兒童文章重複語詞分析自動化指標建置與應用-以名詞、動詞、實詞為例

黃勇娟
國立臺中教育大學教育測驗統計研究所
碩士生
huangyc@webmail.ptes.tc.edu.tw

蔡亞韋
國立臺中教育大學
教育測驗統計研究所
碩士生
barlow47@yahoo.com.tw

郭伯臣
國立臺中教育大學
教育測驗統計研究所
教授
kbc@mail.ntcu.edu.tw

廖晨惠
國立臺中教育大學
特殊教育學系
教授
chenhueiliao@gmail.com

白鎧誌
國立臺中教育大學
教育測驗統計研究所
博士生
minbai0926@gmail.com

摘要

本研究目的為建置文本詞彙重複出現自動化分析指標以進行自動化文本特徵分析，並探討國小學童教科書-國語科、社會領域、自然領域中，詞彙重複出現(動詞、名詞及實詞)對文章的共同參照凝聚力的影響，本研究根據國外的線上文本分析器 Coh-Metrix 發展文本自動化分析指標並發展線上文本自動化分析系統，其文本自動化分析之研究結果如下：

一、 一年級國語科重複指數較高，由於一年級學童先備知識不足，所以較高的重複性指數能幫助理解。

二、 五年級自然領域，重複指數分數較高，顯示在重複語詞行程較強的連貫性，可以加強較不熟悉或是較專業的階段領域課程。

六年級所有重複性指數皆低，顯示文本聯貫性低，因此文本較難理解。

關鍵詞：詞彙重複出現、Coh-Metrix、文本自動化分析。

Abstract

The present study developed the computer analyses of texts for the characteristics of texts. We provided 4 validated indices: verb overlap, noun overlap and content words overlap, and analyzed the characteristics of texts on different grade level. The results are summarized as follows:

The level of overlap is higher in grade1 because of their poor prior knowledge. The higher overlap can help them to comprehend. Grade3 and grade5 had higher level of overlap in natural science filed. It means that the overlap has the stronger connection and it can strengthen unfamiliar or professional lessons. The level of overlap is almost lower in grade6. It means the texts were difficult to comprehend because of their low connection.

Keywords: computer analyses of texts、verb overlap、noun overlap、content words overlap.

1. 前言

閱讀能力與國家競爭力成正相關，閱讀力越強的國家，國家就越具有競爭力[4]。世界各國評比無不以「閱讀素養」為檢視核心。2012年底，國際教育學習成就調查委員會(IEA)，公佈最新「促進國際閱讀素養研究」(PIRLS)，與「國際數學與科學教育成究趨勢調查」(TIMSS)的國際評比，發現台灣中小學生的閱讀素養排名，從 2007 年的二十二名，進步到全球第九名。數學與科學教育的評比成績，台灣則較 2007 年時排名第一略退步，排名全球第三，僅次於韓國與新加坡。但分數略有進步。雖然台灣學生學習成就進步了，但是「高成就、低興趣、低自信」，這種情況，日益嚴重。師大教授林福來指出「人類學習的興趣，來自於探究之後的滿足」，如果教學以每天重複考試來學習，學生對知識就會失去興趣。

近年台灣教育政策與國際接軌，利用各種閱讀策略指導學生閱讀，提升學生的閱讀力及興趣是教師現行指導學生的要務。在國中小積極辦理規劃精進閱讀教學計畫，各種閱讀策略因應而生，舉凡晨間閱讀，讀報，數位閱讀推展活動等多元登場。然而閱讀文本的選擇需考慮其連貫性，有四種不同的知識對有效的閱讀理解是必備的：內容知識、文章結構組織、策略知識、後設認知知識。尤其是在文章結構組織的知識建構中，編撰教科書需注意連貫性。在文章整體結構或組織上，必須對重要概念加

以統整，以達到整體連貫性（Global coherence）；對文字間或句子間的觀念應產生有效聯結，以促進局部連貫性（Local coherence）[7]。

本研究應用中央研究院開發之斷詞系統並參考美國曼菲斯大學開發的線上文本分析器Coh-Metrix，建置文本重複詞分析的自動化指標，並針對現行國小教科書進行文本分析以了解國小教科書文本之特徵。

2. 文獻探討

2.1 中文詞彙的特性

黎錦熙 1933 年在「國語文法」有意用漢語的「詞」 對等表示英語的「word」,「字」和「詞」兩個述語在漢語中逐漸分工[5]。

中文分詞(Chinese word segmentation)的目的就是將文本的句子切分成詞，使其成為詞序列的形式[7]。

漢語的詞是由語素構成。語素是最小的語音與語義結合體，是最小的語言單位[1]。把一個語言片段，切分到不能再分的最小單位，就是語素。例如：

遵｜守｜交｜通｜安｜全｜生｜命｜有｜保｜障

詞是比語素高一級的語言單位,詞的數量非常龐大,可使用詞性(part-of-speech)來概括一個詞在一個句子中所展現的句法功能和意義，詞性標註的目標就是在產生中文分詞的詞序列時，給每個產生的詞標註一個詞性[3]。

「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[1]，將實詞、虛詞分類如表 1 。

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

從自然語言的處理 (Natural language Processing)應用和理解的角度來看，中文分詞的詞性標註是許多應用的基礎。包含句法分析、信息提取、機器翻譯等，這些都可以從好的中文分詞詞性標註模型中獲得較正確的結果[2] [3]。

以自然語言理解(Natural language Understanding)的角度看來，中文分詞詞性的標記是中文理解的基礎步驟。

2.2 Coh-Metrix

Coh-Metrix 是一種智慧型工具，提供文和論述的語言數據索引[9]。這些數據可以用許多不同的方式，明確的分析文本的凝聚力與心理表徵的連貫性。

Coh-Metrix 的核心是以凝聚力(cohesion)為最重要的評估，卻在可讀性評估方法上常被忽略。在 Coh-Metrix 中，凝聚力的定義是：經由明確的文本特點發揮作用，幫助讀者連接文本想法[11]。凝聚力可連結文字間想傳達的結果和概念，也可結合文章中詞和句子關係，更可聯繫句子、文本和讀者的想法。具有高凝聚力的文本，需包含詞和想法的重複，如此不僅在句子間，也在全文間形成明確的線索，可幫助讀者加快理解或推論文本間的關係。然而凝聚力較低之文本，若讀者先備知識夠多，則可以刺激讀者產生推論極更多的解釋，反之，會因缺乏線索，較難連接文本與讀者的想法。

Coh-Metrix 會依據版本和工具，評估特定的指標。目前發展至 3.0 版本，指標數增加狀況如下：

表 2 Coh-Metrix2.0 與 3.0 版本比較

Coh-Metrix 版本	指標個數
2.0	60
3.0	106

本研究是以 Coh-Metrix 中，共同參照凝聚力(Co-Referential Cohesion)研究為主。在早期，兩個句子裡，有一個共同的參數(如：名詞，動詞..等)，這兩個句子就具有共同參照凝聚力。一個明確有力的凝聚力來源是參照(referential)和語意的重複，其出現在相鄰句中，段落中或相鄰段落中的句子，句中的詞、概念或想法重複，構成了句子之間的聯繫。當文字、概念或想法重複於句型中時，便可以形成銜接多個句子的連結[12] 。共同參照指數關係著文章中語意是否連接的一個重要指標[11]。若凝聚力指數過低則會出現理解斷層或增加閱讀時間[8]。

共同參照凝聚力包含了實詞重複指標（content word overlap）、名詞重複指標（noun overlap）、動詞重複指標（verb overlap）、參數重複指標（argument overlap）、詞幹重複指標（stem overlap）。其計算方式又分為兩部份：相鄰句凝聚力指標(Co-Referential Cohesion

local)及文章整體凝聚力指標(Co-Referential Cohesion global)。

以上所述共同參照指標項目皆以英文文本研究發展而成[8],其中參數重複指標(argument overlap)和詞幹重複指標(stem overlap)在中文文法上無共同處,因此本研究針對實詞重複指標(content word overlap)、動詞重複指標(verb overlap)與名詞重複指標(noun overlap)進行發展與探討。

3.研究方法

本研究主要目的為發展自動化文本重複性指標,其指標說明如下:實詞重複指標(content word overlap),動詞重複指標(verb overlap),名詞重複指標(noun overlap)。其指標計算方式皆包含相鄰句凝聚力指標計算(公式(1))與文章整體凝聚力指標計算(公式(2)) [10]。其公式說明如下。

3.1 文本自動化重複指標建置

3.1.1 相鄰句凝聚力(adjacent sentences)

相鄰句之文本凝聚力計算是利用運用重複得分(Repetition score)計算,假若相鄰兩句中有相同詞彙出現,則有共同參數,表示值為 1,否則為 0,此為測量相鄰兩句之間的共同參照凝聚力。文本中相鄰句的比對方式為:第 1 句 v.s.第 2 句,第 2 句 v.s.第 3 句..至文本結束。其範例如下,S_1~ S_4 為文本中的句子。

S_1 今天是個晴天,
S_2 老師喜歡利用晴天的日子帶我到操場玩,
S_3 通常會踢球、盪鞦韆,還能捉迷藏
S_4 我愛晴天

其名詞的共同參照凝聚力矩陣詳見表 3,S_1 與 S_2 有相同的名詞詞彙,得分為 1,S_2 與 S_3 無相同的詞彙,得分為 0。

表 3 共同參照凝聚力文本矩陣

	S_1	S_2	S_3	S_4
S_1	1	1	0	1
S_2	1	1	0	1
S_3	0	0	1	0
S_4	1	1	0	1

相鄰句之文本凝聚力計算公式如下,其值越大,凝聚力高,值越小,凝聚力低:

$$\text{Co - reference cohesion local} = \frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1} \quad (1)$$

n=句子數 ,i 為文章中欲比對的句子

3.1.2 文章整體凝聚力(all sentences)

本研究使用矩陣來描述文本段落之間的重複。$n \times n$ 矩陣中,共同參照凝聚力被稱為 R。文本中有 n 個句子,定義為 S_1,S_2,….S_n,兩個句子用 S_i 和 S_j 表示, R_{ij} 為二者的參照凝聚力。如果兩個句子至少有一個共同參照的關係,表示值是 1,否則為 0。

文本中文章整體凝聚力的比對方式為:第 1 句 v.s.第 2 句,第 1 句 v.s.第 3 句.. 第 1 句 v.s.第 11 句;第 2 句 v.s.第 3 句..以此類推至文本結束。

文章整體凝聚力考慮文章中每一句對文章整體的凝聚力的影響程度,兩個句子的遠近將影響凝聚力的高低。距離近,兩句子之間的凝聚力關係緊密,則權重考量較高;反之,距離遠,兩句子之間的凝聚力關係較疏遠,則權重考量較小,因此以兩句子之間的距離的倒數做為加權依據,如表 4。文本中兩句的距離為 1,2,3, . . . , k, 其距離倒數則為 1, 1/2, 1/3, . . . , 1/k,.若兩句有語詞重複,則加權分數為句子距離的倒數,例如:S_1 與 S_4 有名詞重複,距離為 4-1=3,加權分數為 1/3;S_2 與 S_4 有重複,距離為 4-2=2,加權分數為 1/2。

表 4 共同參照凝聚力文本矩陣

	S_1	S_2	S_3	S_4
S_1	1	1	0	1/3
S_2	1	1	0	1/2
S_3	0	0	1	0
S_4	1/3	1/2	0	1

其計算文本整體凝聚力是將比對值帶入 Global 公式,如公式(2),其值越大,凝聚力高,值越小,凝聚力低。

$$\text{Co - reference cohesion global} = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n} R_{ij}, i<j}{n \times \frac{n-1}{2}} \quad (2)$$

其中 n=句子數,i、j 分別為文章中欲比對的句子。

3.2 線上文本自動化分析系統

本研究發展文本自動化分析之重複性指標,並發展線上文本自動化分析系統,使用者可以透過系統進行文本自動化分析以了解文本的重複性指標,圖 1 到圖 3 為系統使用介

面，使用者可以輸入文章標題、文章來源語文章內容，並勾選欲分析之指標即可以進行文本特徵的自動化分析。

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

3.3 文本分析

本研究使用之文本資料來源為廖晨惠(2010)國科會「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」計畫(編號：100-2420-H-142-001-MY3)所建置的國小語料庫[6]，本研究挑選國小一至六年級課文中國語、自然與社會科目，以比較不同科別、年級之詞彙，分析文本重複語詞與凝聚力的趨勢。

4.研究結果

本研究為發展自動化文本重複性指標並探討國小教科書國語、自然與社會不同年級文本自動化分析之趨勢，其研究結果如下。

圖 4 為國語、自然、社會動詞重複相鄰句之文本凝聚性在不同年級之文本自動化分析之趨勢，其結果顯示在國語文本，一年級動詞重複凝聚性相對比二年級高，是因為讀者先備知識不足，需要語詞重複出現幫助理解。自然領域動詞的局部重複凝聚性顯示，一~五年級並無太大差距，而六年級曲線下降，凝聚力分數低。社會科文本分析則呈現隨著年級增高而下降。

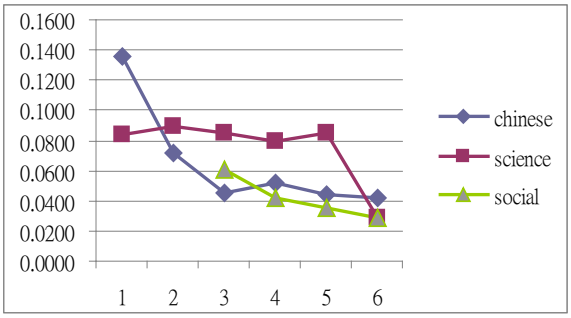

圖 4 不同年級文本凝聚性(重複相鄰句-動詞)

圖 5 是不同年級文本整體凝聚性(動詞)之趨勢分析，其結果顯示國語科在一~二年級重

複凝聚性下降明顯，是因為動詞文章整體間重複凝聚力距離加權，使一年級凝聚力指數更突顯。自然領域動詞文章整體間重複凝聚性稍提升，是因為一、二年級自然包含生活領域課程，而三年級獨立出自然領域，所以需要較多重複語詞形成連貫性來幫助理解。社會科動詞文章整體間重複凝聚性則隨著年級升高而下降。

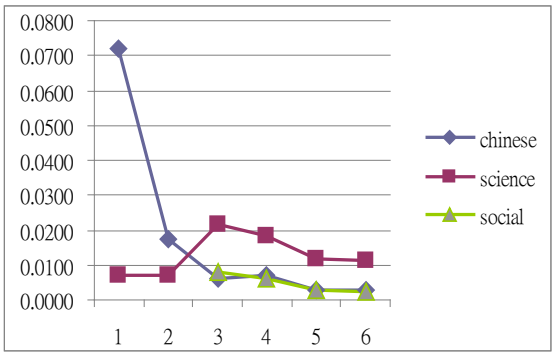

圖 5 不同年級文本整體凝聚性(動詞)

圖 6 為不同年級文本整體凝聚性(名詞)之趨勢分析，分析結果顯示一~二年級中，國語科及自然領域的名詞相鄰句重複凝聚性下降明顯，是因為讀者先備知識較低，所以需要較高重複來幫助理解。自然領域名詞相鄰句重複凝聚性稍提升，是因為一、二年級自然涵括於生活領域課程，三年級獨立出自然領域，所以需要較高重複來幫助理解。

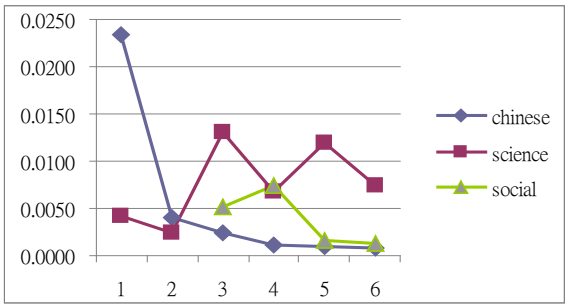

圖 6 不同年級文本凝聚性(重複相鄰句-名詞)

圖 7 為文本整體凝聚性(名詞)之趨勢分析，結果顯示國語科名詞文章整體重複凝聚性隨級別漸高而降低。自然科以三年級名詞文章整體重複凝聚性相對較高。

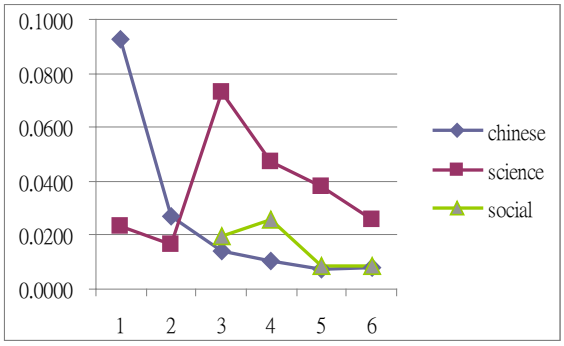

圖 7 不同年級文本整體凝聚性(名詞)

圖 8 為不同年級文本凝聚性(重複相鄰句-實詞)之趨勢分析，研究結果顯示實詞的相鄰句凝聚性分數皆高於名詞與動詞的相鄰句凝聚性。自然領域的相鄰句凝聚性高於社會領域及國語科，顯示出自然領域的學習藉由實驗操作及結果歸納重複描述，協助學生理解習。六年級重複指數降低，凝聚性低。從社會領域的實詞相鄰句凝聚性中，可看出三、四年級高於五、六年級。

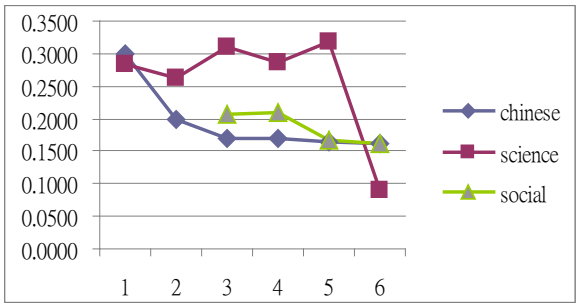

圖 8 不同年級文本凝聚性(重複相鄰句-實詞)

圖 9 是文本整體凝聚性實詞指標分析，分析結果顯示實詞的文章整體凝聚性皆高於名詞與動詞的文章整體凝聚性。

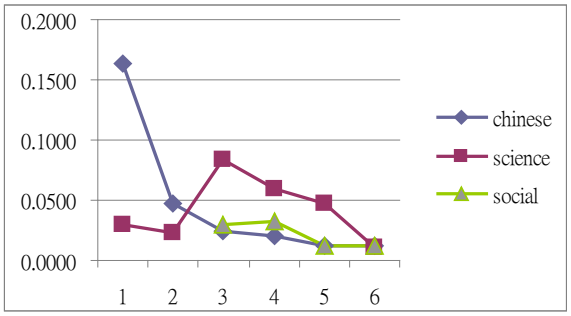

圖 9 不同年級文本整體凝聚性(實詞)

5. 結論

本研究主要目的為建置自動化文本重複性指標，包括實詞重複指標，動詞重複指標，名

詞重複指標，以分析現有國小教科書中。國語科、自然科和社會科詞彙重複現況，了解教科書中凝聚性趨勢。其結論如下：

（1）一年級國語科重複凝聚性較高，對先備知識較不足的一年級可以協助理解閱讀文本。

（2）三、五年級自然領域，重複凝聚性較高，而九年一貫年段階段中，三、五年級各為一新階段，重複語詞形成較強的連貫性，可以加強不熟悉的階段領域課程，以加強舊經驗的連結。

（3）六年級所有科別重複凝聚性皆低，顯示文本連貫性低，因此文本難度相對較高。然而凝聚力較低之文本，若讀者先備知識夠多，則可以刺激讀者產生推論極更多的解釋，反之，會因缺乏線索，較難連接文本與讀者的想法。

(4)名詞重複平均值<動詞重複平均值<實詞重複平均值，名詞若於文章中過度重複，並不非是篇好文章。

本研究建置電腦自動化文本分析指標，可觀察出國小教科書國語、自然與社會科詞彙重複特徵和趨勢，提供學生在課程學習中，掌握凝聚性較低之閱讀關鍵，加強複習或增加閱讀量，擴充先備知識，以提高學習能力。

參考文獻

[1] 胡裕樹，***現代漢語***，新文豐出版公司。

[2] 侯呈風，”基於 HMM 的哈薩克語詞性標注研究”，***新疆大學論文***，2011。

[3] 張開旭，”使用壓縮表示的中文分詞詞性標注研究”，***清華大學工學博士論文***，2012。

[4] 張瓊元，”國際性學生閱讀能力評量之分析”，***暨南國際大學教育政策與行政研究所碩士論文***，2003。

[5] 彭澤潤，***詞和字研究-中國語言規劃中和漢語個性***，中國文史出版社。

[6] 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***，行政院國家科學委員會專題補助研究計畫，2010。

[7] 鄧守信，***對外漢語教學法***，文鶴出版社。

[8] 顏若映，”教科書內容設計與閱讀理解之認知研究”，***教育與心理研究***，第 15 期，pp.101-128，1992。

[9] Coh-metrix2.0(http:cohmetrix.memphis.edu/CohmetrixWeb2/HelpFile2.htm)

[10] Graesser. A. C. McNamara, D. S., Louwerse, M. M., & Cai, Z. ***Cometrix: Analysis of text on cohesionand language.*** *Behavior Research Methods, Instruments & Computers, 36(2), 193-202. 2004*

[11] Halliday, M. A., and Hasan, R. ***Cohesion in English.*** London: Longman. *1976.*

[12] McNamara, D. S., Graesser, A. C., & Louwerse, M. M., (in press). ***Sources of text difficulty: Across the ages and genres.*** *In J. P. Sabatini & E. Albro (Eds.). Assessing reading in the 21” century: Algning and applying advances in the reading and measurement science. Lanham, MD: R&L Education.*