The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 71-86

可讀性預測於中小學國語文教科書及優良課外讀物之研究

A Study of Readability Prediction on Elementary and Secondary Chinese Textbooks and Excellent Extracurricular Reading Materials

劉憶年　Yi-Nian Liu
國立臺灣師範大學資訊工程學系
60247056s@ntnu.edu.tw

陳冠宇　Kuan-Yu Chen
中央研究院資訊科學研究所
kychen@iis.sinica.edu.tw

曾厚強　Ho-Chiang Tseng
國立臺灣師範大學資訊工程學系
ouartz99@gmail.com

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
berlin@ntnu.edu.tw

摘要

可讀性（Readability）是指閱讀材料能夠被讀者理解的程度。可讀性高的文章較容易被讀者理解。文章的可讀性與很多因素有關，如：文長、字詞難度、句法結構、內容是否符合讀者的先備知識等，然而表淺的語言特徵無法反映這些複雜的成分。本論文以先前的研究為基礎，更深入的探討不同種類的特徵，包括句法分析（Syntactic Analysis）、詞性標記（Part-of-Speech, POS）、詞表示法（Word Embedding）、語意資訊（Semantic Information）與寫作程度（Well-written）等特徵，分析比對不同類型的特徵與可讀性高低的關聯性。實驗資料分為二部分：其一為中小學國語文教科書，選自 98 年度台灣三大出版社所出版的 1~9 年級（共 18 冊）審定版國中小國語文教科書；其二為優良課外讀物，選自文化部歷屆「中小學生優良課外讀物」獲選書籍。本論文嘗試透過逐步迴歸與支持向量機等兩種方式建立可讀性模型，比較兩者之效能優劣；最後，再將兩者加以結合，以提升預測之正確率。實驗結果顯示，本論文所提出的可讀性特徵相較於傳統所使用的表淺特徵，在文本難易度評估的任務中，能有顯著的效能提升。

關鍵詞：可讀性、文本特徵、逐步迴歸、支持向量機

Abstract

Readability is basically concerned with readers' comprehension of given textual materials: the higher the readability of a document, the easier the document can be understood. It may be affected by various factors, such as document length, word difficulty, sentence structure and whether the content of a document meets the prior knowledge of a reader or not. However, simple surface linguistic features cannot always account for these factors in an appropriate manner. To cater for this, we explore in this study a variety of extra features, including syntactic analysis, parts of speech, word embedding, semantic role features and well-written features. The experimental datasets are composed of two parts: one is textbooks of the Chinese language for elementary and junior high schools (K1 to K9) in Taiwan, compiled from three publishers in the academic year of 2009; the other is excellent extracurricular reading materials for students of elementary and junior high schools, collected by the Ministry of Culture in Taiwan. Two readability prediction models, viz. stepwise regression and support vector machine, are evaluated and compared, while the combination of these two models is also investigated so as to further enhance the accuracy of readability prediction. Experimental results reveal that our proposed approach can yield consistently better performance than traditional ones merely with simple surface linguistic features in evaluating text difficulty.

Keywords: Readability, Textual Features, Stepwise Regression, Support Vector Machine

一、 緒論

可讀性（readability）是指閱讀材料能夠被讀者理解的程度[1]。可讀性高的文章較容易被讀者理解。文章的可讀性與很多因素有關，如：文長、字詞難度、句法結構、內容是否符合讀者的先備知識等，然而表淺的語言特徵並無法完全反映這些複雜的成分。英文文本的可讀性研究行之有年，或以詞彙頻率列表，評量文章難度、或將詞表作為參照，建置可讀性公式、或發展線上多文本特徵分析器[2]，計算影響文章難易度的各類型指標，並提供數值化的結果；中文的可讀性研究則屈指可數，或選用表淺的語言特徵建置可讀性公式[3, 4]，或將可讀性指標等當成預測變項，以教科書的年級值當成效標，透過逐步迴歸（Stepwise Regression）建置公式、或結合特徵選取方法與支援向量機（Support Vector Machine, SVM）建立預測模型預測文本等級[1]。可讀性研究除了傳統的語言特徵，心理學上的因素亦是值得考量之因素[5]。可讀性較高的文章除了能讓讀者較容易理解外，亦應有較高的趣味性，增強閱讀印象，加快閱讀速度，令讀者有意願持續閱讀，進而達成如輔助教學、文本推薦等特定目標。文本可讀性預測可依據讀者提供合適的文本閱讀，以提高其理解程度，進而培養從小閱讀的習慣。而可讀性預測的特徵仍有許多探討空間，結合不同模型以提高預測正確性亦為一研究面向。現今資訊來源多元，非傳統文字文件，如圖片、音訊、影片等，皆可成為接收新知的管道，故其可讀性預測亦是未來研究趨勢。然而因多媒體文本所包含的內容形式與純文字文本之特性差異甚大，如何結合既有概念以探討新興領域之可讀性，所面臨之挑戰將更加艱困。

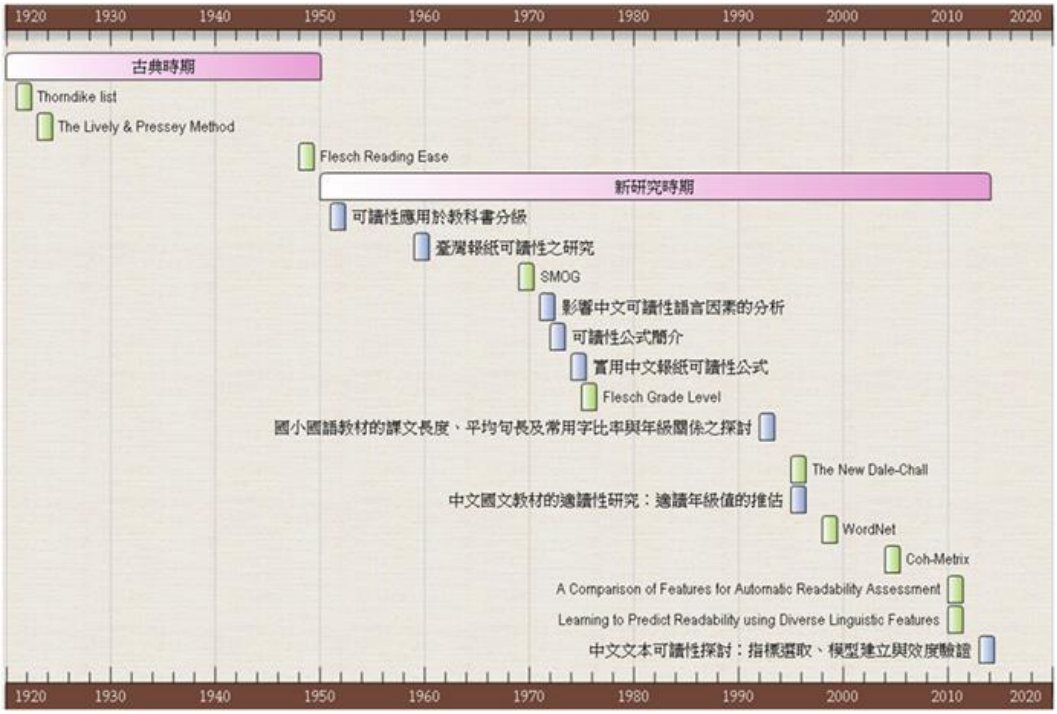
由於中英文字在語言特徵上的差異極大，過去西方研究者在可讀性研究所採用的特徵，是否適合中文可讀性評估有待商榷[1]。有鑑於可讀性研究的重要性，以及可能發展的多元應用，本論文提出使用句法分析（Syntactic Analysis）、詞性標記（Part-of-Speech, POS）、詞表示法（Word Embedding）、語意資訊（Semantic Information）與寫作程度（Well-written）等特徵，分析不同類型的特徵所代表之意義，比對各類特徵與可讀性高低的關聯性，並將特徵彼此結合以提升可讀性預測之正確性。藉由這些特徵，本論文透過逐步迴歸與支持向量機等兩種方式建立可讀性模型，比較兩者用於測試國中小教科書及優良課外讀物之效能優劣，並期望找出可讀性分類之重要因素。

本論文的後續安排如下：第二節說明可讀性的基本概念、回顧可讀性的歷史與公式、分析可讀性的模型、探討可讀性的發展趨勢、介紹可讀性的應用層面。第三節除解釋先前研究的特徵外，亦分別論述本論文所使用的各類特徵。第四節為實驗資料與實驗結果的呈現。第五節為全文總結與未來研究展望。

二、 文獻探討

(一)、可讀性基本概念介紹

可讀性是指閱讀材料能夠被讀者理解的程度（Dale & Chall, 1949; Klare, 1963, 2000; McLaughlin, 1969）。Klare（1984）認為可讀性的定義為：易識別性



圖一、可讀性研究發展歷史

（Legibility）、易閱讀性（Ease of Reading）、易理解性（Ease of Understanding）等任何一種關於材料的特徵。可讀性的概念中，易理解性是在閱讀領域中比較通用的用法[1]。

語言專家藉由不斷修正而得出的「可讀性公式」來計算可讀性的分數，並將這些公式廣泛應用於對文本與讀者群體的閱讀水準加以匹配，然而可讀性公式無法準確反映文本難度，只是給出一個「不錯的粗略估計」[6]。

(二)、可讀性之歷史與公式

西方可讀性研究行之有年，早於 1950 年代時可讀性公式已百家爭鳴，近年來更嘗試探討與文本更相關的凝聚性指標，及各指標間的關係；中文的可讀性研究相對而言則屈指可數，早期僅運用表淺指標，發展一系列中文適讀性公式，近期則有將小學教科書進行可讀性分類之探討[1]。可讀性研究概略發展歷史可參照圖一。西方可讀性研究以發展測量公式為大宗，然而侷限於技術僅納入文本的表淺語言特徵。第一個可讀性公式 The Lively & Pressey Method 利用詞表當成參照，篩選出不同等級難度的詞彙當成文章難度指標，對後來的可讀性研究有重大的影響。另外也有不少的可讀性公式將詞長與句長當成難度指標，納入可讀性公式之計算。由表一可以看出可讀性公式著重於利用如詞彙與句長等淺顯的語言特徵作為指標，有學者因此認為以這些語言特徵預測文本可讀性，並沒有強而有力的證據。

公式名稱	計算公式	採用指標
Flesch Reading Ease （Flesch, 1948）	Reading ease = 206.876 - (1.015 × 平均句長) - (84.6 × 平均音節數)	句長、音節數
New Reading Ease （Flesch, 1951）	Reading ease = 1.599 × 每百詞之單音節詞比率 - 1.015 × 每句平均詞數 - 31.517	單音節數、詞數
Gunning FOG （Gunning, 1952）	Grade level = 0.4 × (平均句長 + 100 × $\frac{\text{難詞}}{\text{總詞數}}$)	句長、難詞比率
Spache （Spache, 1953）	Grade level = 0.839 + (0.086 × 難詞百分比) + (0.141 × 平均句長)	句長、難詞比率
Powers-Summer-Kearl （Power et al., 1958）	Grade Level = -2.2029 + 0.0778 × 平均句長 + 0.455 × 音節數 Reading Age = -2.7971 + 0.0778 × 平均句長 + 0.455 × 音節數	句長、音節數
Fry Graph （Fry, 1968）	計算 3 篇 100 詞文章的平均句數與音節數；將數值在 Fry Graph 中做記號找出閱讀年級	句數、音節數
SMOG （McLaughlin, 1969）	SMOG Grade = 1.0430 × $\sqrt{\text{三音節以上的詞數} \times (\frac{30}{2}) + 3.1291}$ + 3.1291)	多音節詞數、句數

FORCAST （Caylor et al., 1973）	Grade Level = 20 - ($\frac{\text{單音節的詞數}}{10}$) Reading Age = 25 - ($\frac{\text{單音節的詞數}}{10}$) years → 150 詞 Reading Age = 25 - ($\frac{\text{單音節的詞數}}{6.67}$) years → 100 詞	音節數
Flesch Grade Level （Kincaid et al., 1975）	Grade Level = -15.59 + (0.39 × 平均句長) + (11.8 × 平均音節數)	句長、音節數
The New Dale-Chall （Chall and Dale, 1995）	Grade Level = (0.1579 × $\frac{\text{難詞}}{\text{總詞數}}$) + (0.0496 × 平均句長) + 3.6365	難詞比率、句長

表一、西方常見的可讀性公式與採用指標

中文可讀性研究以迴歸分析法發展可讀性公式，將可讀性指標逐一刪去，最後只留下少數影響最大的指標。另外，亦有研究使用支援向量機建置之模型來預估文章適合閱讀的年級（宋曜廷等人，2013）。由表二則可看出研究者多採用較為表淺之指標建立公式。因此，傳統中文可讀性研究，在指標的選取上與拼音文字系統常見的指標並無顯著差異。

公式名稱	計算式	採用指標
Yang（1970）	年級 = 0.1788 × 筆劃數超過 10 劃百分比 + 0.1432 × 平均句長 + 0.6375 × 難字百分比	筆劃、難字比率、句長
	學期 = 14.95961 + 39.07746 × 詞彙數 - 2.48491 × 平均筆劃數 + 1.11506 × 句數	詞彙數、句數、筆劃數
陳世敏（1970）	年級 = （每句平均字數 + 難字數） × 0.7	句長、難字數
荊溪昱（1992）	年級 = 5.43035627 + 0.00657347 × 課文長度 + 0.02443016 × 平均句長 - 5.56746245 × 常用字比率 + 1.38315091 × 詩歌體 - 1.07299966 × 對白文體	課文長度、句長、常用字比率、文體
荊溪昱（1995）	年級 = 8.76105604 + 0.00272438 × 課文長度 + 0.07866782 × 平均句長 - 8.9311010 × 常用字比率 + 0.42920182 × 詩歌體 + 3.23677141 × 文言文體	課文長度、句長、常用字比率、文體
宋曜廷等人（2013）	年級 = 4.53 + 0.01 × 難詞數 － 0.86 × 單句數比率 － 1.45 × 實詞頻對數平均 + 0.02 × 人稱代名詞數	難詞數、單句數比率、實詞頻對數平均、人稱代名詞數

表二、中文常見的可讀性公式與採用指標

(三)、可讀性模型分析比較

傳統可讀性公式多為線性迴歸模型，納入不同的特徵為自變項，估算文章難度，或提供公式估算文本適合閱讀的年級。迴歸分析（Regression Analysis）是一種統計學上分析數據的方法，目的在於了解兩個或多個變數間是否相關，並建立數學模型以便觀察特定變數來預測研究者感興趣的變數[7]。更明確地，迴歸分析是利用依變數 Y 與自變數 X 之間的關係所建立的模型，期望找出一條最能夠代表所有觀測資料的函數（迴歸估計式）[7]。而多元迴歸即為探討一個依變數和多個自變數間的關係，如：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$，其中 β_0 為常數，$\beta_1, \ldots, \beta_n$ 為迴歸係數[8]。

近年來，許多研究開始將可讀性議題視為一種機械學習的問題。藉由抽取自文本的各類可讀性特徵，透過支援向量機建立預測模型後，就可用於預測測試資料集文件之可讀性。支援向量機將原始資料轉換到更高的維度，利用在訓練資料集中所謂的小樣本資料（Support Vectors）找到超平面，用以分類資料[1]。支援向量機主要是在尋找具有最大邊界的超平面，因為其具有較高的分類準確性[9]。目前支援向量機相關研究常使用由台大林智仁教授所開發的 LIBSVM[10]開放原始碼軟體為工具，經由準備資料集、訓練模型、預測新資料所屬之類別等步驟，得到測試之準確率。

(四)、可讀性近來研究趨勢

隨著技術的進步，納入更多複雜的可讀性指標變得可行。Graesser 等人為了改良傳統教科書的寫作方式，並提供符合學生閱讀能力的教材，發展了線上多文本特徵分析器（Coh-Metrix）[2, 11]，可抽取多項文本特徵。

「中文文本自動化分析系統」[12]為 Coh-Metrix 之中文版本，由國立臺中教育大學教育測驗統計研究所與特殊教育學系合作，參考 Coh-Metrix 分析建置的指標應用於中文領域，結合中文詞彙與文章之特性，發展中文文本自動化分析指標，以幫助使用者分析文章的特性作為讀本選擇之參考。

許多研究亦嘗試根據認知理論來分析文本的難度，積極探討與文本更相關的進階指標，並發展新的方式自動化地處理文本，像 WordNet（Fellbaum, 1998）[13]，即分析詞、句子、段落及篇章等較大範圍的文本多層次之凝聚特性與文章難度的關係[1]。相較於 WordNet，中文亦有類似的詞庫。中文詞彙網路（Chinese Wordnet）計畫（黃居仁、謝舒凱，2010）[14]，目的是在提供完整的中文詞義（Sense）區分與詞彙語意關係知識庫。

三、 特徵探討

(一)、基礎特徵

本研究以〈中文文本可讀性探討：指標選取、模型建立與效度驗證〉[1]中之指標為基礎，且經由宋曜廷等人發展的文本可讀性指標自動分析化系統

（Chinese Readability Index Explorer, CRIE）[15]擷取文章可讀性指標的數值。其所包括的指標請參閱表三。其中負向連接詞如「然」、「卻」、「否則」等。

上述特徵為參考中西方文獻回顧，所發展適合中文特性的可讀性指標。然而其所包含之深層類型指標仍較為稀少，故本研究以此為基礎，另外結合其他指標，以期達到考慮文本難易度更深層次因素之目的。

(二)、句法分析與詞性特徵

此節探討由 Feng 等人[16]所提出的句法分析（Syntactic Analysis）特徵及詞性標記（Part-of-Speech, POS）特徵。其所包括的指標請參閱表四。

語法（Gram-mar）是語言單位的結構規則；也可以說：語法是詞、詞組、子句、句子的結構和運用法則[17]。語法特性只有分析句子含意時才得以揭露，因此句法分析就顯得相當重要。

詞性是以個別詞彙為對象，根據其語法作用，兼顧其意義，所分類得到的結果[18]。由於中文語法特性的緣故，同一詞彙可能有不同詞性，如「縱橫交錯」與「稍縱即逝」中的「縱」字因詞性不同，其意義也不同，故此種情況容易造成理解上的困難。

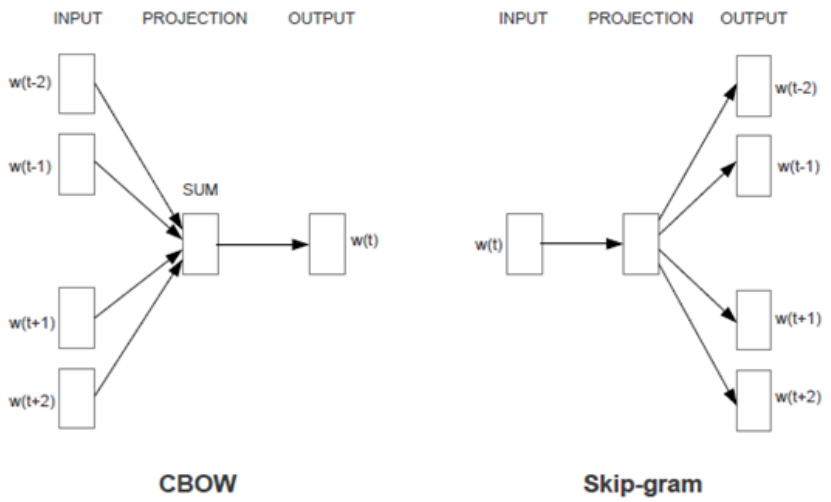
類別	指標編號與指標名稱	定義
詞彙類指標		
詞彙數量	1. 字數	加總文章中的字數
	2. 詞數	計算文章中的詞數
詞彙豐富性	3. 相異詞數比率	相異詞數除以詞總數
	4. 實詞密度	實詞總數除以詞總數
詞彙頻率	5. 實詞頻對數平均	計算文章的實詞在整個資料集出現的頻率取對數後平均
	6. 難詞數	加總文章中不在常用詞表的詞數
詞彙長度	7. 低筆劃字元數	加總文章中筆劃數介於 1~10 筆劃的字元數
	8. 中筆劃字元數	加總文章中筆劃數介於 11~20 筆劃的字元數
	9. 高筆劃字元數	加總文章中筆劃數介於 21 筆劃以上的字元數
	10. 字元平均筆畫數	計算文章中的字元平均筆劃數
	11. 二字詞數	加總文章中的二字元詞
	12. 三字詞數	加總文章中的三字元詞

語意類指標	13. 實詞數	加總文章中的實詞數
	14. 否定詞	加總文章中的否定詞數
	15. 複雜語意類別句子數	加總文章中複雜語意句數
句法類指標	16. 句平均詞數	詞數除以句數
	17. 單句數比率	計算文章中的單句數比例
	18. 名詞片語修飾語數	計算文章中名詞片語的修飾語平均數
	19. 名詞片語比率	計算文章中每句中名詞片語數與詞數比之平均
文章凝聚性指標		
指稱詞	20. 代名詞數	加總文章中的代名詞
	21. 人稱代名詞數	加總文章中的人稱代名詞
連接詞	22. 連接詞數	加總文章中的連接詞
	23. 正向連接詞數	加總文章中的正向連接詞
	24. 負向連接詞數	加總文章中的負向連接詞

表三、本研究採用之基礎特徵名稱與定義

類別	指標編號與指標名稱
Parsed Syntactic Features	1. Number of the NPs
	2. Number of NPs per sentence
	3. Number of the VPs
	4. Number of VPs per sentence
	5. Number of non-terminal nodes per parse tree
POS-based Features	6. Fraction of tokens labeled as noun
	7. Fraction of tokens labeled as preposition
	8. Number of noun tokens per sentence
	9. Number of preposition tokens per sentence

表四、本研究採用之句法分析與詞性特徵名稱與定義



圖二、CBOW 與 Skip-gram 模型示意圖[21]

(三)、表示法特徵

要將自然語言的問題轉變成為機器學習的問題，首先便須把這些符號數學化。傳統的做法為把每個詞表示成一個很長的向量，向量的維度是全部詞的數目，其中除了該詞的維度值為 1，其餘皆為 0，這個向量就代表了當前的詞（One-hot Representation）[19]。

深度學習（Deep Learning）領域中則利用分散式表示法（Distributed Representation）的方式，將每一個詞以一個低維度的實數向量表示之，稱為詞表示法（Word Representation or Word Embedding）[19]。此表示法向量中各維度皆有值，因此讓兩個意思相近的詞在向量空間上的距離縮短。

Google 在 2013 年公開的 Word2Vec 工具[20]，即是用於求取詞向量表示法。常見的詞向量表示法模型有兩種：連續型詞袋模型（Continuous Bag-of-Words, CBOW）與跳躍式模型（Skip-gram）。連續型詞袋模型的訓練目標是給定一個詞的上下文，以預測這個詞出現的機率；在跳躍式模型中，訓練目標則是給定一個詞，預測其上下文中的詞。由於許多研究指出跳躍式模型的效果較佳，故本研究利用跳躍式模型訓練詞向量表示法及詞性向量表示法作為特徵。

(四)、語意資訊特徵

本研究參考〈句結構樹中的語意角色〉[22]中之語意角色為指標，並利用中研院之中文剖析系統將文章進行語意角色的擷取。其所包括的指標請參閱表五。

(五)、寫作程度特徵

此節探討由 Louis 等人[23]所提出的優良寫作概念（Great Writing），並將其應用於可讀性研究。其所包括的指標請參閱表六。其中 Visual nature of articles 類別是經由將 ESP Game Dataset 英文標記資料，隨機抽取五十個單字並轉譯成中文作為描述生動的詞彙。

類別	指標編號與指標名稱
修飾物體名詞	1. apposition
	2. possessor
	3. predicator
	4. property
	5. quantifier
修飾事件動詞	6. companion
	7. comparison
	8. goal
	9. topic
	10. addition
	11. alternative
	12. complement
	13. conclusion
	14. contrast
	15. reason

表五、本研究採用之語意資訊特徵名稱與定義

類別	指標編號與指標名稱
Visual nature of articles	ESP Game Dataset（指標 1-50）
Beautiful language	自行蒐集之優美詞彙及成語（指標 51-100）
Affective content	台灣地區華人情緒與相關心理生理資料庫—中文情緒詞常模研究[24]（指標 101-150）

表六、本研究採用之寫作程度特徵名稱與定義

四、 實驗設置與結果

(一)、實驗資料

中小學國語文教科書選自 98 年度台灣 H、K、N 三大出版社所出版的 1~9 年級（共 18 冊）審定版國中小國語文教科書。各版本在各年級的文章數詳見表七。優良課外讀物選自文化部歷屆「中小學生優良課外讀物」獲選書籍[25]，以書單中標示之適讀年齡為分類正確答案。各級別的文章數詳見表八。

(二)、實驗設定

以下兩節實驗各分為兩部份：第一部份實驗以逐步迴歸方式，以年級當成效標變項，24 個中文可讀性指標為預測變項，以 SPSS 22.0 軟體建立可讀性數學模型計算各篇課文可讀性分數，以預測其屬於哪個年級。第二部份實驗中運用支援向量機學習並預測資料類別。

(三)、國語文教科書實驗

$$年級 = 11.701 - 5.362 \times 領域實詞頻對數平均 + 0.176 \times 負向連接詞數 + 0.167 \times 句平均詞數 + 0.024 \times 代名詞數 \quad (1)$$

式(1)為國語文教科書之迴歸公式。在此先比較以不同區間劃定年級值之預測正確性，結果如表九所示。其中之 0.0 意指若逐步迴歸之分數為 0~1 間即定為 1；0.1 意指若逐步迴歸之分數為 0.1~1.1 間即定為 1，依此類推。由表九得知以 0.9~1.9 為區間劃分正確率最高，故以此測試文章所屬年級的正確性，預測結果如表十所示。由表十可看出以三至六年級之預測正確性較高，且各年級分類結果皆偏向較低年級，尤以七至九年級最為明顯，造成此結果的原因可能為所使用的特徵較為表淺，對國小高年級與國中文章差異不大，故無法有效分類較高年級之文本。

年級 / 出版社	一	二	三	四	五	六	七	八	九	總數
H	22	28	28	28	33	27	31	32	23	252
K	22	28	28	29	36	27	28	29	25	252
N	20	28	24	29	30	29	31	30	24	245
總數	64	84	80	86	99	83	90	91	72	749

表七、國語文教科書各年級文章數

低年級	中年級	高年級	國中	總數
20	20	20	20	80

表八、優良課外讀物各級別文章數

區間	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
正確性	21%	22%	23%	24%	26%	26%	26%	27%	28%	**32%**	31%

表九、逐步迴歸分數以不同區間劃分所預測之文章年級結果

預測年級 / 正確年級	一	二	三	四	五	六	七	八	九	正確性（%）
一	4	8	6	0	1	1	0	0	0	20.00%
二	2	10	9	6	1	0	0	0	0	35.71%
三	0	0	11	5	8	0	0	0	0	45.83%
四	0	0	3	14	9	0	2	1	0	**48.28%**
五	0	0	2	6	9	11	2	0	0	30.00%
六	0	0	2	4	9	11	2	1	0	37.93%

七	0	0	1	3	9	9	8	0	1	25.81%
八	0	0	1	2	5	9	10	3	0	10.00%
九	0	0	0	2	5	5	6	4	2	8.33%

表十、逐步迴歸預測文章年級結果

實驗	使用特徵	正確性（%）
1	基礎特徵	48.57%
2	句法分析與詞性特徵	42.04%
3	詞表示法 256 維	**53.88%**
4	詞表示法 512 維	50.20%
5	詞表示法 1024 維	53.47%
6	詞性表示法 256 維	34.69%
7	詞性表示法 512 維	31.02%
8	詞性表示法 1024 維	31.84%
9	語意資訊特徵	37.96%
10	寫作程度特徵	11.84%

表十一、支援向量機使用各特徵預測文章年級之結果

接著，我們探討使用支援向量機的預測效能，各組實驗設定與結果如表十一所示。上述各項特徵中以 256 維的詞向量表示法效果最佳，且使用詞向量表示法當作特徵測試時，結果皆優於基礎實驗（即實驗 1），原因為其將詞的上下文代表該詞，故當詞用法接近時，表示法也會相似，而年級層越接近時，某詞之用法應較為類似。寫作程度特徵之測試結果取決於其中所使用之各項指標，故須考慮更能區別年級層之詞彙。

最後，我們嘗試比較與結合支援向量機與逐步迴歸模型，其實驗設定與結果如表十二所示。實驗 1 採用與逐步迴歸模型相同特徵，相較之下，支援向量機模型的正確率提升 2%，可見其預測效能較好，然因特徵較少，正確性依然不高。但當使用的特徵數目增多，正確率大多時候也會相對提升，唯實驗 3 退步不少幅度，其原因可能為如名詞片語（NP）、動詞片語（VP）、名詞（N）、介係詞（Prep.）等指標在小學高年級與國中之文本中差異並不明顯。

(四)、優良課外讀物實驗

$$\textit{年級} = 1.871 + 0.052 \times \textit{負向連接詞數} \quad (2)$$

式(2)為優良課外讀物之迴歸公式。同樣地，在此先比較以不同區間劃定年級值之預測正確性。結果如表十三所示。由實驗結果可知，以 1.0~2.0 為區間劃分正確率最高，故以此測試文章所屬年級的正確性，預測結果如表十四所示。由表十四可看出以中年級之預測正確性最高，其結果與國語文教科書實驗一致。

接著，我們探討使用支援向量機的預測效能，各組實驗設定與結果如表十五所示。實驗結果與國語文教科書實驗結果呈現相同趨勢，唯其年級層劃分較少，

故正確性相對較高。同樣地，我們嘗試比較與結合支援向量機與逐步迴歸模型，其實驗設定與結果如表十六所示。

實驗	使用特徵	正確性
1	領域實詞頻對數平均 ＋ 負向連接詞數 ＋ 句平均詞數 ＋ 代名詞數	33.47%
2	基礎特徵 ＋ 逐步迴歸分數	49.80%
3	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵	43.67%
4	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵 ＋ 詞表示法 256 維	53.47%
5	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵 ＋ 詞表示法 256 維 ＋ 詞性表示法 256 維	53.88%
6	基礎特徵 ＋ 逐步迴歸分數 ＋ 詞表示法 256 維 ＋ 詞性表示法 256 維 ＋ 語意資訊特徵	56.33%
7	基礎特徵 ＋ 逐步迴歸分數 ＋ 詞表示法 256 維 ＋ 詞性表示法 256 維 ＋ 語意資訊特徵 ＋ 寫作程度特徵	53.06%

表十二、支援向量機結合各式特徵與逐步迴歸分數於預測文章年級結果

區間	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
正確性	25%	24%	26%	31%	31%	28%	29%	28%	29%	36%	**38%**

表十三、逐步迴歸分數以不同區間劃分所預測之文章年級結果

預測年級 / 原始年級	低年級	中年級	高年級	國中	正確性
低年級	4	6	0	0	40%
中年級	2	8	0	0	**80%**
高年級	2	8	0	0	0%
國中	0	8	1	1	10%

表十四、逐步迴歸預測文章年級結果

實驗	使用特徵	正確性（%）
1	基礎特徵	40.00%
2	句法分析與詞性特徵	42.50%
3	詞表示法 256 維	42.50%
4	詞表示法 512 維	45.00%

5	詞表示法 1024 維	**47.50%**
6	詞性表示法 256 維	45.00%
7	詞性表示法 512 維	40.00%
8	詞性表示法 1024 維	37.50%
9	語意資訊特徵	**47.50%**
10	寫作程度特徵	25.00%

表十五、SVM 使用各特徵預測文章年級結果

實驗	使用特徵	正確性（%）
1	負向連接詞數	40.00%
2	基礎特徵 ＋ 逐步迴歸分數	37.50%
3	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵	37.50%
4	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵 ＋ 詞表示法 1024 維	52.50%
5	基礎特徵 ＋ 逐步迴歸分數 ＋ 句法分析與詞性特徵 ＋ 詞表示法 1024 維 ＋ 詞性表示法 256 維	52.50%
6	基礎特徵 ＋ 逐步迴歸分數 ＋ 詞表示法 1024 維 ＋ 詞性表示法 256 維 ＋ 語意資訊特徵	55.00%
7	基礎特徵 ＋ 逐步迴歸分數 ＋ 詞表示法 1024 維 ＋ 詞性表示法 256 維 ＋ 語意資訊特徵 ＋ 寫作程度特徵	52.50%

表十六、SVM 結合各特徵預測文章年級結果

五、 結論與未來展望

本論文提出句法分析與詞性、詞表示法、語意資訊、寫作程度等特徵用於文本可讀性預測，並將特徵彼此結合以提升預測之正確性。亦分別透過逐步迴歸與支持向量機等兩種方式建立可讀性模型，比較兩者個別用於測試國中小教科書及優良課外讀物之效能優劣。從實驗比較中可以發現，使用的指標數目越多時，預測正確率通常較高，故盡可能的採計多種特徵是顯而易見的策略。

未來研究方向將利用特徵抽取等工具達到增加指標多樣性的目的。而若能找出對於不同年齡層皆具影響力的指標，將可提升預測高年級文本的正確率。此外，可讀性研究仍有許多可以應用之處，如輔助第二語言學習者、有閱讀障礙之讀者選取閱讀文本與多媒體文件之可讀性預測等[5]，這些亦是值得努力的方向。

參考文獻

[1] 宋曜廷、陳茹玲、李宜憲、查日穌、曾厚強、林維駿、張道行、張國恩, “中文文本可讀性探討：指標選取、模型建立與效度驗證”, *中華心理學刊*, 55 卷, 1 期, 75–106, 2013.

[2] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, “Coh-Metrix: Analysis of Text on Cohesion and Language,” *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 193–202, 2004.

[3] 陳世敏, “中文可讀性公式試擬”, *新聞學研究*, 8 卷, 181–226, 1971.

[4] 楊孝濚, “中文可讀性公式”, *新聞學研究*, 8 卷, 77–101, 1971.

[5] K. Collins-Thompson, “Computational Assessment of Text Readability: A Survey of Current and Future Research,” *Recent Advances in Automatic Readability Assessment and Text Simplification*. *Special issue of International Journal of Applied Linguistics*, vol. 165, no. 2, 97–135, 2014.

[6] “可讀性 - 維基百科，自由的百科全書”, available at: https://zh.wikipedia.org/wiki/%E5%8F%AF%E8%AF%BB%E6%80%A7.

[7] “迴歸分析 - 維基百科，自由的百科全書”, available at: https://zh.wikipedia.org/wiki/%E8%BF%B4%E6%AD%B8%E5%88%86%E6%9E%90.

[8] 多變量分析最佳入門實用書：SPSS+LISREL(SEM)（2007）。台北：碁峰資訊。.

[9] 祁亨年, “支持向量機及其應用研究綜述”, *計算機工程*, 10 期, 6–9, 2004.

[10] “LIBSVM - A Library for Support Vector Machines,” available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html?js=1#svm-toy-js.

[11] “Coh-Metrix Web Tool,” available at: http://tool.cohmetrix.com/.

[12] “中文文本自動化分析系統”, available at: http://210.240.188.161/Chinese_CohMetrix/index.html.

[13] “About WordNet - WordNet - About WordNet,” available at: http://wordnet.princeton.edu/.

[14] “中文詞彙網路 Chinese Wordnet”, available at: http://lope.linguistics.ntu.edu.tw/cwn/.

[15] “文本可讀性指標自動化分析系統 2.3”, available at: http://www.chinesereadability.net/CRIE/?LANG=CHT.

[16] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, “A Comparison of Features for Automatic Readability Assessment,” *23rd International Conference*

on Computational Linguistics (COLING 2010), Poster Volume, pp. 276–284, 2010.

[17] 陳惠玉, “認識語法單位”, *台中市國教輔導團電子報*, 2004.

[18] “詞類 - 維基百科，自由的百科全書”, available at: https://zh.wikipedia.org/wiki/%E8%A9%9E%E9%A1%9E.

[19] 張劍、屈丹、李真, “基於詞向量特徵的循環神經網絡語言模型”, *模式識別與人工智能*, vol. 28, no. 4, pp. 299–305, 2015.

[20] “word2vec - Tool for computing continuous distributed representations of words. - Google Project Hosting,” available at: https://code.google.com/p/word2vec/.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space,” *In Proceedings of Workshop at ICLR*, 2013.

[22] 詞庫小組。「句結構樹中的語意角色」。技術報告 13-01。民 102 年。

[23] A. Louis and A. Nenkova, “What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain,” *Transactions of the Association for Computational Linguistics*, 1, pp. 341–352, 2013.

[24] 卓淑玲、陳學志、鄭昭明, “台灣地區華人情緒與相關心理生理資料庫—中文情緒詞常模研究”, *中華心理學刊*, 55 卷, 4 期, 493–523, 2013.

[25] “文化部中小學生優良課外讀物推介評選活動 - 第 37 次”, available at: http://book.moc.gov.tw/book/.