大雅語文智能平台

1. 計劃簡介

1.1 現況

1.2 計劃目標：釐定篇章的難度等級、主題和運用的語文知識

1.3 未來方向：字卡、默書、建立文庫、推薦文章、學生個人檔案

1. 分析目標及準則

2.1 釐定文章難度等級

2.2 釐定文章主題

2.3 釐定文章運用的語文知識

2.4 利用大數據修訂準則

1. 預期功能

3.1 釐定難度等級

3.1.1 字：字數統計

3.1.2 字和詞：按學習階段認字和詞

3.2 釐定主題

3.2.1 詞：分詞

3.2.2 字和詞：統計字詞的出現頻率

3.2.3 字和詞：按字詞分析主題或文體

3.3 釐定語文知識

3.3.1 句：語文知識(句式、複句、標點符號和部份修辭)

3.4 大數據：使用者概況

3.4.1 文章內部資料

3.4.2 用家額外資訊

1. 語文知識重點分類
   1. 分層概念
   2. 例子舉隅
2. 不同年級文本的分析方法

5.1 配合Codex分層記錄方法

5.2 統計數據

5.3 提出疑問

5.3.1 比喻喻詞 與 假設複句及其他：以「像」和「若」為例

5.3.2 關於「段」、「篇」的識別方法

5.3.3 分析的先後次序、數量、深度

1. 文本的分析方法

6.1 情況1：有特定文字和詞語作條件

6.2 情況2：以規則(Rules)來作條件

6.3 情況3：需要編輯者判斷

6.4 情況4：由用家(文章上載者)提供資料

1. 分析數據如何安排分配練習

8.1 程度劃分：學習階段和年級

8.2 現有練習程度劃分

1. 附件

附件1 字表

附件2 小學文章篇幅表

附件3 詞語表 (臺灣、香港)

附件4 大雅語文 語文知識重點分類

附件5 繪本《做自己最快樂》嗒嗒版 - 分析

附件6 繪本《做自己最快樂》魁魁版 - 分析

附件7 文本的分析方法：教科書文章：〈學與問〉

1. 計劃簡介

1.1 現況

語文課本和補充練習五花八門，坊間對於來自不同出版社、不同地區、不同難度的課本通常只能靠主觀感覺或自身經驗來評價和分辨。市面一直缺少了一套能夠量度語文課本篇章程度的標準和系統。是次計劃正是希望能建立統一量度標準，用來比較和檢視不同出版社、不同地區、不同難度的語文課程的準則和指標。

1.2 目標

計劃目標為開發一個利用人工智能科技分析中文篇章的網上平台，透過評定篇章中的潛在語文知識（字／詞／句／段／篇）和主題，比對不同地區的語文課程來測量篇章程度。

1.3 發展方向

**學校**

教師可利用此平台分析語文課程中的篇章，以圖解儀表板的形式查看篇章中的語文知識、主題和難度等級，以不同地區的課程比對難度等級。根據資料分析，教師可檢視和修訂現有課程，分配適當的篇章、補充練習、題目和學習資源來自訂校本課程。

**家長和學生**

收集學生研習過的篇章後，平台可為學生建立個人學習檔案，評估學生在語文知識、主題和難度等級的學習成果，進而推薦適當的篇章作個人化學習。

**出版、教育和研究**

網上平台收集不同地區語文課程的篇章後，可建立語文學習的文庫。文庫的潛在價值在於，出版社能夠依據本計劃數據審視作品；教育者又可安排不同的教學策略；研究者亦可以利用文庫所提供的素材，研究不同年代、地區的中國語文科教材。

**作者**

作者可利用此平台評鑑作品，安排銷售策略，透過撰寫作品獲取利潤。

1. 分析目標及準則

大雅語文智能讓不同持份者在小學中國語文科教學過程中，各取所需。不論是學校想查看教科書的數據、作者想依照教學進程提供教材，抑或家長學生想知道學習進度，「分析文本」都是首要任務。有見及此，本計劃旨在從以下三方面分析文章，並記錄及標籤文章，以作參考。該三方面分別是「難度等級」(Level Of Difficulty)、「文章主題」(Theme)和「語文知識」。

2.1 釐定文章難度等級

現有香港小學教育採用六年級制度來區分學生的學習進程，依照這個制度，本計劃同樣希望將上載文章以六年級制記錄並標籤。

本計劃建議用以下三個方法釐定文章難度等級：

第一、字數統計 (詳見3.1.1)

第二、按「字」分學習階段 (詳見3.1.2)

第三、按「詞」分學習階段 (詳見3.1.2)

2.2 釐定文章主題

除了學習程度，學生和學校普遍關注文章的主題和體裁。分析教材時，能夠得出文章的主題和體裁等具參考價值的資料，對教材發展的幫助不少。

本計劃建議用以下兩個方法釐定文章主題和體裁：

第一、按字詞分析主題或文體 (詳見3.2.3)

第二、由文章上載者提供資料 (詳見6.4)

2.3 釐定文章運用的語文知識

學習語文知識，是小學程度學習中國語文必經過程。現階段，本計劃將涵蓋句式、標點符號和部份修辭。一來可反映文章的程度，二來可以讓用家分析文章的學習重點，配合補充練習等額外教材加以應用。

因為不同的語文知識，也有它不同的特色、獨有模式或關鍵字詞。本計劃建議按字詞、語法特色、句法等分析不同語文知識。 (詳見3.1.1及部份4)

2.4 利用大數據修訂準則

除了2.1-2.3的分析功能以外，本計劃亦希望能夠記錄用戶的資料，包括曾經上載的篇章、用家修訂文章難度或對應年級的編輯歷程等，藉此透過真實用戶的需要，修改以上部分(2.1-2.3)背後的分析準則。務求能夠提供更貼近學生，更精準的語文教學教材。

1. 預期功能

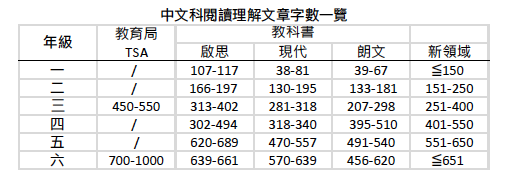
3.1 釐定難度等級

此部份希望透過三種東西，評估文章所屬的程度。首先是文章的長度(篇幅)，第二是文章內的「字」在既定的準則(例如：三千字表)內屬於哪一個年級。第三，與「生字」相同方式，改成詞語。

3.1.1 字：字數統計

大雅語文智能，為上載文章統計字數，然後根據附件2(下表)，以年級為文章分類，評定文章所屬的難度等級。

日後，根據用家輸入的年級和實際字量，可再對照當時狀況，調整以下表格。



圖I：參考教科書後得出文章篇幅數字

3.1.2 字：按學習階段認字

按照附件1(〈小學中文科常用字表〉)，識別文章中每個文字，並按「學習年級」、「常用度」、「學習難度」統計文字分佈。預期不同程度的文章，其不同「學習年級」字詞分佈的百分率有一定規律，用來評定文章的難度等級。



圖II：〈小學中文科常用字表〉(別稱〈三千字表〉)節錄

字和詞的機理大致相同，在此不再重覆。只是「詞語對應年級」的準則尚有待提供和調整。

3.2 釐定主題

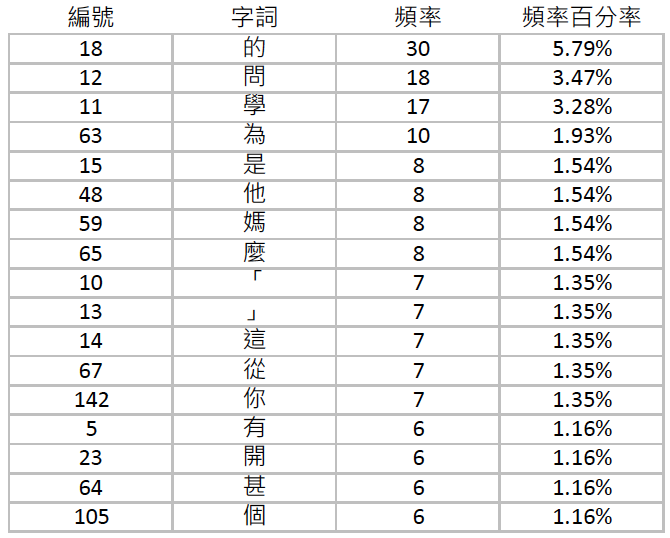
為了抽取文章的主題，提前需要「分詞」處理，然後運用統計頻率、搜字等方式，將文章歸類。

3.2.1 詞：分詞

本計劃在分析自然語言時，均需要先運用中文分詞，將中文字句預先分拆成語義獨立的組合後，才能深入分析字、詞、句等部分。處理中文自然語言時，可按照詞性或語法，將篇章中的詞語區分出來，成為文本掘挖的第一步。

3.2.2 字和詞：統計字詞的出現頻率

透過輸入文章，統計文章內所有字或詞語出現的頻率。記錄各種字詞的出現頻率，用來評估文章的特色、文體或文章主題。同時成為中文教科書的基本研究素材。



圖III：模擬詞頻統計

3.2.3 字和詞：按字詞分析主題或文體

承接3.2.2部份，利用詞頻統計，便初步知道文章的重點。若果更進一步，配合不同主題的字詞表，期望能夠得出某部分主題詞語的出現頻率，或許可有效推測文章的主題或體裁。比如說，文章出現了超10%有關顏色的形容詞，大概可以推測為描寫景物或說明顏色的文章。

3.2 釐定語文知識

3.3.1 句：語文知識(句式、複句、標點符號和部份修辭)

按照大雅圖書中國語文科語文知識重點分類，利用關聯詞等，通過掃瞄和搜尋，記錄並統計文章中所出現的語文知識。再者，根據語文知識的分層，透過語文知識的分佈和困難程度，來判斷文章的難度等級。

3.4 大數據：使用者概況

3.4.1 文章內部資料

本計劃平台具備學習能力。根據不同使用者所輸入、修訂、調整的歷史，與所得出的資料作比對。普遍採用使用者的修正，然後按照修訂，安排「難度等級」(Level Of Difficulty)、「文章主題」(Theme)和「語文知識」三方面的修正。

3.4.2 用家額外資訊

除了本次計劃目標以外，未來希望能夠完成中文選書平台。屆時，平台會記錄使用者選取文章時的數據，譬如：瀏覽人數、使用人數、使用者年級分佈、使用者的學校等數據。這些數據也可以幫助平台分析文章和文章使用概況。

1. 語文知識重點分類

**4.1 分層概念**

大雅圖書將中國語文科語文知識以「字、詞、句、段、篇」分五層，之下再分開四個層次，依次：(1) Domain (2) Area (3) Division (4) Learning Objectives

(5) Level Of Difficulty

附件4

大雅語文 語文知識重點分類

綠色為 (1) Domain

藍色為 (2) Area

黃色為 (3) Division

**4.2 例子舉隅**

以「着色詞」為例(表1)：

| (1) Domain | (2) Area | (3) Division | (4)  Learning Objectives | (5)  Level Of Difficulty |
| --- | --- | --- | --- | --- |
| 2 詞 | A 詞性 | 8 着色詞 | 1 單字 | 1A 認識不同顏色  2A 利用不同顏色形容事物 |
| 2 表達深淺程度的着色詞 | 1A 認識表達深淺程度的着色詞  2A 應用表達深淺程度的着色詞 |
| 3 比喻式着色詞 | 1A 認識比喻式着色詞  2A 應用比喻式着色詞 |
| 4 重疊式着色詞  AA式 | 1A 認識AA式重疊着色詞  2A 應用AA式重疊着色詞 |
| 5 重疊式着色詞  ABB式 | 1A 認識ABB式重疊着色詞  2A 應用ABB式重疊着色詞 |
| 6 重疊式着色詞  AABB式 | 1A 認識AABB式重疊着色詞  2A 應用AABB式重疊着色詞 |
| 7 四字詞着色詞 | 1A 認識四字詞着色詞  2A 應用四字詞着色詞 |

1. 不同年級文本的分析方法

**5.1 配合Codex分層記錄方法**

掃瞄文章後，配合Codex分層記錄第(1)至(4)層，即(1) Domain (2) Area (3) Division (4) Learning Objectives。例如說，當電腦掃瞄文章後，取得詞組「紅色」和「紅彤彤」，希望電腦能記錄如下：

| 事例(A)  文字：  「紅色」  記錄：  詞：詞性：着色詞：單字 | 事例(B)  文字：  「紅彤彤」  記錄：  詞：詞性：着色詞：重疊式着色詞ABB式 |
| --- | --- |

本社以下列兩個附件作示範，呈現從文章分析語文知識分層的概況，望能對日後掃瞄工作提供粗略情況和格式。

(附件5) 繪本《做自己最快樂》嗒嗒版

(附件6) 繪本《做自己最快樂》魁魁版

**5.2** **統計數據**

除了分析語文知識的階段，本社期望能統計每篇文本使用不同語文知識的數量。好讓提供各種語文知識的數量、頻密程度等資訊。

**5.3 提出疑問**

以下是執行時遇到的疑問和困難，未知能如何配合以解決。如有能協助之處，惟望不吝賜教。

5.3.1 語文知識文字特點——以喻詞與假設複句關聯詞「若」為例。

部份語文知識在文字的特徵相似，例如：「若」字，同時是「假設複句」關聯詞，也是「比喻」喻詞。未知在以「若」字識別兩者時，會不會有困難？

5.3.2 關於「段」、「篇」的識別方法

「段」和「篇」的語文知識的特徵比較隱晦，或者比較龐大，未知在識別兩者時將如何操作，或者本社能如何協助。

5.3.3 分析的先後次序、數量、深度

任何句子都能從「字、詞、句、段、篇」的結構中，分析海量的資料和分類。但是並非所有成份皆有作用，未知能否提供一個篩選素材的形式和方法，好讓能夠配合分析結果提供適切答案。

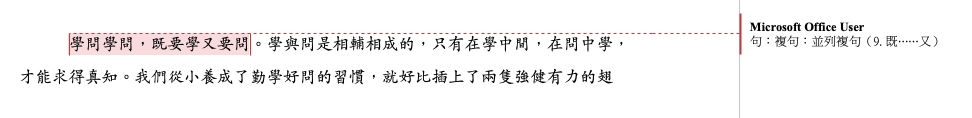
1. 文本的分析方法

**6.1 情況1：有特定文字和詞語作條件**

以「並列複句」為例，用以區分「並列複句」的字詞有如下，詳見附件7：

| 區分「並列複句」的字詞（關聯詞) | |
| --- | --- |
| 1 | 又 |
| 2 | 也 |
| 3 | 還 |
| 4 | 一會兒⋯⋯一會兒 |
| 5 | 有時⋯⋯有時 |
| 6 | 一方面⋯⋯一方面 |
| 7 | 一方面⋯⋯另一方面 |
| 8 | 一邊⋯⋯一邊 |
| 9 | 既⋯⋯又 |
| 10 | 既⋯⋯也 |
| 11 | 又⋯⋯又 |
| 12 | 是⋯⋯不是 |
| 13 | 不是……而是…… |

以新編啟思中國語文六年級單元七，文章《學與問》為例，以下標記的句子，因吻合上表第9項「既⋯⋯又」，所以歸入「並列複句」。



**6.2 情況2：以規則(Rules)來作條件**

以「引號」為例，用以區分「引號」的條件(Rules)有如下，詳見附件7：

| 區分「引號」的條件(Rules)  分為：  I. 特殊含義  II. 引用  III. 對話  IV. ［需要編輯者判斷］ | | |
| --- | --- | --- |
|  | 歸類為 | 條件 |
| 1 | I. 特殊含義 | 若⋯⋯少於4個字元，且獨立見於字詞表  「⋯⋯」 |
| 2 | II. 引用 | 若⋯⋯多於4個字元，不獨立見於字詞表  「⋯⋯」 |
| 3 | 配合冒號出現  ：「⋯⋯」 |
| 4 | III. 對話 | 配合冒號(：)和「說」出現  A) 說：「⋯⋯」 或  B) 「⋯⋯」XXXX說，「」 |
| 5 | 配合冒號(：)和「道」出現  A) 道：「⋯⋯」 或  B) 「⋯⋯」XXXX道，「」 |
| 6 | 配合冒號(：)和「曰」出現  A) 曰：「⋯⋯」 或  B) 「⋯⋯」XXXX曰，「」 |
| 7 | 配合冒號(：)和「講」出現  A) 講：「⋯⋯」 或  B) 「⋯⋯」XXXX講，「」 |
| 8 | 配合冒號(：)和「喊」出現  A) 喊：「⋯⋯」 或  B) 「⋯⋯」XXXX喊，「」 |
| 9 | 配合冒號(：)和「叫」出現  A) 叫：「⋯⋯」 或  B) 「⋯⋯」XXXX叫，「」 |
| 10 | IV. ［需要編輯者判斷］ | 若⋯⋯少於4個字元，但不獨立見於字詞表  「⋯⋯」 |

**6.3 情況3：需要編輯者判斷**

以上述例子為例：

[ 人們常把有知識說成「有學問」 ]一句中，

[有學問]三字是由 [有]+ [學問] 組成的詞組，整個詞組並不會獨立出現在字詞表內。

歸入[需要編輯者判斷]的詞條，代表需要編輯者額外判斷。所以，需要額外標記，提示編輯者。日後，當編輯者依靠人手歸類後，可考慮成為歸入「例外清單」或「採納清單」之中。

**6.4 情況4：由使用者(文章上載者)提供資料**

類似於6.1，同樣是在分析資料之外，由用家輸入相關資料。在平台上提供標籤(Tagging)功能，讓用家在上載文章後，能夠以選擇題或直接輸入兩種方式，提供關於文章的資料，例如文章題目、作者、主題、體裁或出處(或年級)等。

1. 分析數據如何安排分配練習

**7.1 程度劃分：學習階段和年級**

本社建議利用字數統計、小學字詞表等工具，釐定篇章適合的年級，以便分配練習和配套。

1. 文本篇幅長度可參考本社的準則：

| 年級 | 字數上限 或 範圍 |
| --- | --- |
| 小一 | 詩歌：80  文章：70 - 150 |
| 小二 | 詩歌：100  文章：150 - 250 |
| 小三 | 300 - 450 |
| 小四 | 400 - 550 |
| 小五 | 500 - 600 |
| 小六 | 600 - 700（亦有長達900） |

1. 小學字詞表建議參考由前教育統籌局於2003年委託香港理工大學中文及雙語學系進行「香港小學學習字詞研究」研究成果《香港小學學習字詞表》。或，臺灣國家教育研究院推出的《國教院三等七級詞表》(見附件三)。

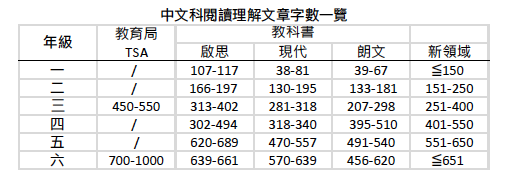
| 小學用字一覽表所得3,171字  第一學習階段（小一至小三）2,169字  第二學習階段（小四至小六）1,002字 | 小學詞語一覽表所得9,706詞語  第一學習階段（小一至小三）4,914詞語  第二學習階段（小四至小六）4,792詞語 |
| --- | --- |

1. 附件

附件1 〈小學中文科常用字表〉

<https://docs.google.com/spreadsheets/d/1GZ6WRFGm87qh7Itt5QosJeXOVBElwGp-/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>

附件2 〈小學文章篇幅表〉



(文字原稿)

<https://docs.google.com/spreadsheets/d/1unG6IQs7yWVR7jyJamZ0LzdLLPyDY-ex/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>

附件3 - 〈國教院三等七級詞表〉

<https://drive.google.com/file/d/1Dfa34ydd_8sPzhwM0E9K3PqZYPjDIRW-/view?usp=sharing>

附件4 - 大雅語文 語文知識重點分類

<https://drive.google.com/file/d/1rMTyPclcdR2pavvvN3gU6GXqA_nHlvi4/view?usp=sharing>

附件5 - 繪本《做自己最快樂》嗒嗒版 - 分析

<https://drive.google.com/file/d/1_I9JgOLuiy6fkoJUr-fwbZTAVHX4QLhn/view?usp=sharing>

(文字原稿)

<https://docs.google.com/document/d/1KvPJGP_QjiQHSqL7cdBtz0SpPjSwp6kF/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>

附件6 - 繪本《做自己最快樂》魁魁版 - 分析

<https://drive.google.com/file/d/1ml9Fh9bKP9UZcam4pukMMvshjZB-T0UX/view?usp=sharing>

(文字原稿)

<https://docs.google.com/document/d/1fRWWwtPHtEJkt_klUEuikRHyCu7VZ2dL/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>

附件7 - 文本的分析方法：教科書文章：〈學與問〉

<https://drive.google.com/file/d/1Li5Apts2UZXFi_11xgmVQfxXkJLWil6w/view?usp=sharing>

(文字原稿)

<https://docs.google.com/document/d/1STOwqBjTekf9Yh4v0bPIjPIpcnFsAcbC/edit?usp=sharing&ouid=112300334093909079389&rtpof=true&sd=true>