第二章

背景知識和相關技術

本章節將探討英語作文輔助的背景知識，並論及本研究引用的排序技術。其中背景知識分為相關研究、語料庫、語料庫語言學、詞性標籤、concordance、詞語搭配和推薦式系統。相關研究介紹與本研究相關的學術研究；語料庫為規劃過的範例資源；語料庫語言學為研究語言語料庫的一門學問；詞性標籤為語料庫中各個詞彙的詞性分類，良好的詞性標籤有利於語料庫的分析和應用；concordance為資訊檢索時的一種初級排序功能；詞語搭配是描述特定詞彙間的共同(co-occurred)關係，也是 ESL/EFL 學習者於學習上較為困難的部分。推薦式系統能適當依據學習者的需求而推薦相關資訊，論文將引用其評估技術於分析推薦方法的成效。而本研究採行的排序技術為多重序列排列，該技術能有效率地處理英語不限定子句問題，並能依據單字結構提供有效率地排序結果。

2.1 相關研究

本小節將介紹與本研究相關的學術研究，分為 Steffan Corley, Martin Corley, Frank Keller, Matthew Crocker, and Shari Trewin 共同開發的 Gsearch 系統[35]，以及 Carnegie Mellon University(CMU)所制訂的 REAP 系統[27]，和雲林科技大學所研發的「線上英語學習環境」[43][44]。

Gsearch系統主要目的為在語料庫中建立句法準則(syntactic criteria)，而學習者可以依據建立的句法準則有效率地從語料庫中選取合宜的例句，Gsearch系統規劃的句法準則較為彈性，對於多樣的學習者能依據不同的句法準則有效率地查

詢。本研究將延續其理念，建立一個彈性的表達元素模組供多樣的學習者依其認知程度輸入。

REAP 系統主要目的為提供學習者英語閱讀方面的協助，目前著重於英語詞彙的教學。REAP 系統將學習者依認知程度分類，以測驗的方式評估學習者在使用 REAP 系統前和使用 REAP 系統後，學習者的詞彙程度是否有顯著性的差異，並以問卷調查和統計數值整合分析 REAP 系統的成效。REAP 系統雖為英語閱讀上的協助系統，然閱讀為寫作的前驅，其規劃的英語學習理念和評估方法，和本研究是有關連的，本研究將參照 REAP 系統的評估方式評估系統的成效。

雲林科技大學所研發的「線上英語學習環境」，建立一個以 concordance 為基礎的學習網站，針對學習者在英語詞語搭配上所衍生的相關問題，以 concordance 的格式呈現相對應的例句並提供相關的統計資訊給學習者參照。本研究將延續其理念使用例句的呈現供學習者於英語寫作上參照學習。然「線上英語學習環境」除使用較為初階的 concordance 檢索技術而不適用於寫作參照外，其對於例句的推薦方式以統計學的頻率為主，並非針對學習者的需求給予相對應的排序，導致推薦的例句常不符合學習者的語意需求。除此之外，雲林科技大學的研究論文並沒有提及任何嚴謹評估系統的方法，難以衡量其系統有一定的協助成效。本研究將改進上述缺失，建立一個文句推薦方法，推薦合適的例句供學習者參照，並以嚴謹的方式審慎評量系統的成效。

2.2 語料庫 (Corpus/Corpora)

語料庫是一個超大型的語料數據庫，用以儲存語料分析的文本。而市面上用於英語語言上有兩種常見的語料庫，一個是由 HarperCollins 出版的 Cobuild，所建立的語料庫為 Bank of English，共有 5.2 億個英語單字。

另外一個語料庫資源為 BNC(British National Corpus)，是由牛津大學以及

Addison-Wesley Longman 出版社、British Library's Research and Innovation Centre 等機構共同研究的成果，整個語料庫約有 1 億個英語單字。

許多研究指出，語料庫的適當使用對於學習英語上是相當有助益的[22][26]，目前語料庫在學習上最有助益的工具為 concordance，也有相關的系統以 concordance 為基礎，靈活運用語料庫的資源給予學習者協助。然而認知程度較好的學習者，對於語料庫的使用頻率、利用程度遠高於認知程度稍差的學習者[22]，因此本研究擬定簡便的彈性表達模組鼓勵學習者活用語料庫資源。

2.3 語料庫語言學

語料庫語言學指藉由電腦輔助技術分析大量自然語言的一門學問(the empirical study of language relying on computer-assisted techniques to analyze large, principled databases of naturally occurring language.)[34]，換言之，語料庫語言學著重於從真實生活(real life)體驗中讓學習者學習語言[36]，而現有語料庫語言學用於英語協助上以 concordance 和詞語搭配最為顯著。相關研究指出，語料庫語言學用於 ESL/EFL 學習者於學習英語上，已有一定的成效[12]。然語料庫語言學目前著重於研究英語字彙和搭配兩大議題，如針對英語寫作，僅能協助英語單字、英語搭配的兩個前提知識提供協助，對於英語文法句型的協助，語料庫語言學還有許多可以研究發展的空間。

2.4 詞性標籤 (Part-of-Speech)

在自然語言中，一個詞往往具有多個詞性，詞性標籤的目的就是透過詞語的上下文將句子中的詞性標注唯一。然詞性標注本身也具有模糊性。以 Brown 語料庫[8]中的英語句子為例，11.5%的詞句具有多個詞性判定標準，整個語料庫中含有多重詞性的字詞約佔了總詞數的 40%。這表明了詞性標注工作的困難度，因

此，出現了多種標注方法和標注集。

關於英語語言上較為有名的詞性標注方法其中一個為 TOSCA (Tools for Syntactic Corpus Analysis) [30]，為荷蘭 Nijmegen University 所研發的一套系統，共計有 16 個特定標籤(扣除無法標籤的 TAG? 和不確定的 HEUR)，其標籤如下[7]：

ADJ	Adjective
ADV	Adverb
ART	Article
CONJUNC	Conjection
EXTHERE	Existential there
GENM	Genitive marker
HEUR	(unknown)
MISC	Miscellaneous
N	Noun
NADJ	Nominal adjective
NUM	Numeral
PREP	Preposition
PROFM	Proform
PRON	Pronoun
PRTCL	Particle
PUNC	Punctuation
TAG?	Word unable to tag
VB	verb

圖 2.1　TOSCA 的詞性標籤

另一個較為著名的標注工具為 Tree - tagger，是德國 Stuttgart University 所研發的一套軟體，共有 48 個特定標籤，並以決策樹(Decision Tree)的分析方法決定詞性標籤[20]。和 TOSCA 所不同的是，Tree - tagger 對於動詞的變化有更詳盡的定義，例如 eat 一詞，Tree - tagger 分為原式 eat (VB)、過去式 ate (VBD)、現在進行式 eating(VBG)等。

2.5 Concordance

Concordance 一詞，最早出現於 13 世紀的拉丁聖經(Vulgate Bible)中，一直到 17 世紀才在英國由 Cruden 出版第一本相關書籍[14]。

Concordance 最早的用途是替詩集做相關的索引集，例如查詢某個單字在詩集中的用法。現今對於 concordance 一詞的定義如下：Concordance is the display of words or simple grammatical items with their surrounding text.[33]。簡而言之，concordance 是以特定字詞為基準的上下文例句排列。

舉例而言，對於 concordance 此一英語單字而言，在 Cobuild[5]語料庫中出現的結果如下：

Surely when Bernie Grant is seen to be in ***concordance*** with Winston
investment did seem to indicate some kind of ***concordance***, it was easily
paying formations" characterized by a ***concordance*** or compatibility
observant artist, who savors the unexpected ***concordance*** linking say, a
a real sort of flexible way of approaching a ***concordance*** like that

圖 2.2　concordance 一詞在語料庫中的 concordance

Concordance 運用於英語輔助上的文獻相當多[24][42]，主要是結合語料庫資源，對於使用者輸入的關鍵字以相對應的例句呈現。值得注意的是，concordance 是初級的檢索功能，提供的例句並非針對寫作而設計，所以往往無法讓學習者得到良好的參照，而 concordance 對於選取的例句並無詳盡地篩選和排序，無法進一步滿足 ESL/EFL 學習者在參照協助上的需求。

2.6 詞語搭配 (Collocation)

詞語搭配為特定詞彙彼此出現的頻率大於期望值的詞彙組合(Collocations are strings of specific lexical items that co-occur with a mutual expectancy greater than chance.)[23]。舉例而言，英語單字 problem 常和 cause、create、solve 一同搭配，可是較少人會用 make 一詞和 problem 搭配，原因為詞語搭配為一種特定的

結合方式，類似於中文的「成語」，用詞搭配是具有特殊意義的。

Ilson, Morton, and Evelyn 是最先試圖將複雜的英語詞語搭配分門別類[31]，他們將英語詞語搭配分為兩類：(1)文法搭配(grammatical collocation)；(2)詞彙搭配(lexical collocation)。

所謂的文法搭配定義如下：Grammatical collocation is a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause. 舉例來說，關於英語句型decide on a boat，若我們以choose (to buy) a boat來解釋這句英語句型的意義，則此句型含有文法搭配的字詞decide on；但若以make a decision while on a boat詮釋英語意義，則此英語句型屬於自由結合(free combination)不屬於文法搭配，所謂的自由結合，根據Ilson et al. 的定義為：Free combination consists of elements that are joined in accordance with the general rules of English syntax and freely allow substitution. 換而言之，符合一般英語規則的結合，不屬於文法搭配，而何謂一般的英語規則，Ilson et al. 並無明確的定義，然而對於文法搭配的組合，Ilson et al. 有明確分為以下八種組合：

（G1）名詞 + 介詞：ex: a chance against
（G2）名詞 + to + 動詞：ex: an attempt to do
（G3）名詞 + that + 子句：ex: an agreement that
（G4）介詞 + 名詞：ex: on somebody's advice
（G5）形容詞 + 介詞：ex: angry at
（G6）形容詞 + to + 動詞：ex: necessary to work
（G7）形容詞 + that + 子句：ex: be afraid that
（G8）動詞的詞語搭配，共有十九種搭配摸式：
AB: 動詞 + 直接受詞(direct object) + to + 間接受詞 (undirect object) ex: She sent the book to him.
C: 動詞 + 直接受詞 + for + 間接受詞 ex: She bought a shirt for her husband.
D: 動詞 + 介詞 + 受詞 ex: They came by train.
E: 動詞 + to不定式(to infinitive) ex: She continued to write.
F: 動詞 + 原形不定式(bare infinitive) ex: Mary had better go.
G: 動詞 + 動詞ing(V-ing) ex: They enjoy watching TV.
H: 動詞 + 受詞 + to不定式 ex: We forced them to leave.
I: 動詞 + 受詞 + 原形不定式 ex: She heard them leave.
J: 動詞 + 受詞 + 動詞ing ex: He felt his heart beating.
K: 動詞 + 所有格(possessive) + 動詞ing ex: I cannot image their staling apples.
L: 動詞 + that + 子句: ex: The doctor suggests me that …
M: 動詞 + 受詞 + to be + 受詞補語(complement) ex: We consider her to be well trained.
N: 動詞 + 受詞 + 受詞補語 ex: She dyed her hair red.

圖 2.3(a) Ilson, Morton, and Evelyn **文法搭配的八種形式**

O: 動詞 + 受詞A + 受詞B
ex: The teacher asked the pupil a question.
P: 動詞 + 受詞(可有可無)+ 副詞
ex: He carried (himself) well.
Q: 動詞 + 受詞(可有可無) + wh-子句
ex: She asked (somebody) why we had come.
R: It + 動詞 + 受詞 + to不定式 或
It + 動詞 + 受詞 + that + 子句
ex: It surprised me to learn.
It surprised me that …
S: 動詞 + 形容詞 或 動詞 + 名詞
ex: He is a teacher.
The food tastes good.

圖 2.3(b) Ilson, Morton, and Evelyn **文法搭配的八種形式**

相對於文法搭配，詞彙搭配是學習上較為困難的部分。而詞彙搭配和文法搭配最不同的地方是詞彙搭配沒有 dominant word，舉例而言，compose music 為一個詞彙搭配(L12)，然而不像 decide on 有很明確的 decide 單字，可以讓學習者知道 decide 後很有可能接介係詞 on 搭配，這也是往往詞彙搭配會比文法搭配更加複雜的原因。同樣的，詞彙搭配也需考慮自由結合(free combination)問題，舉凡像 drink tea，因 drink tea 符合一般性英語規則的結合，是不屬於詞彙搭配的範圍的。Ilson et al. 對於詞彙搭配的組合有明確分為以下七種組合：

（L12）動詞 + 名詞（或代名詞、介詞片語）: ex: compose music
（L3）形容詞 + 名詞: ex: strong tea
（L4）名詞 + 動詞: ex: bombs explode
（L5）名詞 + of + 名詞: ex: a pride of lions
（L6）副詞 + 形容詞: ex: deeply absorbed
（L7）動詞 + 副詞: ex: argue heatedly

圖2.4 Ilson, Morton, and Evelyn **詞彙搭配的七種形式**

詞語搭配於英語寫作上已有相當的成效，並有相關的研究案例指出 ESL/EFL 學習者於詞語搭配上是較為薄弱的一環[16][18][17]，論文將延續其研究知識，對於詞語搭配問題予以針對性的協助。

2.7 推薦式系統 (Recommender System)

隨著網際網路的普及，使用者可以輕鬆從網路中取得多方面的資訊，然而眾多的資訊中，使用者常無法有效率地篩選出有實用價值的資訊，亦即發生所謂資訊超載(Information Overload)的問題。有鑑於此，Goldberg 提出了第一套資訊過濾系統(Information Filtering system)[19]，並取名為 Tapestry；而後 Renick and Varian 並認為一般的資訊過濾系統亦泛稱推薦式系統[40]，然推薦式系統除須具備資訊過濾的功能，應更著重於發掘使用者感興趣的資訊(finding something interesting)並推薦給使用者。一般而言，推薦式系統主要分為三個推薦步驟：第一步驟為收集使用者的資訊，並加以分析；第二步驟為依據分析結果推薦使用者相關的資訊；第三步驟為使用者對於推薦的結果予以喜好程度的回饋，而系統藉此更新使用者的資訊以利於後續的推薦成效。

推薦系統的評估端看學習者對於推薦結果的評斷，往往評估的方法是由使用者觀察系統推薦的資訊，和使用者原本預期得到的資訊，兩者之間的差異程度當推薦指標。然除了上述的推薦指標，Saracevic 於研究中亦考量使用者主觀的感受當作系統評估的準則[37][38][39]；Sarwar, Karypis, Konstan, and Riedl 更明確地指出評估推薦系統的績效[15]，不論從哪一個角度評估，但底線應該是使用者的滿意度(User Satisfaction)和使用者感受系統的有用度(Usefulness)。

本研究的推薦方法，是依據學習者表達元素模組提供的資訊，經過比對選取，並排序而回傳例句供學習者參考，其中排序的觀念和推薦式系統篩選的觀念是異曲同工的，同是對於資訊超載的問題，系統能先經過一定程度的篩選，而回

傳值得推薦的資訊給學習者。本研究將採推薦式系統的評估指標用以評估推薦方法的成效。

2.8 多重序列排列（Multiple sequence alignment (MSA)）

多重序列排列最早應用於生物資訊學(Bioinformatics)，是蛋白質或DNA的一種比對方式，可用來分析不同物種間的相似程度。一個MSA定義如下[32]：多個長度不同的序列(sequence) S_1、S_2、… S_n($n \geq 3$)，一個多重序列排列為相同長度的序列A_1、A_2、… A_n，其中A_1對應S_1、A_2對應S_2，其餘依此類推。而A_1、A_2、… A_n序列中允許元素「間隔」(gap)，以"－"表示，間隔和A_1、A_2、… A_n以及S_1、S_2、… S_n的關係為若去除間隔，則A_1等於S_1、A_2等於S_2，依此類推。(Given a set of sequences with different length $S=\{S_1, S_2, ..., S_n\}$ ($n \geq 2$), $\Sigma=\{e|\forall e \in S_i, i=1 \ldots n\}$, A multiple sequence alignment is a set of sequences with same length $A=\{A_1, A_2, A_3 \ldots, A_n\}$, $\forall A_i \in \Sigma', \Sigma' = \Sigma \cup \{"-"\}$, $S_i = A_i \backslash \{"-"\}$ $\forall i=1 \ldots n$.)

簡便的例子如下，假設有兩個序列S_1= CCAATA、S_2= CCAT，序列元素為Σ ={A,C,T}：

S_1. C C A A T A

S_2. C C A T

則排列的結果可能有以下的結果：

A_1. － － － － C C A A T A

A_2. C C A T － － － － － －

或者有以下的結果：

A_1. C C A A T A

A_2. C C A − T −

以上兩種結果均為S_1、S_2的MSA，然而如何在上述兩種結果中取捨較為合宜的排列，一般是以分數的高低來表示，較高的分數代表較為貼切的結果。例如以sum-of-pair score來計算分數，計算的方式為計算結果中所有位於同一相對位置元素的比對分數總和。如第一個結果的分數為C_1，而第二個結果的分數為C_2：

C_1=S(−,C)+S(−,C)+ S(−,A)+S(−,T)+ S(C, −)+S(C, −)+S(A,−)+S(T,−)+S(A,−)

C_2=S(C,C)+S(C,C)+S(A,A)+S(A,−)+S(T, T)+ S(A,−)

其中 S(x,y) x,y∈Σ' 代表 x 元素比對 y 元素的分數。習慣上，會以一個統合的置換矩陣(Substitution matrix)來代表上述所有元素間的比對分數，若兩元素比對的結果相同，置換矩陣以較大的分數表示；相反的，若兩元素比對的結果不相同，因兩元素須要取代，相對應置換矩陣分數會較小。例如上述的例子我們可以下表的矩陣來表示，其中第 i 列 j 行的值代表由 i 列元素比對 j 行元素的比對分數，例如第二列第三行代表 C 和 T 的比對分數(-3)：

	A	C	T	–
A	8	0	-2	-8
C	0	8	-3	-8
T	-2	-3	8	-8
–	-8	-8	-8	0

將上述置換矩陣套入例子中可得C_1=-8-8-8-8-8-8-8-8-8=-72，而C_2=8+8+8-8+8-8=16，代表第二個排列方法較為優異。然而，如何找出「最優」的排列結果(optimal MSA)，對於兩個序列的MSA，一個可行的方式是採用Needleman-Wunsch algorithm。

2.8.1 Needleman-Wunsch algorithm

Needleman-Wunsch algorithm 為 Needleman and Wunsch 於 1970 年提出的動態規劃(dynamic programming)演算法，試圖求出兩個序列在 MSA 中最佳的排列結果。

承上例，若要比對的序列為S_1和S_2，對於S_1中第i個元素我們以X_i表示；對於S_2中第i個元素我們以Y_i來表示，依據Needleman-Wunsch algorithm計算，會將結果以矩陣的方式呈現，我們定義此矩陣為F，而矩陣中的數值由以下的規則計算：

F(1,1)=0 //initiate 0

F(i , j)=max{ F(i-1,j-1)+S(X_i,Y_j),

F(i-1,j)+d,

F(i,j-1)+d };// d=gap-value (in last example: -8)

S(X_i,Y_j)= Substitution matrix(X_i→Y_j)

根據此一法則，將S_1、S_2兩序列套入的結果如下：

	–	C	C	A	A	T	A
–	0	-8	-16	-24	-32	-40	-48
C	-8	8	0	-8	-16	-24	-32
C	-16	0	16	8	0	-8	-16
A	-24	-8	8	24	16	8	0
T	-32	-16	0	16	16	24	16

事實上，當計算完 F 矩陣，兩序列的最佳排列結果已呼之欲出，試圖將 F 矩陣的結果回溯，矩陣中的對角線代表兩序列相對應的元素需經過取代才能得最

佳的排列、水平線代表兩序列相對應的元素需經過刪除才能得最佳的排列、而鉛直線代表兩序列相對應的元素需經過增加才能得最佳的排列。逐一回溯 F 矩陣過程，可得知兩序對的最佳排列方式。例如上述的例子其最佳排列的結果如下：

A_1. C C A A T A

A_2. C C A – T –

2.8.2 The Center Star algorithm

Needleman-Wunsch algorithm僅適用於兩個序列比對，然而若我們有n個序列，比對的時間會相當耗時。如果對於比對的結果，只需要較佳的解而不是最佳的解，同時我們有一個參考序列當基準可用以比對其他的序列，那麼我們可以使用the Center Star algorithm大大改善比對排列之效能。對於長度均為k的n個序列，the Center Star algorithm的效能為$O(k^2n^2)$[41]。演算法的流程如下：

I. Choose sequence S_c and make it the center of the star.

II. Produce a multiple alignment M such that, for every sequence S_i in M, the induced pairwise alignment of S_c and S_i is the same as the optimum alignment of S_c and S_i.

例如我們要計算S_1=CCAATA、S_2=CCAT、S_3=CAACA的MSA，我們取S_1當中心點，將S_1和S_2依據Needleman-Wunsch algorithm取最佳排列(optimum alignment)，並將S_2的結果記錄為A_2，此時A_2視為已排列完成的序列，再將S_1和S_3取最佳排列，所得的最後集合即是相對應的MSA結果。結果的示意圖如下：

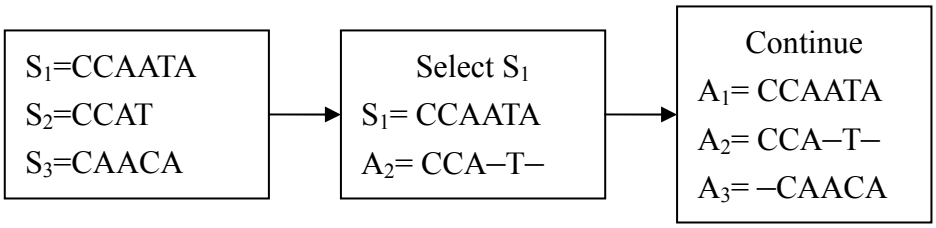

圖 2.5 The Center Star algorithm

本研究採行 MSA 技術在於 MSA 有良好的排序流程，並能以 gap 方式合適描述英語中含有不限定子句的查詢。除此之外，本研究將學習者的語意需求視為 the Center Star algorithm 的中心點，如此可在大量選取的例句中，迅速有效率地推薦學習者感興趣的例句。