實詞比、一詞多義、筆畫數指標建置中文文本自動化分析系統

張琇涵	倪雅真	廖晨惠	郭伯臣	白鎧誌
國立臺中教育大學教育測驗統計研究所碩士生	國立臺中教育大學教育測驗統計研究所碩士生	國立臺中教育大學特殊教育學系教授	國立臺中教育大學教育測驗統計研究所教授	國立臺中教育大學教育測驗統計研究所博士候選人
aaahannah@gmail.com	sunnyni37@gmail.com	chenhueiliao@gmail.com	kbc@mail.ntcu.edu.tw	minbai0926@gmail.com

摘要

本研究旨在建立實詞比、一詞多義、筆畫數中文文本自動化分析指標並發展線上自動化文本分析系統。本研究使用國小教科書進行文本分析，依三大領域各年級文本之指標值進行趨勢性分析，研究結果如下：

一、國語與社會科除六年級外有隨著年級增加而實詞比指標降低的現象。原因應為低年級所學的詞彙包含較多實詞，隨著年級增加，開始學習虛詞所致。

二、一詞多義指標在各年級國語>自然>社會。在國語及自然科上有隨著年級增加指標值降低的趨勢。

三、國小課文平均筆畫數介於 8 至 10 劃，低筆畫字數佔全文六至七成。國語及自然科平均筆畫數及中筆畫數比隨年級增加而升高，低筆畫數比則隨年級增加而降低。國語科 1-3 年級及社會科高筆畫數比亦隨年級增加而升高。

關鍵詞：文本分析、實詞比、一詞多義、筆畫數

Abstract

The presented study aimed to develop the computer analyzes of texts for the characteristics of texts and established the web computational system. The study provided three validated indices to measure the characteristics of texts: the percentage of content words, polysemy and numbers of strokes. We analyzed the cohesions of the text on different genres for grade1 to grade6 and investigated the trend of the cohesions on different grade levels. The results are summarized as follows:

First, the percentage of content words enlarges with the grade levels except grade 6 in Chinese and Social Science. Because of low grade students learned more content words. Second, the values of polysemy index in Chinese is greater than Science greater than Social Science. Polysemy decreases with the grade levels in Chinese and Science. Third, the average number of strokes in elementary texts is between 8-10 and the low strokes account for 60%~70% in all texts.

Keywords: text analysis, percentage of content words, polysemy, numbers of strokes

1. 前言

1.1 研究背景與動機

前教育部長曾志朗教授曾說:「閱讀是教育的靈魂，一切知識的基礎都從閱讀開始。」前總統陳水扁先生也曾提出:「在知識經濟的年代，影響國力的最關鍵因素，不在於土地的大小、武力的強弱、自然資源的多寡，而在於知識的累積、流通與應用。」聯合國經濟合作發展組織（Organization for Economic and Co-operation Development, OECD）所主導的國際學生評量計畫（Programme for International Student Assessment, PISA）便將閱讀素養定義為:「個體對書面文本的理解、運用和省思的能力，透過這種能力得以達成個人目標、發展個人知識和潛能與參與社會。」[19]綜合以上可知透過閱讀我們不僅得以站在巨人的肩膀

上，在這個資訊爆炸的年代，閱讀更成為了我們不可或缺的基礎能力。

在剛出爐的 2012 年 PISA 的結果報告，臺灣 15 歲學生的閱讀素養從 2009 的第 15 名進步到第 8 名，且在 2011 年國際教育成就調查委員會主辦的的促進國際閱讀素養研究」(Progress in International Reading Literacy Study, PIRLS)的調查研究中，亦顯示臺灣在國小學生閱讀素養方面從 2006 年的第 22 名進步到第 9 名。這兩項重要的國際閱讀評比的進步皆顯示出教育部近年致力推展中小學閱讀教育的成效。然 PIRLS 調查結果亦指出臺灣學生對閱讀素養喜歡的正向態度偏低。Chall & Conard(1991)指出唯有透過提供難易適中的文本方能提起學生的閱讀動機[16]，是故如何透過挑選符合的學生能力的閱讀文本，即高可讀性文本，以增進學生的閱讀動機成為一個值得探討的議題。

在西方國家有關可讀性的研究早已相當蓬勃，近年亦發展出線上文本分析系統 Coh-Metrix[18]，但反觀中文方面的相關研究卻相對較少，而中文文本分析的線上系統發展又更是缺乏。因此，本研究即是透過建立文本描述性指標-筆畫數及詞彙訊息相關指標-實詞比及一詞多義，以期未來能結合其他指標發展出一多樣性的中文文本自動化分析系統，為學生挑選出適讀的文本。

1.2 研究目的

根據上述研究動機，本研究目的如下：

(1) 建置中文文本實詞比、一詞多義、筆畫數文本自動化分析指標。
(2) 發展自動化中文文本線上分析系統。
(3) 國小文本實詞比與年級之趨勢探討與分析。
(4) 國小文本一詞多義與年級之趨勢探討與分析。
(5) 國小文本筆畫數與年級之趨勢探討與分析。

2. 文獻探討

2.1 線上文本分析系統 Coh-Metrix

Coh-Metrix 為美國曼菲斯大學所開發的多樣化自動化線上文本分析系統，自 2002 年開始其發展，至今已開發至 3.0 版(http://tool.cohmetrix.com/)，共包含 11 個面向及 106 項指標。其中 11 大面向包含了描述性指標、文本適讀性分數、參照凝聚力、潛在語意分析（LSA）、詞彙多樣性、關聯詞、情境模式、句法複雜度、句法密度、詞彙訊息及可讀性。[18]

目前本研究團隊已根據 Coh-Metrix 3.0 系統，並參照中文文本特性陸續發展了 33 項中文文本分析指標[4][5][9][11][14][15]。而本研究則延續前述研究，參考 Coh-Metrix 描述性指標及詞彙訊息這兩個面向所涵蓋的實詞比、一詞多義、字元數指標，並依照中文語法特徵作修改，作為中文文本分析指標的建置的依據。

2.2 實詞

詞彙是最小能獨立運用且有意義的語言單位[13]。而根據胡裕樹現代漢語增定本華語詞彙又可區分為實詞與虛詞兩類。實詞能充當句法成分且具有明確的詞彙意義，可以表示事物、動作行或狀態，而虛詞（另稱功能詞）通常不具實質意義無法充當句法成分的詞彙，其功能主要是和其他詞或語句建立關係。因實詞及虛詞分類標準，各家定義分歧，本研究根據現代漢語增訂本[2]，將實詞、虛詞分類如表 1 所示。

表 1 **實詞虛詞分類表**

實詞 **Content words**	名詞、動詞、量詞、形容詞、代詞、數詞、副詞
虛詞 **Function words**	介詞、連詞、助詞、嘆詞

2.3 一詞多義

語言學家 John Lyons (1977)根據語義的關聯性與否與詞源是否相同，將其劃分成兩種類別：同形異義(homonymy)與一詞多義(polysemy)[17]。同形異義指的是意義(meaning)上沒有相關聯、詞源不同且字形、發音都相同的詞彙。英文同形異義的例子如‘bark’，可以當作「吠叫」之意，亦能夠當作「樹皮」解釋；而中文的例子則如「開天窗」一詞，可以視為單純「打開天窗」的動作，也可以當作「做事出紕漏」的意思來解釋。相對而言，一詞多義指則是一個詞彙具有數個不同的詞義(sense)，其詞義彼此有關聯且詞源相同。英文中一詞多義的例子如‘man’，可以解視為「男人」，亦可以指稱「全體人類」，而中文的「掉漆」一詞，可以指稱「油漆脫落」，現在又常延伸為「遜掉了」

或「很瞎」的意思。

2.4 筆畫

西方的可讀性公式常以音節數作為判斷文章難度的指標。因為通常辨識長度較長的詞需要比辨識長度短的詞耗費較多的時間，而此現象又稱為詞長效果[8]。但是中文字不同於拼音文字，中文字是由筆畫與部件組成的方塊字，故我們可以視筆劃為影響中文認字的重要因素之一。楊孝濚曾以筆劃數為指標，發現筆劃數可以預測文本的難度，筆劃數越多，則閱讀的難度越高[20]。陳如玲與蘇宜芬在小學學童的實驗中發現以筆劃數為分析單位時，會有字元複雜度效果的結果[8]。胡夢軻的研究亦將中筆畫數視為中文可讀性分析前五重要的指標[3]。

3.研究方法

3.1 研究流程

本研究流程為：

(1) 指標建置：參考 Coh-Metrix 中描述性指標及詞彙訊息相關指標，選取與中文相關之部分指標，修改其計算方式使其適用於中文文本分析。

(2) 建立系統使用介面，置入研究指標，以供使用者操作、分析文本。

(3) 分析現有文章：取兒童語料庫文章，以建置指標分析，了解國內現行學童學習教材內容在各指標上的趨勢及意義，並考慮年級及科目上的差異。

3.2 自動化文本分析指標建置

3.1.1 實詞比

本指標計算係以中研院斷詞系統[1]先進行斷詞，並以斷詞系統的詞性分類為基礎，將名詞、動詞、形容詞、副詞、代詞、量詞視為實詞，予以計算單篇文章實詞數佔總詞數的比例。公式如下：

$$\text{實詞比}=\frac{\text{實詞數}}{\text{總詞數}} \quad (1)$$

3.1.1 一詞多義

本指標計算係以中研院斷詞系統[1]先進行斷詞，並以中文詞彙網路[10]所建立的詞彙意義做計算。當文本出現中文詞彙網路未建立詞彙，其詞義以最小詞義 1 計算之。公式如下：

$$\text{一詞多義}=\frac{\text{單詞總字義數}}{\text{總詞數}} \quad (2)$$

3.1.1 筆畫數

本指標之筆畫數建置係以教育部 1979 年所公布之常用國字標準字體表[6]所收錄之 4808 字為依據，當文本出現常用國字標準字體表所未收錄之國字時，則以教育部國語辭典簡編本[7]收錄之國字筆畫為計算之依據。筆畫數之計算又以筆畫數 10、20 為切割點，區分為低筆畫字數、中筆畫字數及高筆字畫數三類，並將其除以該文本總字數計算為低筆畫數比、中筆畫數筆及高筆畫數比。又平均筆畫數之計算即為總筆畫數除以總字數。公式如下：

$$\text{平均筆畫數}=\frac{\text{總筆畫數}}{\text{總字數}} \quad (3)$$

$$\text{高筆畫數比}=\frac{\text{筆畫數}>20\text{的字數}}{\text{總字數}} \quad (4)$$

$$\text{中筆畫數比}=\frac{20\geq\text{筆畫數}>10\text{的字數}}{\text{總字數}} \quad (5)$$

$$\text{低筆畫數比}=\frac{\text{筆畫數}\geq 10\text{的字數}}{\text{總字數}} \quad (6)$$

3.3 線上自動化文本分析系統

本研究發展文本自動化分析指標實詞比、一詞多義、筆畫數指標外，並以此為基礎發展線上自動化文本分析系統。使用者可以自行線上輸入文本標題、資料來源及文本內容，並勾選欲分析之指標進行自動化文本分析。詳如圖 1 至圖 3 所示：

圖 1 系統登入介面

圖 2 **文本輸入與指標勾選介面**

圖 3 **文本自動化分析結果**

4.研究結果

以下為應用本研究所建置之中文文本自動化分析系統，分析國小現行教科書探討國小教科書國語、自然與社會科不同年級文本自動化分析之趨勢，結果討論如下。

表 2 為國語、自然、社會科在不同年級實詞比的指標值，而圖 4 為其趨勢分析。其結果顯示國語及社會科之趨勢甚為接近，在三到六年級中社會科實詞比指標數值又皆略高於國語科實詞比。國語與社會科除六年級外皆隨年級增加而降低，顯示國語與社會課文隨著年級增加而實詞所佔比例降低。造成此趨勢的可能原因為，低年級所學的詞彙包含較多的名詞、動詞、形容詞、副詞等基本詞彙，而隨著年級增加，開始增加連接詞、感嘆詞、介詞、助詞等虛詞的學習，以使文意的表達更加完整。而自然科各年級指標值皆低於國語與社會科許多，且自四年級起隨著年級增加，實詞比指標有增加的趨勢。

表 2 **兒童語料庫各年級各科目實詞比指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	0.8185	0.8107	0.7951	0.7894	0.7842	0.7872
自然			0.7691	0.7582	0.7689	0.7763
社會			0.7987	0.7922	0.7862	0.7915

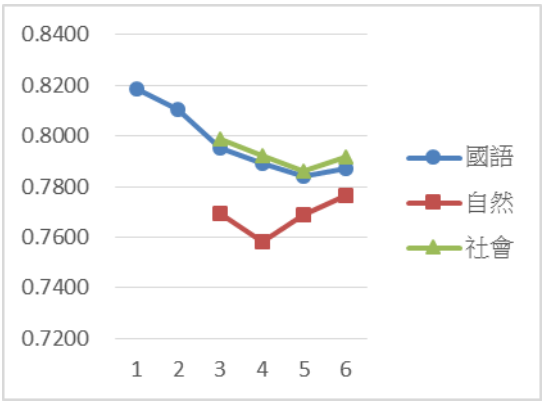

圖 4 **兒童語料庫各年級各科目實詞比指標趨勢**

表 3 為國語、自然、社會科在不同年級一詞多義的指標值，而圖 5 為其趨勢分析。其結果顯示在一詞多義這個指標上，各年級國語科皆為最高，自然科次之，社會科則為最低。造成此現象的可能原因為自然科的文本多為說明文，在用字譴詞上意義較為明確，而國語科恰好與之相反之故。並且在國語及自然兩個科目上有隨著年級增加，指標值漸趨降低的趨勢。相對於此，社會科的指標值則維持在一相差不大的數值上。造成低年級的文本明顯較高年級一詞多義指標為高的可能原因為，在低年級所學的詞彙，通常較為簡單，卻可能同時含有不同的意思和用法，而我們會在學習的過程中慢慢學到這個字詞的其他意思。而較為生難的詞彙，可能其含意反而較為單一而特定。以下舉「好像」及「範疇」這兩個詞在中文詞彙網路中的意思為例：

好像 $_1$：表與事實不符，但感覺逼真的狀況。
好像 $_2$：表以與前述事物有共同點的後述事物作為比喻。
好像 $_3$：表事件有可能不確定，用於委婉語氣。
好像 $_4$：表說話者主觀的印象。
範疇 $_1$：比喻特定對象所涉及到的所有相關事物。

表 3 **兒童語料庫各年級各科目一詞多義指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	6.9840	6.3693	5.5362	5.0325	4.8391	4.7050
自然			4.8399	4.8827	4.3325	4.2818
社會			3.7975	3.4925	3.7065	3.7313

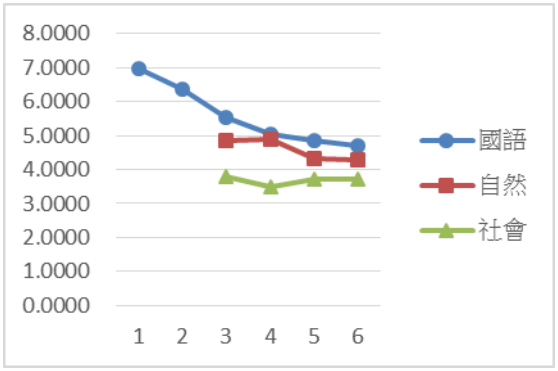

圖 5 **兒童語料庫各年級各科目一詞多義指標趨勢**

表 4 為國語、自然、社會科在不同年級平均筆畫數的指標值，而圖 6 為其趨勢分析。其結果顯示在各年級的文本中，平均筆畫數大約介於 8 至 10 劃之間。其中國語科的平均筆畫數皆低於自然與社會科。國語科及自然科平均筆畫數指標皆隨年級增加而有上生的趨勢，與此相對，社會科的平均筆畫數指標則維持在一相差不大的數值上。

表 4 **兒童語料庫各年級各科目平均筆畫數指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	8.2713	8.8204	9.0524	9.1633	9.1588	9.2442
自然			9.1324	9.3121	9.5049	9.6605
社會			9.4501	9.4357	9.3966	9.4396

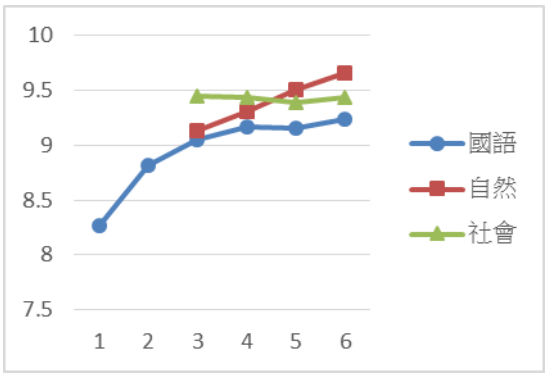

圖 6 **兒童語料庫各年級各科目平均筆畫數指標趨勢**

表 5 為國語、自然、社會科在不同年級高筆畫數比的指標值，而圖 7 為其趨勢分析。其結果顯示高筆畫數比指標在各年級的文本中，所佔比例僅 1%至 3%。其中各年級指標數皆以自然科為最高，社會科次之，而以國語科為最低。國語科一至三年級隨著年級增加高筆畫數的字數比也隨之增加，但三年級後此趨勢轉為持平。自然科亦有隨著年級增加高筆畫數的字數比增加的現象。相對而言社會科的高筆畫數字數比則並未呈現一定趨勢。

表 5 **兒童語料庫各年級各科目高筆畫數比指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	0.0096	0.0127	0.0165	0.0158	0.0157	0.0166
自然			0.0282	0.0316	0.0327	0.0327
社會			0.0208	0.0197	0.0259	0.0229

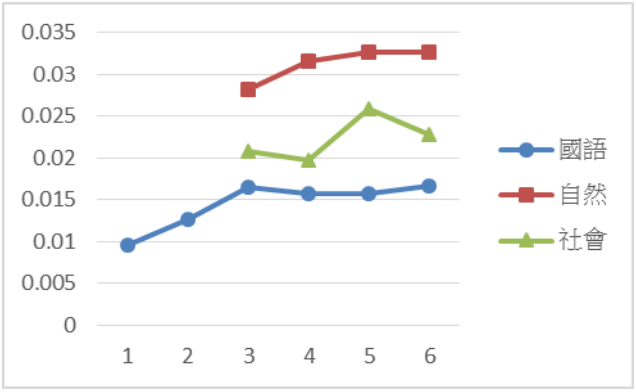

圖 7 **兒童語料庫各年級各科目高筆畫數比指標趨勢**

表 6 為國語、自然、社會科在不同年級中筆畫數比的指標值，而圖 8 為其趨勢分析。其結果顯示中筆畫數比指標在各年級、所佔比例

約在 26%至 36%之間。各科目的文本中，除了四年級社會科外，皆呈現隨著年級增加中筆畫數比亦呈現升高的趨勢。且在各年級的文本中，中筆畫數比皆以社會科指標數為最高，國語科次之，自然科為最低。

表 6 **兒童語料庫各年級各科目中筆畫數比指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	0.2600	0.3162	0.3281	0.3355	0.3389	0.3422
自然			0.2939	0.3195	0.3301	0.3341
社會			0.3396	0.3589	0.3474	0.3491

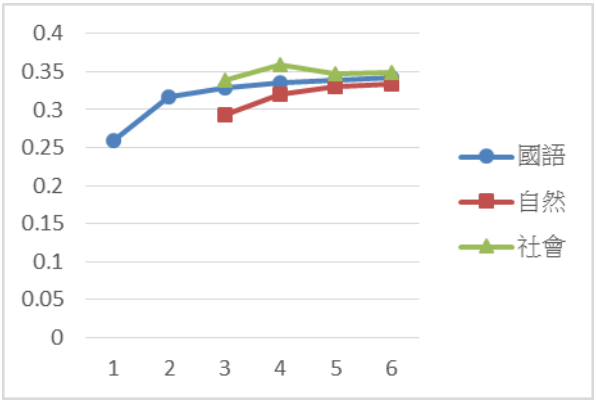

圖 8 **兒童語料庫各年級各科目中筆畫數比指標趨勢**

表 7 為國語、自然、社會科在不同年級低筆畫數比的指標值，而圖 9 為其趨勢分析。其結果顯示低筆畫數比指標在各文本所佔比例約在 62%至 73%，表示國小文本字詞的筆畫數以 10 劃以下比例最高，此結果與平均筆畫數的結果相同。此外，在國語與自然科文本中皆有隨著年級增加而指標值隨之降低的趨勢。而社會科在三、四年級間低筆畫數比指標亦有降低之趨勢，與此相對在四年級後則呈現微幅上升之情形。

表 7 **兒童語料庫各年級各科目低筆畫數比指標數值**

	一年級	二年級	三年級	四年級	五年級	六年級
國語	0.7304	0.6712	0.6553	0.6487	0.6454	0.6412
自然			0.6779	0.6489	0.6372	0.6332
社會			0.6395	0.6214	0.6267	0.6280

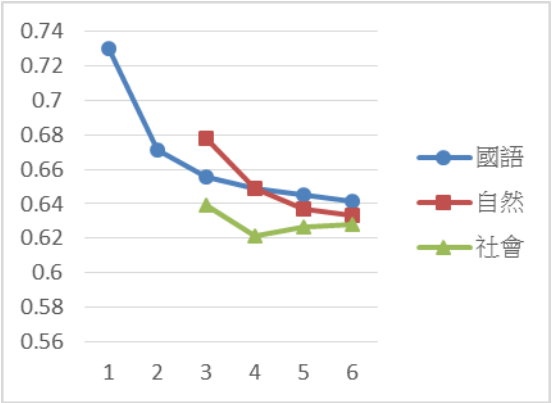

圖 9 **兒童語料庫各年級各科目低筆畫數比指標趨勢圖**

5. 主要內容

本研究主要目的為建置自動化文本實詞比、一詞多義、筆畫數指標，以分析現有國小教科書中。國語科、自然科和社會科相關詞彙現況，了解教科書中凝聚性趨勢。其結論如下：

(1) 國語科與社會科課文除六年級外有隨著年級增加而實詞比指標降低的現象。造成此現象的可能原因為，低年級所學的詞彙包含較多有實質意義的實詞，而隨著年級增加，開始增加虛詞的學習所致。

(2) 一詞多義指標，各年級國語科皆為最高，自然科次之，社會科則為最低。並且在國語及自然兩個科目上有隨著年級增加指標值降低的趨勢。而社會科的指標值在各年級文本則差異不大。

(3) 全部文本的平均筆畫數大約介於 8 至 10 劃之間，低筆畫字數佔全文六至七成。國語科及自然科的筆畫數相關指標有以下趨勢，平均筆畫數、中筆畫數比隨年級增加而升高，低筆畫數比隨年級增加而降低。國語科 1-3 年級及社會科全年級高筆畫數比亦有隨年級增加而升高的現象。相對而言，社會科的文本較為呈現固定的趨勢。

綜合上述研究成果，本研究未來目標為結合過去已建立的詞類、詞頻、關聯詞、重複語詞、潛在語意、句子結構等不同面向的分析指標作更進一步的整合分析，目的希望能提供一個易於操作的自動化文本分析系統以供教師或家長們使用，以期藉此挑選出適合學童閱讀的中文文本。

參考文獻

[1] ***中央研究院數位典藏國家型科技計畫中文斷詞系統***，取自：http://ckipsvr.iis.sinica.edu.tw/

[2] 胡裕樹，***現代漢語詞類（增訂本）***，上海：上海教育出版社，1994。

[3] 胡夢珂，***使用支援向量機進行中文文本可讀性分類-以國小國語課本文為例***，國立臺灣師範大學資訊教育學系碩士論文，2010。

[4] 倪雅真、張琇涵、廖晨惠、白鎧誌，”中文文本自動化指標建置與探討-句子最小編輯距離與結構相似度”，***十九屆資訊管理暨實務研討會***，2013。

[5] 陳文蘭，”兒童文本關聯詞指標分析系統建置與應用”，***國立台中教育大學教育測驗統計研究所碩士論文***，2013。

[6] 教育部，***常用國字標準字體表***，正中書局，1979。

[7] 教育部，***教育部國語辭典簡編本***，取自：http://dict.concised.moe.edu.tw

[8] 陳茹玲、蘇宜芬，”國小不同認字能力學童辨識中文字詞之字元複雜度效果與詞長效果研究”，***教育心理學報***，41(3)，2010。

[9] 陳建宏，”兒童文本詞類指標分析系統建置與應用”，***國立台中教育大學教育測驗統計研究所碩士論文***，2013。

[10] 黃居仁、謝舒凱，***跨語言知識表徵基礎架構—面向多語化與全球化的語言學研究***，國科會專題補助計畫 (NSC 96-2411-H-003-061-MY3)，2007-2010。

[11] 黃勇嫻，”兒童文本重複指標分析系統建置與應用”，***國立台中教育大學教育測驗統計研究所碩士論文***，2013。

[12] 廖晨惠，***閱讀研究議題八：以 LSA 為基礎之電腦化閱讀認知測驗及 AutoTutor 建置***，國科會專題補助研究計畫(NSC 100-2420-H-142-001-MY3)，2011。

[13] 劉月華、潘文娛、故韡（2001）。實用現代漢語語法（增訂本）。北京：商務印書館。

[14] 蔡亞韋，”兒童文本潛在語意指標分析系統建置與應用”，***國立台中教育大學教育測驗統計研究所碩士論文***，2013。

[15] 蔡筱倩，”兒童文本詞頻詞彙指標分析系統建置與應用”，***國立台中教育大學教育測驗統計研究所碩士論文***，2013。

[16] Chall, J. S., and Conard, S. S., ***Should textbooks challenge students? The case for easier or harder textbooks***, New York: Teachers College Press, 1991.

[17] Lyons, J, ***Semantics***, Cambridge: Cambridge University Press, 1977.

[18] McNamara, D. S., Graesser. A. C., McCarthy, P. M., and Cai, Z.. ***Automated evaluation of text and discourse with coh-metrix,*** Institute for Intelligent Systems, University of Memphi s, 2012.

[19] OECD, ***PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy***, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

[20] Yang, S. J., ***A readability formula for Chinese language***, Unpublished doctoral dissertation, University of Wisconsin, Wisconsin, 1971.