應用詞嵌入技術訓練中文可聽性模型以預測口語文本難度
(Training a Chinese Listenability Model Using Word2Vec to Predict the Difficulty of Spoken Texts)

錢彥翔 Yen-Hsiang Chien[1], 曾厚強 Hou-Chiang Tseng[1], 陳冠宇 Kuan-Yu Chen[2], 宋曜廷 Yao-Ting Sung[3]

[1]Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology
[2]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
[3]Department of Educational Psychology and Counseling, National Taiwan Normal University
wwww43992001@gmail.com, tsenghc@mail.ntust.edu.tw, kychen@mail.ntust.edu.tw, sungtc@ntnu.edu.tw

摘要

隨著數位學習的普及，越來越多學習者會接觸到以影音為主的教材。對於學齡前至小學低年級的學生來說，受限於有限的識字能力，更仰賴以聲音與影像為主的教材來獲取知識。早期的可讀性模型主要針對書面語設計，其在口語材料上的適用性尚待驗證。為解決此問題，本研究針對中文不同年級的口語材料，探討不同斷詞工具、語言模型，對自動分級模型效能的影響。本研究以支持向量機進行年級分級，旨在自動預測教材對應年級，協助學習者選擇適性化教材。結果顯示，詞嵌入維度較高的語言模型在分級效能上有較佳表現，準確率最高可達61%，鄰近準確率則為76%。此結果有助於未來在數位學習平台或教學資源推薦系統中，自動化地為學生選擇合適的聽力教材，提升學習的成效。

Abstract

With the proliferation of digital learning, an increasing number of learners are engaging with audio-visual materials. For preschool and lower elementary students, whose literacy skills are still limited, knowledge acquisition relies more heavily on spoken and visual content. Traditional readability models were primarily developed for written texts, and their applicability to spoken materials remains uncertain. To address this issue, this study investigates the impact of different word segmentation tools and language models on the performance of automatic grade classification models for Chinese spoken materials. Support Vector Machines were employed for grade prediction, aiming to automatically determine the appropriate grade level of learning resources and assist learners in selecting suitable materials. The results show that language models with higher-dimensional word embeddings achieved better classification performance, with an accuracy of up to 61% and an adjacent accuracy of 76%. These findings may contribute to future digital learning platforms or educational resource recommendation systems by automatically providing students with appropriate listening materials to enhance learning outcomes.

關鍵字：可聽性、預訓練語言模型、自然語言處理

Keywords:Listenability, Pre-Trained Language Models, Natural Language Processing

1 緒論

「適性學習」核心宗旨在於針對每位學習者的程度與需求，對其教學方式提供最適合建議以優化其學習效率 (Fariani et al., 2023)，根據 Chall (1983) 著作《閱讀發展階段》中提出，閱讀能力的成長可以分為六個階段——「閱讀準備期」(Pre-Reading)、「基礎識字與破譯期」(Initial Reading and Decoding)、「鞏固與流暢期」(Confirmation & Fluency)、「閱讀學習新知期」(Reading for Learning the New)、「多角度閱讀期」(Reading at Multiple Viewpoints) 和「建構與重組期」(Construction-Reconstruction)。每個階段對應不同的年齡與能力。基於此理論，學習材料的難度選擇便格外關鍵，若材料過於簡單，學習者難以獲得新知識；若過於困難，

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 1–10
November 20-22, 2025

則可能造成過重的認知負荷（Cognitive Load）(Cambria & Guthrie, 2010)。因此，事先評估閱讀材料的難易度並針對學習者能力來提供適切內容，將有助於提升整體學習成效(Bahmani & Farvardin, 2017)。文本可讀性（Text Readability），指文本可以被理解的程度，可讀性越高，表示文本越能夠被理解(Dale & Chall, 1949)。為了評估閱讀材料的難易度，許多學者自 20 世紀中葉起便開始研究文本可讀性，以評估書面文本難易度 (Dale & Chall, 1948; Flesch, 1948; Gunning, 1952; Kincaid et al., 1975; Mc Laughlin, 1969)。隨著時代進入 21 世紀，機器學習技術的快速發展，評估文本可讀性的方式也日益自動化與精確化，越來越多研究將文本可讀性公式計算得到的分數作為特徵，再結合其他語言特徵後來訓練分類模型，以自動評估文本可讀性。例如，Petersen and Ostendorf (2009) ，採用支持向量機（Support Vector Machine，SVM）(Cortes & Vapnik, 1995)，結合平均句法樹高度（Parse Tree）、平均名詞數、Flesch－Kincaid 分數…等等進行英語文本的閱讀等級判別。Liu et al. (2021) 則比較多種機器學習模型在線上健康資訊可讀性分級的表現，模型包括 XGBoost (Chen & Guestrin, 2016)、Random Forest (Breiman, 2001)、Bayes Net (Pearl, 2014)…等等，結果顯示，整合多個模型方法表現最佳。這些研究驗證了結合語言特徵的機器學習模型，能顯著提升文本分級的準確性。

由上述可知，許多研究在探討文本可讀性，然而，隨著時代演變，在網際網路與寬頻的快速進步下，學習型態也產生強烈改變，學習型態從以往的書面「文字為主」逐漸轉向影音拓展，如：YouTube、Podcast、有聲書，再到線上教學平台，如因材網（教育部因材網, 2020)等，以「聲音與影像」為主的數位內容，使知識獲取的方式更加多元化。學習者不再僅透過閱讀文字，還可以透過「聆聽講述」與「觀看影像」來獲取知識。而對於年齡較小的兒童而言，他們往往尚未具備理解複雜書面文本的能力 (Hogan et al., 2014)。根據 Chall (1983) 的閱讀發展階段理論，從 6 個月大到 8 歲這段期間，處於「閱讀準備期」到「鞏固與流暢期」，也就是學齡前至小學 3 年級這一段期間，學童的閱讀能力尚在發展，識字能力有限。對於這樣的學習者，他們可能更仰賴聽覺來理解世界，甚至透過語音與歌謠式的學習，來提高學習動機與理解程度。換言之，對於學齡前至國小低年級的學生而言，在「聽」和「讀」的理解能力上，可能會出現差異 (Catts et al., 2006)；根據 Ehri and Wilce (1985) 研究發現，學齡前及初學閱讀的兒童常常能夠正確理解與說出某些詞彙，但未必能夠準確辨認這些詞彙，顯示口語理解與閱讀辨識之間存在明顯落差。因此，在學習型態改變之下，傳統的可讀性公式雖然能夠有效評估書面文本難易度，但是否合適評估學齡前及初學閱讀兒童的口語材料的難度，便成為值得關注的議題。

事實上，針對口語材料難度的分析，稱之為可聽性（Listenability）模型，並已有多項相關研究 (Alghamdi et al., 2022; Fang, 1966; Yoon et al., 2016)。可聽性指語音材料被聽者理解的程度 (Harwood & Cartier, 1952)。其評估有助於教師與學習者快速篩選適合其聽力程度的語料，進而提升學習動機與聽力成效 (Alghamdi et al., 2023)。在教學實務與相關研究中，常見的語音材料大多為教學錄音、新聞稿與線上課程，這些內容多會轉換為口語文本，作為可聽性分析與模型建構的基礎。因此，口語文本的難度評估也逐漸成為可聽性研究的重要方向。例如：在美國 Kayam (2018) 以三種可讀性公式分析 2016 年美國總統候選人的演講與訪談文本，發現語言結構較為簡單、難度較低的文本，更容易觸及廣泛的受眾。Bayona et al. (2023) 等人採用四種可讀性公式，評估線上口語教學素材的難度，結果顯示這些公式雖能在一定程度上區分語料難度，但當教材難度提高時，僅依賴可讀性公式，可能難以充分反映口語材料的真實難度。該研究也進一步指出，對於高階或不同類別的口語材料，其可聽性差異往往無法僅以文字特徵來評估，顯示出未納入語速、斷句等語音特徵，可能導致難度分級的錯誤。值得一提的是 Leal et al. (2024) 等人針對同為兒童設計的語料，兒童電影字幕與兒童非小說書面文本，使用包含 200 個語言特徵的 NILC-Metrix 工具進行比較分析，發現兒童電影字幕與兒童非小說書面文本在語言複雜度、詞彙豐富度、句法結構、語篇銜接等面向均發現顯著差異。其

中，兒童非小說書面文本具有較高的語法複雜度與詞彙多樣性，語篇結構更具連貫性，而兒童電影字幕則更趨向簡化、重複。這說明即使語料來自同一個年齡層，書面語料與口語材料在語言特徵上仍存在顯著差異。這與 Louwerse et al. (2004) 等人的研究結果相似，該研究從詞彙、句法、語篇結構和凝聚性等多個語言特徵進行分析，證實當分析層次提升到語篇與凝聚性特徵時，口語和書面語之間的差異明顯。這不僅體現在語音中的語速、停頓等語音現象，更在詞彙、句法、語篇結構等語言特徵上有本質差異。由此可見，單純仰賴書面文本所開發的可讀性公式，難以直接應用於口語材料。因此，發展可聽性模型時，除了可考慮納入語音特徵外，更應重視口語材料獨有的語言特徵，以提升分級的準確性和實用性。整體而言，雖有部分研究者嘗試以可讀性公式與部分語言特徵應用於評估口語材料的難度，但大多數研究忽略了口語材料中獨有語言特徵與語音特徵，因此可能無法反映其真實難度。

隨著機器學習技術的成熟，語言模型在自然語言處理（Natural Language Processing，NLP）的任務中應用廣泛，有研究者試圖利用語言模型，如 Word2Vec (Mikolov et al., 2013)、BERT (Devlin et al., 2019)應用於文本分級任務。應用語言模型進行文本分級的可行性已被多項研究證實 (Uçar et al., 2024; Zhang, 2024)。然而，目前利用語言模型來作為特徵，以訓練出可聽性模型來評估口語文本難度的相關研究仍相對有限。因此，鑑於網路上現有的大量預訓練語言模型與豐富語料資源，本研究將採用二種不同語言模型：中央研究院 Word2Vec (Chen & Ma, 2018)、奧斯陸大學 Word2Vec (Fares et al., 2017)，訓練學齡前至高中三年級（K-12）中文口語文本可聽性模型，並比較不同模型準確率。本研究的內容如下：第二節將闡述可聽性模型相關研究，第三節講述研究設計與資料，第四節分析研究結果，第五節將總結以及未來展望。

2 相關研究

早期可聽性模型研究主要將各類語言特徵應用於口語文本來驗證模型效能，如：Rogers (1962) 收集美國 12 個年級共 480 段口語錄音為依變項，並以多元迴歸分析（Multiple Regression Analysis）對多個語言特徵作為自變項，用以預測聽眾所需的年級水平。結果顯示，句法複雜度和詞彙難度能夠有效預測年級水平，其預測年級與實際年級相差不超過兩個年級。Fang (1966) 則使用相關性分析（Correlate Analysis），探討多音節詞比例與平均句長等語言特徵，對書面新聞稿與口語新聞稿之間差異的影響。研究結果顯示，在口語新聞稿中，每句中包含越多多音節詞，會顯著增加聽眾的理解負擔。換句話說，句子中若包含越多發音較長、超過一個音節的詞，會讓內容聽起來更難懂。由上述可知，利用語言特徵來訓練可聽性模型以評估口語文本的難度具有可行性。隨著科技的發展，開始有研究者考量到口語中的語音特徵，例如 Yoon et al. (2016) 等人收集來自聽力測驗、新聞、訪談等口語樣本；先由不同程度英語學習者評定口語樣本的難度，並分為初級、中級與高級做為口語樣本難度標籤，再以語音特徵，包括語速、停頓頻率…等等，以及語言特徵來進行相關性分析，並選出了 12 項重要特徵建立多元迴歸模型。結果顯示，模型若僅納入語音特徵，預測準確度略高於僅使用語言特徵，而語音特徵與語言特徵同時納入時，準確度最高。Alghamdi et al. (2022) 則收集了涵蓋人文、商管、理工等領域的大學講座影片，以 EFL（English as a Foreign Language）B1 水準的學生針對每部影片的理解難度進行評分，最後取所有學生的平均分數作為該影片的難度標籤。研究在特徵方面考慮了共計 168 項語音特徵與語言特徵，包括發音比例、音節平均時長、平均停頓時長…等等，透過相關性與多元共線性（Multicollinearity）檢驗從 168 個特徵中選出 130 項特徵來訓練出可預測影片難度的偏最小二乘迴歸（Partial Least Squares Regression，PLS）模型。研究結果顯示，經過篩選後語音特徵與語言特徵，用於PLS模型時，在測試集上可解釋52%的總變異，表現優於只納入句長與音節長度特徵的模型。綜上所述，從早期的可聽性模型僅考慮語言特徵，到納入語音特徵，讓模型預測能力有所提升。然而，這些研究多聚焦於句長、音節數、語速、停頓等語言特徵和語音特徵。

隨著機器學習與自然語言處理技術的進步，文本可讀性研究已逐漸從分析語言特徵的方法，開始結合詞嵌入（Word Embedding）技術，例如 Word2Vec、BERT 等語言模型，以自動捕捉詞彙間的語義關聯。例如，Uçar et al. (2024) 等人的研究聚焦於多語言的科學教育文本，利用基於 BERT 與 Longformer (Beltagy et al., 2020)的深度學習模型，對英語、西班牙語及巴斯克語的科學教材進行文本可讀性評估。他們建立專門的教育語料庫，通過比較傳統機器學習方法與深度學習模型的效能，發現深度學習模型在準確率與 F1 分數上皆顯著優於傳統方法，尤其是 BERT 模型表現突出。而 Zhang (2024) 在國際漢語教學領域中開發了基於 BERT 的 RCS-CSLT 系統，並結合中國教育部頒布的「國際漢語教育中文能力分級標準」（Chinese Proficiency Grading Standards，CPGS）以及特定的中文語言特徵，包括詞彙豐富度、句法複雜度…等等。進行以漢語作為第二語言文本的可讀性分級。他們收集並標記大量國際漢語教學文本，訓練並微調 BERT 模型，結果顯示使用基礎 BERT 模型的平均準確率 84.1%，而使用結合語言特徵的 RCS-CSLT 模型後，平均準確率可提升至 89.8%。

由上述研究可知，近年來，可讀性模型已經普遍利用語言模型來進行訓練，以分析書面文本的難度。然而，相較於書面可讀性模型，現有的可聽性模型研究較少基於語言模型進行訓練。此外，對於語言模型分級任務而言，詞嵌入維度的選擇常被認為是影響分級效能的重要因素 (Yin & Shen, 2018)。一般來說，詞嵌入的維度越大，模型能夠捕捉到的語義資訊就越豐富，但同時維度過高也會導致泛化能力下降，必須在表徵能力與效能之間取得平衡 (Melamud et al., 2016)。因此，若能進一步比較不同詞嵌入維度的語言模型，對於 K-12 年級口語文本的分級任務中的效能，也將是一個值得研究的議題。綜上所述，本研究將比較不同維度語言模型對 K-12 年級口語文本分級效能的影響，期望為中文口語文本可聽性模型的建立提供實證依據與新思路。

3 研究設計

本研究的整體流程如圖 1 所示，依序為資料收集、資料前處理、特徵提取（Feature Extraction），以及分級器訓練與驗證等階段。在資料收集階段，本研究收集 2865 篇中文口語材料，涵蓋學齡前至高中三年級。其中，學齡前階段語料來自多個為 3 至 8 歲幼兒教育的 YouTub 頻道 (Peppa-Pig-Chinese-Official, n.d.; PTSKIDS, n.d.; XIAOXINGXING-樂樂 TV, n.d.; 北鼻故事屋 YouTube 頻道, n.d.)，1 至 12 年級則取自臺灣臺北酷課雲 (臺北酷課雲 Taipei CooC-Cloud, n.d.)的官方教學影片。所有影片皆由具教師資格之教師錄製，並依據課程難度編排，內容具有公信力。各年級影片數量詳見表 1。

在語音資料取得後，本研究統一採用微軟 Azure 語音辨識 (Microsoft, 2025)進行語音轉文字（Speech to Text，STT）處理，並進行人工校對，完整保留口語中的冗詞，以維持資料的原始性與真實口語特徵。在口語文本取得後，因有別於英文在語言結構上詞彙之間留有空格，中文在語言結構上詞彙之間缺乏空格作為分隔，因此需要斷詞（Word Segmentation）來劃分詞彙，做為後續輸入語言模型的前置準備。現行的斷詞工具在斷詞策略上亦有優劣，斷詞的精確度不僅影響詞彙的劃分，也會進一步影響語言模型在詞嵌入建構上的表現。因此，本研究分別採用 Jieba (fxsjy, 2012)與 CkipTagger (Li P-H, 2019) 兩套中文斷詞工具進行斷詞，以評估斷詞對於詞嵌入建構與模型效能的影響。

斷詞完成後，斷詞結果將作為語言模型的輸入。本研究在語言模型部分採用 Word2Vec，而 Word2Vec 常見的兩種訓練策略為 Skip-gram 與 Continuous Bag-of-Words（CBOW），兩者特性和適用情境並不相同。Jang et al. (2019) 指出，Skip-gram 架構在處理短篇、語境稀疏的文本分類上表現更佳。CBOW 則較適合長文本與語境完整的資料。綜上所述，本研究所採用的 Word2Vec 模型皆基於 Skip-gram 架構。主要考量到 K-12 口語材料多為句子短、語境分散，Skip-gram 將有助於提升語義向量於文本難度分級上的效能。本研究分別選中央研究院開發之 Word2Vec 模型，詞嵌入維度為 300，其餘參數為預設，訓練語料來自 Chinese Gigaword 與 ASBC 語料庫，詞彙量約 517,015 詞；另一為奧斯陸大學提供之 Word2Vec 模型，同樣採用 Skip-gram 架構，詞嵌入維度為 100，視窗大小為 10，其餘為預設，訓練語料來自 ChineseT CoNLL17 語料庫，詞彙量約

1,935,503 詞。經由不同斷詞工具與語言模型的組合，本研究最終產生四種特徵組合。表 2 所示分別為：（1）Jieba 斷詞搭配中央研究院 Word2Vec 後稱 J-AS、（2）Jieba 斷詞搭配奧斯陸大學 Word2Vec 後稱 J-OS、（3）CkipTagger 斷詞搭配中央研究院 Word2Vec 後稱 C-AS，以及（4）CkipTagger 斷詞搭配奧斯陸大學 Word2Vec 後稱 C-OS。每一組皆以詞嵌入平均（Mean Pooling）的方式生成文本語義向量，作為後續分級模型的特徵輸入。

最終，所有組合產生的文本語義向量，皆統一作為特徵輸入支持向量機。支持向量機採用 Scikit-Learn (Pedregosa et al., 2011)套件，核函數（Kernel）設定為徑向基函數（Radial Basis Function，RBF）其餘參數維持預設，並以五折交叉驗證（5-Fold Cross-Validation）(Arlot & Celisse, 2010)訓練 K-12 年級口語文本的可聽性模型。本研究藉由系統性比較各組特徵組合的分級效能，以探討斷詞策略及語言模型對於分級模型表現的影響。

年級	總數
K	715
1	2
2	14
3	43
4	31
5	49
6	20
7	209
8	222
9	233
10	460
11	597
12	270

表 1. 各年級影片總數

斷詞工具	語言模型	模型名稱
Jieba	中央研究院 Word2Vec	J-AS
	奧斯陸大學 Word2Vec	J-OS
CkipTagger	中央研究院 Word2Vec	C-AS
	奧斯陸大學 Word2Vec	C-OS

表 2. 斷詞工具與語言模型組合及對應模型名稱

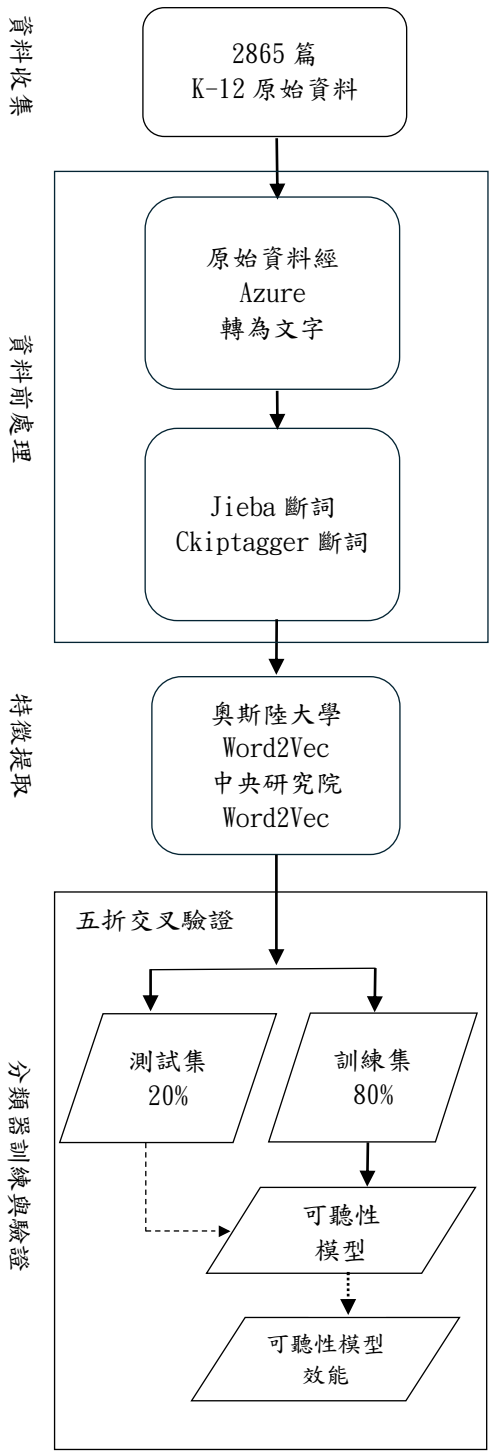

圖 1. 基於不同斷詞及 Word2Vec 可聽性模型訓練流程圖

4 研究結果

本節將分析不同斷詞工具與不同語言模型組合下，可聽性模型於 K-12 年級分級任務的效能表現。為檢驗模型效能的穩定性，所有效能指標均採用五折交叉驗證所得結果。評估

指標包括準確率（Accuracy）、鄰近準確率（Adjacent Accuracy）及混淆矩陣（Confusion Matrix）。其中，準確率衡量模型預測完全正確的比例；鄰近準確率則放寬標準，將預測年級與實際年級落差在前後一個年級內亦視為正確，有助於觀察模型分級錯誤的情形。混淆矩陣則能以直觀方式呈現模型在不同年級間的分級混淆狀況，幫助檢視哪些年級最易被誤判為其他年級，進而分析模型分級的困難點。

4.1 Jieba 斷詞工具與不同與語言模型之分級效能比較

本小節比較了 Jieba 斷詞下， J-AS 與 J-OS 兩組模型於K-12年級分級任務的整體表現。表3中呈現的指標包括：準確率與鄰近準確率。這兩項指標可用以整體評估模型在 K-12 年級分級任務中的效能，亦能反映模型對於相近年級語料的實際預測能力。從數據結果來看，J-AS 的整體準確率與鄰近準確率分別為 61%和76%，皆略高於 J-OS 的 59%與 74%。顯示出較高維度的詞嵌入有助於提升分級效能。不過，兩組模型整體表現相近，這可能與奧斯陸大學 Word2Vec 模型的詞彙量較高（約 1,935,503 詞），而中央研究院 Word2Vec 僅約 517,015 詞有關。較高的詞彙量可能使奧斯陸大學模型在詞嵌入維度較低的情況下，依然能維持一定的分級效能。整體來看，在 Jieba 斷詞下，詞嵌入維度與詞彙量皆可能對模型表現產生影響。

為進一步剖析模型分級，圖 2 展示了 J-AS 與 J-OS 兩組模型五折交叉驗證後的平均混淆矩陣。混淆矩陣可直觀呈現模型在各年級間的分級結果，縱軸為實際年級，橫軸為預測年級，對角線數值越高表示預測準確度越佳，對角線以外則代表誤判情形。藉由觀察混淆矩陣，不僅能了解模型在整體上的分級效能，也能發現模型容易混淆的年級區間。從圖 2 兩組模型的混淆矩陣結果可以發現，對於學齡前（K）年級的辨識表現最為突出，多數 K 年級語料均能正確分級。然而，國小 1 至 6 年級語料較常被誤分至 K 年級或高年級。顯示該區間語料分級錯誤可能被 K 年級和高年級樣本主導，此種誤判現象也可見於部分 K 年級與高年級資料。此外，自 7 年級起，模型預測結果顯著集中於對角線，10 至 12 年級的分級成效較為穩定，唯 9 及 12 年級仍有部分誤判。整體來看，混淆矩陣反映出樣本數較多或分布較集中的年級，其分級準確度較高，此外，低年級與 K 年級或高年級之間的誤判現象也較為明顯，顯示資料本身的分布特性會影響模型的分級效果。

模型	J-AS	J-OS
準確率	0.61	0.59
鄰近準確率	0.76	0.74

表 3. Jieba 斷詞下不同語言模型之分級效能比較

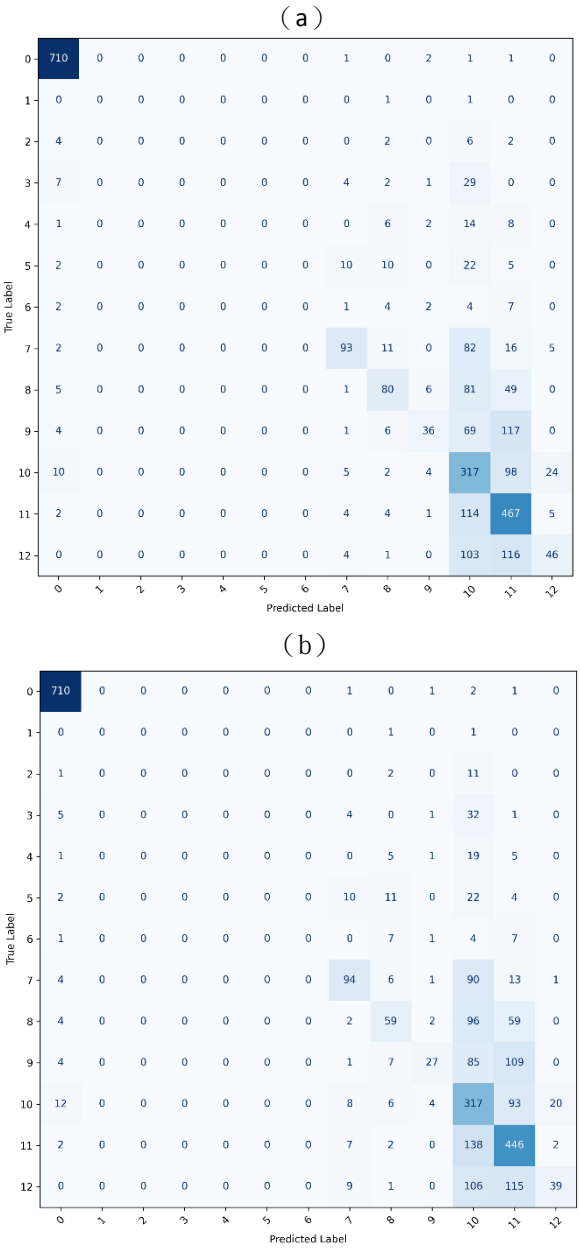

圖 2. 實驗 1 之各年級語料混淆矩陣 （a）J-AS 模型；（b）J-OS 模型

4.2 CkipTagger 斷詞工具與不同語言模型之分級效能比較

本小節比較了 CkipTagger 斷詞下，C-AS 與 C-OS 兩組模型於 K-12 年級分級任務的整體表現。在進一步比較 C-AS 與 C-OS 分級效能前，先說明表 4 中的 OOV（Out of Vocabulary）指標。OOV 指的是斷詞結果中，未能在 Word2Vec 詞彙表中找到向量的比例。OOV 比例越低，代表斷詞結果中的詞彙能夠被語言模型覆蓋。根據表 4 所示，CkipTagger 斷詞搭配其他 Word2Vec 模型時 OOV 比例都低於 Jieba 斷詞，顯示 CkipTagger 在中文口語材料的詞彙覆蓋表現較佳，能提升語義向量的完整度。

然而，根據表 5 兩組模型數據，C-AS 與 C-OS 有著更低的 OOV 比例，但準確率與鄰近準確率卻沒有提升，反而些微降低。此現象顯示，以本研究而言，降低 OOV 並未提升可聽性模型的效能。在現有的模型設計下，即使詞彙覆蓋率增加，模型對年級特徵的區辨力依然有限。因此，分級效能的提升或許還需結合更多語言特徵或更細緻的向量表示設計，方能發揮 CkipTagger 斷詞詞彙覆蓋率的優勢。此外，圖 3 二組混淆矩陣也顯示，低年級語料仍頻繁誤判為 K 年級或高年級，誤判形勢與 Jieba 斷詞結果高度相似，進一步說明提升詞彙覆蓋率對分級困難區間的改善有限。綜合來看，CkipTagger 斷詞能顯著降低 OOV 比例，但在目前模型設計下，對整體分級效能提升有限。因此，本研究建議未來可嘗試結合更進階的語言模型或分級器，以探討是否能進一步提升模型整體效能，以及加強對個別年級的預測能力。

模型	OOV 比例
J-AS	9.92%
J-OS	15.19%
C-AS	3.25%
C-OS	12.01%

表 4. 各組模型之平均 OOV 比例

模型	C-AS	C-OS
準確率	0.60	0.58
鄰近準確率	0.75	0.74

表 5. CkipTagger 斷詞下不同語言模型之分級效能比較

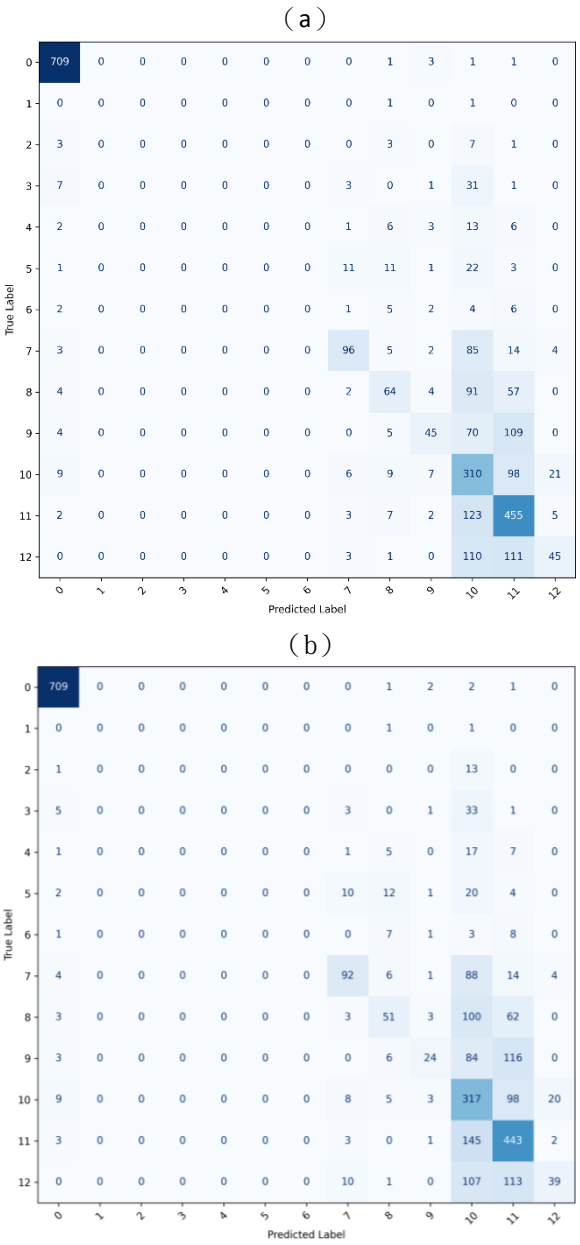

圖 3. 實驗 2 之各年級語料混淆矩陣 （a）C-AS 模型；（b）C-OS 模型

5 結論與未來發展

本研究針對中文 K-12 年級口語材料，系統性探討不同斷詞工具與語言模型組合下，口語可聽性自動分級模型之效能表現。主要研究內容包括：收集涵蓋學齡前至高中三年級的中文口語材料，並以 Jieba、CkipTagger 兩套中文斷詞工具及中央研究院與奧斯陸大學 Word2Vec 語言模型，組合成四組分級特徵進行效能比較。所有模型效能均採用五折交叉驗證進行評估，指標包括準確率、鄰近準確率以及混淆矩陣。實驗結果顯示，在所有組

合中，以 J-AS 表現最佳，整體準確率達 61%、鄰近準確率高達76%。同時，本研究發現詞嵌入維度較高的語言模型，在年級分級任務上表現普遍優於維度較低的模型。相較之下，CkipTagger 斷詞工具雖能有效降低 OOV 比例，但在現有設計下，並未顯著提升分級準確率。此外，混淆矩陣結果也揭示，低年級語料較容易被誤判為K年級或高年級，而高年級分級成效較為穩定。

在學術貢獻上，本研究彌補了中文口語可聽性模型領域中，針對斷詞工具與語言模型組合比較之研究空白，提供系統性實證結果，亦為未來中文可聽性模型帶來一定參考價值。然而，值得注意的是，現行大多國小至高中的教材內容多採用螺旋式設計（Spiral Curriculum），即同一主題會在不同年級反覆出現，但難度、深度與用詞豐富度逐步提升(Bruner, 2009)。例如，「水循環」這一主題在低年級的教材中，可能僅以簡單的圖畫介紹水的蒸發與降雨，而在高年級則會進一步說明蒸發、凝結、降水等科學原理，甚至討論水資源管理等議題。若分級模型僅依賴「水循環」這一關鍵詞，便可能將內容較淺顯的教材誤判為高年級，進而降低分級準確度。這種設計雖有助於學生循序漸進地建立知識體系，但當分級模型僅應用語言模型進行分級時，過度依賴主題關鍵詞可能導致分級不準確，容易將內容簡單的教材誤分為高年級。

回顧相關文獻，過去可聽性模型多結合語言特徵與語音特徵，本研究則以語言模型為主要特徵。未來建議可以進一步結合語言特徵與語音特徵，強化模型在部分年級分級上的區分能力。另外，本研究雖以較早期的語言模型 Word2Vec 為主，然而透過系統性比較不同斷詞工具與詞嵌入模型的組合效能，彌補現有口語可聽性模型分級領域的不足，未來亦可在此基礎上，導入如 BERT、Longformer 等語言模型，提升中文口語可聽性模型的分級效能。

6 致謝

本研究承國科會研究計畫「114-2628-H-011-002-MY3」、國立臺灣科技大學教育部高教深耕計畫特色領域技職賦能研究中心及國立臺灣師範大學教育部高教深耕計畫華語文科技中心補助，並感謝中央研究院詞庫小組提供語言資源，謹此致謝。

參考文獻

Alghamdi, E. A., Gruba, P., Masrai, A., & Velloso, E. (2023). The use of lexical complexity for assessing difficulty in instructional videos. https://hdl.handle.net/10125/73524

Alghamdi, E. A., Gruba, P., & Velloso, E. (2022). The relative contribution of language complexity to second language video lectures difficulty assessment. *The Modern Language Journal*, *106*(2), 393-410. https://doi.org/10.1111/modl.12773

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. https://doi.org/10.1214/09-SS054

Bahmani, R., & Farvardin, M. T. (2017). Effects of Different Text Difficulty Levels on EFL Learners' Foreign Language Reading Anxiety and Reading Comprehension. *Reading in a foreign language*, *29*(2), 185-202. https://files.eric.ed.gov/fulltext/EJ1157550.pdf

Bayona, M. G. A., Hines, A., Gilmartin, E., & Dhonnchadha, E. U. (2023). An Evaluation of the Use of Text-Based Comprehensibility Measures on Online Spoken Language Learning Materials. In 2023 34th Irish Signals and Systems Conference (ISSC), https://doi.org/10.1109/ISSC59246.2023.10162065

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. https://doi.org/10.48550/arXiv.2004.05150

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Bruner, J. S. (2009). *The process of education*. Harvard university press. https://doi.org/10.4159/9780674028999

Cambria, J., & Guthrie, J. T. (2010). Motivating and engaging students in reading. *The NERA journal*, *46*(1), 16-29.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. https://doi.org/10.1044/1092-4388(2006/023)

Chall, J. S. (1983). Stages of reading development. https://www.academia.edu/download/56874484/stages_of_reading_development.pdf

Chen, C.-Y., & Ma, W.-Y. (2018). Word embedding evaluation datasets and wikipedia title embedding for Chinese. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), https://aclanthology.org/L18-1132/

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54. https://www.jstor.org/stable/1473669

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19-26. https://www.jstor.org/stable/41383594

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), https://doi.org/10.18653/v1/N19-1423

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic? *Reading Research Quarterly*, 163-179. https://doi.org/10.2307/747753

Fang, I. E. (1966). The "Easy listening formula". *Journal of Broadcasting & Electronic Media*, *11*(1), 63-68. https://doi.org/10.1080/08838156609363529

Fares, M., Kutuzov, A., Oepen, S., & Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In Proceedings of the 21st nordic conference on computational linguistics, https://aclanthology.org/W17-0237/

Fariani, R. I., Junus, K., & Santoso, H. B. (2023). A systematic literature review on personalised learning in the higher education context. *Technology, Knowledge and Learning*, *28*(2), 449-476. https://doi.org/10.1007/s10758-022-09628-4

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221. https://doi.org/10.1037/h0057532

fxsjy. (2012). *Jieba*. Retrieved July from https://github.com/fxsjy/jieba

Gunning, R. (1952). The technique of clear writing. *(No Title)*. https://cir.nii.ac.jp/crid/1971149384740811428

Harwood, K., & Cartier, F. (1952). On definition of listenability. *Southern Journal of Communication*, *18*(1), 20-23. https://doi.org/10.1080/10417945209371245

Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International journal of speech-language pathology*, *16*(3), 199-207. https://doi.org/10.3109/17549507.2014.904441

Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, *14*(8), e0220976. https://doi.org/10.1371/journal.pone.0220976

Kayam, O. (2018). The readability and simplicity of Donald Trump's language. *Political Studies Review*, *16*(1), 73-88. https://doi.org/10.1177/1478929917706844

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. https://doi.org/10.21236/ADA006655

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S., & Aluísio, S. M. (2024). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, *58*(1), 73-110. https://doi.org/10.1007/s10579-023-09693-w

Li P-H, M. W.-Y. (2019). *CkipTagger*. Retrieved July from https://github.com/ckiplab/ckiptagger

Liu, Y., Ji, M., Lin, S. S., Zhao, M., & Lyv, Z. (2021). Combining readability formulas and machine learning for reader-oriented evaluation of online health resources. *IEEE Access*, *9*, 67610-67619. https://doi.org/10.1109/ACCESS.2021.3077073

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In Proceedings of the Annual Meeting of the Cognitive Science Society, https://escholarship.org/content/qt7d8631cr/qt7d8631cr.pdf

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646. https://www.jstor.org/stable/40011226

Melamud, O., Goldberger, J., & Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In Proceedings of the 20th SIGNLL conference on computational natural language learning, https://aclanthology.org/K16-1006.pdf

Microsoft. (2025). Azure AI 語音. Retrieved 8 from https://azure.microsoft.com/zh-tw/products/ai-services/ai-speech

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier. https://books.google.com/books?hl=zh-TW&lr=&id=mn2jBQAAQBAJ&oi=fnd&pg=PP1&dq=Probabilistic+reasoning+in+intelligent+systems:&ots=4tFU2E8M90&sig=ww1O97ittFH8heGbsDUlmuGFRRY

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://doi.org/10.5555/1953048.2078195

Peppa-Pig-Chinese-Official. (n.d.). 小豬佩奇中文官方 Peppa Pig Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@PeppaPigChineseOfficial

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, *23*(1), 89-106. https://doi.org/10.1016/j.csl.2008.04.003

PTSKIDS. (n.d.). 小公視 Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@PTSKIDS

Rogers, J. R. (1962). A formula for predicting the comprehension level of material to be presented orally. *The journal of educational research*, *56*(4), 218-220. https://doi.org/10.1080/00220671.1962.10882926

Uçar, S.-Ş., Aldabe, I., Aranberri, N., & Arruarte, A. (2024). Exploring automatic readability assessment for science documents within a multilingual educational context. *International Journal of Artificial Intelligence in Education*, *34*(4), 1417-1459. https://doi.org/10.1007/s40593-024-00393-2

XIAOXINGXING-樂樂 TV. (n.d.). 樂樂 TV Youtube 頻道. Youtube. Retrieved 7 from https://www.youtube.com/@XIAOXINGXING-%E6%A8%82%E6%A8%82TV

Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *Advances in neural information processing systems*, *31*. https://papers.neurips.cc/paper/2018/file/b534ba68236ba543ae44b22bd110a1d6-Paper.pdf

Yoon, S.-Y., Cho, Y., & Napolitano, D. (2016). Spoken text difficulty estimation using linguistic features. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, https://doi.org/10.18653/v1/W16-0531

Zhang, C. (2024). RCS-CSLT: a language feature-driven readability classification system for international Chinese education. *Computer Assisted Language Learning*, 1-27. https://doi.org/doi.org/10.1080/09588221.2024.2430723

北鼻故事屋 YouTube 頻道. (n.d.). Retrieved 8 from https://www.youtube.com/channel/UCHNtUji8P8yz645YCgs8GUA

教育部因材網. (2020). Retrieved 0804 from https://adl.edu.tw/HomePage/home/

臺北酷課雲 Taipei CooC-Cloud. (n.d.). Retrieved 0804 from https://cooc.tp.edu.tw/