國立交通大學

多媒體工程研究所

碩 士 論 文

以 CLEC 語料庫為基準設計實作與改良一個英語文法檢查系統

The Design and Implementation of an Enhanced English Grammar Checker Based on Chinese Learner English Corpus

研 究 生：阮慕芬

指導教授：陳登吉　教授

中 華 民 國　九 十 九　年 八 月

以 CLEC 語料庫為基準設計實作與改良一個英語文法檢查系統

The Design and Implementation of an Enhanced English Grammar Checker Based on Chinese Learner English Corpus

研 究 生：阮慕芬　　Student：Un Mou Fan Fanny

指導教授：陳登吉　　Advisor：Deng-Jyi Che

國 立 交 通 大 學

多 媒 體 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Multimedia Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

August 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年八月

以語料庫為基準設計與實作一個英語作文錯誤檢查系統

學生：阮慕芬　　指導教授：陳登吉 博士

國立交通大學多媒體工程研究所

摘要

在現在的社會中，小學生便開始學習英語了，而且英語的檢定考試越來越普及、歐盟委員會亦提出將英語、法語、德語作為歐盟“共同專利體係”的官方語言，學習英語已蔚為風行。因應這些發展，學生從小就開始學習英語，以及從小就開始參加英語相關的檢定考試。可是，在學習英文時，以英文作文為例，當學習者在完成一篇作文的時候，只能以人工的方式去批改。人工方式批改作文，不但費時，而且因為作文的批改沒有標準答案，有可能出現評分標準不一的情況。即使目前有少數系統可以批改作文，正常率還是有待提升。

因此，本研究試著去實作一個低出錯率、高檢查率的文法檢查系統。希冀先以基本文法規則為設計藍本，設計出不同的文法檢查模組後，再以 CLEC 語料庫的錯誤類型分類去增加模組的反向的規則並加以驗證。期望以一致的角度，先將作文當中的拼字，文法錯誤的情形清楚的標示出來，以提供閱卷者在閱卷時評分的參考。

另外，對我們華人來說，由於英語不是我們的母語，我們在文法和習慣用法上也會發生一些與以英語為母語的學習者不同的錯誤。這些錯誤不一定能被一般以英語為母語的英語檢查系統檢查出來。因此我們希望建立一個專門以華人為背景的英語文法錯誤檢查系統。

關鍵字：語料庫、文法檢查、英語、作文、學習者。

The Design and Implementation of an Enhanced English Grammar Checker Based on Chinese Learner English Corpus

Student：Un Mou Fan Fanny　　Advisor：Dr. Deng-Jyi Chen

Institute of Multimedia Engineering College of Computer Science
National Chiao Tung University

Abstract

Nowadays, elementary students start to learn English, taking English test becomes more and more popular, and The European Union proposed to make English, French and German as the official language of the "Common Patent Policy" of the EU patent. It is a trend for people to learn English due to the above reasons. Listening, Speaking, Reading ad Writing are the four basic skills in learning a language. But it is not easy for people to learn how to write. After wring a passage or a composition, the only way to check its correctness is to check it manually by a teacher. Checking a composition is time-consuming. And since there is no standard answer to the composition, the grading standard way vary. Even though some system can find the grammar errors in passages, their correctness still need improvement.

Therefore, we try to implement a grammar checker with low error rate and high error checking rate. We first take the basic grammar rules as the base to design several modules with different functions. Then we continue adding other rules to those modules according to the error classification of CLEC. Finally we examine our modules with passages in CLEC. By using the corresponding approaches, spelling errors and grammar errors can be marked and served as references for the graders.

Also, for Chinese people, since English is not out mother tongue, we will make

some errors that native speakers would not. And these errors may not be found out by those grammar checkers which were implemented by native speaker. Therefore, we hope to implement an English grammar checker which was specially designed for Chinese.

Keyword: corpus, grammar checker, English, composition, learners

誌 謝

謝謝指導教授陳登吉老師與曾建超老師兩年來的照顧。兩位老師除了在學術上一直幫助我們，教會我們專業智識外，平常也很關心我們的生活，又與我們分享很多寶貴的人生經驗，教會我們待人處世的態度，真是讓我們獲益良多。謝謝您們！

另外也要感謝快樂 613 的大家！謝謝實驗室學長學弟學妹平常的照顧，大家相處這一兩年中，常常一起修課，一起吃飯，一起玩樂，謝謝你們陪我走過這一段既歡樂又痛苦的時光。當然少不了戰友們詩婷和尹暄了。我們一直一起奮鬥，一起解決問題，一起口試。最後我們也一起畢業了！真是高興！

最後要感謝的是男友星閔，在這段期間一直支持我，聽我訴苦抱怨，一直包容我在遇到問題時的壞脾氣，當兵放假時還要一直陪我在圖書館用功，真是辛苦你了。謝謝！

最後的最後，當然是要感謝爸爸、媽媽、妹妹們了！謝謝爸爸、媽媽一直默默的關心，雖然遠在澳門卻常常打電話來關心我的近況和給予支持。還有跟動物園裡面團團圓圓一樣可愛的楹楹和君君，感謝楹楹常常提供君君鬧的笑話給我聽，也謝謝君君常常鬧笑話。因為有妳們我才能天天都保持開朗的心情去面對學業上的壓力。當然也謝謝楹楹一直幫我測試我的系統，回報 bug 給我，讓我順利完成我的論文，君君熬夜幫我把論文校正也幫了我一個大忙！謝謝我的家人！我愛你們！

謝謝各位幫助過我，跟我說過加油或是默默在心中支持我的人。謝謝你們。

目錄

圖目錄

表目錄

一、緒論

1.1 研究動機與目標

在現今的社會中，英語是國際性的語言之一。是其中一種與國際接觸的工具。在國家之間的溝通：不管商業上、教育上、社會文化交流上，英語都扮演著很重要的角色。2010 年七月，歐盟委員會亦提出將英語、法語、德語作為歐盟“共同專利體係”的官方語言。所以學習英語已經是一件刻不容緩的事情了。

教育部為了提升台灣人民的英語能力，強化國際競爭力，在近年來一直積極推行英語相關的教學。自民國 89 年開始推出「全民英語能力分級檢定測驗」至今已過了十年，總報考人數已超過 380 萬人次[9]。在這樣的情況下，英語學習與英語考試已經是一種全民運動了。

在英語的學習中，包括了聽、說、讀、寫。其中“英文作文”是大家學習的重點之一，亦是全民英檢考試的中一個項目。通常學習者在自在練習的環境中練習寫作，或是在考試時所完成的寫作測驗，最後都需要人工的方式來閱卷。人工閱卷比較費時，而且因為寫作測驗沒有標準答案，結果可能會因為不同閱卷者的要求而略有出入。所以我們希望可以提供一個英文文法檢查工具。讓學習者或考試者在英語寫作完成時可以把英文文法的錯誤先找出來。這樣一來可以減輕閱卷者的工作量，增快他們的閱卷速度；而且因為文法部分的錯誤已經先標示出來了，評分標準的差異可以縮小、一致性也能提升。

目前市面上存在著一些英文文法檢查相關的工具，可是其正確性還是有待提升。加上這些工具大多數是以英語為母語的使用者所開發的。可是英語系國家與華人的英語學習背景不同，一般以英文為母語的國家所制作的文法檢查工具，針對非英語系國家的文法檢查比較沒那麼有效。

所以我們利用一個華人的學習者語料庫－Chinese Learner English Corpus (CLEC)[6][7][8]。針對 CLEC 中的已經整理好的錯誤類型及 CLEC 中大量的學習者錯誤句型加以分析，協助我們設計文法模組以及驗證文法模組的正確性。同時亦把華人的常用慣用語錯誤標示出來，加入資料庫中。以達到我們實作一個能提升錯誤檢查率、降低出錯率、及能找出部分華人慣用語錯誤的適合華人所使用的英文作文文法檢查系統的目的。

1.2 研究範圍

本系統的研究範圍主要是在對詞性標記的工具(Apple Pie Parser)跟 CLEC 語料庫的研究。APP 的部份，我們主要是找出 APP 的限制，即 APP 所用到的 Penn Tree Bank 的不足的地方。再重新定義一些新的 tag 去取代原來的 tag 或新增一些新的 tag。CLEC 的部份，因為我們在設計一個文法錯誤的模組，所以我們主需要完整而清晰錯誤分類；同時，也需要大量的英語作文為測試樣本，以驗證文法模組，這也是 CLEC 對我們重要的地方。

1.3 研究方法與步驟

我們的系統是建構在實驗室學長顏佐宇學長的系統上。所以我們一開始先找大量的範例去測驗原系統，把需要改進的模組與不足夠的地方記錄起來。之後再以基本文法規則設計我們的模組，再以部份的 CLEC 樣本作驗證，總計其錯誤率，並根據 CLEC 的反例把反例的規則新增到文法模組中。步驟如下：

1. 了解原系統所能檢查出的文法、拼字錯誤類型與原系統的限制。
2. 以基本文法模組為樣本，設計文法模組。
3. 以系統之文法模組，與 CLEC 語料庫相互驗證。
4. 以 CLEC 語料庫分類方式為樣本，新增反向規則到文法模組。
5. 統計與 CLEC 語料庫相互驗證之後的錯誤種類及數量。

1.4 章節概要

本論文共分為八個章節，依序如下：

第一章 敘述本研究的研究背景、動機、目的、以及研究方法與步驟

第二章 系統的技術探討，對本系統所用到的一些工具，如 APP、APPP、語料庫、CLEC 等作一些背景的技術探討

第三章 介紹第二章所使用的工具有什麼限制。為了克服這些限制，對系統的元件做了些什麼改變。以及對各元件的設計

第四章 本章為系統實作的介紹。敘述各元件跟模組的實作方法

第五章 比較原系統與本系統的正確性，本系統與市面上的文法檢查器的檢查率

第六章 系統的限制說明

第七章 系統畫面展示

第八章 說明本研究的結果與本系統未來可延伸開發或改進的地方

二、系統技術探討

本系統的目的是以基本文法規則與 CLEC 語料庫建構出一個英文作文文法檢查系統。本系統包括了拼字的檢查、文法的檢查與習慣用語的檢查。以下章節我們先對目前有公開 open source 的文法檢查器 – After the deadline 做相關的背景研究，再對每一個所使用的工具作一個技術的探討。

2.1 背景研究

After the deadline[15]是其中一個 open source 的文法檢查器，他的前身是 Language Tool -- OpenOffice 的一個 plugin。After the deadline 是一個 rule based 的文法檢查器，他先對句子的詞性作標記，標記的方法是以 mixed statistical
和 rule-based approach。如果一個 word 是已知的，而已經有關聯的 tag，tagger 會去找以下機率中最大的 tag

$$P(tag_n|word_n) * P(tag_n|tag_{n-1}, tag_{n-2})$$

如果那個 word 是 unknown 的，另外一個 model 會用來 tag 這個 word。這個 model 會利用到此 word 的最後 3 個 character。

當 tagger 把 word tag 好之後，這系統會使用到一個 rule engine。當一個 phrase match grammar rule 時，便會回傳 matched rules 的轉換型態。

以下爲一個 tagger 與 rule base engine (grammar checker)的使用方法，以下句爲例:

```
I wonder if this is your companies way of providing support?
```

其中錯誤的位置在 “your companies way” 應該是 “your company’s way”
一開始系統先對此句句子 tag 上詞性:

```
I/PRP wonder/VBP if/IN this/DT is/VBZ your/PRP$ companies/ NNS way/NN
of/IN providing/ VBG support/NN
```

當 tag 上詞性後便會應過很多的 rule，其中已下這條 rule 便會把文中的錯誤找出來:

```
your .*/NNS .*/NN
```

這個 rule 會去找一個 phrase 是以 your 開頭的，之後跟到一個 NNS 再跟一個 NN。如果發現這種 phrase 的 pattern(即 your companies way)，即會回傳以下的 suggestion:

your \1:possessive \2

這個 suggestion 爲 your 的位置後面的第一個字跟第二個字出錯了，其中 your 後面的第一個字應該是一個所有格代名詞。

在 After the deadline 的方法中，必需要先把所有錯誤的 phrase pattern 找出來，再以 phrase 去找出句子中的文法錯誤。本論文所使用的方法是以一整句 sentence 去檢查，在 module 設計上參考了一些此系統的 grammar rule。

2.2 GNU Aspell

在一個作文錯誤檢查工具中，找出錯誤的第一關就是要把錯誤的拼字找出來。雖然找出拼字錯誤的方法很簡單，只要把一本大型字典中的所有單字都輸入到資料庫，當使用者所輸入的句子中的單字不在資料庫，即發生拼字錯誤。拼字錯誤的檢查雖然簡單，卻非常地重要。因為這是作文錯誤的最基本的地方。我們希望系統除了把錯誤的拼字找出來之外，還可以給使用者一些建議的字。

所以我們選擇了 GNU Aspell[1]的拼字錯誤檢查工具。Aspell 是 Ispell 的改良板，他改良了 Ispell[2]的一些缺點，讓使用者使用起來更方便。原本 Ispell 是一個支援很多歐洲語系的 Unix 程式。它可以使用在 Emacs 編輯器下，但是大部分的人發現，在發現錯誤後的建議字上並不十分理想，所以開發者 Kevin Atkinson 才開始致力研究於 Aspell，目前 Aspell 也是由 Kevin 在維護。除了 Aspell 的建議字比 Ispell 準確外，Aspell 還允許每一個使用者有自己的個人字典。所以我們選擇使用 Aspell 為本系統的拼字檢查工具。

很多其他的軟體也使用 Aspell 當拼字檢查工具，如: Opera、Notepad++。Aspell 可以以一個獨立的程式去執行，亦可以以 function call 的方式去執行。在本系統中我們使用的是 function call 的方式。
以下為使用 scenario:

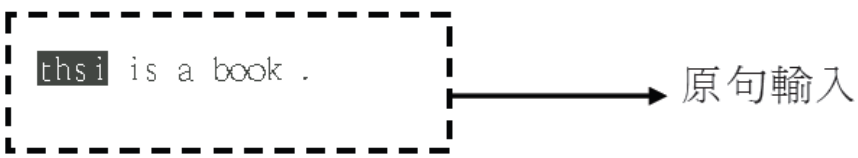

圖 1 Aspell 使用介面

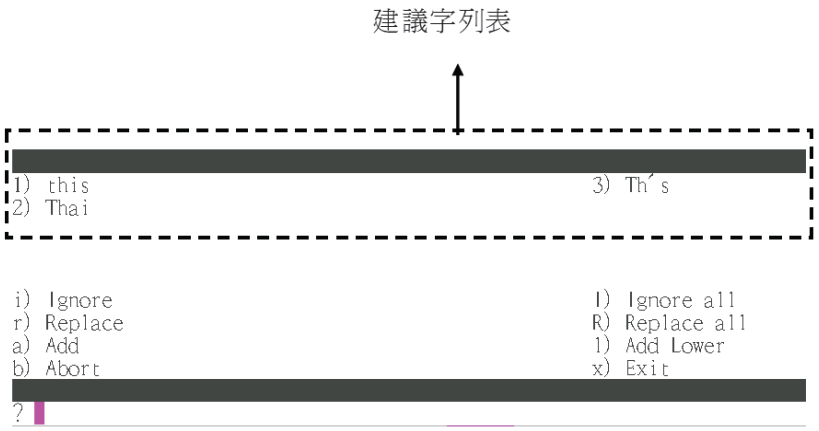

圖 2 Aspell 使用介面，建議字列表

使用者輸入了錯誤的字: this。 Aspell 程式會列出 this, Thai, Th’s 等建議字。

2.3 Apple Pie Parser

在拼字檢查結束之後，進入文法檢查之前，我們要先對句子作詞性與結構的標記。之後的文法檢查才能針對已經標記好詞性與結構的句子作一些檢查。

我們對句子作詞性的標記（POS(part of speech) tagging）所使用的工具是 Apple Pie Parser(APP)。此 parser 是由美國紐約大學 Satoshi Sekine 與 Ralph Grishman 所開發的。至於詞性的部分則使用到美國濱洲大學的 Penn Tree Bank。

圖 3 Penn Tree Bank 詞性標記表

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

此 parser 是一個 bottom up probabilistic chart parser。以 best first search 的方法找出最高分數的 parse tree。也就是說 APP 是以 Penn Treebank 中所統計的文法規則以及字典，利用演算法所找出來的最佳解析詞性。

以下為 APP 的執行畫面:

```
[#> using app to parse : [She is a beautiful girl . ]
[#> app parsed : (S (NPL (PRP She)) (VP (VBZ is) (NP (NPL (DT a) (JJ beautiful)) (NPL (NN girl) (. -PERIOD-)))))
```

圖 4 APP 執行畫面

當使用者輸入了句子: She is a beautiful girl.
APP 便會把句子的最佳詞性 tag 出來。好讓之後的文法檢查模組能對句子的詞性作一些文法上的檢查。要注意的是，無論句子中的拼字以及文法是否正確，APP 都會將句子做上標記。

2.4 Apple Pie Parser Parser (APPP)

當句子作完詞性的標記之後，我們要對句子作結構的標記。為了系統可以清楚的辨別各個詞性與句子結構，我們將剛才 APP 輸出的結果，經過 APPP 再標記一次，把我們想要的資料結構加註進去。我們以 APP 切出 token，以 lex&yacc 的方式，也就是 APPP 中輸出成具有 XML 語法標記的樹狀結構標記句型：

1. LEAF：tree 的末端，格式為(tag 單字)
   例：(NN book)
2. NODE：可以是 LEAF 或(tag NODE_LIST)
   例：(NPL(DT This)(NN book))
   (tag (NODE_LIST))
3. NODE_LIST：NODE 或 NODE_LIST+NODE
   例：(DT This)(NN book)
4. TREE：最頂端的結構。
   例：(S)

我們也可以藉由樹狀圖的表示方式來說明上列的結構，以及相關的標記方式，由下兩圖可以更清楚的呈現此部分定義的結構。

```
01 <Tree>
02 <CNode tag="S" >（簡單敘述子句）
03  <CNode tag="NPL" >（最小名詞片語）
04   <CLeaf tag="NNP" word="Mary"  index="0" />（單數專有名詞）
05  </CNode>
06  <CNode tag="VP" >（動詞片語）
07   <CLeaf tag="VBZ" word="is"  index="1" />（第三人稱單數動詞）
08   <CNode tag="NPL" >（最小名詞片語）
09      <CLeaf tag="DT" word="a"  index="2" />（限定詞）
10      <CLeaf tag="JJ" word="beautiful"  index="3" />（形容詞）
11      <CLeaf tag="NN" word="girl"  index="4" />（單數名詞）
12   </CNode>
13  </CNode>
14  <CLeaf tag="-PERIOD-" word="."  index="5" />（句子的結束）
15 </CNode>
16 </Tree>
```

圖 5 Apple Pie Parser 結果輸出

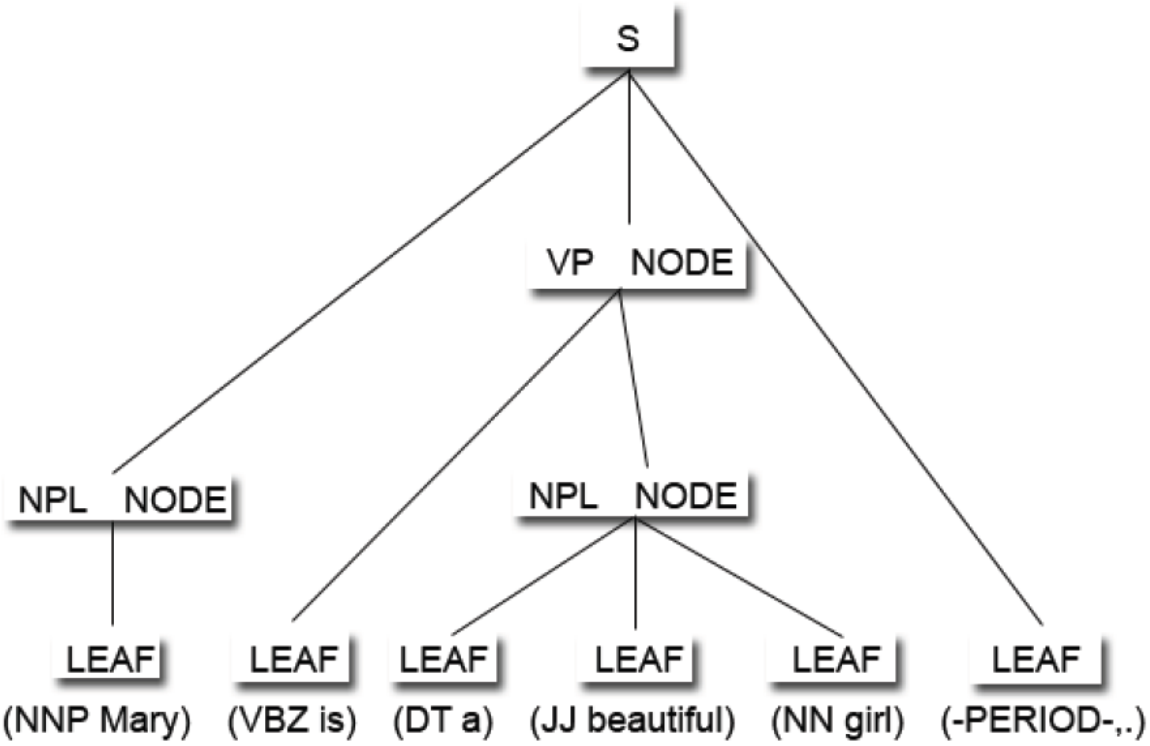

圖 6 Apple Pie Parser 結構

從上面 2 個圖可發現，每一個字都是 leaf，一個以上的 leaf 可組成很多一個 node。這些 node 又可跟其他的 leaf 或 node 再組成另外一個 node。所有 node 合起來就是一個 tree 了。

2.5 語料庫

「整體性的語言資料」，意思是指人類口語或文字紀錄的彙編。語料庫指按照一定的語言學原則，運用隨機抽樣的方法，收集自然出現的連續的語言運用文本或話語片段而建成的具有一定容量的大型電子文庫。

黃希敏教授的論文裡也將語料庫依照內容，分為以下四類：

1. 單一語料庫：
這是為了某種特定的研究所做的語料，收集過程需要非常嚴謹，收入該語料庫的每個字句都必須具有代表性。
例如：文言文語料庫。

2. 複合語料庫：
此資料庫是將兩種或數種語料「集大成」，不過在收集時需要在類型之間加以區隔，讓研究者自行選取某單一類型來做研究或某幾個跨類型語料來做比較。
例如：多語(multilingual)語料庫：包括兩種或兩種以上語言，提供翻譯研究、語言對比研究。

3. 開放性語料庫：
國內中研院製作的現代漢語開放性語料庫「中央研究院平衡語料庫」，可不斷擴張，提供從事語言文字分析的人士使用。

4. 學習者語料庫：
廣泛蒐集學習者的語料，提供外語教學界研究。

對於系統而言，文法錯誤的檢查模組必須以大量的錯誤樣本測試其正確性及完備性。在這樣的情況之下，上述之語料庫就是一個最好的樣本，可以提供給我們系統作為驗證測試之用，以便在撰寫程式碼的時候有所遵循。

我們的系統的使用對象是英語的學習者，所以我們選擇的是學習者語料庫。

2.6 學習者語料庫

Leech[9]也曾在其論文中提到，建立學習者語料庫的目的是：

1. 比較學習者語料庫 LC（Learner Corpus）和欲學習語言為母語的語料庫 ECNS（English Corpus of Native Speakers），比較其中使用過多或過少的語詞。
2. 學習者的母語對於在使用欲學習語言時的影響程度。
3. 學習者使用新語言時，在哪些方面能夠達到所欲表達的目標，在哪些方面無法達成。
4. 學習者在哪些方面無法達到所欲表達的目標，而需要幫助。

從上面的第二點我們可以找出一些使用者的英語習慣用法錯誤的地方，並針對這些地方去設計習慣用法錯誤的檢查規則。

目前語料庫種類繁多，如果我們僅針對學習者語料庫，而且將範圍限制在英語學習者語料庫來看，具代表性的大概有以下幾個:

表 1 英語學習者語料庫列表

語料庫名稱	詞數	研究單位
ICLE（International Corpus of Learner English）	200 萬	比利時 Louvain-La-Neuve 大學
JEFLL（Japanese EFL Learner）Corpus	50 萬	日本明海大學
中國學習者語料庫（CLEC）	100 萬	廣東外貿大學、上海交通大學
英語學習者口語語料庫(COLSEC)	5 萬	上海交通大學
香港科技大學學習者語料庫(HKUST Learner Corpus)	2500 萬	香港科技大學
中國英語專業語料庫(CEME)	148 萬	南京大學
中國英語學習者口語語料庫(SECCL)	100 萬	南京大學
國際外語學習者英語口語語料庫中國部分(LINSEI-China)	10 萬	華南師大
碩士寫作語料庫(MWC)	12 萬	華中科技大學

2.7 Chinese Learner English Corpus

從上一節的不同的學習者語料庫中，我們選出了 CLEC 語料庫作為本系統的文法模組設計工具。選舉此語料庫的原因請見 Ch 3.4。下面我們要介紹 CLEC 語料庫的一些相關背景。

2.7.1 CLEC 的收集文本對象

樣本的收集對象都是學生，分為 5 大類:

表 2 CLEC 語料分佈表

類型	詞次
ST2	208088
ST3	209043
ST4	212855
ST5	214510
ST6	226106
總計	1070602

1. ST2：中學階段，主要是高中生。
2. ST3：大學英語 4 級，大學 1～2 年級非英語科系學習者。
3. ST4：大學英語 6 級，大學 3～4 年級非英語科系學習者。
4. ST5：英語專業科系 1～2 年級學習者。
5. ST6：英語專業科系 3～4 年級學習者。

CLEC 的收集樣本對象，從國一到大學一般生, 甚至英語科系學生都有涵蓋。每類學生二十萬詞, 約 38000 句句子。收集的樣本大多為作業的內容

2.7.2 CLEC 的錯誤分類與標記

CLEC 中的內容主要為學習英語之學生的作文，CLEC 的建立者把所有樣本的錯誤找出來，並標出其中錯誤位置及錯誤類型。錯誤類型共有 11 大類、61 小類。如下表：

表 3 言語失誤分類表

詞形		動詞片語		名詞片語		代詞	
代碼	類型	代碼	類型	代碼	類型	代碼	類型
fm1	Spelling	vp1	pattern	np1	pattern	pr1	Reference
fm2	word building	vp2	set phrase	np2	set phrase	pr2	anticipatory it
fm3	capitalization	vp3	agreement	np3	agreement	pr3	Agreement
		vp4	finite/non-finite	np4	case	pr4	Case
		vp5	non-finite	np5	countability	pr5	wh-
		vp6	tense	np6	number	pr6	Indefinite
		vp7	voice	np7	article		
		vp8	mood	np8	quantifiers		
		vp9	modal/auxiliary	np9	other determiners		
形容詞片語		副詞		介詞片語		連詞	
代碼	類型	代碼	類型	代碼	類型	代碼	類型
aj1	pattern	ad1	order	pp1	pattern	cj1	pattern
aj2	set phrase	ad2	modification	pp2	set phrase	cj2	set phrase
aj3	degree	ad3	degree				
aj4	-ed/-ing confusion						
aj5	predicative/attributive						

詞語		搭配		句子	
代碼	類型	代碼	類型	代碼	類型
wd1	order	cc1	noun/noun	sn1	run-on sentence
wd2	part of speech	cc2	noun/verb	sn2	sentence fragment
wd3	substitution	cc3	verb/noun	sn3	dangling modifier
wd4	absence	cc4	adj/noun	sn4	illogical comparison
wd5	redundancy	cc5	verb/adv	sn5	topic prominence
wd6	repetition	cc6	adv/adj	sn6	Coordination
wd7	ambiguity			sn7	Subordination
				sn8	structural deficiency
				sn9	Punctuation

2.7.3 CLEC 的實例說明：

下表是 CLEC 的一個範例

表 4 CLEC 範例文章

1.	<ST 2> <SEX ?><Y ?> <SCH GDWYWMDXFSWYXX> <AGE ?>
2.	<WAY ?><DIC ?> <TYP 2> <TITLE A Shop>
3.	There is a fruit shop near my home. its [fm3, 1-] name is Many [fm1,-]
4.	fruit shop. It's not very big, but it's clean and bright. There is [vp3,-2]
5.	two women working in it. The women are very friendly [wd2, 1-] and
6.	busy. Every buyer comes to the shop, they both give them smiles
7.	and say [sn9, s] Hello. Can I help you? So every buyer comes to
8.	here are very satisfy. [sn8,s]The shop has many different kinds of
9.	fruit. There are apples [sn9, s]oranges bananes [fm1,-]pears
10.	bananas many [wd6]. [sn8,s] So it aways [fm1,-] give [vp3, 1-] the
11.	buyers a good time. I like the shop very much.

範例一開始是一些樣本的作者的資訊:
<ST 2>: 學生的程度在 ST2 - 中學程度
<SEX ?>: 性別不詳
<SCH GDWYWMDXFSWYXX>: 此為學校代碼
<TITLE A Shop> :文章 title 為 A shop

以下是 CLEC 中錯誤標記的方式，分為錯誤位置、錯誤類型與錯誤範圍標記。錯誤位置是在句子當中使用方括號標記在錯誤單字的後方，括號中左邊標示錯誤類型，右邊則標示出錯誤範圍，如第 3 列後方：its [fm3, 1-] name is Many [fm1,-] fruit shop.，即表示此句中 its、Many 兩個字是錯誤位置，而錯誤類型分別是 fm3（大小寫錯誤）、fm1（拼字錯誤）。

至於針對錯誤範圍標示部分，有以下幾種分類，分別說明如下：

1. 錯誤字前方導致錯誤：
第 5 列：The women are very friendly [wd2, 1-] and busy.
括號中錯誤範圍標示的 1 在－號前，表示是 friendly 是錯誤，是因為往前數一個字的位置的字導致該字發生錯誤，也就是因為 very 也是副詞，而導致錯誤。

2. 錯誤字後方導致錯誤：

第 4 列：There is [vp3,-2] two women working in it.括號中錯誤範圍標示的 2 在一號後，仍表示是 is 發生錯誤，但原因是後方的兩個字 two women 所導致的。

3. 本身錯誤：

第 3 列：its [fm3, 1-] name is Many [fm1,-] fruit shop.此句所示的第二個括號，在一號前後並沒有任何數字標示，此即表示並沒有其他的字導致 Many 發生錯誤，是這個字本身就拼錯，並非原本作者所要使用的字。

4. 句中結構或標點錯誤：

第 9 列：There are apples [sn9,s] oranges bananes [fm1,-] pears…此句的第一個錯誤標示範圍僅用一個 s 表示，此代表是句中的結構或標點發生錯誤，因此這裡發生錯誤的原因，在於表示多個名詞的時候，需要用逗點分開。

三、英語作文錯誤檢查系統設計

3.1 系統架構與流程

下圖為原系統架構流程圖，學習者輸入英文作文。系統會先把英文作文中的句子一句一句送到拼字檢查模組，之後再到 APP 標記詞性、APPP 標記結構、最後進入各個不同的文法模組作檢查。從拼字檢查到文法檢查中，只要其中一個模組檢查出錯誤，系統就會回傳錯誤結果、把錯誤位置標示出來並停止檢查。

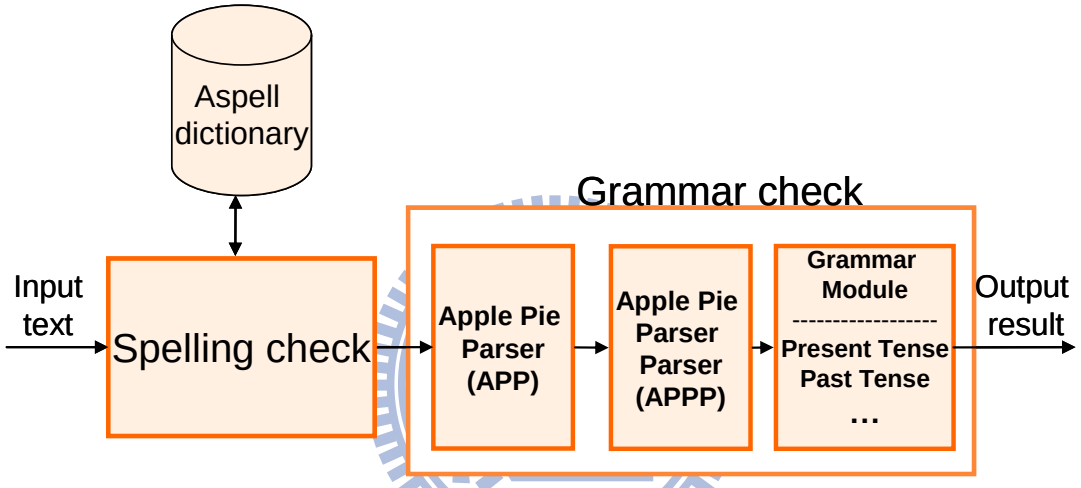

圖 7 原系統架構流程圖

可是由於原系統的錯誤檢查率跟出錯率都不是很理想，所以我們針對原系統再作改良。下圖為本論文所實作的系統架構與流程圖。我們針對原系統的限制新增與修改了一些元件。在下面的章節將會介紹技術限制、新增元件的原因、及設計考量。

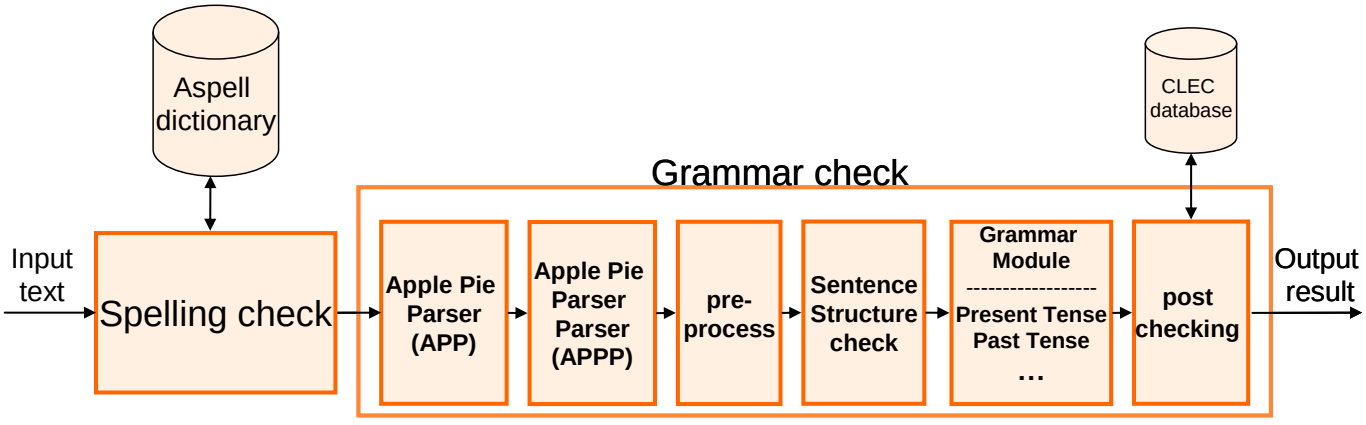

圖 8 本系統最終系統架構流程圖

3.2 技術限制

因為系統所使用的詞性標記工具 - Apple Pie Parser(APP)跟句子結構標示工具 - Apple Pie Parser Parser(APPP)有一些限制，導致文法模組檢查不完善，出現錯誤的結果。以下為工具的限制：

Penn Treebank 的詞性限制

由於 APP 的詞性標記是從 Penn Tree Bank 中取出來的，而 Penn Tree Bank 的詞性為下圖:

表 5 Penn Tree Bank 詞性標記表

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

3.2.1 詞性標記工具 - Apple Pie Parser(APP) 之限制

從上圖可看出 Penn Tree Bank 有以下的限制:

1. 不同的詞性使用相同的 tag

像是 IN 這個 tag，Penn Tree Bank 的定義為 preposition(介詞)或是 subordinate conjunction(從屬連接詞)。沒有把 2 種可同的詞性區分開來，造成文法檢查時不容易實作。

2. 名詞標記不夠細緻

從上圖可看到，名詞只有分為單數名詞/單數專有名詞 (NN, NNP)、複數名詞/複數專有名詞 (NNS, NNPS)。對不可數名詞, 單雙數同型之名詞沒有再細分。

3. 代名詞沒有區分單複數

代名詞只有一般代名詞(PRP)跟所有格代名詞(PRP$)這 2 個分類。可是不能從 PRP 或 PRP$中區分此代名詞是主格、受格，第一人稱、第二人稱，第三人稱和單數、複數。

3.2.2 句子結構標示工具 - Apple Pie Parser Parser(APPP)之限制

APPP 只有標示句子的資料結構，可是沒有再對句子的結構作一些標記，亦沒有檢查句子的結構是正確。導致原系統出現很多正確句子被標成錯誤以及錯誤的結構卻沒有找到的現象。

3.3 解決技術限制的方法

為了解決以上 2 個工具的限制，我們在系統中加入了 2 個元件: preprocess 與 spilt sentence。下面為更詳盡的說明：

3.3.1 Apple Pie Parser(APP)限制的解決方法

加入 preprocess 元件，在 preprocess 中新增一個 post-tagging 的動作。在 post-tagging 時針對上面的的 3 個 tagging 問題把 Penn Tree Bank 再細分成如下表。

1. 解決不同的詞性使用相同的 tag 的問題: 把不同的詞性區分開來

表 6 IN/SC 之詞性標記表

Preposition/Subordinate Conjunctions		
IN	Preposition	In, on, near, …
SC	Subordinate Conjunctions	While, When, Since, …

2. 解決名詞標記不夠細緻的問題:
   把名詞的標記新增 3 種類型:
   不可數名詞、單複數同型名詞、同時為可數或是不可數的名詞。

表 7 細分後的名詞標記表

Noun	
NN	Noun, singular
NNS	Noun, plural
NNM	mass noun
NNI	identical noun
NNB	Both count and mass noun

3. 解決代名詞沒有區分單複數的問題:
   把原來的 PRP 跟 PRP$延伸成下表，分成主格、受格，第一人稱、第二人稱，第三人稱和單數、複數

表 8 細分後的代名詞標記表

Pronoun	I	3rd Singular	Others
Subject	PRPsI	PRPs3	PRPs0
	I	he/she/it	we/you/they
Object	PRPoI	PRPo3	PRPo0
	me	him/her/it	us/you/them
Possessive	PRPI$	PRP3$	PRP0$
	my	his/her/its	our/your/their

3.3.2 Apple Pie Parser Parser(APPP)限制的解決方法

加入preprocess模組，在preprocess中新增一個spilt sentence元件。Spilt sentence元件是為了先把句子切成子句再把子句送到文法模組檢查。這是為了解決原系統直接把一整句句子送到文法模組檢查所造成的錯誤。因為原系統沒有加

入子句的觀念，所以會發生很多正確句子卻標示為錯誤的失誤。如下例:
正確句子: Though I am tired , I help her with the homework .
被錯誤標示成: There are more than two verbs in the sentence.

正確句子: She cooks the meal and they clean the floor.
被錯誤標示成: The tense of verbs on both sides of "and" should be identical.

上面的兩個觀念:
1. 一句句子中如果沒有連接詞的話, 只能有一個 main verb。
2. 在同一句句子中, 動詞要一致。

其實都是正確的。只是原系統沒有子句的觀念，造成了錯誤的結果。所以我們在本系統中加入了 spilt sentence 元件，把原系統的出錯率降低。

另外原系統對句子的結構檢查也有不足，原系統沒有作句子 fragmentation 的檢查。可是從 CLEC 的錯誤分類中(請見上一章)，我們可知句子的片後錯誤在 61 種錯誤分類中佔第 16 位，是前 1/3 常犯的錯誤。所以如果沒有作片段錯誤的檢查，對系統的檢查率有很大的影響。所以我們把原系統原本寫在文法模組的缺少名詞片語，動詞片語的檢查拉出來，加入檢查句子片段錯誤的功能，組成一個新的模組 – Sentence Structure Check。在進入文法檢查前先作句子結構的檢查。

3.3.3 結論

為了解決 APP 跟 APPP 的限制，我們加入了 2 個模組，3 個元件。
1. preprocess : i)post tagging ii)spilt sentence
2. Sentence Structure Check

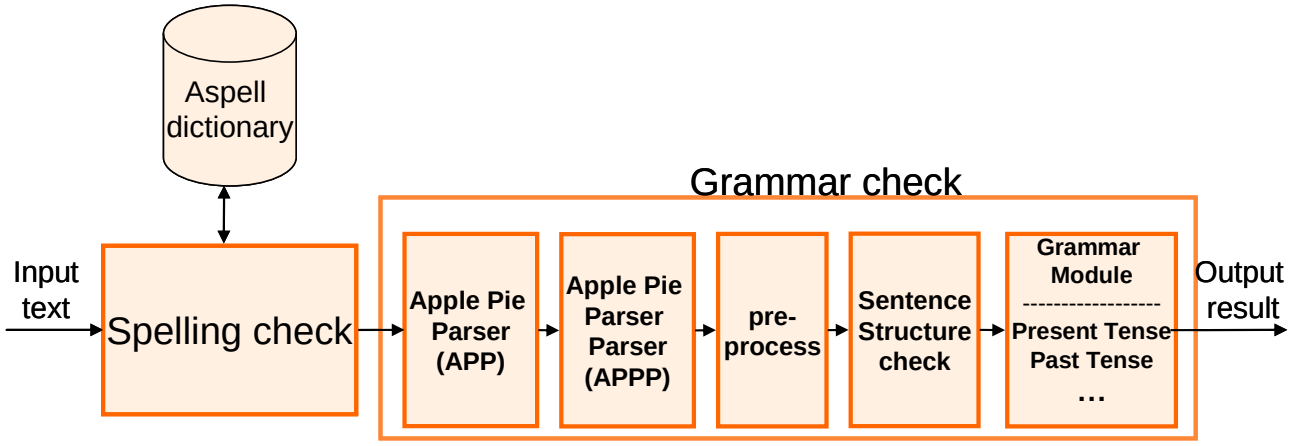

圖 9 加了 preprocess 跟結構檢查的系統架構圖

3.4 文法模組的設計

我們以 2 種方法去設計文法模組:

1. 以正向的方法去設計 module
   例如:現在式的第三人稱單數代名詞之動詞為單數動詞。
   這是根據基本文法規則的設定規則。

2. 以反向的方法去設計 module
   例如: 現在式的第三人稱單數代名詞之動詞，若為複數動詞則是錯誤。
   這種是根據錯誤的範例去設定規則。

我們首先以第一種方法去設定基本的規則，再以錯誤的句子範例去驗證模組的正確性。當發現了一些模組找不出錯誤來的錯誤句子時，修正文法模組(此為方法 2 的反向設計)，再重新驗證模組，重複以上的 loop。

流程如下：

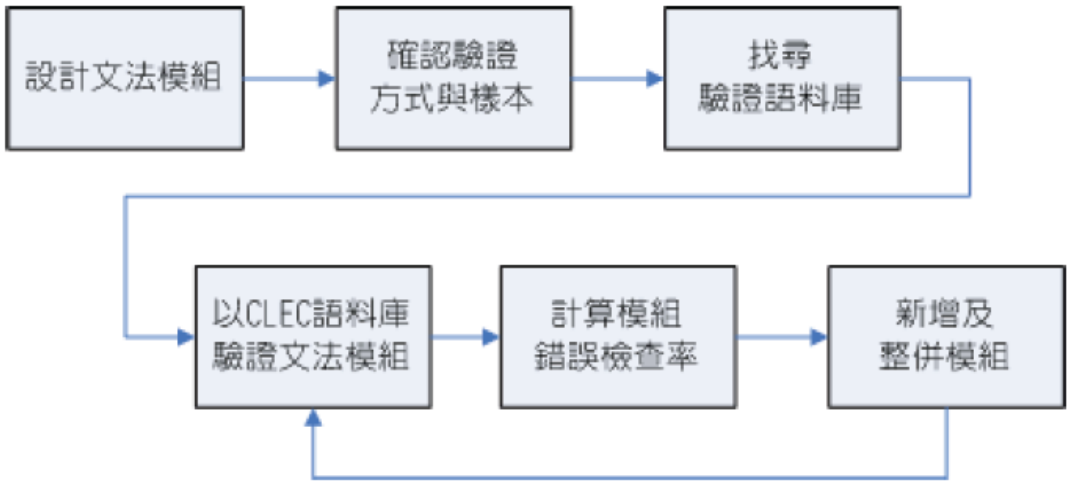

圖 10 文法模組檢查認證方式流程圖

因為以反向方法去設計模組需要很多的錯誤句子樣本，所以我們需要一個語料庫。而且因為本系統的使用對象為學習者，所以我們需要一個學習者語料庫。

學習者語料庫的種類繁多，本系統的選擇語料庫的考量為:

1. **華人語料庫製作背景**

   因為本系統的目的是實作一個符合華人使用習慣的文法檢查系統，所以收集以華人的樣本的語料庫為優先考量。

2. **內容為寫作類型**

   本系統的目的是檢查英文作文類的文章，所以口說類的或是專業分類的語料庫都不作選擇的考量。

3. **文本收集對象範圍比較大**

   本系統的使用對象為一般學習者，所以樣本的對象最好是比較廣。

我們選擇了最符合以上的條件的語料庫 – Chinese Learner English Corpus (CLEC)語料庫作為本論文設計文法模組的樣本。

3.5 系統的檢查順序

系統的檢查順序是根據 CLEC 中的錯誤分類中最常出錯的頻率高低所排例的。

CLEC 最常出現的錯誤為右圖表，類型對照表請見[章節 2.3]

從上表可知，最常出現的錯誤第一名跟第三名分別為 fm1(拼寫錯誤)和 fm3(構詞錯誤)。他們都是拼字錯誤的一種，所以我們在系統的一開始就先作拼字的檢查。

第二名 wd3 是 word substitution 的錯誤，可是他是跟 semantics 相關的。我們在此先不作檢查。

再來是 sn8(結構缺陷)和 sn9(標點符號)錯誤。所以我們完成拼字錯誤之後，就是作句子結構的檢查跟標點符號的檢查。

CLEC 中共有 61 種錯誤類型，我們的系統就是按照此錯誤類型的頻率作為檢查的順序。這樣一來，檢查出錯誤的速度便可提升了。

中國學習者最常見21種的言語失誤

類型	百分比
fm1	17.47
wd3	9.49
fm3	5.83
sn8	5.25
sn9	4.6
wd4	4.5
wd2	4.18
vp6	4.07
vp3	3.82
np6	3.72
wd5	3.31
fm2	3
sn1	2.94
wd7	2.33
vp1	2.32
sn2	2.22
cc3	2.16
np3	1.84
vp9	1.33
np7	1.11
pr1	1.06

表 9 CLEC 中最常見的錯誤

3.6 華人慣用語錯誤檢查模組設計

在所有文法模組檢查完畢之後，本論文在最後再加了一個華人常犯的慣用語錯誤的檢查。慣用語錯誤就是一些習慣的用法，這裡是指一些只有華人才會犯的習慣用法，在文法上，結構上都是正確的。可以在選詞上或是使用上卻讓其他人搞不懂意思。

慣用語錯誤的檢模組的設計主要是根據 CLEC 中的 wd 的錯誤分類中取樣出來的。像以下句子為一句慣用語錯誤的例句：

I want to do a good student.（例句一）

上句的意思是：我想做一個好學生。可是因為“做”的英文通常被寫成”“do”，所以很多學生都會直接翻把“做”譯成“do”。正確句子應該是：

I want to be a good student.（例句二）

可是在例句一中，文法上是正確的。只是在意思上外國人可能就看不懂了。

下圖為本系統最終的結構圖：

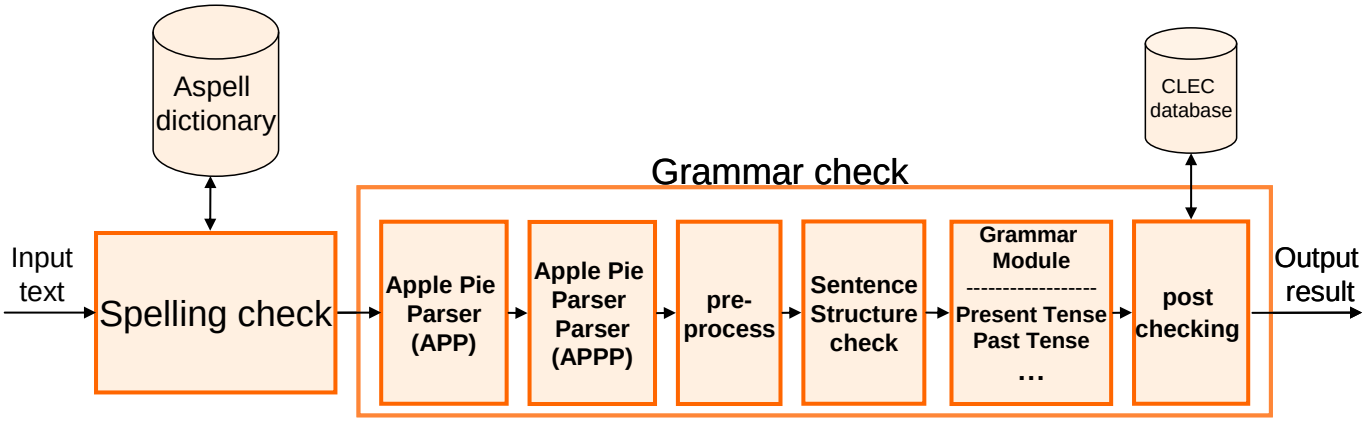

圖 11 本系統最終的系統結構流程圖

3.7 系統流程說明

本章節分別以 3 個 use cases 作範例說明本系統的檢查流程:

3.7.1 Case1: 拼字錯誤

You can improve your speed through a lot of pratice.

此句為拼字錯誤的例子。在句中把 practice 錯拼成 pratice.

系統在一開始的 spelling check 時，會把句子每一個 word 去 Aspell Dictionary 中 query。錯誤字 pratice 會回傳找不到此字的訊息給系統。系統則會回傳錯誤跟建議字給使用者。

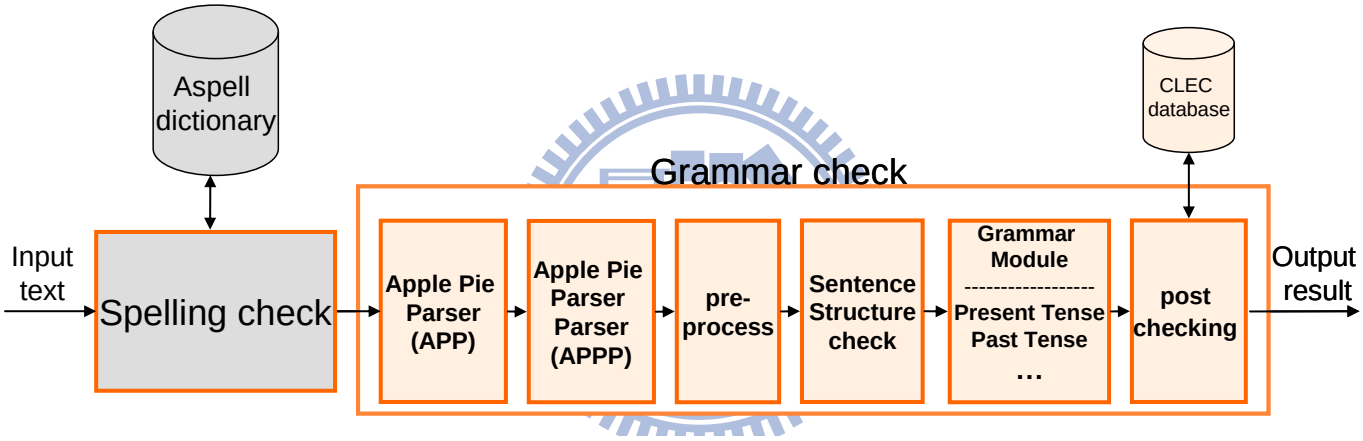

圖 12 拼字錯誤流程說明

3.7.2 Case2: 文法錯誤

She go to school everyday.

此句為文法錯誤，正確句子應該是 She goes to school everyday.

句子一開始先通過 spelling check，進入 grammar check. 首先 APP 跟 APPP 會標記句子的詞性與結構，如下:

(S (S (NPL (PRP She)) (VP (VBP go) (PP (TO to) (NP (NPL (NN school)) (ADJP (JJ everyday))))))

之後進入 preprocess 會做 post-tagging 跟 spilt sentence 的動作。此句為 simple sentence 所以不需要 spilt。

不過(PRP, She)是一個需要重新 tag 的字。所以句子結構會變成:

```
(S (S (NPL (PRPs3 She)) (VP (VBP go) (PP (TO to) (NP (NPL (NN school))
(ADJP (JJ everyday))))))
```

之後句子通過 Structure check 進入各個文法模組的檢查。因為此句為現在式的錯誤，所以會在 Present Tense 模組中回傳錯誤給使用者。

流程圖如下:

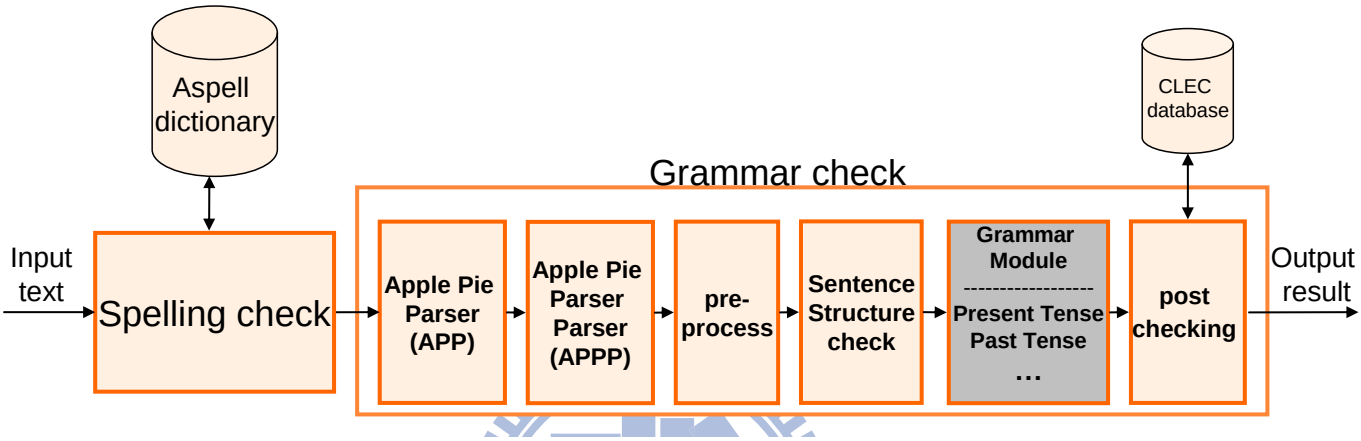

圖 13 文法錯誤流程說明

3.7.3 Case3: 習慣用語錯誤

Usually English hearing is a big problem.

此句句子拼字，文法皆正確，可是在選詞方面卻用錯了。正確句子應該為: Usually English listening is a big problem.

所以此句在 post-checking 前都會很順利地通過檢查。可是因為在 CLEC 中有標示此錯誤，所以我們可以從 database 中找出此錯誤。

流程圖如下:

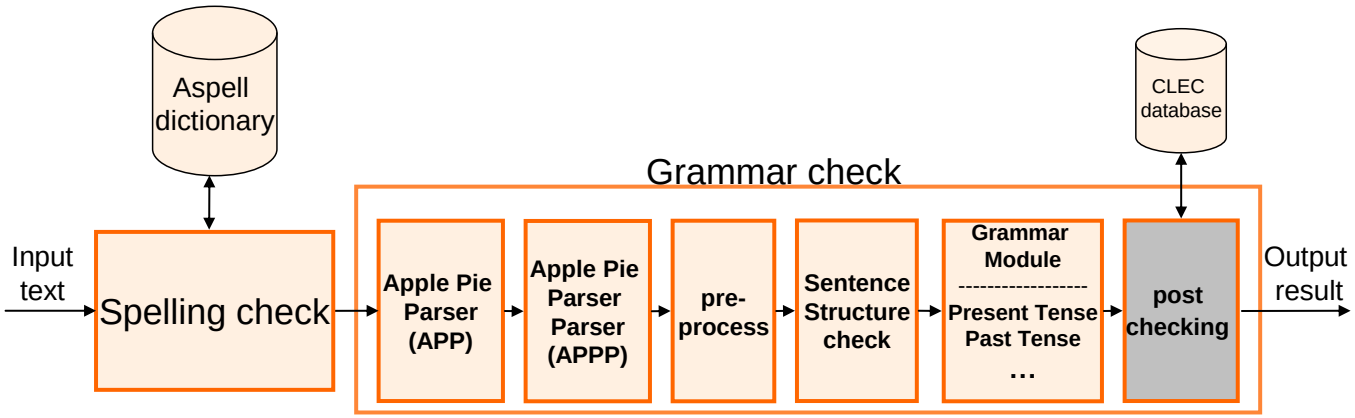

圖 14 習慣用語流程說明

四、英語作文錯誤檢查系統實作

以下為本系統的實作部分，本章節會詳細地把上一章系統設計所提及的元件的實作方法詳細地描述。

4.1 Preprocess 模組 - post-tagging 元件的實作

Step1. 把已經標記好詞性與結構的句子當成 input，進入 post-tagging 模組。
Step2. 檢查句子中每一個 word 的 tag 是否有需要再細分的 tag，即 PRP，PRP$，NN，NNS，SC 和 IN。
Step3. 如果是以上的 tag，則根據相對的 word 去 database query，置換成 database 中指定的 tag。

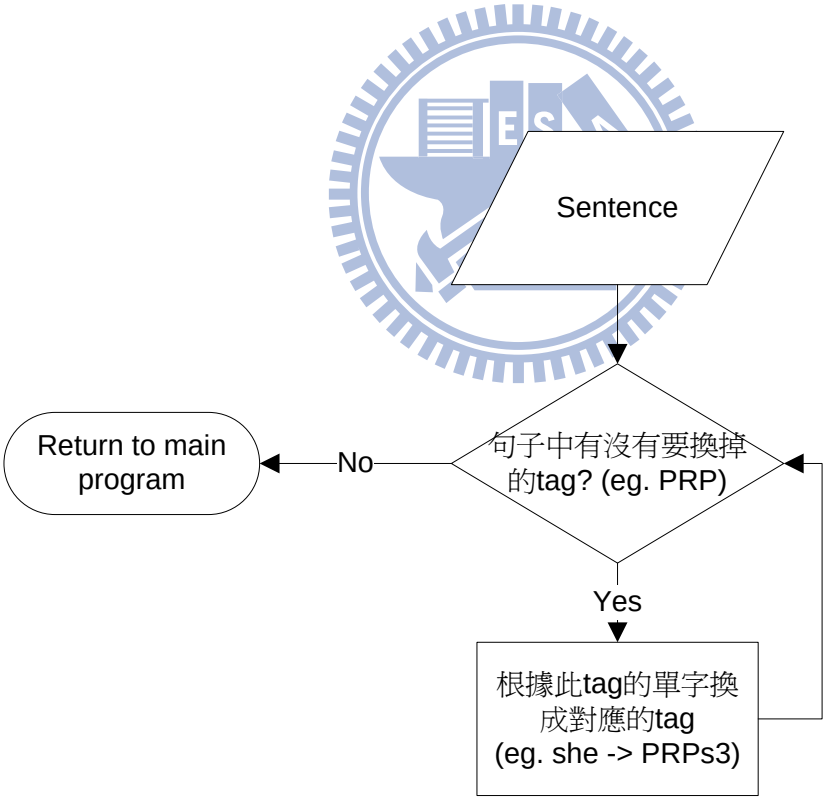

圖 15 Post-tagging 元件流程圖

4.2 Preprocess 模組 - SpiltSentence 元件的實作

Step 1. if 句中有連接詞
　　記錄位置為 CCpos
　　go to Step 2
　　else return
Step 2. if CCpos 前面跟後面都有 verb
　　go to Step 3
　　else if CCpos 前面沒有 verb
　　找下一個連接詞並記錄位置
　　Go to Step 2
　　else return
Step 3. if CCpos 與 CCpos 後的 verb 之間有 noun
　　設定 spilt range 為[start,CCpos-1]
　　else 找下一個 CCpos
　　go to Step 2

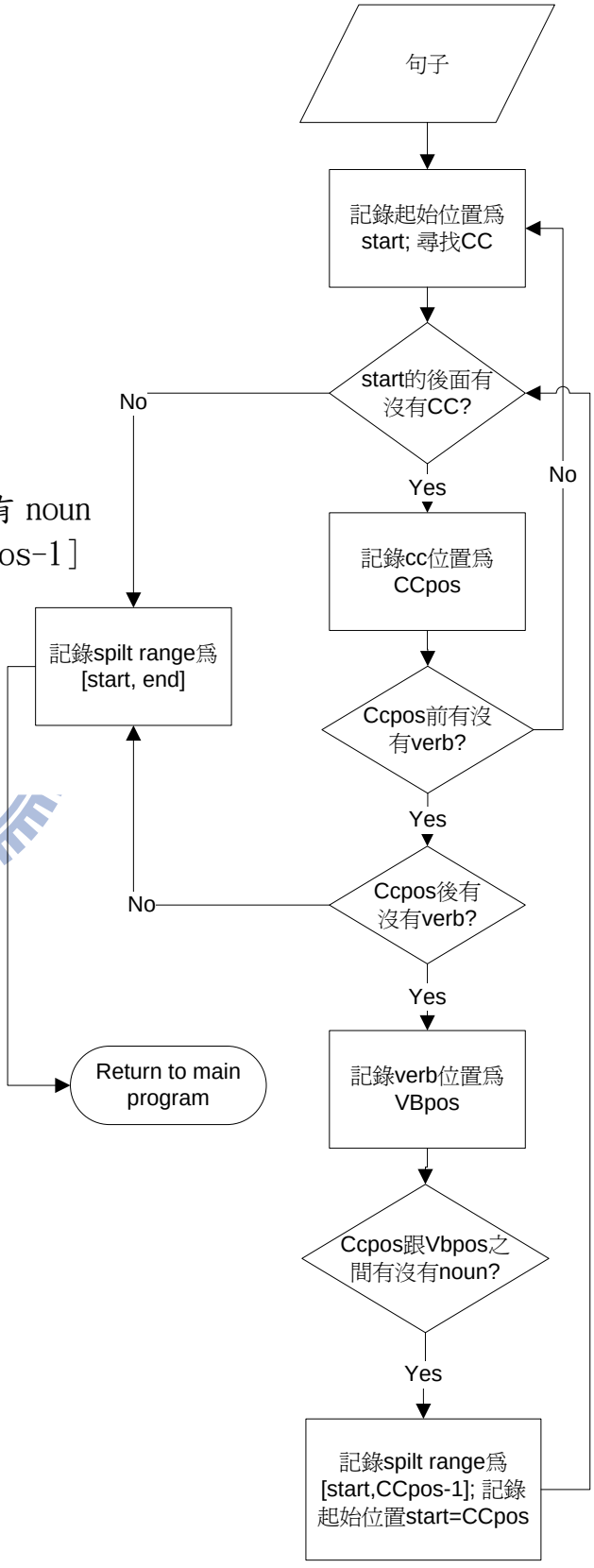

圖 16 Spilt Sentence 元件流程圖

4.3 Sentence Structure Check 的實作

4.3.1 檢查句子結構是否完整

因為一句句子最基本是由主詞跟動詞所組成的，所以我們一開始先檢查句中有沒有 NPL(最小名詞片語)跟 VP(動詞片語)或是不是 UH(感嘆詞)。如果都沒有，則回傳句子結構不完整。

4.3.2 檢查句中有沒有片段的錯誤

片段的錯誤即代表句子中只有 dependent clause。所以我們去找如果句子中有 SC(subordinate clause)的 tag，如果有的話，代表句子有 dependent clause。所以要再去檢查句子有沒有其他子句可當 independent clause。如果沒有，則回傳錯誤。

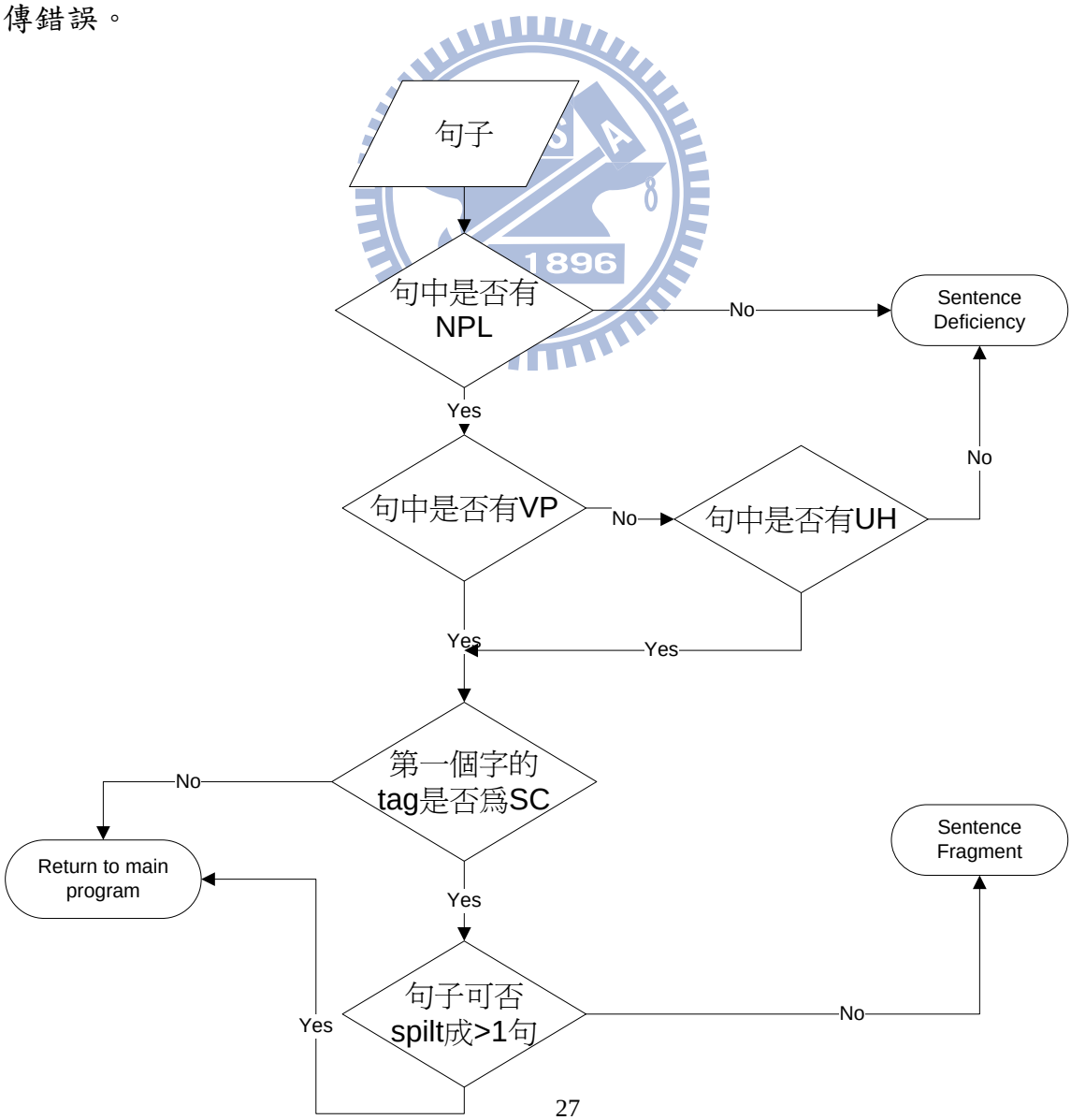

圖 17 檢查句子結構流程圖

4.4 文法模組的實作

表 10 本系統所有文法模組列表

	沿用以前的module		修改及新增過的module

模組名稱	模組功能	模組名稱	模組功能
COMPARATIVE	比較級的用法	NOT A QUESTION	非疑問句
SUPERLATIVE	最高級的用法	PRONOUN NOT REFLEXIVE	代名詞不是反身代名詞
USE A OR AN	A 和AN 的用法	NUMBER ERROR	數量詞/名詞不一致
SEQUENCE OF TENSES	時態一致	TENSE ERROR	時態錯誤
PUNCTUATION	需要正確的標點	PRESENT TENSE	現在式
USE OR INSTEAD OF NOR	or, nor 的使用	PAST TENSE	過去式
USE NOR INSTEAD OF OR	or, nor 的使用	MODAL VERB	時態動詞
SENTENCE CAPITALIZATION	句首大寫	TAG QUESTION	附加問句
DOUBLE NEGATIVE	雙重否定句	INTERROGATIVE SENTENCE	疑問句
REPEAT WORDS	重覆詞性或單字	RELATIVE CLAUSE	關係子句的用法
NEED SUBJECTIVE PRONOUN	須要主格代名詞	QUESTION WORD	WH問句的用法
GERUNDINFINITIVES	動名詞的用法	PREPOSITION	介詞的使用
OBJECTAGREEMENT	接詞檢查	MISCELLANEOUS	其他

上表為本系統所有的文法模組與各模組的功能。表中分成沿用原系統的模組跟新增及修改過的模組兩部分。
以下為對各模組的詳細說明:

1. comparative 模組
   此模組主要是找出比較級形容詞的錯誤。當句子中出現［more + 一般形容詞］pattern 時，便去檢查形容詞的音節。如果形容詞不是多音節，則回傳錯誤。如果句中出現［more + 比較級形容詞］亦回傳錯誤。

2. superlative 模組
   此模組主要是找出最高級的錯誤。當句子中出現［most + 一般形容詞］pattern 時，便去檢查形容詞的音節。如果形容詞不是多音節，則回傳錯誤。如果句中出現［most + 最高級形容詞］亦回傳錯誤。

3. Use A or An 模組
此模組主要是檢查名詞跟其冠詞之間有沒有符合子音，母音的規則。

4. Sequence of Tense 模組
此模組主要是檢查同一句句子中以連接詞連接的動詞時態有沒有一致。

5. Punctuation 模組
此模組主要是檢查句中的結尾標點符號是否正確或有沒有缺少。

6. Use or Instead of nor 模組
此模組主要是檢查 either…or 跟 neither…nor 的使用方法。如果句中有 either…nor 或是 neither…or 時，剛回傳錯誤。

7. Sentence Capitalization 模組
此模組主要是檢查句首有沒有大寫。此錯誤佔了 CLEC 最出現的錯誤中的第 12 位，是很常出現的錯誤。

8. Double Negative 模組
此模組主要是檢查句中有沒有雙重否定的使用。如下句:
We could not never find him. 則會回傳錯誤。

9. Repeat Word 模組
此模組主要是檢查有相同的字 consecutive 地出現。如下句
He is very very happy. 則會回傳錯誤。

10. Need Subjective Pronoun 模組
此模組主要是檢查主格代名詞，用成了所有格代名詞的錯誤。如下句：
Them are very happy. 則會回傳錯誤。

11. GerundInfinivite 模組
此模組主要是檢查 Gerund 跟 Infinitive 的使用方法。若句中出現某些後面只能接 gerund 的動詞(如: enjoy)，可是他後面接的卻是 infinitive，則回傳錯誤。另外一種是，當句中出現後面只能接 infinitive 的動詞(如: agree)，可是他後面接的卻是 gerund，亦回傳錯誤。

12. Object Agreement
此模組主要是檢查 object 跟 be 動詞有沒有 agree。如下句:
She is students. 則會回傳錯誤。

13. Not A Question
此模組主要是檢查句子是不是疑問句，如果不是疑問句卻用了“?”則回傳錯誤。

14. Number Error
此模組主要是檢查可數名詞、不可數名詞、單複數名詞與其限定詞之間有沒有 agree。如[a, books]、[many, money]都是錯誤的 pattern。

15. Tense Error
此模組主要是檢查很多不能出現的 pattern，如[does+plays] 、2 個 verb 連在一起跟一些不用 tense 的使用錯誤。

16. Present Tense
此模組主要是檢查現在式中，肯定句、否定句跟疑問句的 subject verb agreement。

17. Past Tense
此模組主要是檢查過去式中，肯定句、否定句跟疑問句的 subject verb agreement。

18. Modal Verb
此模組主要是檢查情態動詞的使用，確定不會出現以下的 pattern: [modal verb + VBZ] 或是 modal verb 單獨出現。

19. Tag Question
此模組主要是檢查 tag question 時，前後有沒有一致。如下句:
They are happy, isn’ t she ? 則回傳錯誤。

20. Relative Clause
此模組主要是檢查關係子句間的 subject verb agreement。

21. Question Word
此模組主要是檢查 WH-word 的 subject verb agreement。

22. Preposition
此模組主要是檢查一些 preposition 的用法。

4.5 華人慣用語錯誤檢查模組實作

```
Step 1. 對句子中的每一個 word 搜尋
Step 2. If keyword==found
            搜尋前後文有沒有 CLEC 慣用語錯誤
             if error==found
                  return error
              else return
          else return;
```

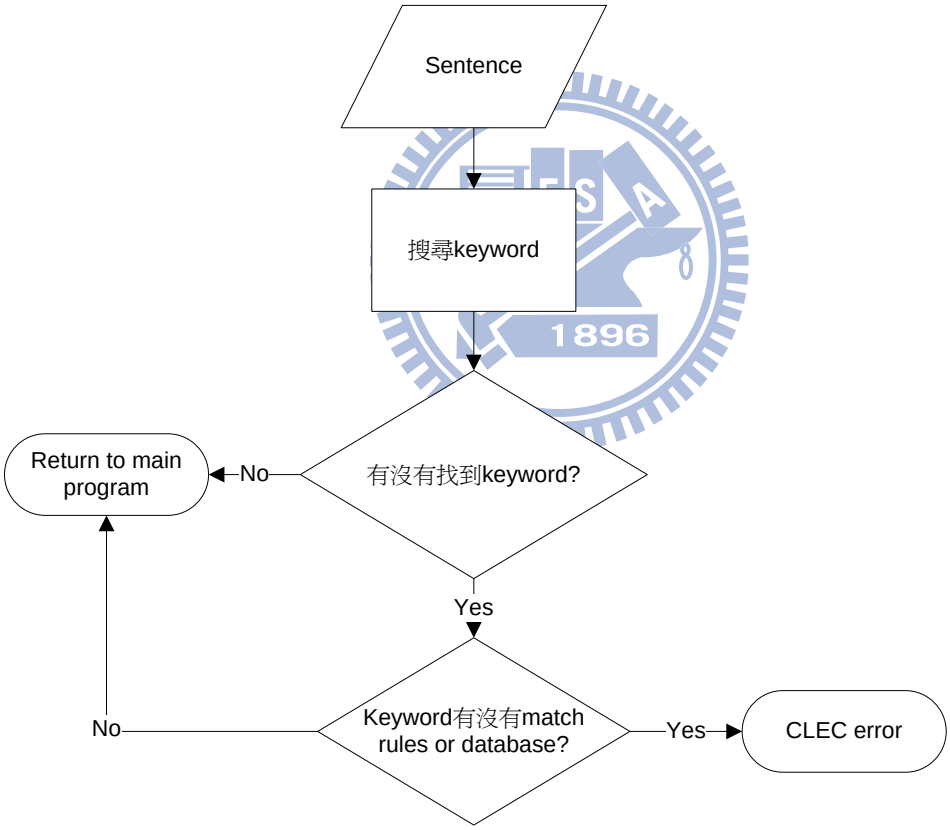

圖 18 華人習慣用語模組流程圖

五、英語作文錯誤檢查系統正確率比較

此章節對本系統作兩方面的正確率探討:**出錯率與錯誤檢查率**

出錯率就是比較系統在檢查正確文章時，出錯的百分比。我們希望出錯率能夠越低越好，這證明系統比較不容易出錯。定義如下：

$$\text{出錯率} = \frac{\text{誤標成錯誤的句子}}{\text{句子總數}} \times 100\%$$

圖 19 出錯率之定義

另一個是錯誤檢查率，即找出錯誤句子的百分比。我們希望錯誤檢查率能夠越高越好，這證明系統的找錯能力很好。我們希望定義如下：

$$\text{錯誤檢查率} = \frac{\text{找出錯誤的句子數}}{\text{錯誤的句子總數}} \times 100\%$$

圖 20 錯誤檢查率之定義

原系統的錯誤檢查率是 84%，可是本系統的錯誤率只有 80%。可是比較一個系統的正確率，除了錯誤檢查率外，還要比較出錯率。原系統沒有對系統作出錯率的探討，所以我們嘗試去比較 2 個系統的出錯率。

出錯率的探討:

我們以 Apple Diary 的英文專欄中 45 天的文章作為測試的 benchmark，分別放到原系統跟我們的系統作比較，比較如下：

表 11 原系統與本系統之出錯率比較表

出錯率	修改前的系統	修改後的系統	降低幅度
368 句句子	189 (51.2%)	12 句 (3.26%)	47.94 ↓

從上表可發現，原系統的出錯率有到 5 成以上。這代表原來 100 句正確的句子，有 51 句被誤判成錯誤的句子。而原系統中有 3 句被錯判，降低幅度達到約 48%。

雖然原系統的錯誤檢查率比本系統高 4%，可是原系統的出錯率竟然到 51%，在錯誤檢查率與正確率的 tradeoff 下，本系統的錯誤檢查率(80%)與出錯誤(3.26%)算是在可以接受的範圍。

然後我們再比較本系統與市面上的文法檢查器之間的錯誤檢查率，benchmark 為 CLEC 中錯誤的句子。比較表如下：

表 12 不同文法檢查器之錯誤檢查率比較表

文法檢查器	After the Deadline	Microsoft word	本系統
錯誤檢查率	45.6%	57%	80%

從表上可看出，本系統的錯誤檢查率比其他的文法檢查器的檢查率都要來得高。這代表本系統在 100 句錯誤句子中，可把其中 80 句的錯誤挑出來。

其中一個本系統的錯誤檢查率比較高的原因是，我們的 benchmark 都是華人學習者所寫的作文，而本系統就是針對華人所常犯的錯誤所設計的，所以錯誤檢查率比一般的檢查器高。

六、系統限制與系統展示

6.1 系統限制

本系統的系統限制分為下列三點：

1. **文法問題**

本系統還是有少量的文法問題，最主要是因為本系統所使用的詞性標記工具不是 100%標記正確。當 APP 標記錯誤時，文法檢查模組就可能會出錯了。另外本系統亦沒有作前後文的時態檢查。以下為這 2 個問題的詳細內容:

APP 標記詞性錯誤導致文法檢查錯誤

例句 1: （正確句子）This dress suits you.

上面這句的詞性標記為:

```
(S (NPL (DT This) (NN dress)) (VP (NNS suits) (NPL (PRP you))))
```

我們可以看到 dress 跟 suits 都被標記作名詞了，所以此句在進入 SentenceStructure 模組檢查時，會出現“沒有動詞”的錯誤。

例句 2:（錯誤句子）I want to learn play violin.

這句被標記成:

```
(S (NPL (PRP I)) (VP (VBP want) (TOINF (VP (TO to) (VP (VB learn) (NPL (NN play) (NN violin)))))))
```

原本這是一句錯誤的句子，句子中 2 個動詞 learn 跟 play 連在一起。這種錯誤應該會被 TenseError 模組檢查出來的，可是因為 APP 把 play 誤標成名詞了，所以 TenseError 模組就沒有把錯誤找出來了.

前後句時態的檢查

前後句的時態檢查的問題，跟語意的有部份的關係。如下例：

```
I was sick last week. Susan looked after me carefully. After two days, I was fine. I was glad that I had a good friend.
```

雖然前文中的其他句子都是過去式，可是上例中的最後一句“I was glad that I had a good friend.”卻是錯誤的。正確句子應該是 “I am glad that I have a good friend.”可是這種錯誤，一定要跟語意配合才能檢查出來的。所以本系統沒有作前後句的時態檢查。

2. **語意問題**

本系統沒有對文章的語意作檢查，如一些介系詞的使用或是一些用字遣詞。如下面 2 句都用了介系詞片詞 – in the same time.

例句一：I ran a mile in the same time as you. 例句二：I could study and play in the same time.

例句一是正確的句子，意思是“我跑一米的**時間**跟你一樣”
例句二是錯誤的句子，意思是“我可以同時唸書跟跑步”，應該要使用 at the same time 才對。

上例從文法上來看是沒有錯的，只是語意錯了。本系統沒有作這一類的錯誤檢查。

3. **特定名詞錯誤**

本系統所使用的是 Aspell 的字典，可是在字典中，可能有些專業用語、 地名、人名或國名都沒有包括在字典中。當句子出現這些字跟詞時，系統就會顯示成單字錯誤了。

6.2 系統展示

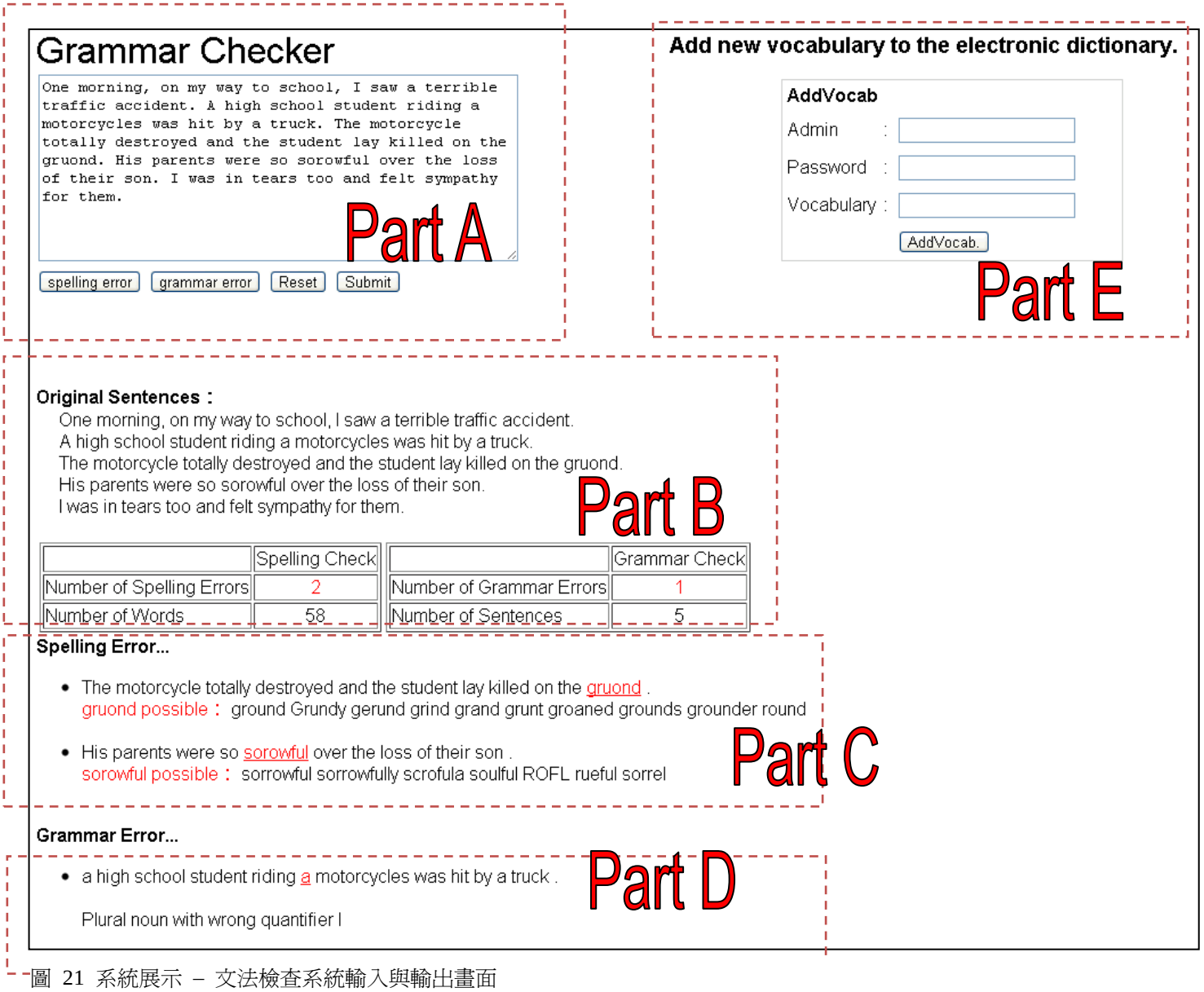

圖 21 系統展示 – 文法檢查系統輸入與輸出畫面

本系統是以網頁的介面跟使用者溝通的，從上圖可看到本系統的使用介面。一共分為 5 個部份：

Part A:

使用者輸入介面，使用者把要檢查的文章或句子輸入在文字方塊中，接 submit 送出，文章會送到後端的 server 中。當 server 接收到 request 便會把句子送到文法檢查系統中檢查。

Part B:

檢查完的結果會送回網頁中。圖中可看句子已經作了切割跟字數、句數的統計，並顯示文章中的拼字、文法錯誤的統計。

Part C:

拼字錯誤的句子顯示在這裡，系統把錯別字以紅色標示出來並把建議的可能的字列出。

Part D:

文法錯誤的句子則顯示在這，系統亦把錯誤的位置與錯誤的訊息回傳給使用者。

Part E:

這個是增加單字的介面。當遇到句子中的單字電子詞典沒有定義而顯示為錯誤時，我們可以把此單字新增到字典中。

七、結論與未來展望

7.1 結論

我們希望改進由實驗室顏佐宇學長所開發的文法檢查系統，降低原系統的出錯誤率、提升錯誤檢查率與加入一些華人習慣用法的檢查。經過對原系統的研究，我們發現原系統出現了 2 個問題。

第一個問題在於系統所使用的詞性標記工具 APP，APP 以 Penn Tree Bank 來標記的詞性對文法檢查模組是不足夠的。所以我們針對這些不足，新增了一些 tag，供文法模組的使用。

第二個問題是原系統只對句子結構做了少量的檢查，亦沒有把一句 compound 或 complex 的句子作切割，分成子句來檢查。這樣會造成很多原來正確的句子被誤判成錯誤。為了解決這個問題，我們新增了句子結構檢查的模組與句子切割的模組，讓原本一整句句子進入文法模組作檢查變成以子句進入模組。

然後我們亦針對文法模組的部份，刪除了一些不合理的模組、沿用部分的模組、新增與修改了其他的模組。把所有以上的模組組合起來組成了一個文法模組子系統。

最後我們針對華人習慣用語的部份建立一個資料庫，把一些 CLEC 中華人的習慣用法錯誤加入資料庫中，讓系統能檢查部分的華人習慣用語錯誤。

本系統的出錯率降至 3.3%且錯誤檢查率達 80%。雖然錯誤檢查率沒有原系統的 84%高，可是出錯率的下降幅度卻是 47.94%。本系統可標示出拼字、文法、以及一些習慣用法的錯誤，以提供參考。同時亦可讓批改作文考題的閱卷者先挑出文法錯誤的地方，以增快他們的改題速度及提升一致性。亦達到了提供學習者一個自我學習的作文練習環境的目的。

7.2 未來展望

針對系統的限制，本系統仍有一些地方需要改進。改進部份分為以下 5 點:

1. 提升標記詞性的正確性

   本系統所使用的詞性標記工具,正確率還是有持提升。我們可以對 APP 的 open source 著手，把 APP 中的一些 grammar rules 修改成更正確的 rules。另一方面，APP 的詞性字典是以單字的詞性在 Brown corpus 出現的頻率作為 tagging 的考量之一。所以如果想要改善 APP 的詞性標記，也可以從 APP 的字典著手，以更多的語料庫去 train APP 的詞性字典。

2. 增加習慣用法的資料庫

   本系統新增了一個華人習慣用語的資料庫,把 CLEC 部分的華人習慣用語錯誤加入資料庫中。可是本系統只加了部份的錯誤，如果把更多這一類的錯誤都加入資料庫中，系統的錯誤檢查率必定會提升，更多的華人習慣用語錯誤亦可被找出來。

3. 把系統當成互動式教學的一種

   以下句子在系統限制時有提及過，suits 可當名詞亦可當動詞使用。

This dress suits you.

   我們可以實作一個系統，讓使用者可跟系統互動。系統可提供不同的 suits 的詞性供使用者選擇。當使用者選擇 suits 的詞性為名詞時，系統則顯示句子無動詞的錯誤，並作一些文法上的教法與解釋。當使用者選擇 suits 的詞性為動詞時，則顯示文法正確。

4. 增加其他華人語料庫

   我們亦可再增加更多的華人語料庫，以驗證及設計本系統

5. 新增專有名詞的資料庫

   改善專有名詞的部分，我們可以新增一些地名、人名、國名之類的專有名詞資料庫。

參考文獻

1. GNU Aspell，[On-line]. Available: http://aspell.net/
2. International Ispell，[On-line]. Available: http://fmg-www.cs.ucla.edu/geoff/ispell.html
3. Leech, G. Preface, in Granger, S. (1998). Learner English on Computer, XVII, Essex: Addison Wesley Longman Limited.
4. Proteus Project-Apple Pie Parser，[On-line]. Available: http://nlp.cs.nyu.edu/app/
5. Satoshi Sekine, Ralph Grishman , “A Corpus-based Probabilistic Grammar with Only Two Non-terminals,” Fourth International Workshop on Parsing Technology,1995
6. 桂詩春、楊惠中，中國學習者英語語料庫，上海外語教育出版社，上海，2002。
7. 楊惠中，基于 CLEC 語料庫的中國學習者英語分析，上海外语教育出版社，上海，2005
8. 楊惠中、衛乃興，中國學習者英語口語語料庫建設與研究，上海外語教育出版社，上海，2005。
9. 財團法人語言訓練測驗中心，[On-line]. Available: https://www.gept.org.tw/
10. 張鈞凱，「英語文試題檢測與答題驗證系統設計與實作」，國立交通大學，碩士論文，民 94 年。
11. 顏佐宇，「以 CLEC 語料庫為基準來設計實作一個英語作文錯誤檢查系統」，國立交通大學，碩士論文，民 96 年
12. 黃希敏，「語料語言學研究面面觀」，《敦煌英語教學電子雜誌》。2005 年 3 月號。
13. 黃希敏，「語料語言學概述」，《敦煌英語教學電子雜誌》。2004 年 11 月號
14. Mitchell P.Marcus, Mary Ann Marcinkiewicz and Beatrice Santorini, “Building a Large Annotated Corpus of English: The Penn Treebank”, Computational Linguistics, 19(2)
15. Raphael Mudge, “The Design of a Proofreading Software Service”, Workshop on Computational Linguistics and Writing,2010