

國立台灣師範大學  
資訊教育研究所碩士論文

指導教授： 張國恩 博士  
宋曜廷 博士  
張道行 博士

使用支援向量機進行中文文本可讀性分類

-以國小國語課文為例

**Using the Support Vector Machine to classify the  
Chinese text readability**

**– A Case of Elementary Chinese Textbook**

研究生：胡夢珂 撰

中華民國一百年七月

國立臺灣師範大學 資訊教育研究所  
博/碩士論文通過簽名表

系所別：資訊教育研究所

姓名：胡夢珂

學號：698080296

博/碩士論文題目：使用支援向量機進行中文文本可讀性分類-  
以國小國語課課文為例

經審查合格，特予證明

論文口試委員



侯惠澤 博士  
國立臺灣科技大學 工程技術研究所 教授

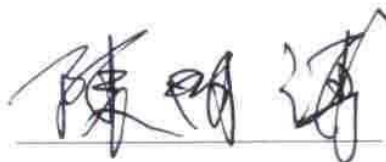


張國恩 博士  
國立臺灣師範大學資訊教育研究所 教授  
論文指導教授



宋曜廷 博士  
國立臺灣師範大學 教育心理與輔導研究所 教授  
論文指導教授

所長簽章：



中華民國 100 年 7 月 18 日

## 國立臺灣師範大學學位論文授權書

本授權書所授權之論文為授權人在國立臺灣師範大學 \_\_\_\_\_ 教育 \_\_\_\_\_ 學院  
\_\_\_\_\_ 資訊教育 \_\_\_\_\_ 研究所 99 學年度第 2 學期取得 碩 士學位之論文。

論文題目：使用支援向量機進行中文文本可讀性分類以國小國語科課文為例

指導教授：張國恩老師 / 宋曜廷老師

### 授權事項：

一、 授權人同意非專屬無償授權本校將上列論文全文資料以微縮、光碟、數位化或其他方式進行重製作為典藏之用。本校在上述範圍內得再授權第三人進行重製。

二、 授權人 ☒ 同意 ☐ 不同意 非專屬無償授權本校及國家圖書館將前條典藏之資料收錄於資料庫，並以電子形式透過單機、網際網路、無線網路或其他傳輸方式，提供讀者基於個人非營利性質之線上檢索、瀏覽、下載、傳輸、列印等利用。本校得將上述權利再授權于第三者。

三、 論文全文電子檔上載網路公開時間：【第二項勾選同意者，以下須擇一勾選】

☐ 即時公開

☒ 自 105 年 7 月 18 日始公開

授權人姓名：胡夢珂 (請親筆正楷簽名)  
學 號：698080296

註：1. 本授權書須列印並簽署兩份，一份裝訂於紙本論文書名頁，一份繳至圖書館辦理離校手續

2. 授權事項未勾選者，分別視同「同意」與「即時公開」

中 華 民 國 100 年 8 月 8 日

## 摘要

語文能力在各方面都扮演著重要的角色。而獲取語文能力最重要、最直接的管道之一就是透過閱讀。可讀性可以評估一個文本是否適合閱讀者的閱讀能力。以往的研究指出可讀性公式是一個工具，可以把對於不同教育程度的讀者所閱讀的文章加以調整。英文文本的可讀性研究很早就出現了，可是中文領域這方面的研究不多，而中文能力在現今社會又是一個很主要的趨勢。因此，一個適合文本可讀性的分類方法是很重要的。過去西方學者因為過去技術的不足多採用線性的可讀性公式對文本做可讀性分類，而線性的可讀性公式對本研究的資料有些限制，因此本研究的目的是在建立一個由支援向量機(Support Vector Machine, SVM)所訓練產生的預測模型，將國小的國語科課文做可讀性的分類。進而觀察預測的課文跟原來實際的課文的年級是否相符，並針對錯誤的課文做分析，以改善與謀求分類上的準確性。

本研究以課程專家編撰，經國家編審單位審定的三個民間版本教科書(H版、K版、N版)，國小一年級至六年級國語科課文刪減掉新詩、絕句、古文、律詩的課文後共計 386 篇為實驗資料，將課文一部分做為訓練資料，另一部分課文為測試資料，透過中文斷詞的處理及資料格式的轉換，最後以 SVM 來對文本的可讀性進行分類。研究結果發現：利用 LIBSVM 預測國小國語科課文冊別的準確率(accuracy)為 47.92%、正確率(fit rate)為 80.31%。最後針對預測錯誤的課文做錯誤分析，了解是甚麼因素造成預測上的錯誤。

**關鍵字：**可讀性、文本分類、支援向量機、中文斷詞

## **Abstract**

Language plays an important part in every reign. And the most efficient way to enhance our ability is to read. Readability can estimate whether an article is suitable for one reader. Past researches claim that readability is a mean to adjust the level of article according to different kinds of educational attainment. The research of English readability has been on its way while Chinese has a little progression. However, Chinese is a trend in nowadays. It is important to find a suitable way to classify text readability.

In the past researches, many western readability formulas do to the lack of technology use linear models on text classification, and linear readability formulas is a limit for the data in my research. Therefore, the purpose of this research is to use the predict model, which trained by the support vector machine, to classify the elementary Chinese textbook's readability. And to check up that whether the text is matched with the predict text. At last, analyze the wrong text to improve the accuracy of text readability.

This research was compiled by course expert and the experience materials( from first to sixth grades deleting the classical Chinese texts of three vision texts of private publish enterprise including vision H, K, and N) total 386 texts were examined by the national compilation organization. Part of the texts are used as training materials and the others are testing materials. Through the Chinese Word Segmentation processing and data format conversion, we at last do the text classification by SVM. The research conclusion is that the accuracy of predicting elementary texts is 47.92% while the fit rate is 80.31%. At the end, analyze the wrong prediction and understand the reason of this wrong prediction.

**Keywords:** Readability, Text Classification, Support Vector Machine, Chinese Word Segmentation

## 誌謝

光陰荏苒，轉眼碩士兩年的時間已經過了。兩年的時間，感謝一路相伴的同學、朋友、老師、停車場的蓮霧樹、教育大樓的老電梯以及 ITS Lab 窗外從日出看到日落到夜深人靜的夜景。我很感動，任何一景一幕、一花一草讓我在寫程式、寫論文的時候還有一些歡笑與樂趣相伴，最後 27,502 個字終於完成。

特別感謝我的指導老師，張國恩老師與宋曜廷老師。張老師親切的笑容以及與學生們貼近的想法另我感到如沐春風；宋老師常給予精闢的分析及意見，使我受益良多。從兩位老師的身上學習到對研究的執著與認真的態度，對兩位老師實在是非常敬佩。

感謝張道行老師的指導與寫作自動批改團隊的資源與支援。感謝日穌不時的給予鼓勵及適時的幫助；感謝厚強忙碌之餘教我 SVM；感謝宜憲給予我好多好多的指導與指正，從你身上我看到對於學習上認真、求知的態度。

感謝我的研究所同學們，謝謝大家這麼義氣相挺，一起做研究，互相打氣加油。謝謝你阿丁，不管是生活上瑣碎的小事或者是求學上的疑問，你總是能給予我解答，每次看到你笑臉迎人心情都跟著好起來；謝謝你陳宥尹，如此體貼及善解人意；謝謝你李酸酸，讓我的碩士生活充滿歡笑與喜餅，有你的實驗室充滿了生氣；謝謝你楊舒嵐，做研究的過程一起發瘋大笑，一起有氧，一起游泳，也一起發胖；謝謝你黃幀祥，失落的時候給予我最大的鼓勵與支持，論文沒有你無法完成。感謝一路幫助我、鼓勵我的大家！

最後，謝謝家人不時的給予一些意見與關心。我終於畢業了！

夢珂 2011 年 8 月

## 目錄

附表目錄.....	ix
附圖目錄.....	x
第一章 緒論.....	1
第一節 研究背景.....	1
第二節 研究目的.....	4
第二章 文獻探討.....	5
第一節 可讀性(Readability) .....	5
第二節 Coh-Metrix.....	11
第三節 中文斷詞.....	12
第四節 支援向量機(Support Vector Machine).....	13
第三章 系統設計.....	17
第一節 系統架構.....	18
第二節 中文可讀性指標分析系統.....	19
第三節 支援向量機的訓練與測試.....	27
第四章 實驗設計.....	29
第一節 實驗工具.....	29
第二節 實驗資料.....	29
第三節 實驗流程.....	30
第四節 實驗結果.....	43
第五節 實驗結果討論.....	50
第五章 結論與未來發展.....	57
第一節 結論.....	57
第二節 未來發展.....	58
參考文獻.....	60

附錄一 Coh-Metrix 2.0 可讀性指標.....	65
附錄二 中研院平衡語料庫詞類標記集.....	70
附錄三 國小國語科刪減的文章.....	72



## 附表目錄

表 1 英文可讀性公式變數整理 .....	8
表 2 中文可讀性公式變數整理 .....	10
表 3 人稱代名詞詞表 .....	23
表 4 正向連接詞表 .....	24
表 5 負向連接詞表 .....	24
表 6 實詞詞性 .....	25
表 7 否定詞詞表 .....	25
表 8 各版本教科書選用課文數 .....	30
表 9 LIBSVM 內部使用資訊 .....	45
表 10 LIBSVM 在重要特徵上的權重 .....	45
表 11 LIBSVM 預測結果在各 FOLD 的預測結果 .....	46
表 12 文本特徵對於 GLM 公式的變異量 .....	48
表 13 SVM 及 GLM 之準確率與正確率比較 .....	49
表 14 中文文本特徵於 SVM 及 GLM 之排序 .....	50
表 15 SVM 分類錯誤之課文 .....	51

## 附圖目錄

圖 1 超平面示意圖 .....	14
圖 2 系統架構圖 .....	18
圖 3 中文可讀性指標分析系統架構圖 .....	20
圖 4 中文可讀性指標分析系統介面圖 .....	27
圖 5 LIBSVM 文件格式 .....	27
圖 6 LIBSVM 參數解說 .....	28
圖 7 實驗流程圖 .....	31
圖 8 字數 .....	33
圖 9 詞數 .....	33
圖 10 句數 .....	33
圖 11 段落數 .....	34
圖 12 每句平均詞數 .....	34
圖 13 每段落平均句數 .....	34
圖 14 人稱代名詞數 .....	35
圖 15 正向連接詞數 .....	36
圖 16 負向連接詞數 .....	36
圖 17 TYPE-TOKEN RATIO .....	36
圖 18 否定詞數 .....	37
圖 19 低筆畫數 .....	38
圖 20 中筆畫數 .....	38
圖 21 高筆畫數 .....	38
圖 22 平均筆畫數 .....	39
圖 23 不存在詞數 .....	39
圖 24 音節數 .....	40

圖 25	二字詞數 .....	40
圖 26	三字詞數 .....	41
圖 27	實詞種類數 .....	41
圖 28	LIBSVM 的訓練 .....	44
圖 29	LIBSVM 的預測 .....	45
圖 30	以 LIBSVM 預測課文冊別之矩陣 .....	47
圖 31	以 GLM 預測課文冊別之矩陣.....	49



# 第一章 緒論

## 第一節 研究背景

在社會與經濟蓬勃發展的現今，語文能力在各方面都扮演著重要的角色。獲取語文能力最重要、最直接的管道之一就是透過閱讀。而閱讀內容本身就有難易度的不同，依據學習者的需求提供最適合他們程度的閱讀文本，可以減少他們花多餘的時間去搜尋或閱讀不適合自己能力的文本。

透過可讀性評估(readability assessment)的方式，可以量化讀者理解文本的難度，提供適當的閱讀教材供不同能力的讀者閱讀(Feng, Jansche, Huenerfauth, and Elhadad, 2010)。在網際網路普遍以及資訊交流越來越為廣泛之下，搜尋引擎常被大家拿來尋找有興趣的文件。儘管搜尋結果符合了主題，文件的可讀性也未必適合搜尋者的閱讀能力。在教育的領域上，老師往往需要從龐大的文本資料中選取適合的教材，而找到的這些教材卻未必依文本的理解難易度去分類(Lin, Su, Yang, and Hsieh, 2010)。老師們所使用的教材跟學生們的理解能力常常是無法相互配合的，可見區分教材難易程度的重要性，此時，文本分類(text classification)的技術便扮演著重要的角色。在資訊成長快速的現今，文本分類已被視為處理及組織文本資料的關鍵技術(Zhang, Yoshida, and Tang, 2008)。文本分類採監督式學習的方式，我們可以將已經標示好理解難易層級的文本當作訓練資料提供給文本分類器做訓練，爾後輸入未知可讀性類別的文本資料，文本分類器便可以指定該讀者所適合的分類層級給他。透過文本的分類，讀者便可以有效率的從大量的資訊中找到符合自己閱讀能力的閱讀資料，增進對文本的理解，進而達到學習的效果。

在國外的研究中，英文文本的可讀性研究很早就出現了。1980 年代超過 200 個可讀性公式已被開發出來，同時更有超過 1,000 篇的相關研究產生 (Lin., et al., 2010)。這些公式多以平均字長、平均句長、平均段落數等文本表面特徵當作可讀性

的指標，並且常被用來評估和區分不同年級所適合的閱讀文本。應用在英文文本的可讀性公式比中文文本的可讀性分類更為廣泛使用且較早出現。綜合幾個常用的英文可讀性公式來看，這些公式主要以文本的基本語言特徵：平均句長、平均字長、平均段落數等作為為難度指標，另外也有些公式會製作常用字表，或是難字表等對文本的可讀性做評估。可讀性探討被很多學者所研究，然而這些研究中所提到的可讀性公式都偏向美式英文，對於非美式英文的篇章寫法可能不是很恰當(Klare, 1963)。所以，我們不能直接將使用在英文文本的分類方法拿來套用在中文文本上。然而，這些已經推行許久的英文可讀性公式對於本研究仍然很有參考價值。中文領域這方面的研究不多，但是中文能力在現今社會卻是一個很主要的趨勢。所以，根據以上所述，一個適合中文文本可讀性的分類方法是很重要的。

用來判斷閱讀難度的工具有許多。本研究使用了曼菲斯大學所發展出來的一個線上文本分析器 Coh-Metrix，當中所提供的可讀性指標(Readability Indices)當作文本的特徵(features)。除了參考 Coh-Metrix 的指標外，還重製了 Coh-Metrix 上的指標，將原本只適合分析英文文本的指標稍微做了修改再拿來使用。本研究參考了 Coh-Metrix 裡的 16 個指標，修改其計算方式以符合中文文本的特性，再增加其他 8 個原本在 Coh-Metrix 裡面沒有的指標，總共包括：人稱代名詞、TTR( type-token ratio)、連接詞數、正向連接詞數、負向連接詞數、實詞數、LOG 實詞數、否定詞數、代名詞數、字數、詞數、句數、段落數、音節數、每句平均詞數、每段平均句數、二字詞數、三字詞數、低筆畫字數、中筆畫字數、高筆畫字數、平均筆畫數、不存在詞、實詞種類數，共 24 個指標來判別一篇中文文本的可讀性。此外，本研究整合了這些中文可讀性指標到研究所開發的系統中，使用者輸入文本後便可得到這些文本的特徵值。將系統產出的文本特徵值利用文本分類器進行訓練產生預測模型(Predict Model)，再對未知類別的文本做分類的預測。最後觀察分類的準確率與正確率，並找出改善的方法。

然而，文本分類的準確性必須考量到訓練的資料適合用於哪一種分類工具，以及選用的文本特徵在文本可讀性上是否能夠代表該文本的難易程度。過去學者多以

線性的可讀性公式對文本做可讀性分類，而本研究所採用的資料量並不大，無法呈現常態的分布，故使用 SVM 作為文本的分類器。林宗勳(民，95)提到 SVM 在解決小樣本、非線性及高維模式識別的問題中表現出許多優勢。SVM 在很多應用上都被廣泛使用，包括文本分類、影像辨識、手寫識別等相關分類應用，且 Joachims(1998)的研究也發現 SVM 在各領域的分類研究上擁有好的分類效果。所以，本研究使用容易上手的分類工具，SVM，做為分類器，針對適合本研究實驗資料的特性做分類。另外，過去許多學者用 SVM 在文本可讀性的分類上都會以準確率(accuracy)這個指標來衡量分類的結果(Larsson, 2006; Feng., et al., 2010; Crossley, Allen & McNamara, 2011)。本研究也考慮到每個冊別間的課文可能因為要在難度上做銜接，使得每個學期間的課文難度會重疊到。所以，將預測的準確範圍放寬上下一冊，稱之為正確率(fit rate)課。Chang, Tsai, Lee & Tam (2008) 在對照人為批改及自動批改文章的方法上也使用正確率的概念，將預測的結果放寬上下一個層級，故本論文採用準確率與正確率這個指標來當本研究分類文本可讀性的效標。

基於上述的背景，本研究使用台灣大學林智仁(Lin Chih-Jen)教授等開發設計的一套簡單、易於使用和快速有效的 SVM 識別模式 (Chang & Lin, 2001)來進行預測，針對國小一至六年級的國語科課文以冊別為單位做分類預測(即區分出一年級上學期、一年級下學期、二年級上學期、二年級下學期等)。最後對 LIBSVM 預測結果中的錯誤進行分析。

## 第二節 研究目的

本研究的目的是開發一個中文可讀性指標分析系統，整合中文可讀性研究常用指標以及 Coh-Metrix 的語言特徵，嘗試找出適合中文的可讀性指標，並以一般線性模式(General Linear Model, GLM)與 SVM 進行文本可讀性分析，比較此二者對文本難度預測之準確度。最後針對 SVM 分類錯誤的文章進行錯誤分析，以改善與謀求分類上的準確性。

## 第二章 文獻探討

本研究的相關文獻可分為四個部分。第一部分先說明可讀性(Readability)研究的定義及歷史。第二部分說明 Coh-Metrix 線上文本分析器及其所提供的可讀性指標。此文本分類器所提供的指標適合用在英文的文本上，一一了解裡面所提供的 60 個指標後，才將適用於中文文本的指標拿來做為本研究的文本特徵。第三部分探討中文斷詞。在文本的可讀性評估上須先將語言經過處理，才能辨別文章中的字詞並且取得文本的特徵值，故這部分探討中文的斷詞處理。第四部分將介紹在本研究中所使用的分類器。在眾多的分類器中，本研究根據實驗資料的性質以及資料量作為考量，選擇使用容易上手的 SVM 作為文本分類的工具。

### 第一節 可讀性(Readability)

#### 一、可讀性的定義

可讀性(Readability)的概念，自古至今不同的學者對其都有不同的定義。Dale-Chall(1948)將可讀性定義為能夠流暢閱讀、了解及發掘其有趣之處的閱讀材料；Klare(1963)定義可讀性為理解或了解其寫作風格的簡單程度；McLaughlin (1969)說可讀性是一群人能找到吸引他們而且能夠讓他們理解讀物內容的程度；宋佩貞(2009)提到可讀性主要的觀念還是從閱讀者及閱讀的素材這兩方面來看，為了能符合讀者的閱讀能力及達到最大的效益，影響閱讀文本的因素仍陸續被拿來分析。Tanaka-Ishii, Tezuka, and Terada (2010)定義的可讀性是在描述一個文本可以被閱讀和理解的簡單程度。在教育的領域上，Lau & King(2006)提到可讀性評估 (readability assessment) 是評估一份文本素材之困難度的方法，廣泛地應用在教育領域，幫助老師們為學生準備適合的教材。本研究所探討的可讀性在於讓使用者了解進行分析的文章是適合哪個年級的學生。



## 二、可讀性公式

許多人都會以公式來測量文本的可讀性。可讀性公式是一種工具，可以把對於不同教育程度的讀者所閱讀的文章加以調整(楊孝濬，1978)。國文教材的難易，有其規律性，可以從量化公式顯現之(荊溪昱，1995)。自十九世紀以來發展了越來越多的可讀性公式。這些公式大部分是基於理解難度的兩大因素：(1) 詞彙或語意的特徵和(2) 句子或句法的複雜性 (Chall and Dale, 1995)。以下列舉幾個較為著名的英文可讀性公式及中文可讀性公式：

### 1. 英文可讀性公式

#### A、Lorge 公式(1939)

Lorge 以新的特徵組合去測量可讀性，其結果比更早的 Gray-Leary 可讀性公式的準確率來得高，所以他強調詞彙對可讀性而言很重要。1948 年，他的公式以平均句長、每 100 個字中不同的難字(不在 Dale 769 字表的字即為難字)及每 100 個字中的介詞片語數當作文本特徵來計算可讀性。

#### B、Dale-Chall 公式(1948)

此公式不同於其他公式使用字長來辨識字的難度，而是以計算”難字”為基礎。Dale-Chall 公式利用難字與句長來計算美國小學生的閱讀程度。這些難字的定義為沒有出現在 Dale 3000 字表的字。公式如下：

$$\text{年級} = .0596sl + .1579wd + 3.6365 \quad (1)$$

$sl$  表示平均句長

$wd$  表示每 100 個字中不同的難字(不在 Dale 3000 字表的字即為難字)

#### C、Flesch Reading Ease 公式(1975)

Flesch Reading Ease 公式會回傳一個 0-100 的數值，分數越高表示該篇文章越容易閱讀。此公式的限制為文本必須超過 200 字才能讓此公式有效的運作。

$$\text{閱讀舒適度} = 206.835 - 1.015sl - .846wl \quad (2)$$

$sl$  表示平均句長

$wl$  表示每個字的平均音節數

#### D、Flesch Kincaid 年級公式(1975)

此公式將文本分類成美國學校的年級程度，改自 Flesch Reading Ease 公式。其數值越高，表示文章越難閱讀。若分數為 8.5 分則表示該文本適合八年級的美國學生來閱讀。公式如下：

$$\text{年級} = .39sl + 11.8wl - 15.59 \quad (3)$$

$sl$  表示平均句長

$wl$  表示每個字的平均音節數

本研究在上述所列出幾個可讀性公式較為普遍使用，另外根據 Dubay, W.H.(2004) 裡所提及的可讀性公式彙整成一個表格，列出可讀性公式所使用的變數並且依變數進行整理，觀察英文可讀性公式常用到的變數有哪些，如表 1。

表 1 英文可讀性公式變數整理

變數	文本指標	研究學者
字彙	平均音節數	Flesch(1948、1975)
		McLaughlin(1969)
	單音節數	Farr、Jenkins 和 Paterson(1951)
	人稱代名詞數	Gray 和 Leary(1935)
	介係詞片語數	Gray 和 Leary(1935)
	不同字彙數	Washburne 和 Morphett(1938)
		Gray 和 Leary(1935)
句子	平均句長	Gray 和 Leary(1935)
		Lorges(1939)
		Flesch(1948、1975)
		Farr、Jenkins 和 Paterson(1951)
		Gunning(1952)
		Bormuth(1966)
		McLaughlin(1969)
	平均句數	Fry(1977)
字表	Dall-Chall 769 字表	Gray 和 Leary(1935)
		Lorges(1939)
	Dall-Chall 3000 字表	Dale 和 Challs(1948)
		Bormuth(1966)
	Thorndike 1500 字表	Washburne 和 Morphett(1938)

從上表可看出英文的可讀性公式很常用到音節數、字數、平均句長、平均句數等，另外更常輔以字表合併做為指標。從此可知這幾個指標對於文本可讀性有一定程度的辨別能力，這對於中文文本可讀性的公式是一個很好的參考方向。但是相較於英文的可讀性公式，國內極少有對於中文文本的可讀性公式。以下將探討幾個中文可讀性的相關研究，並對其使用的變數做整理。

## 2. 中文可讀性公式

A、楊孝滌(1978)分析了 15 個影響可讀性的中文語言因素，評估這些語言因素和

理解程度的相關性。其計算結果為相對的難易值，而非預測出一個適合閱讀的年級或年齡。這 15 個因素包括：(1)平均筆劃數、(2)對稱字的比例、(3)筆畫一到十的字數、(4)筆畫十一到二十的字數、(5)筆畫二十以上的字數、(6)詞彙表、(7)二字詞、(8)三字詞、(9)複字詞(一個或以上單字的詞彙)比率、(10)真詞比率、(11)相似半詞、(12)相異半詞、(13)句長、(14)片語的平均字數、(15)完整句的比例(完整句包含主詞與謂詞)。其所採用的樣本為來自台灣的圖書館的書籍、雜誌、報紙及其他出版物目錄中任意選兩百字的文章共 85 篇，而詞彙表來自六本字、辭典中出現超過 5 次以上的字彙所組成的。

B、荊溪昱(1995)分別以年級與學期為依變項，分別做了兩次不同的預測(其中一次預測不包含詩歌文體與文言文體)。以課文長度、平均句長、常用字比率(以國小 495 字常用表所產的常用字比率)、詩歌文體及文言文文體做為文本指標組合成一個有效的預測公式，適讀年級為國小至高中。公式如下：

$$\text{年級} = 8.76105604 + 0.00272438 * \text{課文長度} + 0.07866782 * \text{平均句長} \quad (4)$$

$$- 8.94311010 * \text{常用字比率} + 0.42920182 * \text{詩歌文體} + 3.23677141 * \text{文言文體}$$

$$\text{學期} = 17.17897155 + 0.0055097 * \text{課文長度} + 0.15834512 * \text{平均句長} \quad (5)$$

$$- 18.14048568 * \text{常用字比率} + 0.85771961 * \text{詩歌文體} + 6.40727935 * \text{文言文體}$$

$$\text{年級} = 17.52547988 + 0.00242523 * \text{課文長度} + 0.04414527 * \text{平均句長} \quad (6)$$

$$- 18.33435443 * \text{常用字比率}$$

$$\text{學期} = 34.53858379 + 0.00491625 * \text{課文長度} + 0.08996394 * \text{平均句長} \quad (7)$$

$$- 3.673710603 * \text{常用字比率}$$

表 2 中文可讀性公式變數整理

變數	研究學者	
	荊溪昱(1995)	楊孝滌(1978)
字數	V	V
平均每句字數	V	V
常用字	V	V
筆畫數		V
二字詞、三字詞		V
對稱字		V
複字詞		V
真詞		V
相似/相異半詞		V
完整句		V

從表 2 中可看到在中文可讀性研究中，學者常用的變數包括字數、平均每句字數及常用字等，故本研究也將其納入為判別中文文本的文本特徵。

上述所提及之可讀性公式屬線性模式，而以前會使用線性模式的公式是因為技術的不足。更有許多研究指出，非線性的可讀性預測比線性方式的準確率來得高。國外研究中，Peterson & Ostendorf (2008)的實驗拿 Weekly Reader 雜誌當作讀物(國小 2~5 年級的等級)，利用 SVM 與同樣是線性模式的 Flech-Kincaid 公式做比較，發現傳統線性公式的準確率(accuracy)比 SVM 來得低。Feng et al.(2010)也提出對文本分類的實驗中，比較 SVM 與線性模式公式的準確率，發現 SVM 預測的準確率比邏輯式回歸(Logistic Regression)來的好。因本研究實驗資料來自三個出版商的國小國語科課文，刪減掉部分課文後共 386 篇，在資料量上較少，且並非呈現常態分佈，故針對適用於本研究的資料群做可讀性分類的分類器 SVM。

從上面中可看到在中文可讀性研究中，學者常用的變數包括字數、平均每句字數及常用字等，故本研究也將其納入為判別中文文本的文本特徵。

## 第二節 Coh-Metrix

Coh-Metrix 是曼菲斯大學所發展出來的一個線上文本分析器，Coh 的意思代表連貫性(Coherence)及凝聚性(Cohesion)，連貫性主要用來表示講話、寫作等各部分之間的連貫性、一致性及條理性；凝聚性主要用來表示文字上面的意思。Coh-Metrix 所提供運算語言的指標從每篇文章間的程度到每個句子間的程度，都能夠評估及分辨不同文本種類，包括理解程度上以及自動化的文本分析(McNamara, Louwerse, McCarthy & Graesser, 2010)。因此，本研究整合中文可讀性研究常用指標以及 Coh-Metrix 的語言特徵，嘗試找出適合中文的可讀性指標。這樣的整合可以試圖解決中文可讀性研究貧乏的問題，即便是在英美語系國家所發展出為數眾多的可讀性公式，也侷限於技術而僅納入少數文本比較表淺的文本特徵，只能反應文本局部連貫的淺層特性，無法觸及深層的文本結構 (Graesser, McNamara, Louwerse, & Cai, 2004)。在 Coh-Metrix 2.0 這個版本上所產生的指標總共有六十項，本研究拿實驗中參考到的幾文本特徵簡單說明如下，詳細的 Coh-Metrix 指標說明如附錄一。

### 1. DENPRPi (Personal pronoun incidence score)

人稱代名詞。高密度的代名詞會造成指涉上連結困難(人和代名詞無法連結)。

### 2. TYPTOKc (Type-token ratio for all content words)

計算字數在文章中出現頻率。如果數值是 1，則表示文章中每個字只出現一次，也代表為文章相對較難理解。只計算實詞(content word)，對於長度類似的文章可以互相比較。

### 3. CONi (Incidence of all connectives)

計算所有連接詞的次數。

### 4. FRQCRacw (Celex, raw, mean for content words)

文章中的實詞數做平均。

### 5. FRQCLacw (Celex, logarithm, mean for content words)

文章中的實詞數取 log 值做平均。

#### 6. DENNEGi (Number of negations)

文章中否定詞的數量。

#### 7. READNW (Number of Words)

文章中的字數。

#### 8. READNS (Number of Sentences)

文章中的句數。

#### 9. READASL (Average Words per Sentence)

平均每句的字數。

#### 10. READNP (Number of Paragraphs)

文章中的段落數。

#### 11. READASW (Average Syllables per Word)

平均每字的音節數。

### 第三節 中文斷詞

斷詞(Word Segmentation)是在連續的字串中按照一定的規定重新組合成有意義的字詞，很多自然語言處理的工作才得以進行，例如語音辨識、資料檢索、機器翻譯等。斷詞結果若不正確，會造成語法及語意表達偏離原本的意思，使得斷詞後的處理工作，如詞性標記、語言分析、資訊擷取等，發生很多的錯誤(張晏晟，2008)。

在英文的文本中，每個字(word)之間以空格隔開，每個字有自己的意思；而中文在語意的基本單位是「詞」而非「字」(許菱祥，1986)，句子間除了有標點符號外，每個詞都是連在一起的。舉例來說，英文句子”Today is Monday”，中文意思為「今天是星期一」。電腦可以很容易的知道 today 為一個單位的字，但若把「今」、「天」分開來看便無法得知他原本的意思。所以，中文的斷詞比英文的斷詞來的複雜，而且是一個相當重要的工作。

#### 一、中文斷詞的方法：

近年來中文斷詞的方法主要以詞彙法為主導，其中陳稼興、謝佳倫、許芳誠(2000)

將過去十餘年間的學者對中文斷詞法的研究統整歸納成三種不同的方式，分別為統計斷詞法、詞庫斷詞法及混和式的斷詞法(結合統計與詞庫式的方法)。

#### (一)、統計法(Statistical-based)

統計法是依機率統計值來決定如何斷詞。計算兩連續的字在語料庫中一起出現的機率是多少，以找出最佳的斷詞組合。

#### (二)、詞庫法(Dictionary-based)

詞庫法主要在找出文件包含哪些存在於詞典內的詞再切割出詞彙，故此種方法非常仰賴詞庫的正確性及完整性。若有未知詞出現可能會降低其斷詞的正確性，但是為了維持其正確性而增加詞典的字詞又有可能降低其斷詞效率。最具代表性的詞庫斷詞法是「長詞優先法(Maximum Matching Method)」(Li et al, 1988 ; Liang, 1990)，其方法是將句子中較長的詞優先斷出來。例如，「歌聲繞樑」在此斷詞法中則會被優先斷詞成「歌聲繞樑」，而不會被斷詞成「歌聲」。

#### (三)、結合統計與法則式的方法

此方法是為了要解決上述兩種斷詞方法的缺點，故合併統計法及詞庫法來解決斷詞的問題。

### 第四節 支援向量機(Support Vector Machine)

SVM 是一種拿來做分類(Classification)或迴歸(Regression)的方法，由波蘭數學家 Vapnik 等根據統計學習理論提出的一種機器學習(Machine Learning)方法，目前已有的應用包括：手寫體辨識、人臉辨識、文本圖像分類等應用，性能優於已有的學習方法，表現出良好的學習能力(林宗勳，民 95)。

SVM 主要概念再尋找類別間的超平面(hyperplane)，將訓練資料區分開來。所謂超平面是指高維度空間裡面的平面，而所謂高維度是指三度空間以上的抽象空間，二度空間的點可以用 $(x_1, x_2)$ 表示，在高維度空間則以 $(x_1, x_2, \dots, x_n)$ 來表示高維度空間裡的一個座標點。以一個二維的例子來說，如圖 1，希望能找到一條線能夠清楚的將圓形與正方形兩類的資料分開且兩個類別的邊界能夠越大越好。在高維模式下將 H1、



H2 這兩條虛線稱為區分超平面(separating hyperplane)，而與兩類別邊界(margin)距離最大的超平面稱之為目標區分超平面(optimal separating hyperplane)。SVM 的目標就是求得具有最大邊界的區分超平面。

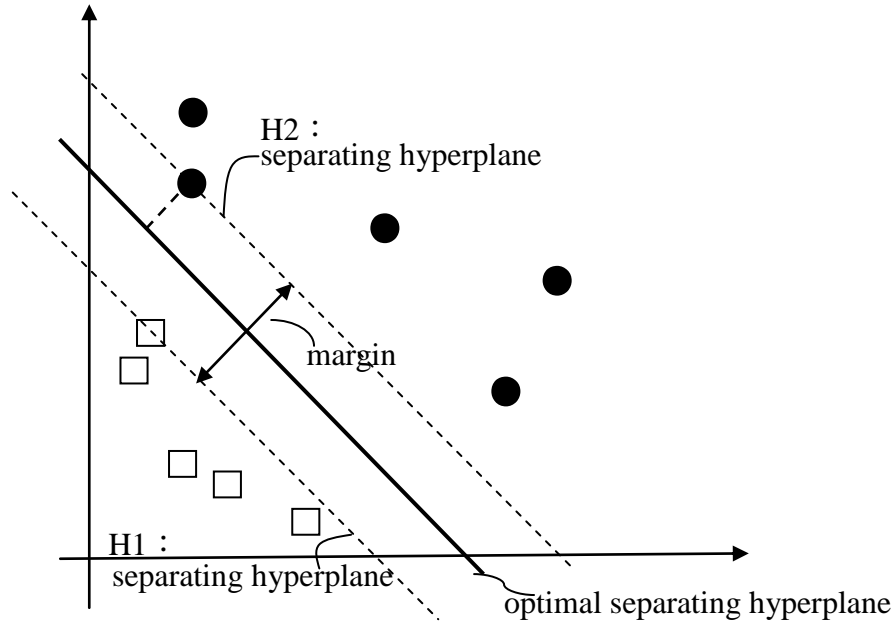


圖 1 超平面示意圖

依據 Cortes & Vapnik(1995)提出如何利用 SVM 做二元分類：假設今天有一群資料  $\{x_i, y_i\}$ ， $i=1, \dots, n$  且  $x_i \in \mathbb{R}^d$ ， $y \in \{+1, -1\}$ ，希望利用這些資料找出  $f(x)=w^T x+b$  這個 optimal separating hyperplane，並將未知的資料做分類。而 SVM 的最終目標就是要找出 optimal separating hyperplane。我們可以將 H1 和 H2 寫成以下式子：

$$H1: w^T x + b = \delta \quad (8)$$

$$H2: w^T x - b = -\delta \quad (9)$$

在此，將  $w, b, \delta$  做尺度調整(scaling)解決過多參數待解的問題可重新將 H1 和 H2 寫成：

$$H1: w^T x + b = 1 \quad (10)$$

$$H2: w^T x - b = -1 \quad (11)$$

H1 和 H2 兩 support hyperplane 的距離為  $2/\|w\|$ ，SVM 為求得最大邊界，換言之就是

要求最小的 $\|w\|$ 。

此類型資料必須滿足以下限制式：

$$w^T x + b \geq 1, y_i = 1 \quad (12)$$

$$w^T x + b \leq -1, y_i = -1 \quad (13)$$

將限制式合併成  $y_i(w^T x + b) - 1 \geq 0$ 。綜合上述，可將 SVM 所要解決的主要問題(the primal problem of the SVM)如下：

$$\begin{cases} \min & \frac{1}{2} \|w\|^2 \\ \text{s.t.} & y_i(w^T x + b) - 1 \geq 0 \end{cases} \quad (14)$$

此問題可用拉格朗日乘數法(Lagrangian Multiplier Method)變換成對偶問題(dual problem)來解決，所謂對偶問題是指把兩個問題可看成一體兩面，解決其中一個問題另外一個問題就會跟著解決，轉換後的對偶問題比較容易解答。把原來的問題轉換成對偶問題解出答案後，就是原來問題的答案。而拉格朗日乘數法所要解決的問題是當尋求的變數受一個或多個條件所限制時求極值的方法。此方法將最佳化問題中的  $n$  個變數及  $k$  個限制條件轉換成極值問題中的  $n+k$  個變數。將此問題轉換成拉格朗日乘數法問題：

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i [y_i(w^T x_i + b) - 1] \quad (15)$$

再找出拉格朗日乘數法問題的對偶問題。原本求最小值的問題，在轉成對偶問題時變成求最大值問題：

$$\begin{cases} \max & L(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j x_i^T x_j \\ \text{s.t.} & \sum_i \alpha_i y_i = 0 \end{cases} \quad (16)$$

上述所說是當資料為線性的情況，Vapnik 最早提出的 SVM 是使用線性核心函數(Linear Kernel Function)，將資料以內積的方式映射到特徵空間中(feature space)。

然而在現實生活中並非所有的資料都是呈線性分布，Boser, Guyon, & Vapnik (1992) 提出一個針對非線性資料分類的解決辦法，將線性核心函數換成其他非線性的核心函數。SVM 的核心函數包含：

1. 線性核心函數(Linear Kernel Function)
2. 多項式核心函數(Polynomial Kernel Function)
3. 輻射基底核心函數(Radial Basis Kernel Function)
4. 多層感知核心函數(Multilayer Perception Kernel Function)

## 第三章 系統設計

本章節將詳細介紹整個系統的架構及其細節。第一節以架構圖說明本研究的系統架構，於各小節詳述系統每個部份。第二節說明中文斷詞的處理。第三節說明本研究中所使用的分類器 LIBSVM 在訓練和測試的階段以及 LIBSVM 所產生的可讀性數學預測模型。

## 第一節 系統架構

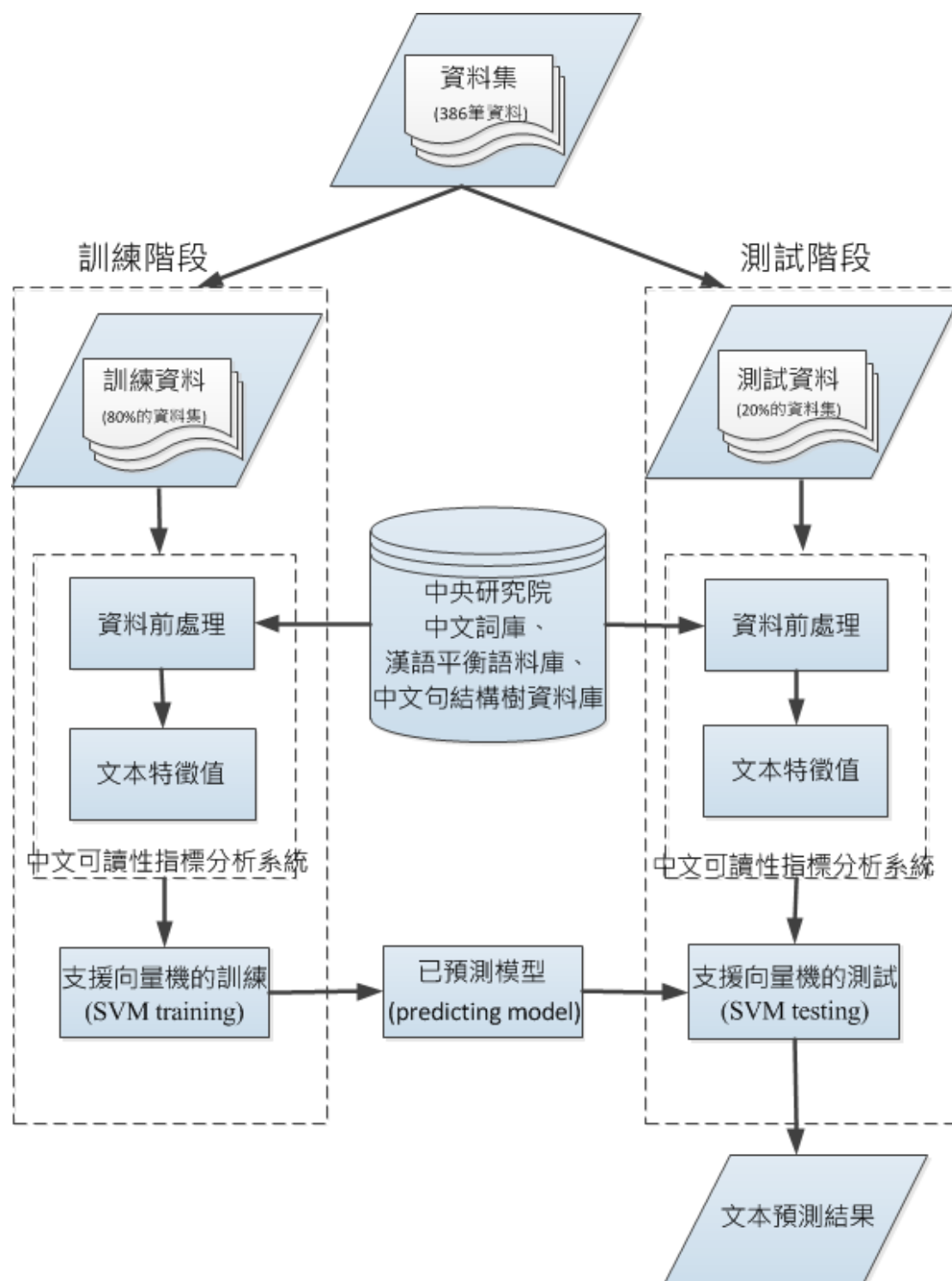


圖 2 系統架構圖

本系統架構圖如圖 2 所示。主要可分為兩部分：訓練階段以及測試階段。將資料分成訓練資料以及測試資料，輸入至本研究所完成的中文可讀性指標分析系統，

由中研院所提供的中文詞庫、漢語平衡語料庫及中文句結構樹資料庫所提供的資訊作資料前處理，並產生文本特徵值。將文本特徵值轉換成 SVM 所能接受的資料格式以 80% 的訓練資料進行 SVM 的訓練(training)，SVM 訓練完資料產生一個預測模型(predicting model)用以供 20% 的測試資料進行 SVM 的測試(testing)使用，最後給予一個預測的分類結果。

## 第二節 中文可讀性指標分析系統

本系統所使用的中文斷詞工具是使用國立高雄應用科技大學資訊工程研究所，由張道行老師主持的智慧型系統實驗室所開發的斷詞系統，處理中文斷詞的問題。斷詞系統所依賴的語料庫是參考中央研究院的中文詞庫、漢語平衡語料庫及中文句結構樹資料庫所建置而成。其所使用的斷詞方法，是以正向長詞優先法(Forward Maximum Matching)配合貝氏機率來實作。系統語言使用 Borland C++ Builder 6.0 撰寫而成。

斷詞(Word Segmentation)在文本分析上是一個很重要的過程。斷詞結果如果不正確，會造成語法及語意表達偏離原本的意思，使得斷詞後的處理工作，如詞性標記、語言分析、資訊擷取等，發生很多的錯誤(張晏晟，2008)。本系統在斷詞後會做詞性的標記，並利用這些資訊製作出文本特徵。系統只要輸入文章便可以產出文本特徵值。

中文可讀性指標分析系統架構如圖 3。

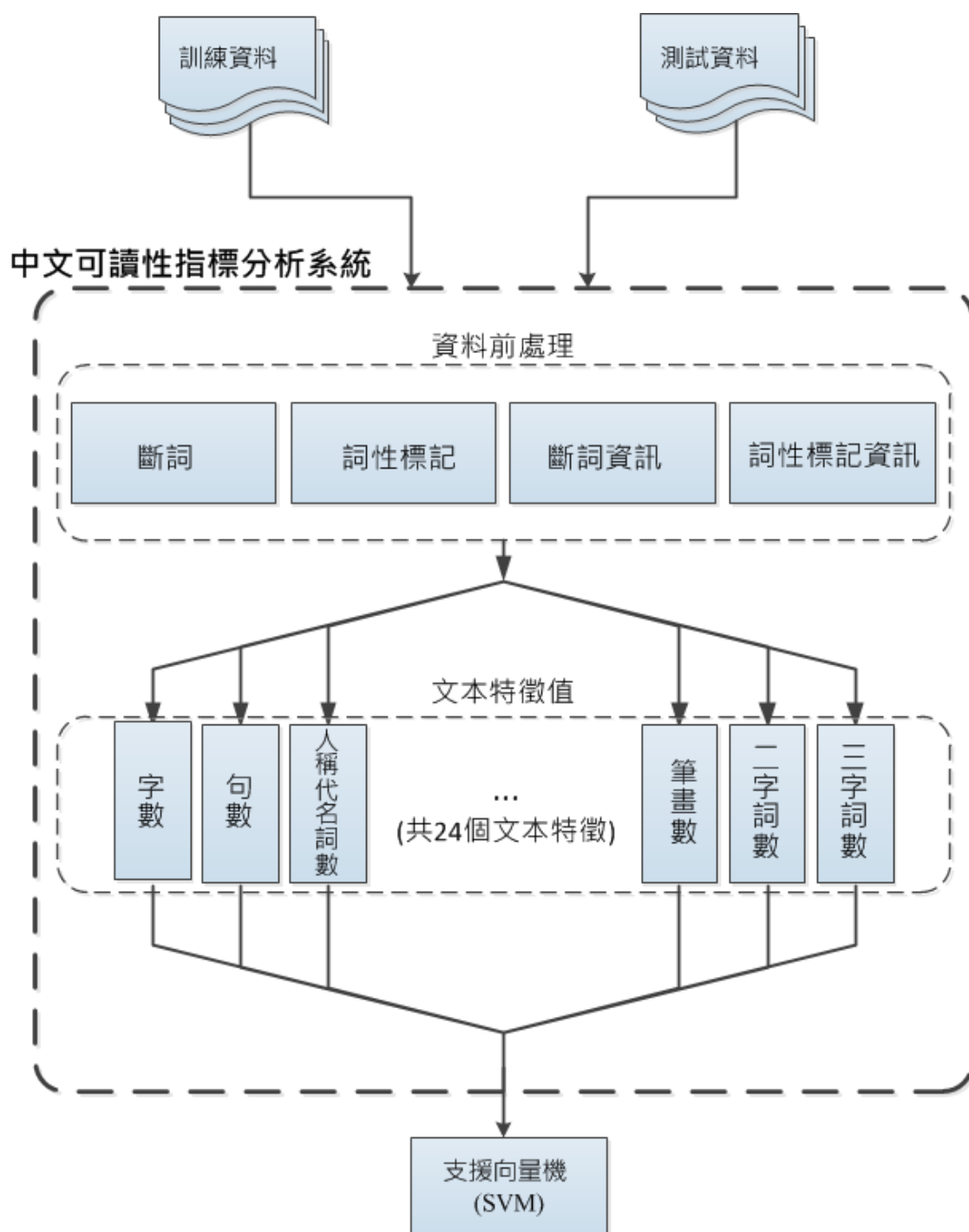


圖 3 中文可讀性指標分析系統架構圖

#### 一、資料前處理

資料在前處理階段包含了四大功能，分別為斷詞(segment)、詞性標記(tagging)、斷詞資訊(information)和詞性標記資訊(part-of-speech info)。利用此四大功能所提供的資訊，製作其餘的文本特徵。其四大功能說明如下：

### 1. 斷詞

「今天天氣很好」，其輸出為「今天 天氣 很好」。

### 2. 詞性標記

「今天天氣很好」，其輸出為「(Nd) (Nd) (D)」。

本斷詞系統所參考的中文字詞詞性來自於中研院平衡語料庫所記錄的中文詞彙，紀錄各詞彙出現頻率與各詞性間彼此共同出現的頻率。詳細的詞性對照表如附錄二。

### 3. 斷詞資訊

斷詞資訊提供總中文字數(不包含標點符號)、斷詞數(包含標點符號)、標點符號數及屬於該詞性代碼的字數。

### 4. 詞性標記資訊

此功能標記一個詞可能出現的詞性之相對比例。

「今天天氣很好」，其輸出為「(1) (1) (VH[0.92]D[1.00])」。(依據中央研究院中文詞庫的詞彙，紀錄各詞彙的詞性與連接的詞彙彼此共同出現的頻率。在這句話中「很好」為詞性 VH 時的機率為 0.92，為詞性 D 的機率為 1。其餘兩個詞「今天」和「天氣」在訓練資料裡只有一種詞性，故只有一種機率值為 1。)

## 二、文本特徵值

本研究利用斷詞系統的四大功能做其他文本特徵的延伸。研究中參考了 Coh-Metrix(<http://cohmetrix.memphis.edu/cohmetrixpr/index.html>)，由曼菲斯大學所發展出來的一個線上文本分析器提供的文本指標(Graesser, McNamara, Louwerse, & Cai, 2004)。在 Coh-Metrix 2.0 這個版本的指標總共有六十項，本系統除了參考部分 Coh-Metrix 所提供的指標外，也增加了新的指標共 24 個 (指標的分析於第四章詳述)。透過斷詞四大功能延伸許多關於文章重要的資訊後，將這些資訊經過處理便能提供我們檢視一篇文本的特徵。本系統利用中文斷詞系統的四大功能產生更多有意義的文本特徵，各功能所產出的文本特徵如下：

#### 1. 斷詞功能所產生的文本特徵：

A、人稱代名詞數：先建立人稱代名詞的詞表，將經過斷詞的文章與人稱代名詞詞表



進行比對，符合的詞即是人稱代名詞數。

B、連接詞數、正向連接詞數、負向連接詞數：建立正向連接詞、負向連接詞的詞表，將經過斷詞後的文章中各個詞語詞表進行比對，即得到正向連接詞數與負向連接詞數，而連接詞數為兩者相加。

C、否定詞數：先建立否定詞的詞表，經過斷詞的文章，與否定詞詞表進行比對，符合的詞數即是人稱代名詞數。

D、二字詞、三字詞數：文章經過斷詞後，檢視每個詞的字元(byte)數，由於中文字一個字有 2 個字元，故 4 字元的詞判定為 2 字詞，6 字元的詞判定為 3 字詞。

E、筆畫數：筆畫數的字典檔建置完後，程式會自動進行比對，算出文章中每個字的筆畫，分析求得整篇文章的平均筆畫數(筆畫總數/字數)、低筆畫數(1-10 畫字數)、中筆畫數(11-20 畫字數)、高筆畫數(20 畫以上字數)

2.詞性標記功能所產生的文本特徵：

A、實詞數：事先定義實詞所涵蓋的詞性。當詞性標記功能標記出各個詞的詞性後，便可以計算符合實詞詞性的字數。

B、實詞數取 log：將計算好的實詞數取 log 值即是。

C、Type-Token-Ratio(TTR)：計算字數在文章中出現頻率。如果數值是 1，則表示文章中每個字只出現一次，也代表為文章相對較難理解。本研究中只計算實詞在文章中出現的頻率。

3.斷詞資訊所產生的文本特徵：

A、字數、詞數：字數與詞數是斷詞核心所內建的功能。此特徵在計算整篇文章的字數以及斷詞後的詞數。

B、句數、每句平均字數：若句子中出現句號、驚嘆號、問號，則將其視為一個句子。每句平均字數是將所有的字數除以句數。

C、段落數、每段平均句數：輸入的資料在每段結束的地方都會用「##」做一個標記，在文章斷詞後，只要搜尋「##」出現的次數，即可知道共有多少段落。每段平均句數即是把文章總句數除以段落數。

D、代名詞數：將文章進行詞性標記後，搜尋詞性為”Nh”出現的次數，即是代名詞數。

24 個文本特徵詳細計算方式如下：

1.人稱代名詞數：建立人稱代名詞的詞表是參考中研院的中文詞類分析(中文詞知識庫小組，1993)，經過斷詞的文章，與人稱代名詞詞表進行比對，符合的詞即是人稱代名詞。以下為事先定義好的人稱代名詞(共 66 個) 如表 3：

表 3 人稱代名詞詞表

我	個人	咱	敝人	人家	本人	鄙人	在下
俺	咱家	你	汝	您	儂	他	她
我們	俺	咱	咱們	我人	我等	你們	您們
汝等	爾等	他們	她們	吾	吾等	我倆	他們倆
他倆	他人	自己	自個兒	個人	本身	自身	己身
別人	大家	閣下	鈞座	大駕	大夥	大伙	大伙兒
大夥兒	筆者	大眾	民眾	伉儷	昆仲	賢昆仲	老夫
老漢	臣	親愛的	老	老衲	貧道	對方	伊
誰人	誰						

2.Type-Token Ratio(TTR)：計算 整篇文章中的實詞種類數 / 實詞總數（針對整篇文章的實詞做運算）。舉例來說：「今天天氣很好。我們大家今天要出去玩。」此時 TTR 值為 6 / 7，實詞種類有 6 個(“我們”、“大家” 不為實詞)，而實詞總數共有 7 個。所以當 TTR 的比值為 1 的時候，可以知道整篇文章的實詞都沒有重複出現，代表文章相對較難理解。

3.連接詞數：正向連接詞數+負向連接詞數

4.正向連接詞數：建立正向連接詞的詞表，經過斷詞的文章，與正向連接詞詞表進行比對，符合的詞即是正向連接詞。事先定義正向連接詞(共 125 個)如表 4：

表 4 正向連接詞表

又	及	以及	且	而且	同	同時	和
與	亦	並	並且	暨	也	抑是	抑或
或是	還有	還是	至	到	既	一手	一方面
若非	要不是	除非	一來	二來	三來	一則	二則
要不然	最後	凡	凡是	舉凡	乃至	乃至於	以至
以致於	致	甚而	甚至	甚至於	甚或	連	不管
無論	於是	爰	不只	不止	不止	不但	不單
不僅僅	只要	另	另外	加上	加以	兼	再者
豈止	豈只	況	況且	何況	由於	因	因為
如果	如果說	如若	若	若是	若要	要	要是
因而	以便	因之	因此	所以	是以	甚且	所以
鑒於	有鑑於	進而	再	再來	既	既然	就是
以便	因之	因此	所以	是以	甚且	兼	跟
或	或者	另一方面	一旦	三則	首先	以至於	以致
不論	任憑	不單單	不僅	此外	何止	蓋	如
就是	是故	之所以	鑑於	是故	因而		

5.負向連接詞數：建立負向連接詞的詞表，經過斷詞的文章，與負向連接詞詞表進行比對，符合的詞即是負向連接詞。事先定義負向連接詞(共 56 個)如表 5：

表 5 負向連接詞表

不然	要不	要不然	不過	可	可是	只有	只不過
就是	然	然而	然則	而	唯有	惟有	反而

而是	另	另外	再者	即令	即使	即便	果
就算	縱使	縱然	否則	萬一	不然	要不然	然而
不如	還不如	毋寧	以免	以防	免得	省得	與其
要麼	雖	雖然	雖說	盡管	儘管	但	但是
反倒	固然	便是	就是	故	故而	要	要不

6.實詞數：若該詞的詞性為表 6 這幾種詞性，則為實詞(詞性對照表請察看附錄二)：

表 6 實詞詞性

Na	Nb	Nc	Nd	VA	VB	VC	VD
VE	VF	VG	VH	VI	VJ	VK	VL
V	A	D	Da	Df			

7.實詞數取 log：將實詞數取 log 值。

8.否定詞數：建立否定詞的詞表，經過斷詞的文章，與否定詞詞表進行比對，符合的詞即是否定詞。事先定義否定詞(共 4 個)如表 7：

表 7 否定詞詞表

不	別	沒	沒有				
---	---	---	----	--	--	--	--

9.代名詞數：經斷詞後的字詞之詞性為「Nh」，則為代名詞。

10.字數：字數是斷詞核心所內建的功能，計算文章進行斷詞後的字數。

11.詞數：詞數是斷詞核心所內建的功能，計算文章進行斷詞後的詞數。

12.句數：若句子中出現「句號」、「驚嘆號」、「問號」，則將其視為一個句子的單位。

13.段落數：本研究在文章中每段結束的地方都會用「##」做一個標記，在文章斷詞後，只要搜尋「##」出現的次數，即可知道文章共有多少段落。

14.音節數：字數/詞數

15.每句平均詞數：詞數/句數

16.每段落平均句數：句數/段落數

17.低筆畫數：筆畫數的字典檔建置完成後，程式會自動進行比對，找出文章中每個字的筆畫，若筆畫數為 1-10 畫，則為低筆畫數。

18.中筆畫數：筆畫數的字典檔建置完成後，程式會自動進行比對，找出文章中每個字的筆畫，若筆畫數為 11-20 畫，則為中筆畫數。

19.高筆畫數：筆畫數的字典檔建置完成後，程式會自動進行比對，找出文章中每個字的筆畫，若筆畫數為 21-30 畫，則為高筆畫數。

20.平均筆畫數：總筆畫數/字數

21.不存在詞數：分析資料庫中所有文章的詞頻，再與教育部國小詞頻總表進行交叉比對，取兩者聯集的前 3000 高詞頻的詞作為常用詞表，再比對文章中有出現及未出現於詞表中的詞，如果不在此表中的詞即為本研究中不存在詞的特徵。

22.實詞種類數：找出所有的實詞，並去除掉重複出現的詞，即為實詞種類數。

23. 二字詞數：文章經過斷詞後，若該詞被判斷為 4 字元(byte)，則為二字詞。

24.三字詞數：文章經過斷詞後，若該詞被判斷為 6 字元(byte)，則為二字詞。

中文可讀性指標分析系統中所使用到的資料庫由中央研究院所提供的中文詞庫、漢語平衡語料庫及中文句結構樹資料庫提供的資訊，以及由台灣師範大學心輔系陳學志老師所提供的筆畫表，本研究將六千餘字的筆畫建置成一個筆畫的字典檔。圖 4 為中文可讀性指標分析系統的介面。

Form1

Eile

請輸入欲分析的文章或句子：
 

發現雨林裡更加昏暗，光線幾乎都不見了，空氣也變得格外濕熱悶熱，我們立刻撐起小營帳。在奇怪的寂靜中，一陣狂烈的疾風由遠而近，吹得樹葉聚聚，緊接著，石礮般粗大的雨珠從樹林上頭傾盆而下。閃電隨之而來，爆裂的雷聲在四周迴盪，震人心弦。然而，不過半個小時左右，暴雨倏然停止，大地一下子靜了下來，彷彿什麼事也沒發生。
 到了晚上，雨林突然熱鬧起來：各種蟲蟻在四周響起，其中有一種蟬鳴，聲如鋼琴，聽來悲涼清切；長臂猿的吼叫聲從遠處傳來，有如人的慘叫！

清除資料
 開始計算
 批次載入計算

新詞(index):
 詞性標記:
 詞性標記資訊:
 

1 395 776 3681  
 12797 7460  
 17093 56653  
 69678 23697  
 31655 52149  
 37171 35825  
 23697 31205  
 31214 9558  
 43390 42667 1  
 23052 45545  
 71853 54137  
 23697 8421

notword  
notword  
Ne  
Ne  
Ne  
Ne  
Nf  
Nd  
notword  
Nh  
P  
VA

notword  
notword  
[D[0.18]Nd  
[0.60]Ne[1.00]  
[Ne[1.00]  
(Nd[0.14]Ne  
(1.00)  
(Nd[0.01]Ne  
(1.00)  
(Na[0.02]Nf  
(1.00)  
(Nd[1.00])

二字詞數: 217  
 三字詞數: 10

新詞資訊:
 1. 人稱代名詞數: 21  
 2. type-token-ratio: 0.7670454382896  
 3. 連接詞數: 24  
 4. positive連接詞數: 15  
 5. negative連接詞數: 9  
 6. 實詞數: 342  
 7. 實詞數取log值: 2.5340261060561  
 8. 否定詞數: 3  
 9. 代名詞數: 24  
 10. 字數: 802  
 11. 詞數: 573  
 12. 句數: 17  
 13. 段落數: 6  
 14. 音節數: 1.3996509589603

新詞:
 一九八五年冬天我像避寒的候鳥搭上馬來西亞航空的飛機飛離正被第一波寒流籠罩的台灣前往位於赤道附近的婆羅洲探險我的嚮導是一個

15. 每句平均詞數: 33.705882352941  
 16. 每段落平均句數: 2.83333333333333  
 17. flesch reading ease: 54.213058207576  
 18. flesch-kincaid: 14.071175443999  
 19. 1-10劃: 521  
 20. 11-20劃: 269  
 21. 21劃以上: 10  
 22. 平均筆劃數: 9.2125

詞頻:
 筆劃:
 

一 出現 10 次  
 九 出現 1 次  
 八 出現 1 次  
 五 出現 1 次

1 劃出現 13 次  
 2 劃出現 25 次  
 3 劃出現 40 次  
 4 劃出現 41 次

建立詞彙母庫

23. 詞頻母率速率: 母體數: 11913  
 條件: 0 次以上 存在詞 294  
 前 0 筆 不存在詞 38

新詞數: 680  
 實詞種類數: 270  
 實詞總數: 352

圖 4 中文可讀性指標分析系統介面圖

### 第三節 支援向量機的訓練與測試

本研究使用台灣大學林智仁(Lin Chih-Jen)教授等開發設計的一個簡單、易於使用且快速有效的 SVM 識別模式工具 LIBSVM 來進行文本的分類(Chang & Lin, 2001)。分類過程可分成三個部分，如下說明：

#### 1. 資料預處理

將中文斷詞系統所產出的文本特徵值轉換成 LIBSVM 所需要的文件格式。其格式固定為<CLASS> <INDEX1>:<VALUE1> <INDEX2>:<VALUE2> <INDEX3>:<VALUE3>...如圖 5 所示。<CLASS>是訓練資料的類別，<INDEX>在本研究中指的是文本的第一個特徵，<VALUE>在本研究中指的是特徵值。拿其中一行來說明：

1    1:0   2:0.666666687    3:0

冒號的前後分別文本特徵的編號及特徵值，在此第一個特徵的值為 0、第二個特徵的值為 0.666666687、第三個特徵的值為 0。此行開頭的 1 代表該資料是屬於類別 1。

1	1:0	2:0.666666687	3:0
1	1:2	2:0.769230783	3:0
2	1:5	2:0.413043469	3:4
2	1:4	2:0.569230795	3:10
3	1:4	2:0.693877578	3:1
3	1:12	2:0.708860755	3:6
4	1:10	2:0.601398587	3:9
4	1:8	2:0.652631581	3:8
4	1:27	2:0.484848499	3:14

圖 5 LIBSVM 文件格式

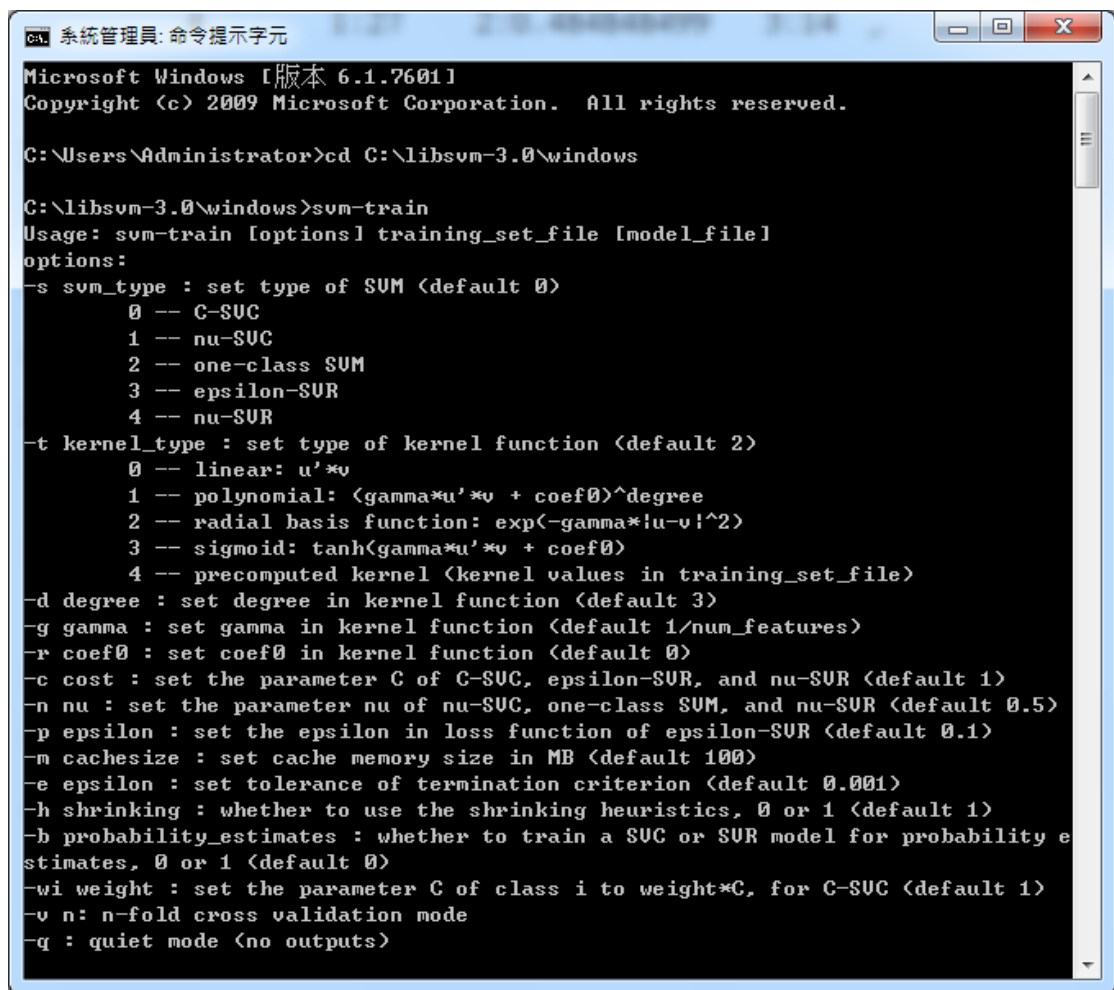
#### 2. 支援向量機的訓練(SVM training)

LIBSVM 提供了 svm-train.exe 的執行程式提供使用者將資料載入作訓練，訓練完會產生一個.model 檔，此模型是給 LIBSVM 的測試階段來使用以進行預測。

### 3. 支援向量機的測試(SVM testing)

LIVSVM 提供 svm-predict.exe 的執行程式，在預測階段系統會依照已經訓練好的模型(.model)，對新輸入的文本做預測(predict)，並輸出預測結果如圖 5 所示。

另外，本研究使用了台大林智仁老師提供的工具 Grid.py 來選擇最佳的參數，因 SVM 的正確率高低有很大的部分取決於參數的選擇。LIVSVM 亦提供了不少參數來讓使用者可依實驗資料型態等的不同可逐一調整，如圖 6 所示。



```
Microsoft Windows [版本 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd C:\libsvm-3.0\windows

C:\libsvm-3.0\windows>svm-train
Usage: svm-train [options] training_set_file [model_file]
options:
-s svm_type : set type of SVM (default 0)
  0 -- C-SVC
  1 -- nu-SVC
  2 -- one-class SVM
  3 -- epsilon-SUR
  4 -- nu-SUR
-t kernel_type : set type of kernel function (default 2)
  0 -- linear: u'*v
  1 -- polynomial: (gamma*u'*v + coef0)^degree
  2 -- radial basis function: exp(-gamma*u'*v)^2
  3 -- sigmoid: tanh(gamma*u'*v + coef0)
  4 -- precomputed kernel (kernel values in training_set_file)
-d degree : set degree in kernel function (default 3)
-g gamma : set gamma in kernel function (default 1/num_features)
-r coef0 : set coef0 in kernel function (default 0)
-c cost : set the parameter C of C-SVC, epsilon-SUR, and nu-SUR (default 1)
-n nu : set the parameter nu of nu-SVC, one-class SVM, and nu-SUR (default 0.5)
-p epsilon : set the epsilon in loss function of epsilon-SUR (default 0.1)
-m cachesize : set cache memory size in MB (default 100)
-e epsilon : set tolerance of termination criterion (default 0.001)
-h shrinking : whether to use the shrinking heuristics, 0 or 1 (default 1)
-b probability_estimates : whether to train a SVC or SUR model for probability estimates, 0 or 1 (default 0)
-wi weight : set the parameter C of class i to weight*C, for C-SVC (default 1)
-v n: n-fold cross validation mode
-q : quiet mode (no outputs)
```

圖 6 LIBSVM 參數解說

## 第四章 實驗設計

本章節將介紹研究中所做的實驗。第一節介紹本實驗所使用到的工具。第二節介紹本實驗所採用的資料。第三節說明整個實驗的流程。第四節探討 GLM 與 LIBSVM 在文本分類上的準確率並對 LIBSVM 的分類結果做錯誤分析。

### 第一節 實驗工具

本研究的研究工具包括中文斷詞工具及 LIBSVM。中文斷詞工具由國立高雄應用科技大學資訊工程研究所，張道行老師所主持的智慧型系統實驗室所提供。LIBSVM 則是使用台灣大學林智仁教授等人所開發設計的一個簡單、易於使用的 SVM 模擬工具。使用詳細內容如第三章所述。

### 第二節 實驗資料

本研究的材料取自課程專家編撰，經國家編審單位審定的三個民間版本教科書 (H 版、K 版、N 版)，國小一年級至六年級國語科課文刪減掉新詩、絕句、古文、律詩的課文後共計 386 篇。各版本教科書在國小一至六年級選用的課文數如表 8 所示。各版刪減掉的課文詳列於附錄三。



表 8 各版本教科書選用課文數

	學期												總計
	一上	一下	二上	二下	三上	三下	四上	四下	五上	五下	六上	六下	
H 版	8	10	12	12	10	13	13	11	11	11	12	10	133
N 版	8	7	13	12	8	8	11	12	11	12	13	9	124
K 版	8	6	12	10	12	11	11	13	13	11	12	10	129
總計	24	23	37	34	30	32	35	36	35	34	37	29	386

### 第三節 實驗流程

本實驗利用中文斷詞工具、LIBSVM 做為實驗工具，將國小一至六年級的國語科課文做可讀性的分類(以冊別當作分類的單位)，並且與過去研究中常用的一般線性模式(GLM)的可讀性預測公式做比較。最後，對 LIBSVM 的預測結果做錯誤的分析及準確率上的改善。實驗流程如圖 7 所示。

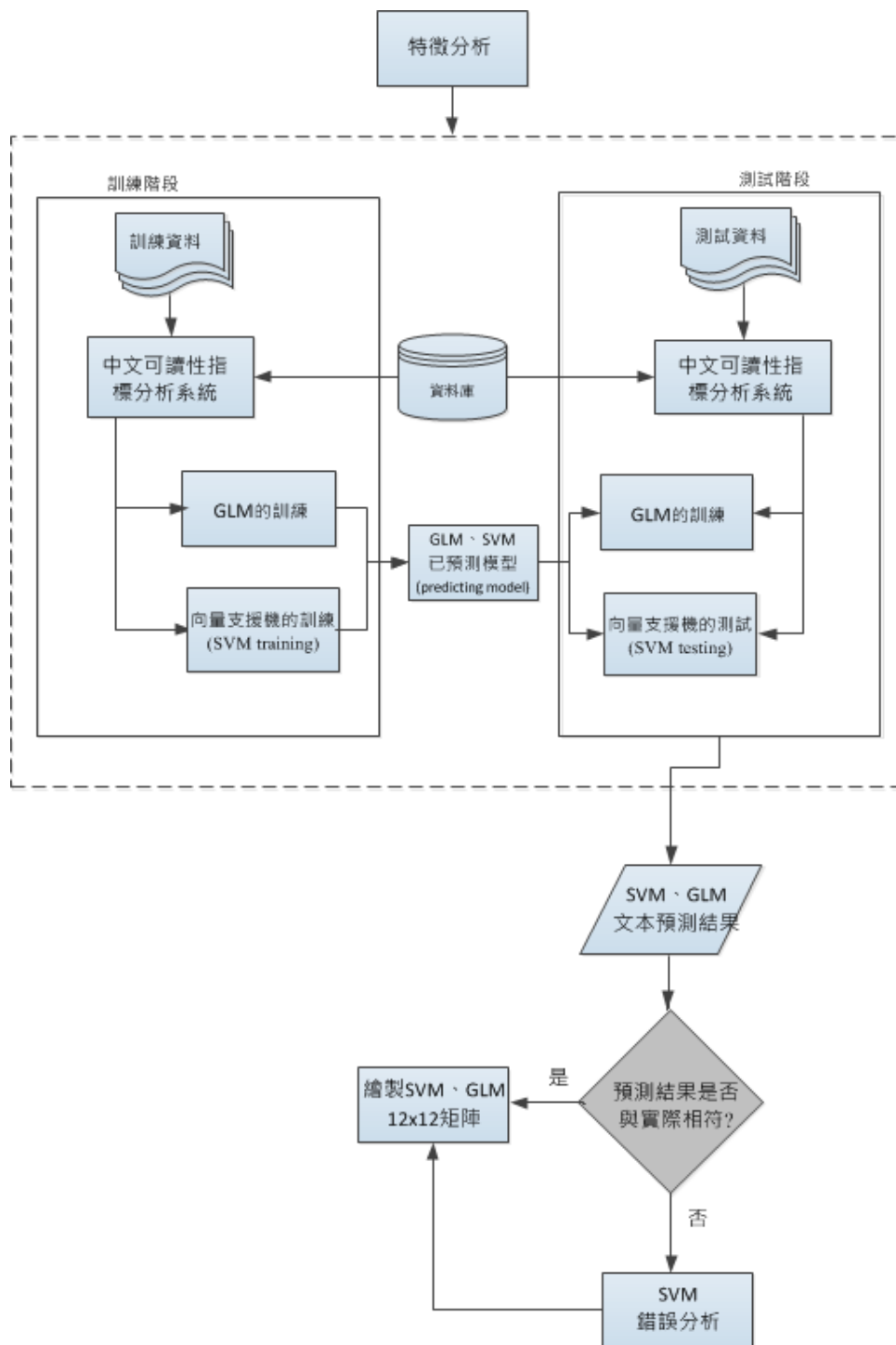


圖 7 實驗流程圖

實驗步驟如下：

### 1.特徵於各冊別特徵分析

本研究從 Coh-Metrix 線上文本分析器所提供的指標中選出 16 個指標做為文本的表面特徵，包括人稱代名詞、TTR (type token ratio)、連接詞數、正向連接詞數、負向連接詞數、實詞數、LOG 實詞數、否定詞數、代名詞數、字數、詞數、句數、段落數、音節數、每句平均詞數、每段平均句數。另外，新增非 Coh-Metrix 裡面的指標 8 個，此 8 個表面特徵為過去學者所拿來當作可讀性指標的特徵(以下將針學者提出的看法對每個文本特徵做說明)，包括二字詞數、三字詞數、低筆畫字數、中筆畫字數、高筆畫字數、平均筆畫數、不存在詞、實詞種類數。總共使用了 24 個特徵來判別一篇中文文本的可讀性。本研究採用之中文文本特徵分析可依據兩部分做探討，其一是將國小國語科三個版本(H 版、K 版、N 版)的課文經中文可讀性指標分析系統處理後，將其特徵值透過 Microsoft Excel 繪製特徵散佈圖，觀察特徵值是否隨年級的增加而呈現某種趨勢。若有，則判定該文本特徵對於國小國語科課文在可讀性分類上能有所辨別。其二是根據專家所指出某些特徵對於中文文本的可讀性上能夠有效區別。

#### A. 字數、詞數、句數、段落數、每句平均詞數、每段落平均句數

這些特徵是文本最基本的訊息。國外使用這些特徵的可讀性公式很早就出現了(詳見第二章文獻探討)，這些公式主要以文本的一般語言特徵：平均句長、平均字長、平均段落數等作為為難度指標，另外也有些公式會輔以常用字比率、常用字表，或是難字表等合併作為對文本的可讀性的數值做出估算。故本研究亦將這些基本文本指標拿來當作中文可讀性的分類特徵。圖8、圖9、圖10、圖11、圖12與圖13為字數、詞數、句數、段落數、每句平均詞數與每段落平均句數於國小國語科各冊間的特徵值散佈圖。

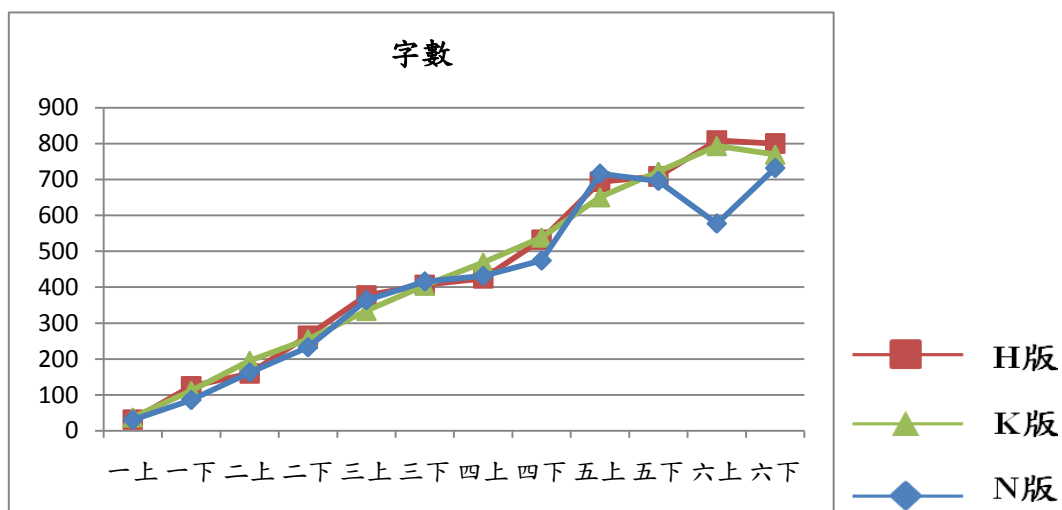


圖 8 字數

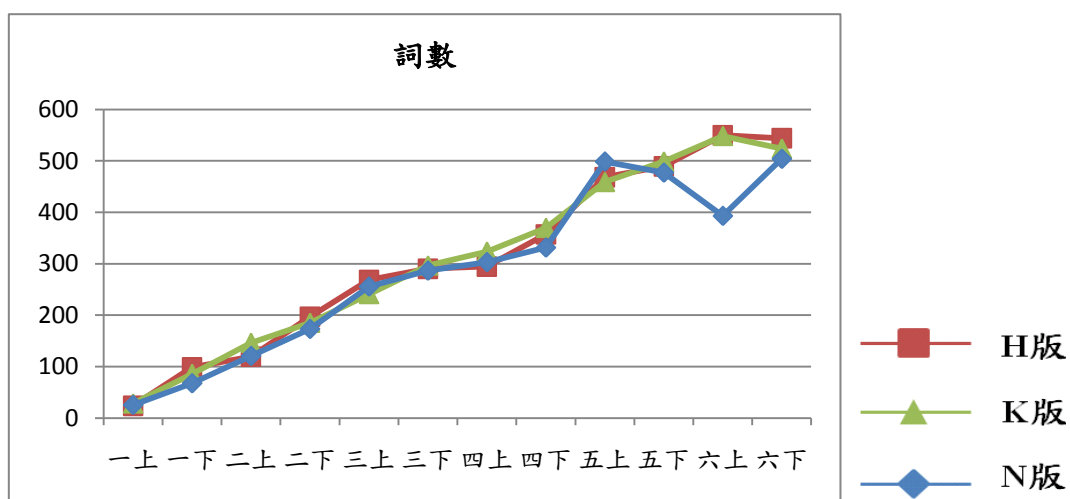


圖 9 詞數

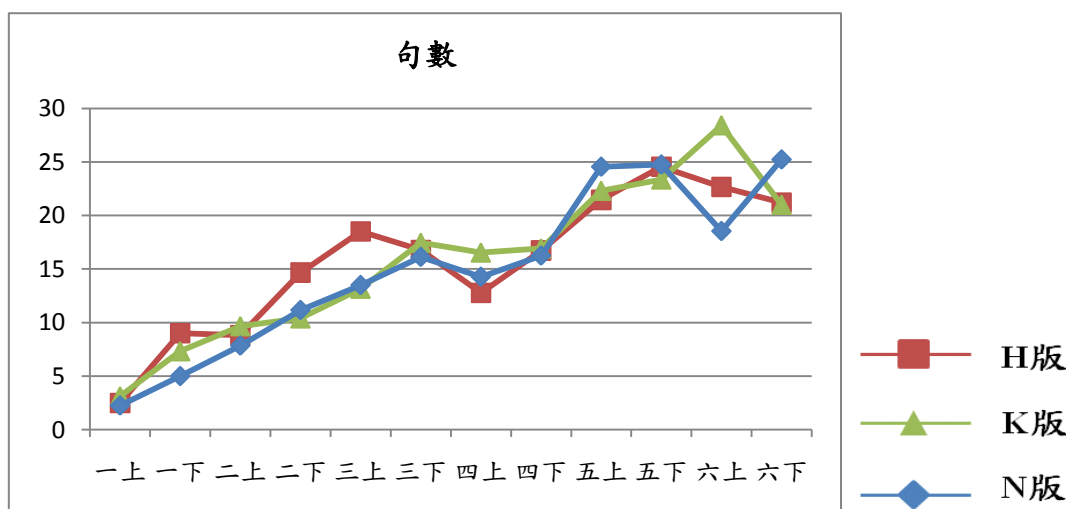


圖 10 句數

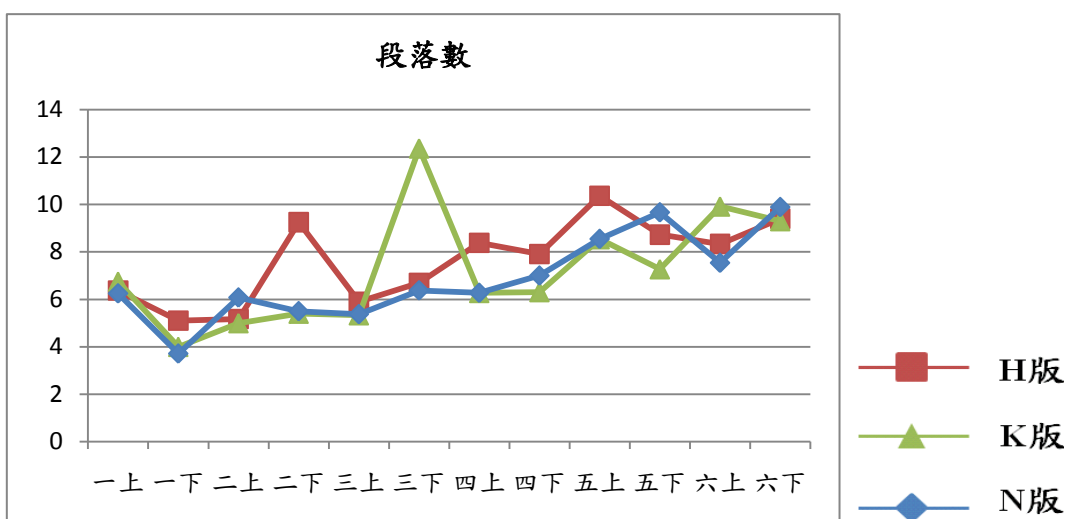


圖 11 段落數

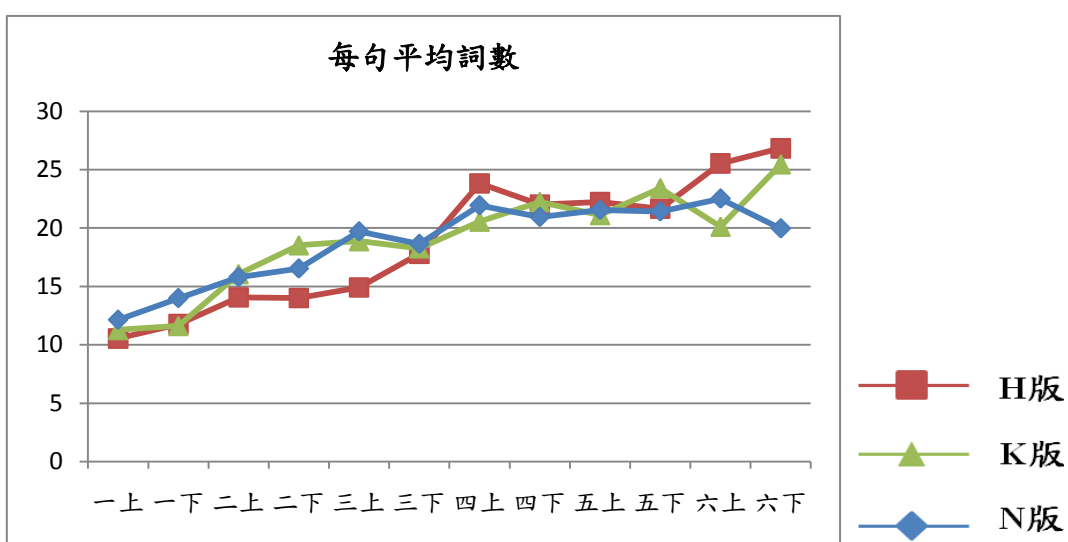


圖 12 每句平均詞數

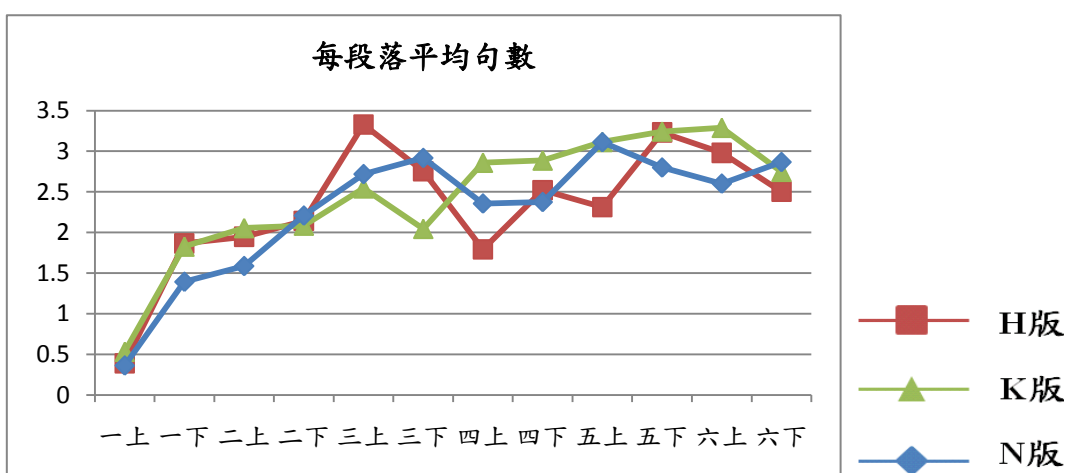


圖 13 每段落平均句數

## B. 人稱代名詞數

在本研究中，人稱代名詞的詞彙是參考中文詞知識庫小組對中文詞類的分類。人稱代名詞在文章中使用過多的話容易造成指涉上的混淆以及閱讀理解上的困難 (Graesser et al., 2004)。故本研究納入人稱代名詞數為中文可讀性的分類特徵。圖 14 為人稱代名詞數於國小國語科各冊間的特徵值散佈圖。

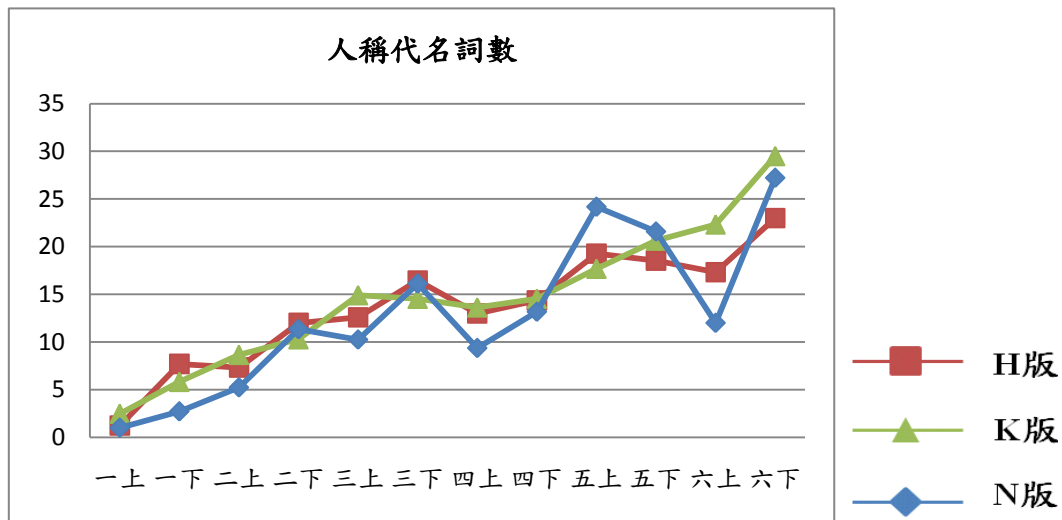


圖 14 人稱代名詞數

## C. 正向連接詞數、負向連接詞數

連接詞是語意連貫的特徵，且是幫助讀者整合文本連貫性的重要因素。Halliday & Hasan(1976)曾提到連接詞在英文文本中是導致文章連貫的主要重點之一。故本研究納入連接詞數為中文可讀性的分類特徵。圖15與圖16為正向連接詞數與負向連接詞數於國小國語科各冊間的特徵值散佈圖。

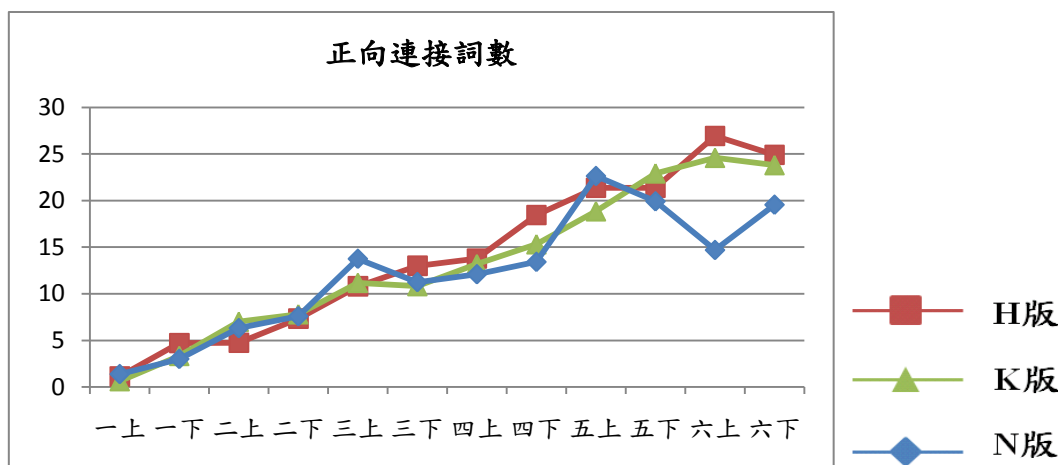


圖 15 正向連接詞數

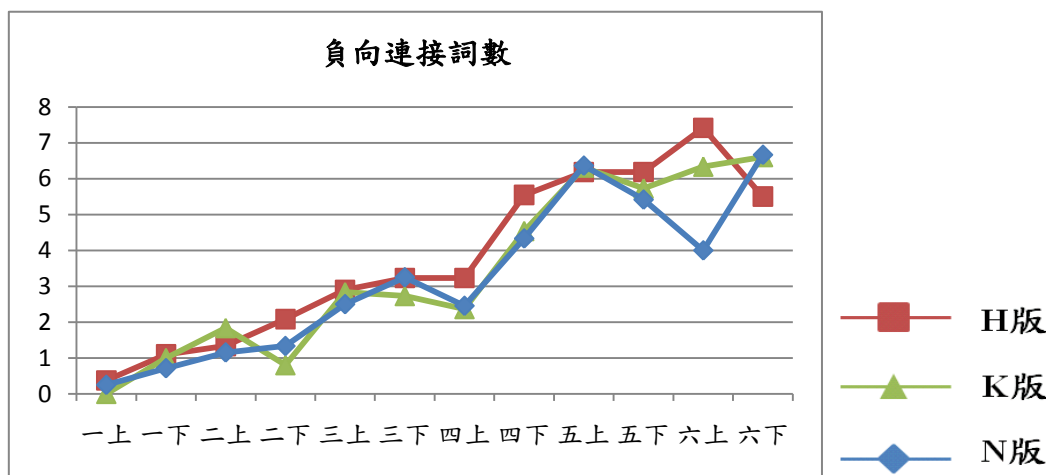


圖 16 負向連接詞數

#### D. Type-Token Ratio

Type-Token Ratio在Coh-Metrix裡面的計算為文句中每一種實詞的出現次數除以總實詞數的值。若比值越接近1，則表示該句由越多不同的實詞所組成。研究中所跑出的TTR散佈圖無法看出任何趨勢，但Graesser et al.,(2004)提到TTR值低的話代表某一個字一直重複的出現，表示讀者去閱讀該文本是很簡單且可以快速閱讀。故本研究參考Coh-Metrix裡的計算，對中文文本分析其TTR值。圖17為Type-Token Ratio於國小國語科各冊間的特徵值散佈圖。

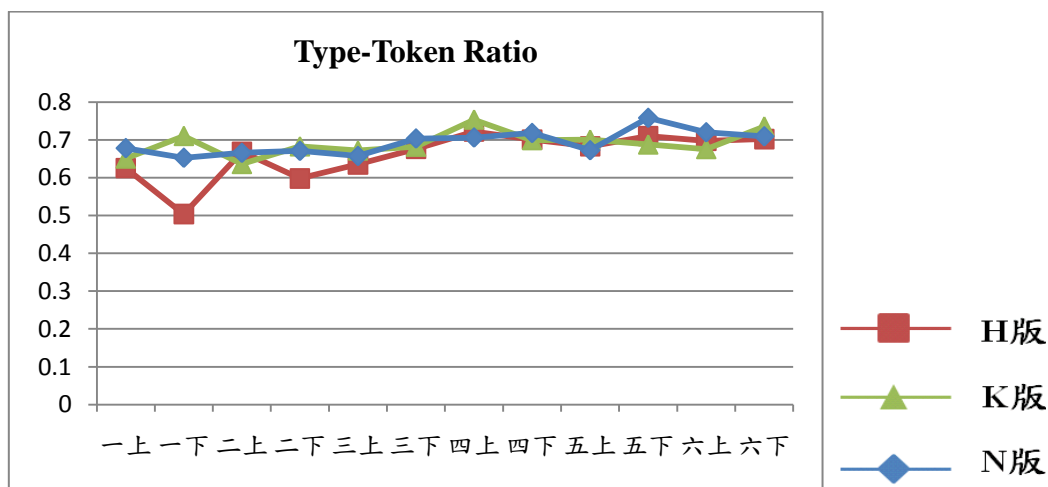


圖 17 Type-token Ratio

#### E. 否定詞數

否定詞在文句中的功能是在語意上表示相反、否定的效果。許多學者也認為否定詞在句子中有一些代表的意義。Hwang (1992)認為否定句在敘述文體之言談中有標記轉折之用。Givón(1979)認為否定句是情境中表示否認的行為。故本研究採用否定詞數為中文可讀性的分類特徵。圖18為否定詞數於國小國文科各冊間的特徵值散佈圖。

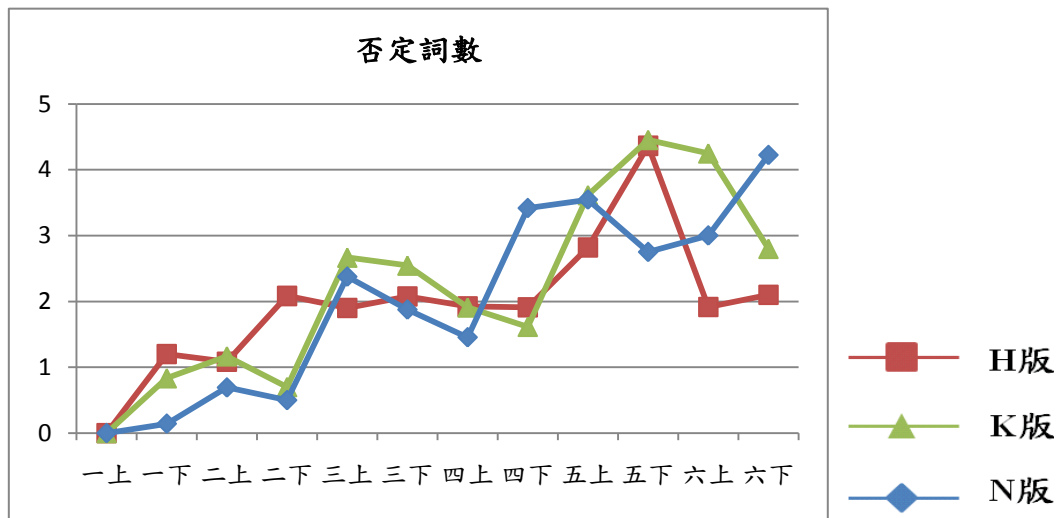


圖 18 否定詞數

#### F. 筆畫數

楊孝滢(1978)指出筆畫數在閱讀理解上有影響力。陳茹玲、蘇宜芬(2010)也提到以筆畫數為分析單位時，發現字元複雜度效果。故本研究將筆畫數分成低筆畫數、中筆畫數、高筆畫數及平均比畫數為中文可讀性的分類特徵。圖19、圖20、圖21與圖22為低筆畫數、中筆畫數、高筆畫數與平均比畫數於國小國語科各冊間的特徵值散佈圖。



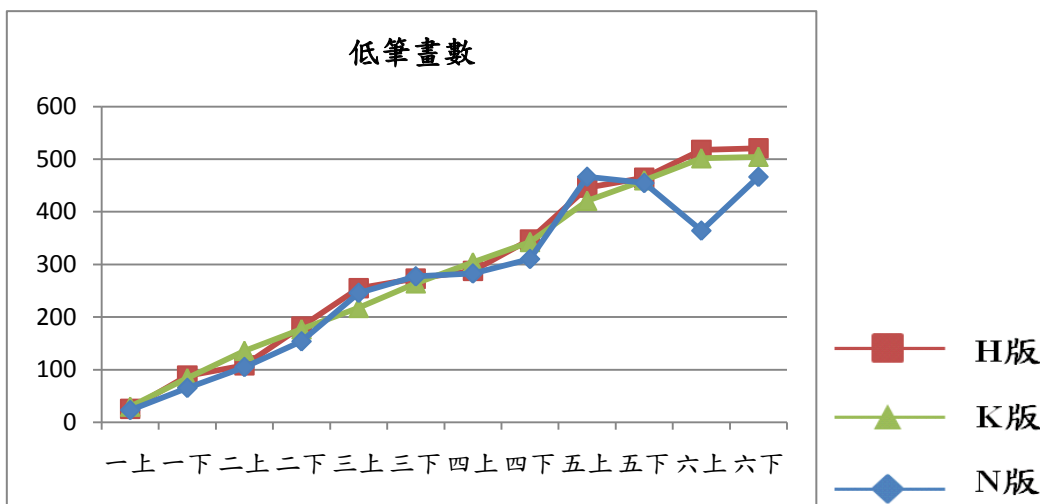


圖 19 低筆畫數

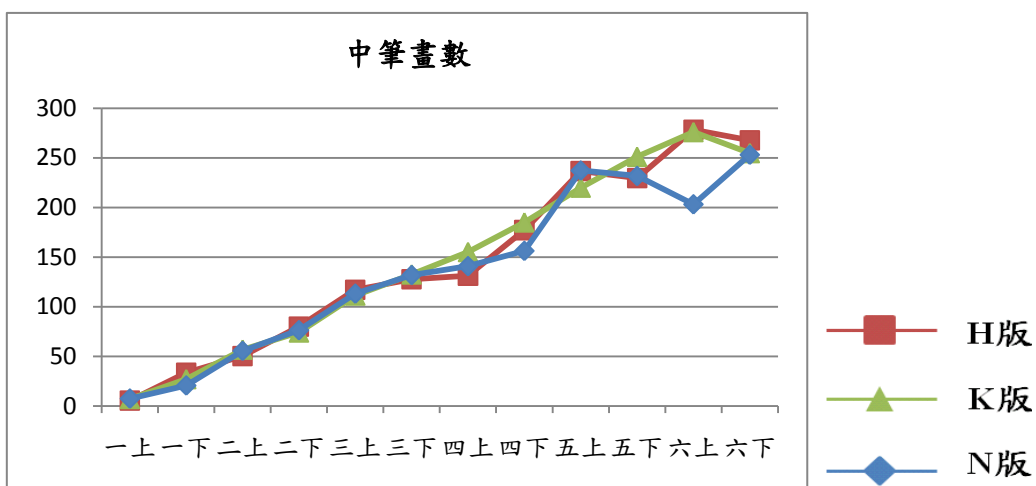


圖 20 中筆畫數

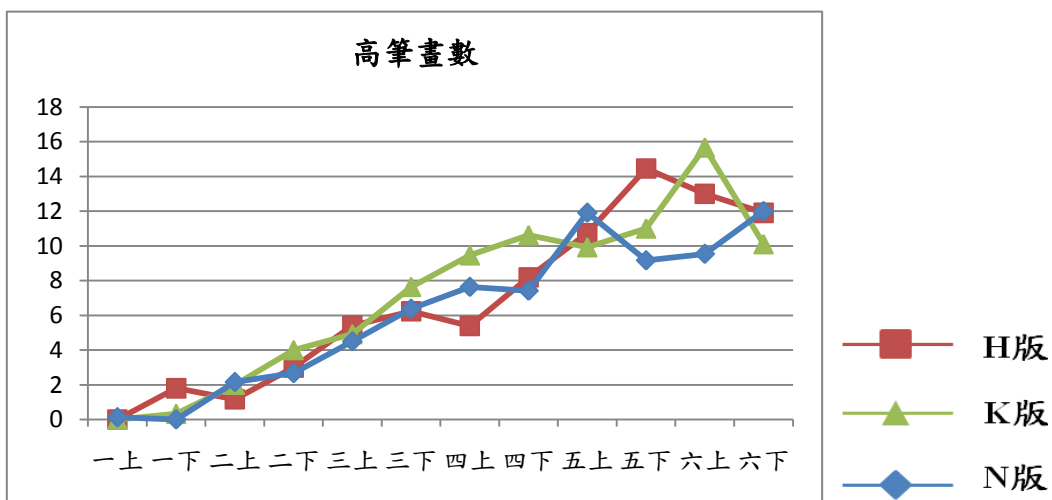


圖 21 高筆畫數

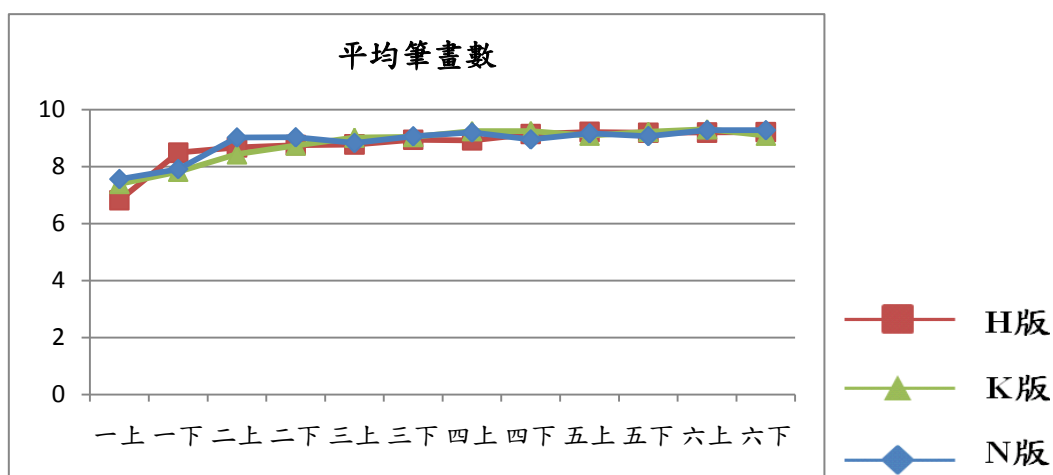


圖 22 平均筆畫數

### G. 不存在詞

圖23為不存在詞數於國小國語科各冊間的特徵值散佈圖。在以前的可讀性公式中許多人也會使用「難字」做為一個指標的參考。Fry, Kress & Fountoukidis(1993)認為詞頻前100高的詞可以組成文章50%的詞彙量，文章內容的65%由詞頻前300高的詞就能組成。由於不在詞表裡的詞通常為頻率相對較低的詞，因此所得到的不存在詞的比率便可代表難詞的概念。文章的難詞越多，表示該文章的詞彙讀者較少接觸，因此，難詞即是根據詞頻而來用以區辨文章難度的指標。本研究利用「難字」的概念建立一個詞表，先分析資料庫中所有文章的詞頻，再與教育部國小詞頻總表進行交叉比對，取兩者聯集的前3000高詞頻的詞作為常用詞表，再比對文章中有出現及未出現於詞表中的詞，如果不在此表中的詞即為本研究中不存在詞的特徵。

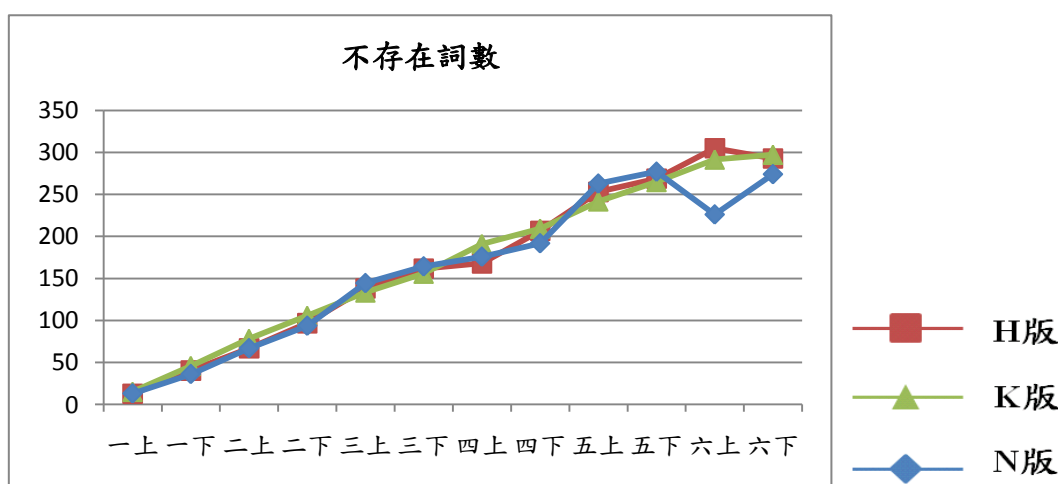


圖 23 不存在詞數

#### H. 音節數、二字詞及三字詞

英文可讀性公式常以音節當作判斷文章難度的指標，Flesch Reading Ease(Flesch, 1948)及Flesch-Kincaid年級公式(Flesch, 1976)等公式均以音節數為可讀性指標。而中文與英文不同，中文的詞中多是以二字詞組成。根據楊淦滢(1978)的研究中顯示，二字詞數具有文章難度的預測能力，因此本研究分別加入音節數、二字詞數與三字詞數為文本的特徵。圖24與圖25為二字詞數與三字詞數於國小國語科各冊間的特徵值散佈圖。

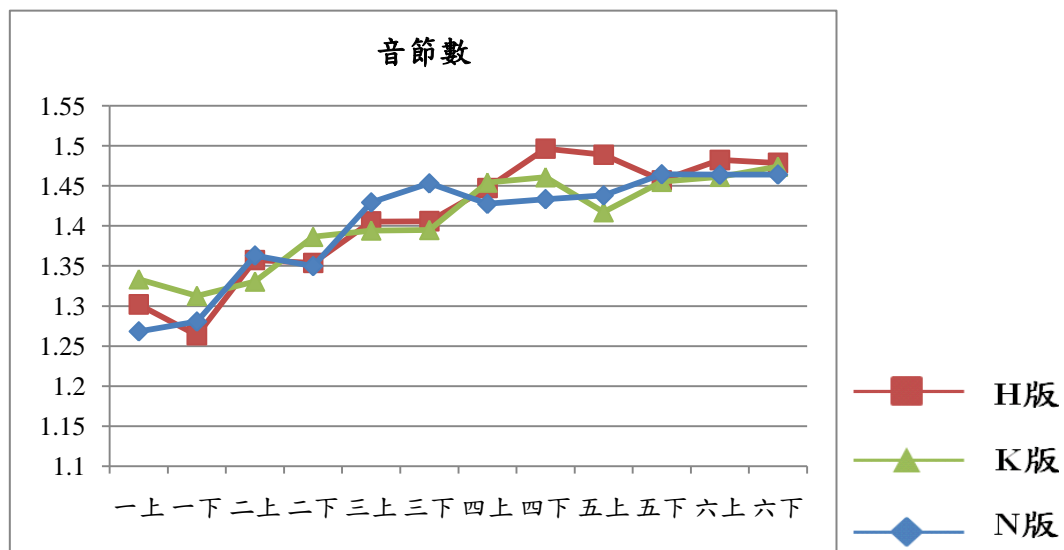


圖 24 音節數

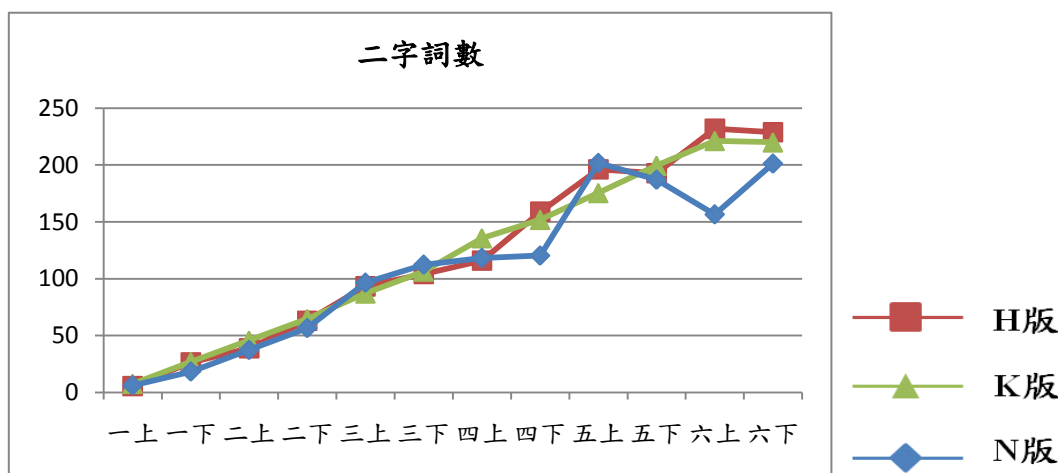


圖 25 二字詞數

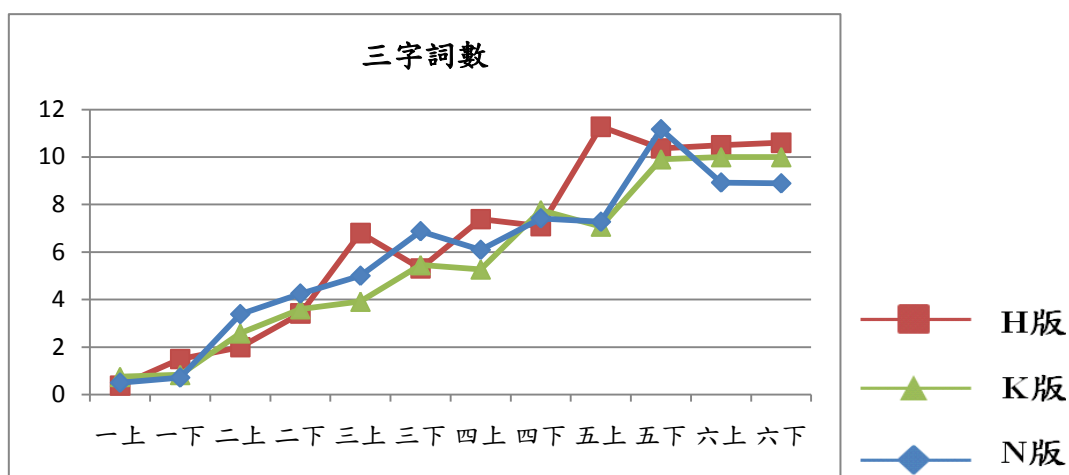


圖 26 三字詞數

## I. 實詞種類數

圖 26 為實詞種類數於國小國語科各冊間的特徵值散佈圖。實詞表示一種具體概念的詞包括了名詞、動詞、形容詞、副詞、數量詞、代名詞。本研究想觀察這些有意義的詞在文中的種類多寡是否能影響文本的可讀性。

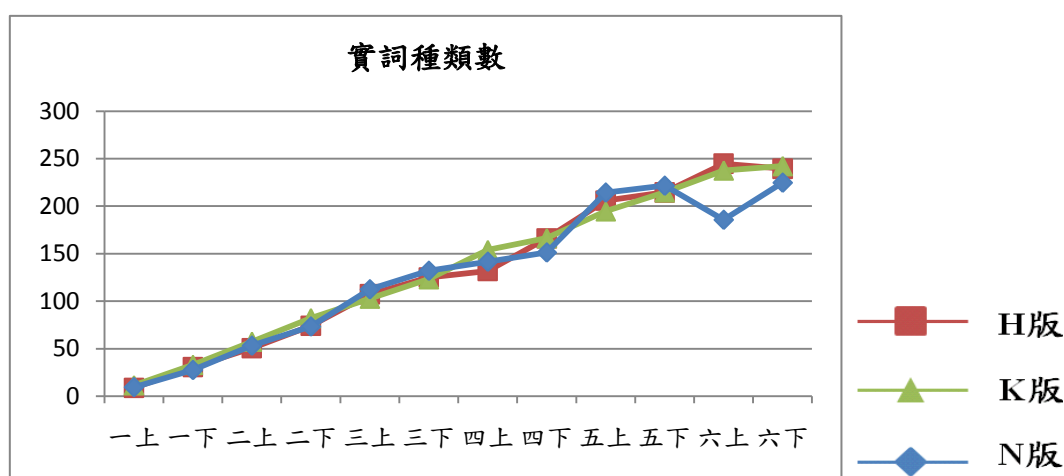


圖 27 實詞種類數

## 2. 訓練與測試階段

本實驗將所有文本輸入至中文斷詞工具做斷詞的處理，經過運算後會輸出文本的特徵值。有了文本的特徵值以 LIBSVM 做為分類的工具，將國小一至六年級的國語科課文做可讀性的分類(以冊別當作分類的單位)，並將 80% 的課文當作訓練資料，再利用預測模型對測試資料做預測。在線性迴歸的實驗上，同樣是以 80% 的資料當作訓練資料以進行訓練來建置線性可讀性預測公式。

### 3.錯誤分析

觀察並統整文本預測結果。對分類錯誤的課文做錯誤分析，觀察使該課文分類錯誤的原因。錯誤分析針對分類準確率較高的工具(SVM)進行分析。本研究將錯誤類型分成三種：實驗樣本的影響、課文編排的影響及課文內容的影響，並統計被高估(SVM 將原本課文的冊別預測到更高的冊別)及低估(SVM 將原本課文的冊別預測到更低的冊別)的課文的共通點等來解釋為什麼這些課文會被分類到錯誤的冊別。最後製成一 12X12 的矩陣清楚觀看分類預測情形。

## 第四節 實驗結果

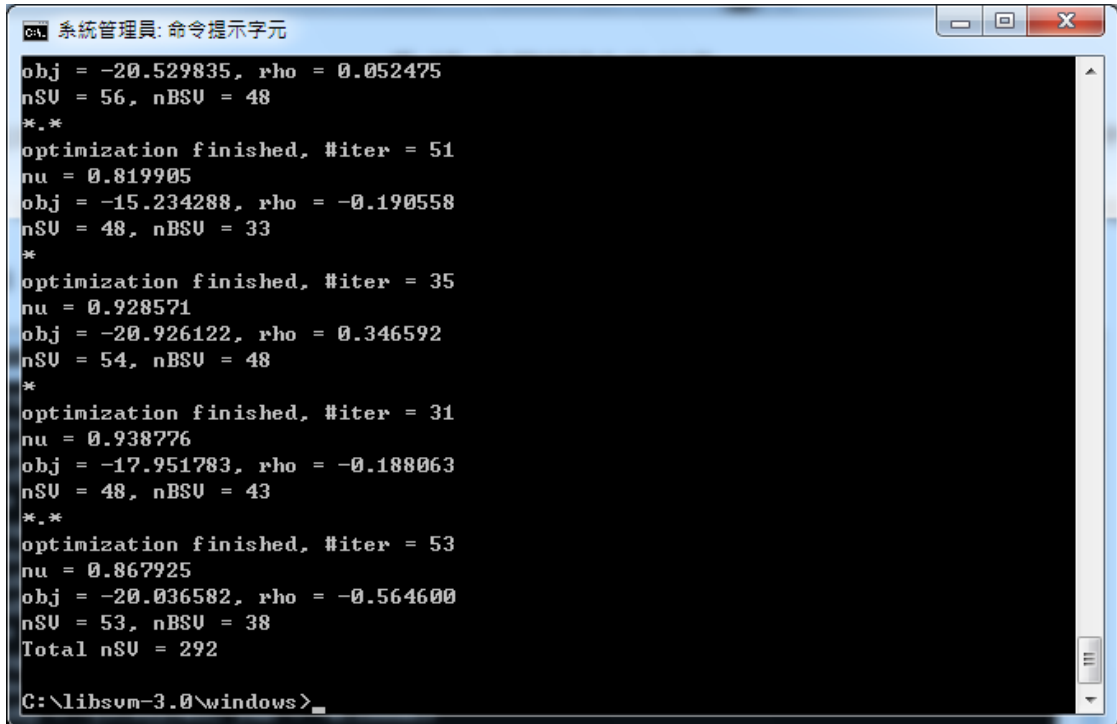
### 一、實驗方法

本研究使用 k-fold 交叉比對(cross-validation)的方式對 SVM 預測出來的正確率做評估。此法是將資料分成 k 個部分。使用此方法，必須進行 k 次訓練和測試，且這 k 群資料彼此之間互斥，這樣便能確保每筆資料皆能當作測試資料，而且全部的測試樣本都是獨立的。Fold 數沒有一定的選擇，大部分的研究多是以試誤法(trial-and-error)去決定。而 fold 數在本實驗中的選擇對準確率的影響差異不大，在不同 fold 數都已達到收斂效果的時候，本實驗以最有效率的方式完成分類的方法。另外，本研究亦參考了許多學者在利用 SVM 進行文本分類時多以 5-fold 交叉驗證的方式來進行實驗(Parrado-Hernandez and Hardoon, 2007; Zhan and Loh, 2009 ; Singh, Murthy, and Gonsalves, 2010)。故本實驗以 5-fold 交叉驗證的分類方法，將三個版本的課文隨機分配至 5 個 fold 裡，使用此種方式進行 5 次的訓練及測試，最後將 5 次的準確率做平均得到最後的準確率。

### 二、LIBSVM 進行可讀性預測

#### 1. LIBSVM 預測每冊課文難度

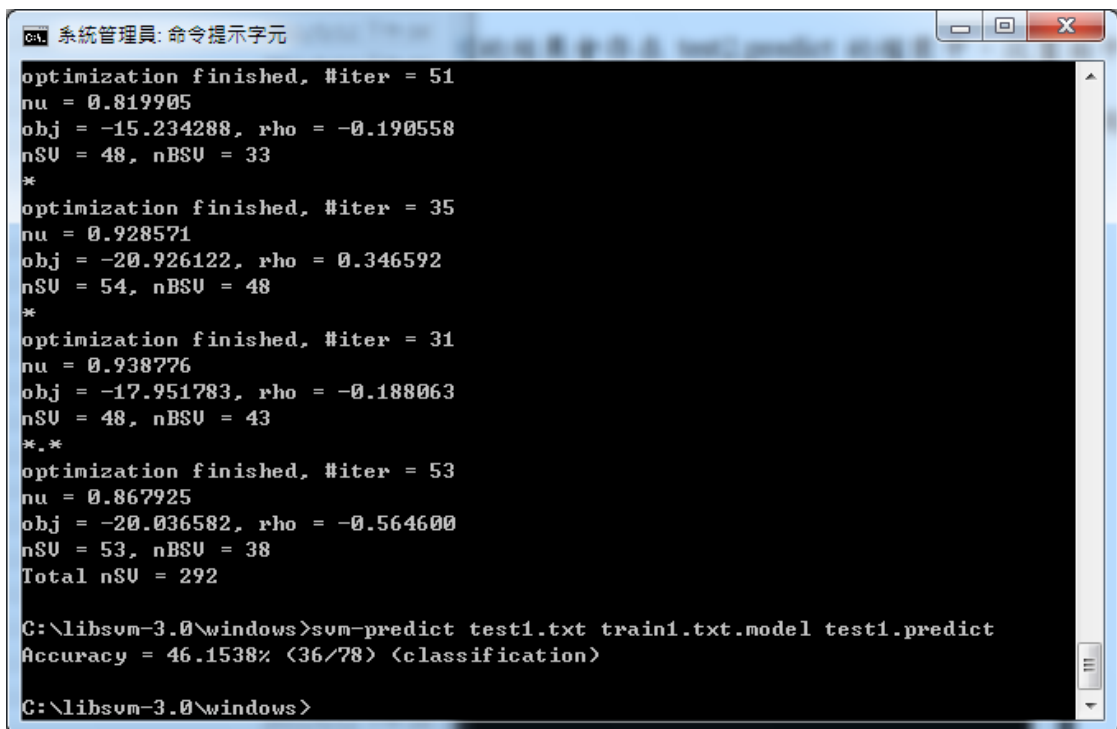
用 LIBSVM 訓練完成的畫面如圖 27 所示，以 fold2 的訓練資料為例。LIBSVM 訓練時的指令為 `svm-train training.txt`，完成後會輸出一個 `training.txt.model` 的檔案，以做為接下來預測時輸入的模型。



```
obj = -20.529835, rho = 0.052475
nSU = 56, nBSU = 48
**
optimization finished, #iter = 51
nu = 0.819905
obj = -15.234288, rho = -0.190558
nSU = 48, nBSU = 33
**
optimization finished, #iter = 35
nu = 0.928571
obj = -20.926122, rho = 0.346592
nSU = 54, nBSU = 48
**
optimization finished, #iter = 31
nu = 0.938776
obj = -17.951783, rho = -0.188063
nSU = 48, nBSU = 43
**
optimization finished, #iter = 53
nu = 0.867925
obj = -20.036582, rho = -0.564600
nSU = 53, nBSU = 38
Total nSU = 292
C:\libsvm-3.0\windows>
```

圖 28 LIBSVM 的訓練

預測完的畫面如圖 28 所示，其預測的結果會存在 test2.predict 的檔案中。從畫面中可得到 fold2 測試資料的準確率為 41.7582% (38/91)，意指在 91 筆的測試資料中，有 38 筆資料是跟訓練資料的類別相符合的。



```
optimization finished, #iter = 51
nu = 0.819905
obj = -15.234288, rho = -0.190558
nSU = 48, nBSU = 33
**
optimization finished, #iter = 35
nu = 0.928571
obj = -20.926122, rho = 0.346592
nSU = 54, nBSU = 48
**
optimization finished, #iter = 31
nu = 0.938776
obj = -17.951783, rho = -0.188063
nSU = 48, nBSU = 43
**
optimization finished, #iter = 53
nu = 0.867925
obj = -20.036582, rho = -0.564600
nSU = 53, nBSU = 38
Total nSU = 292
C:\libsvm-3.0\windows>svm-predict test1.txt train1.txt.model test1.predict
Accuracy = 46.1538% (36/78) (classification)
C:\libsvm-3.0\windows>
```

圖 29 LIBSVM 的預測

由於 SVM 在分類準確率的高低有很大的部份取決於參數的選擇。故本研究直接使用台大林智仁老師提供的工具 Grid.py 來選擇最佳的參數，並以 SVM 的輻射基底核心函數(RBF Function)當做核心函數，Hsu & Lin (2003)提到以 RBF 當作核心函數最適合非線性的資料且可以得到較高的準確率。表 9 為使用 LIBSVM 時的核心函數及在各 fold 的成本參數 C(cost)與  $\gamma$ (gamma)兩參數值。表 10 列出 LIBSVM 在重要特徵上的權重，LIBSVM 內部以 F-score 來計算特徵的重要排序，F-score 的值愈大表示這個特徵越具有辨別性。

表 9 LIBSVM 內部使用資訊

核心函數	輻射基底核心函數(Radial Basis Kernel Function)				
參數:	Fold1	Fold2	Fold3	Fold4	Fold5
cost	0.5	2.0	8.0	32.0	2.0
參數:	Fold1	Fold2	Fold3	Fold4	Fold5
gamma	0.00012207	3.05175781	3.05175781	3.05175781	3.05157781
	03125	25e-05	25e-05	25e-05	25e-05

表 10 LIBSVM 在重要特徵上的權重

	F-score
實詞數取 log	8.965547
不存在詞	5.728489
實詞種類數	5.648761
字數	4.929149
中筆畫數	4.779547
二字詞數	4.531004
低筆畫數	4.276628
實詞數	4.017666
詞數	4.006918



## 2. LIBSVM 預測結果

本研究定義若 LIBSVM 預測的課文冊別與實際上課文的冊別完全符合，則稱之為準確率(accuracy)。另外，考慮到每個冊別間的課文可能因為要在難度上做銜接，使得每個學期間的課文難度會重疊到。所以，將預測的準確範圍放寬上下一冊，稱之為正確率(fit rate)(若實際冊別為 3 上，而 LIBSVM 預測為 2 下或 3 下，本研究就認定其在正確的預測範圍)。由表 11 可知本實驗利用 LIBSVM 預測國小國語科課文冊別的準確率為 47.9275%、正確率為 80.3109%。圖 29 以矩陣的方式統計 LIBSVM 在各冊預測對與錯的課文數。

表 11 LIBSVM 預測結果在各 FOLD 的預測結果

	Folds					平均
	Fold1	Fold2	Fold3	Fold4	Fold5	
課文數	78 篇	78 篇	77 篇	77 篇	76 篇	--
準確率	0.462	0.474	0.468	0.455	0.539	0.479
正確率	0.795	0.821	0.805	0.87	0.776	0.813

		LIBSVM預測冊別												各冊準確率	各冊正確率
		1上	1下	2上	2下	3上	3下	4上	4下	5上	5下	6上	6下		
實 際 冊 別	1上	24												100.00%	100.00%
	1下		19	4										82.61%	100.00%
	2上		8	23	6									62.16%	100.00%
	2下			3	26	4	1							76.47%	97.06%
	3上				2	21	5	1	1					70.00%	93.33%
	3下			1	1	8	9	10	2	1				28.13%	84.38%
	4上				2	5	6	14	5	3				40.00%	71.43%
	4下				1	3	2	8	12	5	2	3		38.89%	69.44%
	5上							2	8	13	6	5	1	37.14%	77.14%
	5下								6	9	4	11	4	11.76%	70.59%
	6上				2	1			4	5	3	15	7	40.54%	67.57%
	6下					1			1	6	5	11	5	17.24%	55.17%

圖 30 以 LIBSVM 預測課文冊別之矩陣

圖 29 中可看出 386 篇的國小課文中預測結果完全符合實際的課文冊別以及為預測的合理範圍(即把預測的準確範圍放寬上下一冊)。從此圖可以發現：

1. 一上及一下課文的正確率都達 100% 。
2. 五下的準確率最低、六下的正確率最低，可知分類錯誤的課文都偏向高年級。

### 三、GLM 進行可讀性預測

本實驗以 STATISTICA 這套軟體進行 GLM 的分析。將中文文本特徵納入線性迴歸分析，以課文冊別為依變項，各文本特徵為自變項。本實驗以 80% 的課文當作訓練資料，以 20% 的資料當作測試資料。陳順宇(2000)提到逐步迴歸法可以選取解釋能力強的預測變數進入迴規模式並排除解釋能力小的變數在外。故研究以逐步迴歸 (Stepwise Regression) 從多個自變數中選擇建立迴歸方程式的變數，顯示本研究建置之可讀性預測公式可有效解釋文章冊別的變異量達 88.0415%，公式如下：

$$\begin{aligned}
 \text{年級} = & -8.01682 + 0.00626 * \text{不存在詞} + 3.02871 * \text{音節數} + \\
 & 1.77152 * \text{實詞數取 log} + 0.1227 * \text{正向連接詞數} + 0.015 * \text{中筆畫數} + \\
 & 2.85746 * \text{Type-Token Ratio} + 0.02509 * \text{人稱代名詞數} + 0.04852 * \text{平均每句詞數} - \\
 & 0.02017 * \text{連接詞數}
 \end{aligned}
 \tag{17}$$

最後以 20% 的訓練資料做效度驗證，其效度驗證的準確率為 39.1192%。以 12x12 矩陣圖統計 GLM 在各冊預測對與錯的課文數，如圖 30。

表 12 文本特徵對於 GLM 公式的變異量

	Multiple R-square	p-level
不存在詞	0.841698	0.000000*
音節數	0.851488	0.000027*
實詞數取 log 值	0.858917	0.000175*
負向連接詞數	0.863649	0.002252*
中筆畫數	0.867366	0.006060*
Type-Token Ratio	0.873829	0.000231*
人稱代名詞數	0.876047	0.028486*
每句平均詞數	0.879710	0.004467*
連接詞數	0.880415	0.208999

\*p < .05.

		GLM預測冊別														
		1上	1下	2上	2下	3上	3下	4上	4下	5上	5下	6上	6下	N/A	各冊準確率	各冊正確率
實際冊別	1上	6	2											16	25.00%	33.33%
	1下	10	12	1											52.17%	100.00%
	2上		7	24	6										64.86%	100.00%
	2下			10	18	6									52.94%	100.00%
	3上				5	13	12								43.33%	100.00%
	3下				2	4	17	8	1						53.13%	90.63%
	4上					2	11	18	3	1					51.43%	91.43%
	4下					1	9	8	13	2	3				36.11%	63.89%
	5上						1	2	12	11	3	4	1	1	31.43%	74.29%
	5下							2	7	12	8	3	2		23.53%	67.65%
	6上				1	2		1	5	7	11	8	2		21.62%	56.76%
	6下						1		2	4	12	6	3	1	10.34%	31.03%

圖 31 以 GLM 預測課文冊別之矩陣

綜合上述兩種預測方式做一整理如表 13，可看出對國小國語課文做分類預測以冊別為單位，SVM 之準確率為 47.9275%、正確率為 80.3109%；GLM 之準確率為 38.82%、正確率為 75.75%。另外，兩種文本可讀性分類方式在重要特徵的排序上不同，不同的特徵排序方式會影響中文文本可讀性分類上的準確性。表 14 列出 SVM 與 GLM 對重要特徵的排序。

表 13 SVM 及 GLM 之準確率與正確率比較

	SVM	GLM
準確率 (以冊為單位)	47.9275%	38.8252%
正確率 (放寬上下一冊)	80.3109%	75.7499%

表 14 中文文本特徵於 SVM 及 GLM 之排序

重要性排名	SVM	GLM
1	實詞取 log	不存在詞
2	不存在詞	音節數
3	實詞種類數	實詞取 log
4	中筆畫數	負向連接詞數
5	實詞數	中筆畫數
6	中筆畫數	TTR
7	字數	人稱代名詞數
8	二字詞	每句平均詞數
9	低筆畫數	連接詞數

## 第五節 實驗結果討論

根據上述實驗結果，研究者分別針對 GLM 與 SVM 進行文章可讀性預測之結果進行討論以及針對 SVM 分類預測的結果進行錯誤分析。

### 一、線性迴歸與 SVM 進行文章可讀性預測之比較

數據顯示 SVM 比 GLM 有較好的準確性。國外研究中，Peterson & Ostendorf (2008) 的實驗拿 Weekly Reader 雜誌當作讀物(國小 2~5 年級的等級)，利用 SVM 與傳統線性公式 Flech-Kincaid 做比較，發現傳統線性公式的準確率(accuracy)比 SVM 來得低。另外，Feng et al.(2010)也提出對文本分類的實驗中，SVM 預測的準確率比線性模式(Logistic Regression)來的好。傳統的線性可讀性公式因須大量的訓練資料且資料呈常態分佈故效能不佳，非線性的分類器確實可以改善以往技術只能利用統計的方式來計算可讀性的方法。

### 二、SVM 與 GLM 的重要變數排序

此兩種可讀性分類方法在挑選重要變數上有差異，因為 SVM 以 RBF 核心函數去做預測，而 GLM 則以線性的模式來預測文本冊別。但可觀察到「不存在詞」這個指標在兩種可讀性分類方法的排名都很前面。其結果也與本實驗所做的「不存在詞散佈圖」吻合，在冊別上呈現隨年級而成長的趨勢。荊溪昱(1995)也以「常用詞」做為影響可讀性的中文語言因素。在此也驗證「難字」的概念在文本可讀性上是一個

重要的指標。

### 三、支援向量機分類預測結果之錯誤分析

根據上述實驗結果，研究者針對 SVM 分類預測的結果進行錯誤分析。以下表 15 為研究者自行整理出預測不正確（即預測結果不在實際冊別上下一冊的範圍內）的課文。課文代號中第一個英文字母分別代表 H 版、K 版及 N 版的課本，C 代表國語科；分類情形中 1 代表 SVM 高估了原本的冊別，0 代表 SVM 低估了原本的冊別。386 篇的課文，有 310 篇分類正確的課文，有 76 篇分類不正確，即有 80.31% 的正確率(fit rate)。總共有 48 篇課文被 SVM 高估、有 29 篇被低估。本研究進行以下幾點觀察可能導致 SVM 分類預測錯誤的原因。

表 15 SVM 分類錯誤之課文

課文代號	課文名稱	冊	SVM 預測結果	相差冊數	分類情形
H098C03201	桃花開了	6	4	2	0
H098C04103	阿里山上看日出	7	5	2	0
H098C041041	山和海的書信(海的神祕)	7	4	3	0
H098C041042	山和海的書信(山的強壯)	7	4	3	0
H098C04111	泰雅族的紋面文化	7	5	2	0
H098C04206	談合作	8	6	2	0
H098C05106	中國結	9	7	2	0
H098C05207	美麗的溫哥華	10	8	2	0
H098C05211	聽！流星的故事	10	8	2	0
H098C06114	用心生活	11	8	3	0
H098C06202	最後一片葉子	12	10	2	0
H098C06204	讀書報告——愛的教育	12	9	3	0
H098C06205	三峽祖師廟	12	10	2	0
H098C06209	我的少年禮	12	10	2	0

H098C06210	禮物	12	9	3	0
H098C06212	畢業生致答詞	12	9	3	0
K098C03201	春天的訊息	6	3	3	0
K098C04102	在空中飛行	7	5	2	0
K098C04202	平溪放天燈	8	6	2	0
K098C04203	兩個和尚	8	5	3	0
K098C05102	秋江獨釣	9	7	2	0
K098C06101	模仿貓	11	9	2	0
K098C06107	馬可·波羅遊中國	11	9	2	0
K098C06112	談辯論	11	8	3	0
K098C06202	雨石花	12	9	3	0
K098C06209	肯定自我	12	8	4	0
K098C06210	寫紀念冊的日子	12	10	2	0
N098C04102	擔仔麵	7	5	2	0
N098C04107	閱讀列車(一) 大自然的雕刻家	7	5	2	0
N098C04210	誰救了我	8	5	3	0
N098C042131	寓言二則(一) 揠苗助長 (原文出自孟子公孫丑上)	8	4	4	0
N098C042132	寓言二則(二) 愚公移山 (原文出自列子湯問)	8	5	3	0
N098C05203	開卷有益	10	8	2	0
N098C05208	秦始皇的地下護衛軍	10	8	2	0
N098C05212	遙遠的友情	10	8	2	0
N098C05214	我們住在地球村	10	8	2	0
N098C06103	大自然的規則	11	9	2	0

N098C06105	翠玉白菜	11	8	3	0
N098C061081	機智過人(一) 學習父親	11	5	6	0
N098C061082	機智過人(二) 等級不同	11	4	7	0
N098C061083	機智過人(三) 狀況不同	11	4	7	0
N098C06109	理直氣和	11	9	2	0
N098C06110	語言與智慧	11	8	3	0
N098C06114	成功的背後	11	9	2	0
N098C06203	魯冰花	12	10	2	0
N098C062041	真正的富有(一) 沒有圍牆的花園	12	9	3	0
N098C062042	真正的富有(二) 有錢人可能很窮 (改寫自古川千勝 其實有錢人可能很窮)	12	5	7	0
N098C06212	誠摯的祝福	12	9	3	0
H098C02213	給李奶奶的信	4	6	2	1
H098C03107	保羅的撿球車	5	8	3	1
H098C03209	小強減重記	6	8	2	1
H098C04106	讀書報告——小恩的祕密花園	7	9	2	1
H098C04209	烏柏巷的故事	8	10	2	1
H098C04212	參觀宜蘭傳統藝術中心	8	11	3	1
H098C05102	帶箭的花鳧	9	11	2	1
H098C05104	邁向低碳生活	9	11	2	1
H098C05111	聆聽天籟	9	11	2	1
H098C05112	創世基金會訪問記	9	11	2	1



H098C05205	紐西蘭的毛利文化	10	12	2	1
K098C03213	狐狸和白鶴	6	9	3	1
K098C03214	神筆馬良	6	8	2	1
K098C04106	永不放棄的愛	7	9	2	1
K098C04108	哥倫布的航海夢	7	9	2	1
K098C04208	以一顆溫柔的心為槳	8	11	3	1
K098C04209	書的王國	8	10	2	1
K098C04214	臺灣昆蟲知己——李淳陽	8	11	3	1
K098C05105	來去都蘭	9	11	2	1
K098C05107	熊與鮭魚	9	11	2	1
K098C05204	湖濱散記	10	12	2	1
K098C05213	筆記四則	10	12	2	1
N098C03112	放牛的發明家	5	7	2	1
N098C05106	勇者鬥惡龍	9	11	2	1
N098C05107	雅典娜與橄欖樹	9	12	3	1
N098C05109	兩兄弟	9	11	2	1
N098C05112	擁抱生命中的每一分鐘	9	11	2	1
N098C05206	父親與我	10	12	2	1

本實驗中，SVM 所預測的分類正確率(fit rate)高達 80.31%，而準確率(accuracy) 47.92%仍有進步的空間。以下討論幾項可能導致準確率下降的因素，做為未來進一步研究的參考。

#### 1. 實驗樣本的影響

本實驗合併三個版本的課文做為樣本，選取部分當作訓練資料，而三個版本在每冊的課文難易度從特徵散佈圖(圖 8 至圖 26)中可看出不一致的情形(但若只使用單一版本會使得樣本數太少)。這樣的情形使得訓練資料中同一冊的難度不一，導致

SVM 在預測的時候有不一樣的標準。若訓練資料更一致，即三個版本對於各冊的難易度有更一致的編排標準，可能可以使得預測的準確率提升。

另外，將課文分到 5 folds 時因採隨機分配，可能導致每個 fold 包含的課文冊別不平均導致有些 fold 裡的課文偏向高年級冊別或低年級冊別，以至於在訓練資料時不準確。

## 2.課文編排的影響

檢視分類錯誤的課文，有些課文的特徵值確實比同冊課文的特徵值平均來的高或者低。故進一步觀察課文實際的內容以及編排，整理以下幾點可能導致的錯誤：

A、有些課文是由幾個小篇故事組成一課。如 H 版課文第七冊的第四課，其課文內容由兩小篇文章(H098C041041 及 H098C041042)所組成。這種情形導致其與字數有相關的特徵都會受到影響。分開計算為單獨的一篇文章其與字數相關的特徵會比同冊別的課文來的少，又若將小文章合併成一篇文章計算則相關的特徵值又會過大，故此種類型的課文都被 SVM 低估。

B、本實驗所使用的文本特徵仍無法將各文體的文本清楚的做分類。如 K 版課文第六冊的第十三課(K09803213)的課文文體屬於劇本。劇本的文章編排方式會出現很多對話式的句子，造成每句平均字數等特徵都會比同冊的其他課文來的少且像是句數等特徵就會比較多，故像此課就會被 SVM 高估。

## 3.課文內容的影響

本實驗所使用的文本特徵皆屬於比較表面的特徵(字數、句數、段落數等)。可以從圖 30 看到預測錯誤的文章都分布在五、六年級居多，且這些文章的可讀性都被 SVM 低估了。本研究從此實驗結果認為五、六年級(高年級)的文章可能比起文章的表面特徵更注重內容所帶給讀者的啟發。雖然有些高年級課文以簡單幾個字句道出整篇故事，其中的意涵卻很深厚。故單純用這些文本的表面特徵在高年級的課文較無法辨別其難易度。

四、目前中文文本分類比起英文文本分類的準確性上還有進步的空間

Schwarm & Ostendorf, 2005; Petersen & Ostendorf, 2008 在過去採用 SVM 對英文文本進行可讀性分類能得到不錯的準確性，因為學者採用的資料以 "Weekly Reader" 的教育雜誌(將近 2400 筆資料)，比起本研究所使用的國小國語科課文資料量(386 筆資料)大量許多。Larsson (2006)也指出影響文本分類方法的效果包括：

- 1.資料集的選擇：大量的資料可以讓分類的效果更好。
  - 2.演算法的開發：演算法的開發是為了能快速且正確的達到分類效果。
  - 3.變數屬性的挑選：選擇最佳的屬性特徵組合可以讓分類模型提升分類上的準確性。
- 本研究認為未來技術的開發可以針對此三點加強，對於如何利用 SVM 提升分類的準確性，達到非常小的分類錯誤。

#### 五、中文文本分類的特徵組合

本研究目前所採用的文本特徵包括參考 Coh-Metrix 裡的指標及研究中自行開發的指標。Coh-Metrix 裡的指標共有 60 項，而本研究採用了目前技術可做到的指標(根據本實驗對文本特徵做的散佈圖趨勢及過去學者對文本特徵的研究)，並捨去"Flesch-Kincaid Grade Level" 及"Flesch Reading Ease"兩個指標，因為其運算方式是以線性迴歸來作計算且此公式是以很早期的英文文本資料做訓練的資料，其訓練得來的迴歸係數不適用於本研究中的中文文本，故沒有加入此兩指標。另外，在現階段技術的考量，本研究沒有加入 Coh-Metrix 裡語意及文法剖析相關的指標。

## 第五章 結論與未來發展

### 第一節 結論

本研究設計一個中文可讀性指標分析系統，經過中文的斷詞後，利用系統產出的文本特徵輸入至 SVM 做分類的預測。過程中先做特徵分析，特徵分析的方法有二。方法一是將 Coh-Metrix 提供的文本特徵拿來先用本實驗的資料做特徵散佈圖分析，觀察每個特徵在各學年的趨勢，並分析該特徵是否適用於中文文本可讀性的分類上；方法二是將學者提出對於中文文本可讀性有辨別力的文本特徵拿來使用。最後本研究將 SVM 產出的分類準確率與線性模式(GLM)的可讀性公式做比較。並針對 SVM 的分類結果做錯誤的分析，將錯誤的課文做一錯誤分析觀察使該課文分類錯誤的特徵、課文內容、使用的文體、課文撰寫的編排等，最後製作一 12X12 矩陣清楚觀看各課分類預測的分布。

本研究以 GLM 及 SVM 預測課文分類的結果，並得到以下結論：

- 1.SVM 對國小國語科課文的可讀性分類準確率為 47.925%、正確率為 80.3109%；GLM 對國小國語科課文的可讀性分類準確率為 38.8252%、正確率為 75.7499%。從此結果發現 SVM 對國小國語科課文的分類準確率與正確率比 GLM 來得高。
2. SVM 所預測的分類結果在低年級部分(國小一、二年級)的預測其準確率較高；而高年級部分(國小五、六年級)則較多課文被預測錯誤。原因在於高年級的課文在乎的並不是表面特徵的難易(字數、詞數、句數、段落數等的多寡)，而是課文內容本身想要帶給學生的啟發與省思。文本的表面特徵固然重要，但是若要深入了解文本所帶給讀者的啟發及意義，除了表面特徵，文本在語意的分析也是很重要的。有學者提到文本理解的過程除了文本淺層語言特徵的接收，文本的深層語意在閱讀理解的過程中也扮演著重要的角色 (Graesser et al.,2004; McNamara et al., 2010)。
- 3.GLM 在低年級(1 下、2 上、2 下、3 上)課文的正確率都達 100%，而越往高年級的

準確率與正確率越低，故本研究推測一般線性模式的可讀性公式比較適合對低年級的文本做分類。

4. 本研究所訓練的資料是取自經國家編審單位審定的三個民間版本教科書(H版、K版、N版)，國小一年級至六年級國語科課文刪減掉新詩、絕句、古文、律詩的課文，但是可能因課文本編排的標準就不一致或不嚴謹，所以導致訓練出來的預測模型不好。

希望在這個與科技整合的時代，透過本研究所建置的中文可讀性指標分析系統能夠讓老師們能快速的搜尋到適合教學上的教材。讓學生依據自己的能力，閱讀適合自己的教材，節省老師及學生的寶貴時間並增進學生的學習效果。

## 第二節 未來發展

目前本研究所使用的文本特徵皆屬於文本的表面特徵，希望未來能夠結合將文本做語意分析的工具，例如潛在語意分析(Latent Semantic Analysis, LSA)。Landauer, Foltz, & Laham(1998)提出 LSA 擁有的特色，使其適合成為分析文本內涵知識的工具：(1)擷取字句間的意義與人類看法類似、(2)從文章中提取的知識如同人類之理解。Graesser et al.,2004; McNamara et al., 2010 也提到文本理解的過程除了文本淺層語言特徵的接收，文本的深層語意在閱讀理解的過程中也扮演著重要的角。張國恩、宋曜廷（2005）曾利用潛在語意分析技術建立一個可以自動評量小六學生閱讀摘要寫作系統，發現以向量餘弦值為計算標準的評分結果，與老師評量的分數，在相關性上都有達到顯著水準。所以，透過 LSA 我們可以建構出一套知識學科的語意空間，進而推測與該文本與哪個年級的語意空間最為接近，幫助學生選擇適合自己程度且有興趣的文本來閱讀。這樣一來除了從文本的表面特徵去檢視一篇文本的可讀性外，亦可輔以文本內容的含意去做可讀性的分類。

其次，目前研究所採用的實驗資料為去除掉新詩、絕句、古文、律詩的課文。因為這些類型的文本在字數、句數、代名詞數等特徵與一般文章都有不一樣的判斷方式，故希望未來在文體的部分也能找出適合的文本特徵來對不同文體的文本做可

讀性分類。並能繼續研究及發展新的文本可讀性指標且對文本的特徵做特徵的篩選，選擇最適合中文文本可讀性的特徵組合，提升可讀性文本分類上的準確性。

最後，希望未來能夠發展出完善教學領域上的中文可讀性指標分析系統，讓教師們能夠節省尋找適合教學教材的時間；讓學生可以依自己的能力，閱讀適合自己能力的文本。透過這樣的方法，學生便可以有效率的從大量的資訊中找到符合自己閱讀能力的閱讀資料，增進對文本的理解，進而達到學習的效果，對於老師的教學也更有幫助。

## 參考文獻

### 一、中文部分

楊孝滌(1978)。影響行文可讀性語言因素的分析。報學，7，58-67。

許菱祥(1986)。中文文法。大中國圖書公司。

CKIP 詞庫小組(1993)。中文詞類分析(三版)技術報告。中央研究院資訊科學研究所技術報告。

荊溪昱(1995)。中文國文教材的適讀性研究：適讀年級的推估。教育研究資訊，3(3)，113-127。

宋佩貞(1998)。台灣審定版國小英語教科書適讀性公式建置與評估。國立台東大學教育學研究所教學科技碩士班碩士論文。

林宗勳，Support Vector Machines 簡介，台灣大學通訊與多媒體實驗室，民 95。

陳稼興、謝佳倫、許芳誠(2000)。以遺傳演算法為基礎的中文斷詞研究。電子商務學報，2(2)，27-44。

陳順宇(2000)。迴歸分析(三版)，華泰書局。

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(編號：NSC93-2520-S-003-011)。台北：行政院國家科學委員會。

張晏晟(2008)。擴展反應型論述反應之自動化評估方法-以教師教學能力為例。國立臺灣師範大學資訊教育研究所碩士論文。

陳茹玲、蘇宜芬(2010)。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。國立臺灣師範大學教育心理與輔導學系教育心理學報，41(3)，579-604。

## 二、西文部分

Boser, B, Guyon, I, & Vapnik, V. (1992). A Training Algorithm for Optimal Margin Classifier. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Chall, J.S., & Dale, E.9 (2000). *Readability revisited: The new Dale-Chall readability formula*. MA: Brookline Books.

Chang, C-H., & Lin, C. J. (2001). *LIBSVM: a library for support vector machines*.

Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

*Coh-Metrix*, <http://cohmetrix.memphis.edu/cohmetrixpr/index.html>.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20.

Crossley, S.A., Allen D.B., & McNamara D.S.(2011). Text readability and inruirice simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1),84-101.

Dubay, W.H. (2004). *The Principles of Readability*. Costa Mesa, CA: BookSurge Publishing.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. *Proceedings of The 23rd International Conference on Computational Linguistics*, 276-284.

Fry, E.B., J. E. Kress, and D. L. Fountoukidis. 1993. *The reading teacher's book of lists: Third edition*. West Nyack, NY: The Center for Applied Research



in Education.

Graesser, A. C., McNamara, D. D., Louwerse, M. L., & Cai, Z. (2004). Coh-Metrix:

Analysis of text on cohesion and language. *Behavior Research Methods,*

*Instruments, & Computers*, 36, 193-202.

Hsu , C.W., & Lin,C.J. (2003). A Comparison of Methods for Support Vector Machine.

*IEEE Transactions of Neural Networks*, 13(2).

Hwang, Shin Ja J. (1992). The Functions of Negation in Narration. In Shin Ja J.

Hwang and William R. Merrifield (eds.), *Language in Context: Essay for Robert E.*

Longacre.321-337. Summer Institute of Linguistics and the University of Texas at

Arlington Publications in Linguistics.

Joachims, T.(1998). Text Categorization with Support Vector Machines: Learning with

Many Relevant Features. *Proceedings of The 10th European Conference on*

*Machine Learning*, 137-142.

Jordan, M. P., The power of negation in English: Text, context and relevance. *Journal of*

*Pragmatics*, 29, 705-752.

Klare, G.R. (1963). The Measurement of Readability: Useful Information for

Communications. *Journal of Computer Documentation*, 24(3).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic

analysis. *Discourse Processes*, 25, 259–284.

Larsson, P.( 2006). *Classification into readability levels*. Master's thesis, Department of

Linguistics and Philology, University Uppsala, Uppsala, Sweden.

Li, G.C., Liu, K.Y., & Zhang, Y. K. (1998). Identifying Chinese Word and Processing

- Different Meaning Structures. *Journal of Chinese Information Processing*, 2, 45-53.
- Liang, N. Y. (1990). Knowledge of Chinese Word Segmentation. *Journal of Chinese Information Processing*, 4, 42-49.
- Lin, S.Y., Su, C.C., Lai, Y.D., Yang, L.C., & Hsieh S.K. (2009). Assessing Text Readability using hierarchical lexical relations retrieved from WordNet. *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 45-84.
- McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47, 292-330.
- Parrado-Hernandez, E., & Hardoon, D. (2008). Text classification with a Primal SVM endowed with domain knowledge. Unpublished. Retrieved from <http://eprints.pascal-network.org/archive/00004968/>
- Petersen, S. & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer, Speech and Language*, 23(1), 106.
- Singh, S.R., Murthy, H.A., & Gonsalves, T.A. (2010). Feature Selection for Text Classification Based on Gini Coefficient of Inequality. *Proceedings of The Fourth Workshop on Feature Selection in Data Mining*, 76-85.
- Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36(2), 203-227.
- Teahan, W. J., McNab, R., Wen Y., and Witten, I. H. (2000). A compression-based

- algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3), 375-393.
- Wu, D. (1998). A position statement on Chinese segmentation. *Chinese Language Processing Workshop*.
- Zhan, J., & Loh, H.T.(2009). Using Redundancy Reduction in Summarization to Improve Text Classification by SVMs. *Journals of Information Science and Engineering*. 25, 591-601.
- Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21(8), 879-886.

## 附錄一 Coh-Metrix 2.0 可讀性指標

No.	Index	Measure	Full description
1	Title	Title	Title
2	Genre	Genre	Genre
3	Source	Source	Source
4	JobCode	JobCode	用來找出之前分析過的文章
5	LSASpace	LSASpace	Latent Semantic Analysis : College Level, Science, Narrative, Encyclopedia, 和 Physics。決定可以用來比較 LSA 的 world knowledge 或 conceptual domain。
6	Date	Date	Date
7	Adjacent anaphor reference	CREFP1u	計算相鄰的兩句話中 anaphor 的數量。
8	Anaphor reference	CREFPau	計算相鄰的每五句話中 anaphor 的比例。
9	Adjacent argument overlap	CREFA1u	計算相鄰兩句話有相同 argument 的比例，{noun, pronoun, NP}
10	Argument overlap	CREFAau	一段文章中，任何兩句話中相同 argument 的比例。
11	Adjacent stem overlap	CREFS1u	計算相鄰兩句話有相同 stem 的比例
12	Stem overlap	CREFSau	一段文章中，任何兩句話中的兩個名詞有相同 stem 的比例。
13	Content word overlap	CREFC1u	計算相鄰兩句話中實詞(content word)重複的比例。
14	LSA sentence adjacent	LSAassa	<p>計算每句話與下一句話概念上的相似程度。Mean LSA cosines。</p> <p>例：</p> <p>(1) The field was full of lush, green grass. The horses grazed peacefully. The young children played with kites. The women occasionally looked up, but only occasionally. A warm summer breeze blew and everyone, for once, was almost happy.</p> <p>(2) The field was full of lush, green grass. An</p>

			elephant is a large animal. No-one appreciates being lied to. What are we going to have for dinner tonight?
15	LSA sentence all	LSApssa	計算每句話與文章中另一句話概念上的相似程度。 Mean LSA cosine。
16	LSA paragraph	LSAppa	計算同篇文章每個段落彼此間概念上的相似程度。
17	Personal pronouns	DENPRPi	特別指人稱代名詞。高密度的代名詞會造成指涉上連結困難（人和代名詞無法連結）。 人稱代名詞/整句字數*1000
18	Pronoun ratio	DENSPR2	使用密集過多的代名詞來代替 NP 時容易造成閱讀上的困擾。代名詞/名詞片語 =pronoun ratio
19	Type-token ratio	TYPTOKc	計算字數在文章中出現頻率。如果數值是 1，則表示文章中每個字只出現一次，也代表為文章相對較難理解。本研究中只計算實詞。每個字/每個字出現的次數 =Type/Token=TTR(Type-Token Ratio)
20	Causal content	CAUSVP	計算文章中 causal verb 和 causal particles 的次數。反應出一個句子的 casual cohesion relation 程度。Causal cohesion relation 指的是文章中的事件（event）和動作（action）是不是 causally related。測量方式是基於 WordNet 中的分類(causality)。文章中 causal verb 越長出現，可以推論出文章有較高的 causal content。但是 causal cohesion 的測量必須要有足夠的文章長度，否則結果較不穩定。  causal particle: as a result, by, due to, enable, for, hence。
21	Causal cohesion	CAUSC	計算 causal verb 和 particles 的比例。
22	Intentional content	INTEi	計算文章中 intentional actions, events, 和 particles 的數量。次數越高，越表示文章是 goal-driven 的。
23	Intentional cohesion	INTEC	計算 intentional particles 和 action/event 的比例。

24	Syntactic structure similarity adjacent	STRUTa	<p>比較相鄰兩句話 tree structure 中 intersection nodes 的比例。</p> <pre> graph TD     S[S] --- NP1[NP]     S --- VP[VP]     NP1 --- N1[N]     N1 --- I[I]     VP --- V[V]     V --- love[love]     VP --- NP2[NP]     NP2 --- N2[N]     N2 --- You[You] </pre>
25	Syntactic structure similarity all-1	STRUTt	比較每個段落中每句話的 tree structure 中 intersection nodes 的比例。
26	Syntactic structure similarity all 2	STRUTp	比較同一個段落中每句話的 tree structure 中 intersection nodes 的比例。
27	Temporal cohesion	TEMPta	時間介係詞/(時間介係詞+時間名詞)
28	Spatial cohesion	SPATC	位置介係詞/(位置介係詞+地方名詞)
29	All connectives	CONi	<p>計算所有 connective 的次數。</p> <p>{positive/negative}</p> <p>{additives/causal/logical/temporal}</p> <p>Additives (AD): also, moreover, however, but</p> <p>Causal (CA): because, so, consequently, although, nevertheless</p> <p>Logical (LG): or, actually, if</p> <p>Temporal (TP): after, before, when, until</p>
30	Conditional operators	DENCONDi	計算條件句的次數。
31	Pos. additive connectives	CONADpi	positive additive connectives 的次數
32	Pos. temporal connectives	CONTPpi	positive temporal connectives 的次數
33	Pos. causal connectives	CONCSp	positive causal connectives 的次數
34	Pos. logical	CONLGpi	positive logical connectives 的次數

	connectives		
35	Neg. additive connectives	CONADni	negative additive connectives 的次數
36	Neg. temporal connectives	CONTPni	Incidence of negative temporal connectives
37	Neg. causal connectives	CONCSni	negative causal connectives 的次數
38	Neg. logical connectives	CONLGni	negative logical connectives 的次數
39	Logic operators	DENLOGi	使用越多的 logical operator 會對大部分的讀者造成閱讀困難，包括布林邏輯。
40	Raw freq. content words	FRQCRacw	文章中實詞的 mean row frequency。
41	Log freq. content words	FRQCLacw	文章中實詞的 log frequency。
42	Min. raw freq. content words	FRQCRmcs	每個句子實詞的最低 mean row frequency。
43	Log min. freq. content words	FRQCLmcs	每個句子實詞的最低 log frequency。
44	Concreteness content words	WORDCacw	文章中所以符合 MRC database 的實詞的平均 concreteness 值。 MRC 測量每個字的特性。提供最多 26 個語言特性。
45	Min. concreteness content words	WORDCmcs	每個句子中實詞的最低 concreteness 值。
46	Noun hypernym	HYNOUNaw	平均名詞上位詞數量。皆和 ambiguity 有關，所以一個字有越多種意思用容易造成理解上的困難。一個字有越多層的 conceptual taxonomic level 則語意越實際。
47	Verb hypernym	HYVERBaw	平均動詞上位詞數量
48	Negations	DENNEGi	否定詞的數量
49	NP incidence	DENSNP	NP 數/總字數*1000
50	Modifiers per NP	SYNNP	平均每個 NP 的 modifier 的數量
51	Higher level constituents	SYNHw	平均每字有的 high level constituent。句子結構上密度越高的有越多 constituent。
52	Words before main verb	SYNLE	平均主要子句動詞前的字數
53	No. of words	READNW	字數

54	No. of sentences	READNS	句數
55	No. of paragraphs	READNP	段落數
56	Syllables per word	READASW	平均每字的音節數
57	Words per sentence	READASL	平均每句的字數
58	Sentences per paragraph	READAPL	平均每段落的句數
59	Flesch Reading Ease	READFRE	<p>結果從 0-100，越高的數值表示越容易讀，平均值在 6-70。</p> $\text{READFRE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})$ <p>ASL = 平均句長 = 字數 / 句數 = READASL  ASW = 每個字的平均音節數目 = 音節數 / 字數 = EADASW</p>
60	Flesch-Kincaid	READFKGL	<p>將 reading ease score 轉成美國 grade-school 的程度。數值越高，表示文章越容易閱讀，範圍從 0 到 12。</p> $\text{READFKGL} = (.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59$ <p>字數必須超過 200 字才能讓上述兩種 formulas 有效成功的運作。</p>



## 附錄二 中研院平衡語料庫詞類標記集

簡化標記	對應的 CKIP 詞類標記	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	Daa	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	Neu	/*數詞定詞*/
Nes	Nes	/*特指定詞*/
Nep	Nep	/*指代定詞*/
Neqa	Neqa	/*數量定詞*/
Neqb	Neqb	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/

VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE		/*的, 之, 得, 地*/
SHI		/*是*/
FW		/*外文標記*/

### 附錄三 國小國語科刪減的文章

檔名	課文	刪減原因
H-098-C-012-01	春雨	新詩
H-098-C-012-03	小草	新詩
H-098-C-012-07	大樹喜歡交朋友	新詩
H-098-C-012-08	家	新詩
H-098-C-021-07	大自然的語言	新詩
H-098-C-021-12	走走聽聽	新詩
H-098-C-022-01	如果可以	新詩
H-098-C-022-05	爬山	新詩
H-098-C-031-01	如果我當了爸爸	新詩
H-098-C-031-08	樹林裡	新詩
H-098-C-031-12	年獸來了	劇本
H-098-C-031-14	快樂過新年	新詩
H-098-C-032-04	鹿港風光	新詩
H-098-C-041-01	瀑布	詩歌
H-098-C-041-10	常常想起的朋友	詩歌
H-098-C-042-01	黑面琵鷺之歌	詩歌
H-098-C-042-13	收藏秋天	詩歌
H-098-C-051-101	詩兩首(觀游魚)	律詩
H-098-C-051-102	詩兩首(贈劉景文)	律詩
H-098-C-051-14	愛的分享	新詩
H-098-C-052-01	玉山之美	新詩
H-098-C-052-121	詩人的心情(一)望廬山瀑布	古詩

H-098-C-052-122	詩人的心情(二)約客	古詩
H-098-C-061-01	飛翔	詩歌
H-098-C-061-081	古詩文選讀（王戎辨苦李）	古詩
H-098-C-061-082	古詩文選讀（九月九日憶山東兄弟）	古詩
H-098-C-062-01	雨，落在高雄的港上	新詩
H-098-C-062-03	狐假虎威	古文
K-098-C-012-01	小樹	新詩
K-098-C-012-02	我要長大	新詩
K-098-C-012-05	花開的聲音	新詩
K-098-C-012-06	小湖邊	新詩
K-098-C-012-08	朋友	新詩
K-098-C-012-09	和你在一起	新詩
K-098-C-012-11	扮家家	新詩
K-098-C-012-12	夏天的海	新詩
K-098-C-021-01	開學日	新詩
K-098-C-021-13	做湯圓	新詩
K-098-C-022-01	橋	新詩
K-098-C-022-05	走過小巷	新詩
K-098-C-022-08	我是行道樹	新詩
K-098-C-031-01	爸爸的相簿	新詩
K-098-C-031-04	淡水小鎮	新詩
K-098-C-032-08	心情日記	新詩
K-098-C-041-01	大地巨人	新詩
K-098-C-041-111	詩二首(一)看月	詩歌
K-098-C-041-112	詩二首(二)賦新月	詩歌

K-098-C-042-01	在春天許願	新詩
K-098-C-051-111	詩二首(一) 竹里館	詩歌
K-098-C-051-112	詩二首(二) 獨坐敬亭山	詩歌
K-098-C-052-011	努力請從今日始(明日歌)	詩歌
K-098-C-052-012	努力請從今日始(今日歌)	詩歌
K-098-C-062-01	過故人莊	新詩
N-098-C-012-02	春天來了	新詩
N-098-C-012-05	來玩球	新詩
N-098-C-012-06	種豆子	新詩
N-098-C-012-08	紅綠燈	新詩
N-098-C-012-09	做卡片	新詩
N-098-C-021-01	秋天裡的舞會	新詩
N-098-C-022-01	小河	新詩
N-098-C-022-04	好朋友	新詩
N-098-C-031-01	我住的家	新詩
N-098-C-031-06	題老人飲驢圖	新詩
N-098-C-031-10	阿瑪迪斯	新詩
N-098-C-032-01	我們這一家	新詩
N-098-C-032-06	手影	新詩
N-098-C-032-10	地球之歌	新詩
N-098-C-041-08	遊子吟	古詩
N-098-C-042-02	雨，落在高雄的港上	新詩
N-098-C-042-05	野薑花	新詩
N-098-C-042-07	閱讀列車(一) 紅豆	古詩
N-098-C-051-01	春天的陽明山	新詩

N-098-C-051-021	詩中有畫(一) 路柴	古詩
N-098-C-051-022	詩中有畫(二) 宿建德江	古詩
N-098-C-051-08	堅持到底	劇本
N-098-C-052-01	絕句選	絕句
N-098-C-052-07	我扶起了一棵小樹	新詩
N-098-C-061-01	漁歌子	古詩
N-098-C-061-02	鵲蚌相爭	古文
N-098-C-061-11	夢想	新詩
N-098-C-062-01	詩選 草	古詩
N-098-C-062-02	媽媽的鏡子	新詩
N-098-C-062-05	讀書筆記	古文
N-098-C-062-11	記憶拼圖	新詩