共計有160位大學生參與本實驗。所有學生皆以中文爲母語。

2. 研究材料

考量學生知識背景對語意評定可能產生的影響，本研究分別使用自然科學及社會科學各八篇，且主題不同的說明文進行語意關係之評定。16篇文章共計有204個句子（見附錄二）。以兩兩呈現的方式請讀者進行語意相似度的評估，呈現的方式包括有：(1) 同一篇文章內的句子，且依文章內容的順序進行比對，例如句一與句二相比，句二與句三相比，依此類推；(2) 同一篇文章內的句子，但以隨機的方式將兩兩句子進行比對，例如句五與句一相比，或是句七與句十相比；(3) 不同文章間的句子相比，以隨機的方式將兩兩句子進行比對例如第一篇文章的第二句話與第三篇文章的第5句話相比。當讀者認爲兩兩句子間的語意關聯性愈高，即勾選高語意相關，反之，評定爲低語意相關。作業表中，5分代表最高語意相關，0分則表示幾乎無語意相關。

3. 實驗程序

本實驗爲團體施測，實驗一開始由主試者說明句子語意相似性的評定方式，確定所有的研究對象都瞭解本評定程序後，即請所有的參與者打開評定材料開始評定兩兩句子間的語意相似性。待全部評定完成，該參與者即可將材料交回。

結果

表7是讀者以及中文語意空間針對三種句子的語意關聯性的評估結果，所謂的「句一/句二」指的是同一篇文章內依句子順序兩兩呈現的方式；而「文章內隨機呈現」指的是同一篇文章內以隨機方式兩兩呈現的方式；「文章間隨機呈現」指的是不同文章的句子以隨機方式兩兩呈現的方式。

一般來說，在同一篇文章的句子裏，不論是以依序呈現，或是隨機呈現的方式，其語意相似度應會高於文章間隨機呈現的方式。因爲，同一篇文章裏的句子通常是扣緊同一個主題，而不同文章間的句子，因主題不同，其語意相似度就應該不高。從表七的結果可以看到，不論是中文讀者的評估結果，或是中文語意空間計算的結果，不同文章間隨機兩兩呈現的句子，其語意關聯性的平均值都低於在文章內逐句呈現或隨機呈現的平均值。重複量數單數變異數檢定的結果顯示，中文讀者的評估結果在「句子呈現方式」上有主要效果，$F(2, 406) = 547.88$，$p < .001$，$\eta^2 = .731$，事後檢定的結果顯示三組間的差異情形皆達統計上的顯著水準。在中文語意空間計算的結果上，「句子呈現方式」上也有主要效果，$F(1, 203) = 35.07$，$p < .001$，$\eta^2 = .159$，事後檢定的結果顯示同一篇文章間，逐句呈現與隨機呈現的方式，其cosine值的差異情形未達顯著差異，$t = .404$，$p > .05$。

但同一篇文章間，逐句呈現與隨機呈現的方式，其cosine值都顯著高於以隨機方式呈現的cosine值（$t = 7.98; 8.51$, $p < .05$）。由中文語意空間所評定的語意相似性，之所以在這兩種呈現方式上沒有差異，可能的原因是中文語意空間的評定過程中，有可能是因爲有些詞彙並未收錄在 ASBC語料庫，例如「斑蝶」或是人名（如霍華德），因此，在文件語意相似性的計算上無法百分之百的反應中文讀者內在的知識表徵。此一限制，在未來的研究中可以嘗試直接選用ASBC語料庫中的篇章，再進一步檢視，在此條件下，中文語意空間的評定結果，是否能更完整的反應中文讀者內在的知識表徵。

進一步以迴歸分析的方式驗證由中文語意空間所評定的結果（包括：文章內逐句呈現、隨機呈現，或是文章間隨機呈現的結果），是否與中文讀者評分的結果一致。結果發現，不論是三種呈現方式，中文語意空間所評定的結果，和讀者的評分結果有一致的組型，其F值分別是：$F(1, 202) = 6.49$、9.93及13.33，*all* ps $< .01$。上述結果顯示，中文語意空間所捕捉到的兩兩文件間的語意關聯性，能適切的反應中文讀者內在知識表徵之兩兩文件的語意關聯程度。

結論

本研究旨在應用LSA技術建置一個可以自動比對詞彙間複雜語意間係的中文語意空間。以LSA技術所建置的語意空間，最大的優勢是，除了可以比對詞彙與詞彙間的語意關係外，同時還能比對不同語言單位間的語意關係，例如詞彙與整篇文章。過去，國內已有一些學者採用LSA技術建置過不同主題的中文語意空間例如張國恩與宋曜廷（2005）以及葉鎮源（2002）的研究，本研究採用更大型的語料來源，期能更完整的捕捉中文讀者內在詞彙間的語意關係。藉由中文語意空間網頁的完成，不僅提供給相關研究者一個能自動化分析句子與句子間或是文章與文章間的語意關聯程度的使用環境，未來也有擴展後續應用的可能性，例如學生摘要寫作自動化評量工具的發展。

現階段所完成的中文語意空間，所呈現的多義詞語意關係，已相當程度反映胡志偉等人（1996）所建置的中文多義詞自由聯想常模。此外，從本次參與研究的大學生所進行的句子語意關係評估作業的結果，也顯示此一中文語意空間對於文件與文件間語意關係的掌握，在一定程度上也能反應出中文讀者內在知識表徵的語意結構。這兩項結果顯示，本研究以大型語料所完成的中文語意空間，所捕捉到的中文詞彙間錯綜複雜語意關係，的確是一個能反映中文讀者內在心理表徵的語意空間。

表7　三類句子比對之讀者評分（Human rating）與LSA之語意關聯值（cosine）的最小值、最大值及平均數

	讀者rating值				中文語意空間cosine值			
	Min.	Max.	*M*	*SD*	Min.	Max.	*M*	*SD*
句一 /句二	0.821	4.775	3.252	0.916	0.001	0.738	0.143	0.166
文章內隨機呈現	0.500	4.950	2.596	0.994	-0.169	0.934	0.149	0.160
文章間隨機呈現	0.025	3.123	0.558	0.582	-0.076	0.659	0.040	0.100

本質上，LSA建置語意空間的語料來源，是影響語彙間彼此意義關聯性的主要因素。因此，當使用的語料來源愈能反應當下社會環境讀者日常生活所接觸的語彙，則所建置出來的語意空間就愈有可能清楚的反應出詞彙彼此間所存在的語意關係。現階段建置完成的中文語意空間，所使用的語料來源，基本上所反應的是，中文環境中，成熟讀者所接觸的語彙資料，因此，本語意空間能否有效的反應中文環境中的兒童讀者的內在心理表徵知識，尚需要有系統的進行檢驗。以科羅拉多大學所建置的英文語意空間爲例，該研究團隊所使用的語料來源除了有英文環境中常用的成人語料，同時也採用了兒童學校裏的教科書以及兒童讀物等教材。此外，英文的語意空間，同時也針對不同的知識主題，在原有的大型語意空間之下，建置一些以知識主題（例如血液循環或是心理學概論）爲主的小型話意空間。不論是增加兒童語料，或是根據知識主題所建置的小型語意空間，這兩者的用意都是要更精準的反應不同背景知識讀者間，內在知識表徵所可能存在的差異性。在目前已有的中文語意空間的基礎上，未來我們除了可以繼續建置不同知識主題的語意空間，在有適當語料來源的前提前，我們也會進一步將現有的語意空間進行擴充，並針對不同年齡層的讀者建置出數個語意空間。最後，在後續的教育應用上，我們將繼續發揮中文語意空間現有的特色，期能發展出學生摘要寫作的自動化評量工具，此一摘要寫作學習系統的開發，將是一個重要的閱讀理解教學研發工作。

註釋

（註一）要使用「multiple Data」的功能，使用者需要先在本網頁設定一組使用者名稱後方能使用。

參考文獻

胡志偉、陳貽照、張世華、宋永麒（1996）。〈中文多義詞自由聯想常模〉。《中華心理學刊》，*38*，67-168。

高千惠、胡志偉、曹昱翔、羅明（2009）。〈從校稿失誤談中文閱讀的單位〉。《中華心理學刊》，*51*，21-36。

張國恩、宋曜廷（2005）。《潛在語意分析及概念構圖在文章摘要和理解評量的應用（3/3）》。國家科學委員會專題計畫成果報告，報告編號NSC93-2520-S-003-011。台北：行政院國家科學委員會。

彭瑞元、陳振宇（2004）。〈「偶語易安，奇字難適」：探討中文讀者斷詞不一致之原因〉。《中華心理學刊》，*46*，49-55。

葉鎮源（2002）。《文件自動化摘要方法之研究及其在中文文件的應用》。國立交通大學資訊科學研究所，碩士論文。

鄭昭明（1981）。〈漢字認知的歷程〉。《中華心理學刊》，*23*，137-153。

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Dennis, S. (2006). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57-70). Mahwah, NJ: Lawrence Erlbaum Associates.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Analysis of text coherence using latent semantic analysis. *Discourse Processes, 25*, 285–307.

Giles, J. T.,Wo, L., & Berry, M. W. (2003). GTP (general text parser) software for text mining. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 455–471). Boca Raton, FL: CRC press.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In N. Ross (Ed.), *The psychology of learning and motivation* (Vol.41, pp. 43 – 84). New York: Academic Press.

Landauer, T. K. (2006). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Quesada, J. (2006). Creating your own LSA spaces. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 71-85). Mahwah, NJ: Lawrence Erlbaum Associates.

附錄一　向量餘弦值之計算公式

若有兩個文件，文件一{$P_1, P_2, \cdots.P_i$}與文件二{$M_1, M_2, \cdots.M_k$}都以n維度向量表現，這兩個文件間向量餘弦值計算方式如下：

$$\cos\theta_{i,k} = \frac{\overrightarrow{P_i} \cdot \overrightarrow{M_k}}{|\overrightarrow{P_i}| \cdot |\overrightarrow{M_k}|}$$

附錄二　實驗二所用之204個句子

篇章	句子順序	句子
1	1	紅樹林大都生長在熱帶型氣候地區、風浪較小的鹹水區域
1	2	由於它們常生長在高低潮線之間
1	3	漲潮時會浸泡於海水之中
1	4	故又有人稱爲潮汐林
1	5	紅樹林生長環境最少有以下兩種特徵
1	6	首先需要有由細粒粉土及黏土所形成的軟泥土壤
1	7	土壤需含多量的有機腐植質
1	8	土壤中的腐植質土層愈深厚
1	9	紅樹林的生長就愈佳
1	10	其次是浪靜的海邊
1	11	通常是彎曲的海灘或河口深處
1	12	這些地方的外緣多有天然屏障
2	1	無法入眠的原因千奇百怪
2	2	包括藥物、酒精、咖啡因、壓力與病痛
2	3	有些人的睡眠問題會轉變爲失眠
2	4	失眠的問題從其型態上來分會有兩種類型
2	5	第一種是入睡困難型
2	6	這類型的失眠者躺在床上
2	7	往往一兩個小時才能睡著
2	8	緊張、焦慮、或身體不舒服引起的失眠常屬此型
2	9	第二種是睡眠維持困難型
2	10	這類型的失眠者睡得不安穩
2	11	時睡時醒
2	12	醒過來就難以入睡

篇章	句子順序	句子
2	13	有些人甚至半夜醒來就未再闔眼
3	1	在十九世紀初期
3	2	英國藥劑師霍華德創造了一個雲的分類法
3	3	並被廣泛採接受
3	4	霍華德的體系主要有兩大原則
3	5	首先依雲的型態
3	6	將雲分爲層層相疊的層雲、分佈的積雲、羽毛狀的卷雲及會下雨的雲
3	7	然後再依雲的高度
3	8	將雲分成高雲、中雲、低雲及直展雲四族
3	9	依據這兩個原則
3	10	可將雲細分爲10個基本型
3	11	並根據它們具有的形狀而命名
3	12	霍華德提出的這一種分類模式
4	1	河流的搬運作用是順流而下的
4	2	被流水搬運的東西稱爲搬運物
4	3	河流搬運東西的能力
4	4	與河流速度及流水量成正比
4	5	流水搬運的形式有四種
4	6	可被溶解的物質
4	7	會以溶液形式的方式搬運
4	8	細小的泥沙顆粒則以懸移形式搬運
4	9	隨波逐流
4	10	較大的碎塊以躍移形式搬運
4	11	偶爾被水流從河床捲起
4	12	巨石及其他笨重東西的搬運
4	13	則只在水流較快、水量較多的情況下
5	1	每個社會都必須面對一項基本的經濟難題
5	2	如何將有限的資源作最佳運用
5	3	以滿足無窮的慾望
5	4	這就是經濟學上稀少性的問題
5	5	由於經濟資源是稀少的
5	6	我們永遠不可能在一定時間內生產出所有人類想要的東西
5	7	稀少的資源就算能增加
5	8	也是要靠努力或付出代價

篇章	句子順序	句子
5	9	而且無法取之不盡用之不竭
5	10	資源的稀少迫使每個經濟體系都必需有所選擇
6	1	台灣我們生活的島嶼
6	2	雖僅有四百餘年的文字歷史
6	3	但據考古學家研究指出
6	4	台灣早在二、三萬年前就有人類的活動
6	5	約在二、三千年前
6	6	屬於南島語系的原住民
6	7	移居入本島過著采獵的生活
6	8	到了十六世紀末雖有華人在台灣活動
6	9	但均只侷限於私人性質
6	10	十七世紀開始
6	11	西方諸國企圖以台灣爲據點與中國進行貿易
6	12	其中以荷蘭人最有進展
6	13	並於一六二四至一六六二年強佔台灣
6	14	佔領期間誘導漢人來台墾耕
7	1	課程變革是一種革新的社會實驗
7	2	因爲嘗試新教育策略之實驗
7	3	乃是從事課程改革實務的重要特質
7	4	由於課程改革具有「社會理解之深度」、「變革時間之長度」與「推廣情境之廣度」等複雜之面向
7	5	因此課程改革的推動者往往在課程革新的理想與目標引導下
7	6	進行教育研究實驗
7	7	在實驗過程中
7	8	課程改革者特別是課程設計者與教師等研究團隊
7	9	可就課程的適切性加以測試與調整
8	1	都市更新是爲了促進都市土地之再開發利用
8	2	復甦都市機能
8	3	改善居住環境
8	4	增進公共利益
8	5	因此都市更新不只是老屋換新屋
8	6	其手段也不只是容積獎勵
8	7	都市更新應以公共利益爲前題
8	8	爲都市實質環境與機能帶來全面性的改善

篇章	句子順序	句子
8	9	甚至更廣泛地帶動社會與經濟環境的改善
8	10	建造符合永續發展的建築
8	11	引入豐富的文化活動等
8	12	在此過程中
8	13	政府應負起主動積極的角色
8	14	除了引入民間投資的資本與企業化管理
9	1	斑蝶住在美洲北部
9	2	牠們整個夏天都在進食
9	3	秋天的時候
9	4	斑蝶可以吃的食物變少
9	5	牠們開始聚集
9	6	並成群往南方遷移
9	7	斑蝶大而強壯的翅膀有助於牠們的速度與長途飛行
9	8	初冬的時候
9	9	斑蝶飛抵南方
9	10	並棲息在某些樹木上
9	11	整個冬天
9	12	牠們都停留在那裡
9	13	呈半睡眠的狀態
9	14	春天的時候
9	15	牠們醒來準備飛回北方
9	16	飛回北方的旅程中
9	17	雌蝶會沿途產卵
9	18	夏天的時候
9	19	新出生的斑蝶來到美洲北部
10	1	蕨類植物不會開花、結果
10	2	它們繁衍下一代的方式
10	3	就和其它的植物不同
10	4	蕨類的繁殖主要以孢子爲主
10	5	一般在蕨葉的背面有褐色的點狀或條狀斑液
10	6	這就是蕨類繁殖的主要構造－孢子囊群
10	7	孢子囊內含有孢子
10	8	孢子成熟後可隨風飄散
10	9	掉在潮濕的土壤上就會萌芽

篇章	句子順序	句子
10	10	長成綠色的原葉體
10	11	原葉體大多非常小
10	12	具精子和卵子
10	13	兩者結合後會在原葉體上
10	14	發育成爲可獨立生活的孢子體
10	15	之後就長出孢子體葉片
11	1	海中所有的哺乳動物中
11	2	海豚是體型較小的一種動物
11	3	小海豚出生的過程
11	4	和一般哺乳動物的出生有很大的不同
11	5	大部份的哺乳動物出生時
11	6	通常是頭先出來
11	7	但海豚出生時
11	8	卻是尾巴先出來
11	9	這種方式可以減少小海豚溺死的危險
11	10	剛出生的小海豚不會呼吸
11	11	如果是頭先出來
11	12	即使在生育過程中沒有任何耽誤
11	13	小海豚的肺也可能會充滿了水
11	14	並因此而溺死
11	15	當小海豚的頭從產道冒出來後
11	16	母親就輕輕地把牠推出水面
12	1	夏日午後的陣雨
12	2	常夾雜著閃電與打雷
12	3	目前最被接受的原因
12	4	是夏天中午後地面溫度高
12	5	使得雲層的垂直對流旺盛
12	6	此種對流現象
12	7	使富含水氣的雲層裡聚積了大量電荷
12	8	當天上有帶電的雲團出現
12	9	地面上也會產生相對應的電荷集中
12	10	一旦雙方電荷差距太大
12	11	連空氣也變成導體
12	12	電荷就會開始中和

篇章	句子順序	句子
12	13	劇烈的中和結果就是溫度急速上升
12	14	造成閃光就是閃電
12	15	而急速升溫造成通路周圍空氣爆炸
13	1	現代記憶的研究可以說有兩個來源
13	2	一是從生物學看神經細胞如何傳遞信息、彼此溝通
13	3	在這裏最主要的發現是神經細胞並不是固定的
13	4	而是受到經驗和神經活動的調節
13	5	第二個來源是研究大腦系統和認知
13	6	此處最重要的發現爲記憶不是單一的
13	7	而是有各種類型
13	8	用著非常不同的邏輯和大腦迴路
13	9	近來有科學家把這兩種完全不同的源流綜合成一個新的合流
14	1	以科學的方法來研究語言的理由最少有兩個
14	2	首先語言是人類有別於其他生物最重要的特徵
14	3	同時也是文明社會精緻組織化直接或間接的推手
14	4	因爲人類學習語言的能力是天生的
14	5	所以當你研究語言時
14	6	就是在研究一個全人類共有的特質
14	7	其次語言也很有趣
14	8	雖然每個人都說著某種特定語言
14	9	但很少人知道自己對於所使用的語言瞭解有多少
14	10	能夠對自己所知道而不自覺的東西有更深一層的瞭解
15	1	政治學者們對負責任的政黨的看法
15	2	有以下兩種主張
15	3	第一種觀點是「競爭團隊」的觀點
15	4	此觀點主張政策形成的工作應由政黨的領導階層來負責
15	5	選民的角色是從兩個競爭團隊中選出一個來執政
15	6	而不是透過選舉來決定政策
15	7	第二種觀點是「政策授權」的觀點
15	8	此觀點認爲選民可以在兩個明確提出政策主張的政黨中作選擇
15	9	選民是選政黨的政策主張
15	10	因此當某一政黨得到多數選民支持而執政時
16	1	從事各項投資活動時
16	2	投資人可能因市場的變動而遭受損失

篇章	句子順序	句子
16	3	此種風險稱爲市場風險
16	4	這種風險由於變動型態的不同又可分爲兩種
16	5	第一種是利率風險
16	6	是指利率的大幅上漲或下跌時所帶來的損失
16	7	例如投資收益型的債券基金
16	8	此種基金的獲利來就是債券票券存款等利息
16	9	當利率升或下跌時
16	10	就會對收益造成影響
16	11	第二種是匯率風險
16	12	乃是指因外匯市場變動引起匯率的變動
16	13	致使以外幣計價的資產遭受損失的風險

The Construction and Validation of Chinese Semantic Space by Using Latent Semantic Analysis

Ming-Lei Chen[1], Hsueh-Cheng Wang[2], and Hwa-Wei Ko[1]

[1] Institute of Learning and Instruction, National Central University
[2] Department of Computer Science, University of Massachusetts at Boston, MA.

Recently, using corpus to create a word net is a new approach in psycholinguistic study. The present study used Latent semantic analysis (LSA) to create a Chinese semantic space. The semantic relationship between words in the Chinese semantic space was estimated by taking the dot product (cosine) between two vectors. The semantic relationship between two sentences or two documents could be estimated in the same way. The results of human data indicate that the Chinese semantic space is a valid way to represent Chinese reader's world knowledge.

Keywords: *latent semantic analysis, semantic relation, semantic space*