Chinese Grammar Intelligence (大雅語文智能)

1. Project Summary

1.1 Background

With numerous textbooks and supplementary exercises published over the years, there has been a lack of systematic and standardized ways to gauge the levels of difficulty of input text in Chinese textbooks and a lack of criteria and indicators to compare and review Chinese curriculums among different regions.

1.2 Aims

Chinese Grammar Intelligence (大雅語文智能) aims to develop a platform capable of diagnosing content of any input texts, identifying grammar items (字 /詞/ 句/ 段/ 篇) employed by the text and reviewing the level of difficulty for a particular passage, collection of passages or a series of publications with the use of artificial intelligence (AI).

1.3 Future Development

School Administrators/ Teachers

By selecting the textbook series used, teachers will get an analysis of the passages on an infographic dashboard. The themes, grammatical items in the text, and the level of difficulty with reference to a specific curriculum will be shown on the dashboard. Statistics of the uses of that passage in comparison with other passages will also be shown. Based on the analysis, school administrators will review and revise existing curriculum, customize their own curriculum, and allocate appropriate questions, exercises and resources to formulate learning experiences that enable progression and enhancement.

Parents/ Students

Learning portfolio of a student can be set up by referring to which passages they have studied, which level they have achieved and what grammatical items they have learnt. Appropriate passages can be assigned for more personalized learning.

Publishing and Education Industry

This platform can serve as a depository to store Chinese passages across curriculums. Publishers can review their publications and educators can research on Chinese teaching materials for different regions over the years.

Authors

By selecting a specific textbook series / curriculum, authors have a better understanding of what constitutes an appropriate passage at a specific level. Authors will also use this platform to evaluate their works and even sell them for royalty.

2. Objectives and Methodology

Chinese Grammar Intelligence (大雅語文智能) aims to diagnose content of any input texts in three aspects, namely level of difficulty (難度等級), theme (文章主題) and grammatical items (語文知識).

2.1 Identifying Level of Difficulty

Student's learning progression has been divided into six grades for primary education in Hong Kong. Passages will be graded with reference to this grading system.

The level of difficulty of the input text will be identified based on:

1. Character count (Refer to Section 3.1.1)
2. "Level of difficulty" of character (Refer to Section 3.1.2)
3. "Level of difficulty" of word (Refer to Section 3.1.2)

2.2 Identifying Theme

Apart from the level of difficulty, the theme and genre of the text are also the concern of this platform. Teachers take this into consideration when they are assigning appropriate learning materials for students.

The theme and genre of the input text will be identified based on:

1. Analysis of characters and words in the text (Refer to Section 3.2.3)
2. User input (Refer to Section 4.4)

2.3 Identifying Grammatical Items

Learning grammatical items is an essential component of Chinese learning in the primary school curriculum. Grammatical items such as sentence structures, punctuations and some rhetorical devices will be identified in the text. This does not only reflect the level of difficulty of the text to a certain extent, but also indicates learning objectives for planning suitable supplementary exercises.

Different grammatical items have different rules and key words. This project will identify grammatical items according to characters and words, syntactic patterns, grammatical rules, etc. (Refer to Sections 3.1.1 and 4)

2.4 Revising the Algorithm with the Use of Big Data

Apart from analyzing items in Sections 2.1-2.3, this platform will capture data from users, including uploaded texts, inputted level of difficulty, inputted grades, etc. If necessary, the above methodology (Sections 2.1-2.3) will be modified after training from user preferences. This will ensure the analysis is accurately reflecting real use cases of input texts.

3. Expected Functions

3.1 Identifying Level of Difficulty

Three criteria will be considered in identifying the level of difficulty of the input text, namely, length of text, “level of difficulty” of the characters in the text, and “level of difficulty” of the words in the text.

3.1.1 Character: Character Count

The number of characters contained in the text will be checked and compared with Appendix 2 (Table Below) to grade the text. When the platform is put into use, user input of actual grade will be considered in revising the table below.

中文科閱讀理解文章字數一覽

年級	教育局	教科書			
	TSA	啟思	現代	朗文	新領域
一	/	107-117	38-81	39-67	≦150
二	/	166-197	130-195	133-181	151-250
三	450-550	313-402	281-318	207-298	251-400
四	/	302-494	318-340	395-510	401-550
五	/	620-689	470-557	491-540	551-650
六	700-1000	639-661	570-639	456-620	≦651

圖I:參考教科書後得出文章篇幅數字

3.1.2 Character and Word: Learning Character and Word According to Grade

The grade, frequency of use and level of difficulty of each character have been ranked in Appendix 1 (〈小學中文科常用字表〉). The input text will be analyzed based on the composition of characters in the text under these three aspects. Different compositions of characters in the text refer to different levels of difficulty of the text.

編號	生字	學習年級	常用度	學習難度
2891	風	1	194	1
280	功	1	424	2
2412	要	1	21	3
1748	田	1	784	4
1723	現	1	68	5
1473	洋	1	731	6
2845	青	1	459	8
2470	誰	1	617	9
608	姐	1	945	10
44	五	1	154	11
1396	歌	1	475	12
986	才	1	199	15
1852	知	1	164	17
1265	朋	1	518	19
283	助	1	552	20

圖II:〈小學中文科常用字表〉(別稱〈三千字表〉)節錄

The algorithm of word analysis is similar to that of character analysis. The ranking of the grade of word will be provided and adjusted.

3.2 Identifying Theme

The input text has to be divided into meaningful units, the frequency of which will be measured, before analyzing the theme of the text.

3.2.1 Words: Text Segmentation

The text has to be automatically segmented into meaningful words or phrases with natural language processing before further analyzing its characters (字), words (詞) and sentences (句). Each segmented word or phrase is categorized and tagged by its part of speech, as the first step of text mining.

3.2.2 Character and Word: Frequency

The frequency of occurrence of each character and word in the input text is analyzed to assess the characteristics, genre and theme of the text.

編號	字詞	頻率	頻率百分率
18	的	30	5.79%
12	問	18	3.47%
11	學	17	3.28%
63	為	10	1.93%
15	是	8	1.54%
48	他	8	1.54%
59	媽	8	1.54%
65	麼	8	1.54%
10	「	7	1.35%
13	」	7	1.35%
14	這	7	1.35%
67	從	7	1.35%
142	你	7	1.35%
5	有	6	1.16%
23	開	6	1.16%
64	甚	6	1.16%
105	個	6	1.16%

圖III：模擬詞頻統計

3.2.3 Character and Word: Analysis of Theme or Genre

With statistics of word frequency shown in Section 3.2.2, the focus of the text can be captured. Utilizing a word list by themes, some themes of most frequent words will be used to infer the theme and genre of the text. For instance, if more than 10% of words in the text are adjectives about colours, the text may probably be a passage describing landscape or describing colours.

3.2 Identifying Grammatical Items

3.3.1 Sentence: Grammatical Items (Sentence Patterns, Compound Sentences,

Punctuations and Some Rhetorical Devices)

The grammatical items present in the input text will be scanned and searched based on their unique features according to "大雅圖書中國語文科語文知識重點分類". The level of difficulty of the text will be determined based on the dispersion and level of difficulty of the grammatical item.

3.4 Big Data

3.4.1 User Input

This platform is expected to exhibit the ability to "learn". Based on comparison of data with records of inputting, editing and adjusting by different users, identification of level of difficulty, theme and grammatical items will be adjusted.

3.4.2 Statistics

Apart from the stated project aims, it is hoped that a digital depository of Chinese texts can be developed. By utilizing the statistical data collected from this platform, such as number of view, number of use, grades of users, demographics of users, etc., the analysis of text and use of text can be improved.

4. Categorization of Grammatical Items

4.1 Hierarchy

Grammatical items in Chinese language has been categorized into five domains, which are character (字), word (詞), sentence (句), paragraph (段) and passage (篇). The codex of grammatical items in Chinese language is made up of five layers, namely, (1) Domain, (2) Area, (3) Division, (4) Learning Objectives, and (5) Level Of Difficulty.

Referring to Appendix 4 - 大雅語文 語文知識重點分類,
(1) Domain is labelled in green;
(2) Area is labelled in blue;
(3) Division is labelled in yellow.

4.2 Example

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

5. Analyzing Text of Different Grades

5.1 Codex

The words and phrases in the text will be categorized by the first four layers of codex, i.e. (1) Domain, (2) Area, (3) Division and (4) Learning Objectives. For instance the phrase「紅色」and「紅彤彤」in the text will be recorded in the following format.

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

Two of the attachments are provided for more examples and further illustration.
Appendix 5 - 繪本《做自己最快樂》嗒嗒版 頁4至8
Appendix 6 - 繪本《做自己最快樂》魁魁版 頁9至14

5.2 Statistical Data

Apart from analyzing the learning progression of each grammatical item, it is hoped that the number of different grammatical items present in the text can also be statistically analyzed to provide information about their frequency of use in relation to level of difficulty.

5.3 Discussion

The following are obstacles which may be encountered in this project.

5.3.1 Comparison Words (比喻喻詞) and Conditional Sentences (假設複句)
Some grammatical items have similar features. For instance, the character "若" is both a correlative conjunction for conditional sentences (假設複句) and a comparison word (比喻喻詞) for simile. There may be difficulty in identifying the function of this specific character in the passage.

5.3.2 Differentiation between Paragraph (段) and Passage (篇)
Features of grammatical items related to paragraph (段) and passage (篇) are more subtle and rely on analysis of a larger amount of characters, words and sentences. The algorithm to identify these items from the text is yet to be developed.

5.3.3 Sampling Methods for Analysis
Any sentence can be divided into the structure of character (字), word (詞), sentence (句), paragraph (段) and passage (篇) and there are multiple ways to divide each sentence. The large number of permutations and combinations may pose an obstacle to the analysis.

6. Textual Analysis

6.1 Case 1 - Character-based or word-based

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」。

學問學問，既要學又要問。學與問是相輔相成的，只有在學中問，在問中學，才能求得真知。我們從小養成了勤學好問的習慣，就好比插上了兩隻強健有力的翅

Microsoft Office User
句：複句：並列複句（9. 既……又）

6.2 Case 2 - Rule-based

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 Case 3 - Determined by the Editor

As in the sentence "人們常把有知識說成「有學問」" in Appendix 7 - 文本的分析方法:教科書文章:〈學與問〉, the phrase "有學問" consists of the words "有" and "學問". This phrase will not be in the word list as an individual item.
Therefore, the option "Determined by the Editor" has to be present in enabling editors to compile a list of exceptional cases manually.

6.4 Case 4 - User Input

Like Case 2, there has to be the function of tagging for user input. After uploading the text, the user can submit information about the text by given multiple options and direct input, such as title, author, theme, genre, source, grade, etc.

7. Assigning Exercises Based on Analytical Data

7.1 Level of Difficulty: Key Stages and Grades

The level of difficulty of the text is identified with the use of character count and word lists to assign suitable supplementary exercises and learning materials for the text.

(A) 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)

(B) 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

8. Appendix

Appendix 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

Appendix 2 -〈小學文章篇幅表〉

中文科閱讀理解文章字數一覽

年級	教育局	教科書			
	TSA	啟思	現代	朗文	新領域
一	/	107-117	38-81	39-67	≦150
二	/	166-197	130-195	133-181	151-250
三	450-550	313-402	281-318	207-298	251-400
四	/	302-494	318-340	395-510	401-550
五	/	620-689	470-557	491-540	551-650
六	700-1000	639-661	570-639	456-620	≦651

Appendix 3 -〈國教院三等七級詞表〉
https://drive.google.com/file/d/1Dfa34ydd_8sPzhwM0E9K3PqZYPjDIRW-/view?usp=sharing

Appendix 4 - 大雅語文 語文知識重點分類
https://drive.google.com/file/d/1rMTyPclcdR2pavvvN3gU6GXqA_nHlvi4/view?usp=sharing

Appendix 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

Appendix 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

Appendix 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