行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　※
※　　智慧型電腦輔助翻譯與教學--　　　　※
※　　網路中英語料庫對應句檢索系統之建構　※
※　　　　　　　　　　　　　　　　　　　※
※　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC 89－2411－H－002－074

執行期間：89 年 8 月 1 日至 90 年 7 月 31 日

計畫主持人：高照明

zmgao@ccms.ntu.edu.tw

本成果報告包括以下應繳交之附件：
　　□赴國外出差或研習心得報告一份
　　□赴大陸地區出差或研習心得報告一份
　　☑出席國際學術會議心得報告及發表之論文各一份
　　□國際合作研究計畫國外研究報告書一份

執行單位：台灣大學外國語文學系

中　華　民　國　　90　年　10　月　15　日

行政院國家科學委員會專題研究計畫成果報告

智慧型電腦輔助翻譯與教學--

網路中英語料庫對應句檢索系統之建構

Building an Intelligent Web-Based Bilingual Chinese-English

Concordancer for Computer-Assisted Language Learning and Translation

中文摘要 本計劃利用中英平行語料庫建構一個可應用於電腦輔助語言教學與翻譯的網路程式. 此程式可從平行語料中自動擷取包含中英文關鍵詞的句子及其對應的翻譯, 並可以擷取與輸入的中英文句子最近似的翻譯.

關鍵詞: 中英平行語料庫, 電腦輔助語言教學, 電腦輔助翻譯, 中英雙語對應句檢索

Abstract： In this project, a web-based bilingual concordancer is developed which can retrieve example sentences containing a keyword and their corresponding translations from parallel Chinese-English corpora. This useful and practical tool is capable of retrieving and ranking relevant examples and their translations in terms of their similarity to an input sentence and can be fruitfully used in computer-assisted language learning and translation.

Keywords： bilingual concordancer, parallel Chinese-English corpora, computer-assisted language learning, computer-assisted translation

A perennial problem for language learners and translators is that they are often confounded by which word they should use to convey an idea. Monolingual and bilingual dictionaries are not very helpful to resolving this problem, because many expressions involve word combinations such as collocations and lexical phrases, which are either difficult to find or not listed in the dictionaries. For second language learners, the task of searching for the right expressions in the target language is like looking for a needle in a hay. It is time-consuming, painful, and often turns out to be in vain.

With the advent of the world wide web and the development of corpus linguistics, a new approach to tackling the problem has emerged. This promising new approach draws on a bilingual concordancer which can retrieve bilingual examples from large bilingual corpora. Unlike a traditional monolingual concordancer, it allows users to query an unfamiliar word in both the source and target language. This greatly facilitates language learners in looking for appropriate expressions.

The origin of such a tool may in part be attributed to a technical report by Martin Kay at Xerox in 1980, which was not published until 1997 (Kay (1997). In that seminal paper, Kay proposed a computer-aided translation system which can access previous translation examples created by professional translators. Similar

idea was reintroduced by Harris (1988), who suggested using "bitexts" (i.e. translation equivalents coupled together unit by unit) as a computer-aided tool for professional translators. It was not until the mid 1990s that Kay's idea was finally realized (cf. Dagan and Church (1997)) Unfortunately, this kind of powerful tool is not widely used in Taiwan. Part of the reason is that it presupposes the correct identification of bilingual correspondences of the source and target language at the sentence level. In computational linguistics, this procedure is called sentence alignment and has been actively researched in recent years. Approaches to sentence alignment can be categorized into statistics-based, dictionary-based, and hybrid method, none of which performs well in Chinese-English bilingual corpora according the experiments reported in Gao (1997).

A simple method to circumvent the problem of sentence alignment is to look for bilingual texts where there are clearly identifiable markers for text alignment. The internal structure of Bible makes it a good resource for aligning texts at paragraph level, because each paragraph (more precisely passage) in Bible is marked by a numerical indicator. The difficult task of text alignment is thus simplified to matching numerical indicators in the English Bible and the Chinese Bible. Figure 1 shows the user interface of the bilingual concordancer using Chinese and English Bible. Figure 2 is the output of a query.

Another important Chinese-English bilingual resource is Sinorama Magazine. Articles in the Sinorama Magazine are written in Chinese and translated into English, Spanish, and Japanese. The English translations are all checked by native speakers of English. Despite its unauthentic nature, the English translations are important resources for Chinese learners of English who wish to learn English expressions related to Chinese culture or events that takes place in Taiwan.

Sentence alignment in the Sinorama bilingual magazine is complicated by several factors. As with most translations, additions and omissions of sentences or even paragraphs are quite common in the Sinorama magazine. Another difficulty is caused by the abundance of nonliteral translations in the Sinorama magazine. In the absence of a highly robust algorithm of text alignment, we perform paragraph alignment using indentation as paragraph markers. We assume that paragraph correspondence can be established if the bilingual texts have the same number of paragraph markers. Our approach is of course subject to errors, because it ignores the fact that there might be additions or omissions involved in the translation process. Furthermore, indentation might not be a reliable clue for paragraph markers. In inspecting the results of our program, we do find quite a few errors in paragraph alignment. To remedy this inadequacy, we provide hyperlinks for users to view the source and target texts. To facilitate browsing the bilingual texts, the keyword is highlighted and the paragraph containing the keyword is highlighted in another color. The proposed corresponding paragraphs are also highlighted for ease of reference (Figure 4). As the program is designed for advanced learners of English, it is reasonable to expect that they are capable of deciding whether the aligned paragraphs are correct. If they find errors in the alignment of paragraphs, they can view the whole texts and find the correct alignment around the paragraph proposed by the program. The tool, therefore, is still very useful even though it is error-prone.

As illustrated, a bilingual concordancer can accept a query word in the source or target language and present example sentences (or paragraphs) containing the input word juxtaposed with their (sentence or paragraph) translations. As expressions might take the form of a phrase or a sentence, it is more desirable to design a bilingual concordancer which can take a sentence as input. But given the small size of Chinese-English bilingual corpora available, the chance of finding examples identical to the input sentence is slight. To make the tool useful, it should have the capability of extracting and ranking examples in terms of

their relevance to the input query. With this powerful tool, translators and language learners can use the intelligent bilingual concordancer to find the closest match to the input sentence and its translation.

The central question of building such a program is how to compute the similarity between a sentence and an input sentence. The following procedure is proposed for an intelligent Chinese-English bilingual concordancer.

1. Language Identification: A program is written to check if the input is Chinese or English.

2. Tokenization: As there is no delimiter between words in Chinese, a word segmentation program is required to process the Chinese input as well as the Chinese corpus.

3. Lemmatization: Through lemmatization, English words with inflection (whether regular or irregular) will be converted into their lemmas (i.e. basic forms).

4. Indexing: All the words in the Chinese and English texts are indexed so that sentences containing a particular word can be retrieved in no time.

5. Term Weighting: Every Chinese and English word is assigned a weighting depending on its frequency. Content words are given a weight based on Inverse Document Frequency (IDF), which assumes that the importance of a word is in inverse proportion to the number of documents in which it occurs. In other words, the more documents in which a word occurs, the lower IDF value it has. IDF captures the generalization that words of low indexing value such as function words occur in most documents.

6. Scoring Function: All sentences in the bilingual texts which contain a matched word with the input sentence are retrieved. The similarity of the input sentence and the retrieved sentences is computed based on the sum of the IDF of the matched terms. The sentence with the highest score is considered most similar to the input query.

The algorithm above proposes a simple measure to calculate semantic similarity, which is used in building an intelligent web-based Chinese and English bilingual concordancer. To test its performance, we create a bilingual database by extracting bilingual sentences from a bilingual dictionary. The interface and the output of the program are shown in Figure 5 – 7. As can be seen from Figure 6 and 7, the program works reasonably well and is a valuable and convenient tool to search for appropriate English expressions. To enhance its functions, more Chinese-English bilingual corpora will be collected. Meanwhile, more robust and sophisticated methods of performing sentence alignment and calculating semantic similarity will be experimented in our next NSC project.

References

Dagan, I., Church, W, and Gale, W. (1993) "Robust Bilingual Word Alignment for Machine Aided Translation." In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp. 1-8, Ohio.

Dagan, I. (1996) "Bilingual Word Alignment and Lexicon Construction." Tutorial paper given at the International Conference on Computational Linguistics, Copenhagen.

Fung, P. and Church, K. (1994) "K-vec: A New Approach for Aligning Parallel Texts." Proceedings of the International Conference of Computational Linguistics, pp.1096-1102, Kyoto.

Fung, P. and McKeown, K. (1994) "Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping." Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation, pp. 81-88.

Fung, P. and KcKeown, K. (1997) "A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups." Machine Translation, Vol. 12, Nos. 1-2., pp. 53-87.

Gao, Zhao-Ming. (1997) Automatic

Extraction of Translation Equivalents from A Parallel Chinese-English Corpora. Ph.D. Thesis. Department of Language Engineering. University of Manchester Institute of Science and Technology.

Hutchins, J. (1998) The Origins of the Translator's Workstation. Machine Translation, Vol 13, No. 4, pp. 287-307.

Haruno, M. and Yamazaki, T. (1996) "High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information." Proceedings of Annual Conference of the Association for Computational Linguistics, pp. 131 -138.

Jones, D. and Somers, H. (1995) "Bilingual Vocabulary Estimation from Noisy Parallel Corpora Using Variable Bag Estimation." In JADT III GiornateInternazionali di Analsi Statistica dei Dati Testuali, pp. 255-262, Rome.

Kay, M. and Röscheisen, M. (1993) "Text-Translation Alignment." Computational Linguistics, Vol. 19, No 1, pp 121-142.

Kay, M. (1997) "The Proper Place of Men and Machines in Language Translation", Machine Translation, Vol. 12, No. (1/2), pp.3-23, 1997

Ker, S.-J. and Chang, J.-S. (1997) "A Class-based Approach to Word Alignment." Computational Linguistics, Vol. 23., No. 2, pp. 313-343.

Kumano, A. and Hirakawa, H. (1994) "Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information." in Proceedings of International Conference on Computational Linguistics, pp. 76-81, Kyoto.

Macklovitch, M. and Hannan, M. (1998) "Line 'Em Up: Advances in Alignment Technology and their Impact on Translation Support Tools", Machine Translation, Vol. 13, No. 1, pp. 41-57.

Melamed, D. (1995) "Automatic Construction of Clean Broad-Coverage Translation Lexicons." In Proceedings of 2nd Conference of the Association for Machine Translation in the Americas, Montreal.
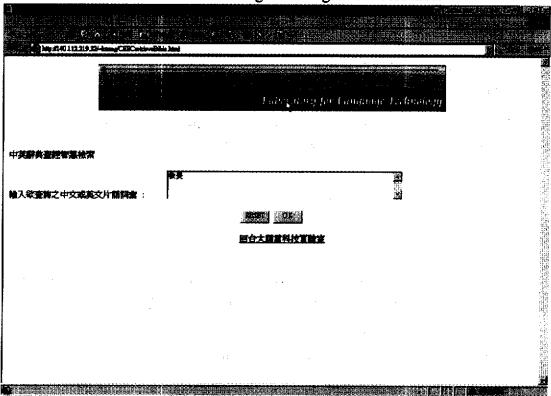
Somers, H. and Ward, A. (1996) "Some More Experiments in Bilingual Text Alignments." In Oflazer, K. and Somers, H.(eds) Proceedings of the Second International Conference on New Methods in Language Methods in Language Processing, pp. 66-78, Ankara.
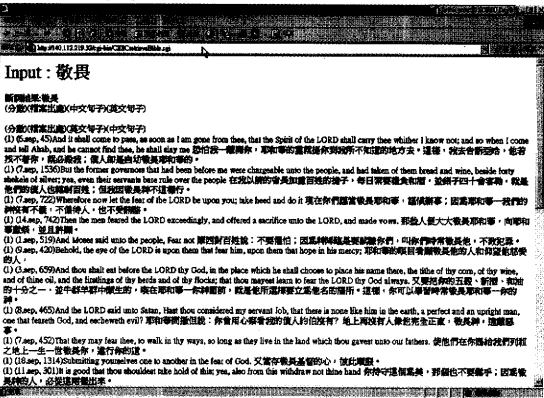
Utsuro, T. et al. (1994) "Bilingual Text Matching Using Bilingual Dictionary and Statistics." in Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. and Xia, X. (1995) "Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon." Machine Translation, Vol. 9, pp. 285-313.
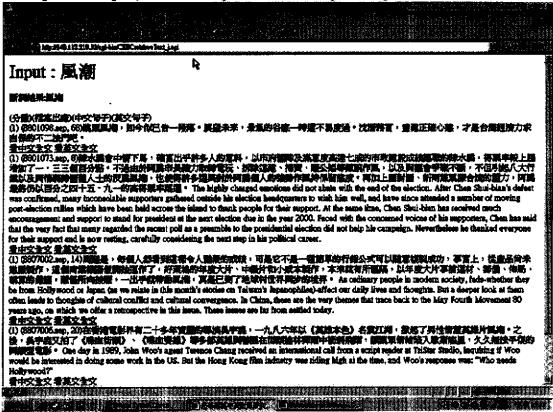
1. The User Interface of the Chinese-English Bilingual Concordancer
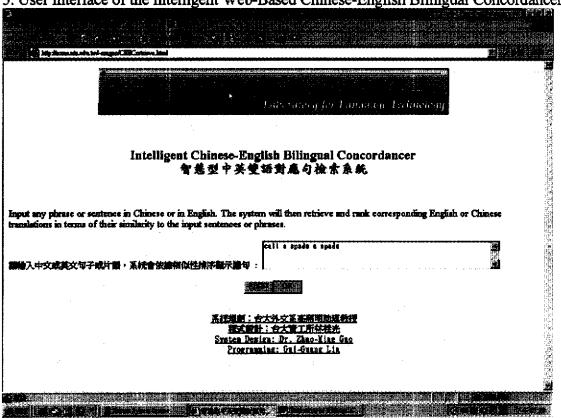

2. Output of the query "敬畏"

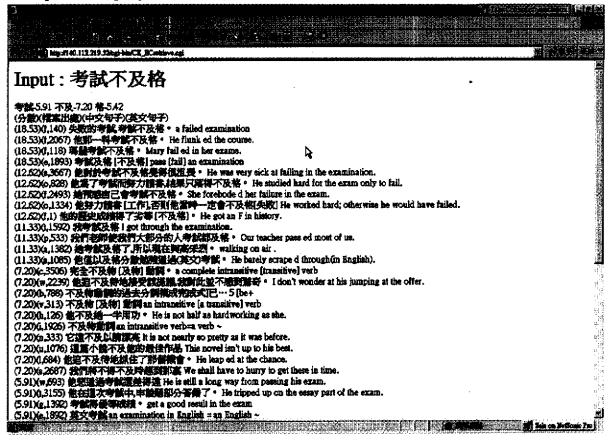## 3. Output of the query "風潮"using Sinorama Bilingual Magazine

4. Users can view the whole texts by clickiing the hyperlinks.

5. User Interface of the Intelligent Web-Based Chinese-English Bilingual Concordancer

Intelligent Chinese-English Bilingual Concordancer
智慧型中英雙語對應句檢索系統

Input any phrase or sentence in Chinese or in English. The system will then retrieve and rank corresponding English or Chinese translations in terms of their similarity to the input sentences or phrases.

請輸入中文或英文句子或片語，系統會依據相似性排序顯示語句：

> call a spade a spade

系統規劃：台大外文系高照明助理教授
程式設計：台大電工所林桂光
System Design: Dr. Zhao-Ming Gao
Programming: Gui-Guang Lin

6. Output of the query "call a spade a spade"

Input : call a spade a spade

After stemming: call a spade a spade
After deleting stop words: call spade spade

(分數)(檔案出處)(英文句子)(中文句子)
(5) (s, 5663)call a spade a spade 直截了當地說, 直言不諱。
(2) (s, 5664)spade a garden 以鏟挖掘花園。
(2) (s, 5665)the seven[queen]of spade s 黑桃七[皇后]。
(1) (r, 1606)call to remembrance 想起
(1) (r, 3145)come back before roll call 在點名之前回來。
(1) (w, 2133)within call 在呼叫能聽得見的範圍內
(1) (t, 3271)call a truce 休戰
(1) (s, 8231)call up suggestion s of the sea 使人聯想到海。
(1) (r, 3138)call the roll 點名。
(1) (s, 3344)call the shots ((口語))命令, 指揮((◆源自「下令射擊」之義))
(1) (s, 5333)I call that some poem. 我認為那首詩很棒。
(1) (t, 3318)a call for trump s 要對方出王牌的信號
(1) (r, 3634)I'll call you in, each in your turn . 我會依次逐一地叫你們進來。
(1) (s, 3672)call in sick 打電話請病假。
(1) (t, 1304)I'm trying to think what to call it. 我正在設法給它取個名字 [考慮要稱它什麼]。
(1) (r, 3143)have a roll call 點名
(1) (r, 3144)skip(the) roll call 省略點名
(1) (t, 3469)call the tune 按自己的意思發號施令[指示]。
(1) (t, 134)All it take s is a phone call. 只要打通電話就行了。
(1) (s, 6788)a state call 正式[官事]訪問
(1) (t, 554)call a person on the telephone 打電話給某人
(1) (t, 555)call a person to the telephone 叫某人聽電話

7. Output of the query "考試不及格"

Input：考試不及格

考試-5.91 不及-7.20 格-5.42
(分數)(檔案出處)(中文句子)(英文句子)
(18.53)(f,140) 失敗的考試 考試不及格 • a failed examination
(18.53)(f,2067) 他那一科考試不及格 • He flunk ed the course.
(18.53)(f,118) 瑪麗考試不及格 • Mary fail ed in her exams.
(18.53)(e,1893) 考試及格 [不及格] pass [fail] an examination
(12.62)(a,3667) 他對於考試不及格是很沮喪 • He was very sick at failing in the examination.
(12.62)(o,828) 他為了考試而努力讀書,結果只落得不及格 • He studied hard for the exam only to fail.
(12.62)(f,2493) 她預感自己會考試不及格 • She forebode d her failure in the exam.
(12.62)(o,1334) 他努力讀書[工作],否則他當時一定會不及格[失敗] He worked hard; otherwise he would have failed.
(12.62)(f,1) 他的歷史成績得了丙等[不及格] • He got an F in history.
(11.33)(i,1592) 我考試及格 I got through the examination.
(11.33)(p,533) 我們老師使我們大部分的人考試都及格 • Our teacher pass ed most of us.
(11.33)(a,1382) 她考試及格了,所以現在興高采烈 • walking on air .
(11.33)(s,1085) 他僅以及格分數勉勉強強通過(英文考試) • He barely scrape d through(in English).
(7.20)(c,3506) 完全不及物 [及物] 動詞 • a complete intransitive [transitive] verb
(7.20)(w,2239) 他迫不及待地接受該提議,我對此並不感到驚奇 • I don't wonder at his jumping at the offer.
(7.20)(b,798) 不及物動詞的過去分詞構成完成式已…5 [be+
(7.20)(v,313) 不及物 [及物] 動詞 an intransitive [a transitive] verb
(7.20)(h,126) 他不及她一半用功 • He is not half as hard working as she.
(7.20)(i,1926) 不及物動詞 an intransitive verb=a verb ~
(7.20)(n,333) 它遠不及以前漂亮 It is not nearly so pretty as it was before.
(7.20)(u,1076) 這篇小說不及他的最佳作品 This novel isn't up to his best.
(7.20)(l,684) 他迫不及待地抓住了那個機會 • He leap ed at the chance.
(7.20)(s,2687) 我們將不得不趕緊到那裏 We shall have to hurry to get there in time.
(5.91)(w,693) 他想通過考試還差得遠 He is still a long way from passing his exam.
(5.91)(o,3155) 他在這次考試中,申論題部分落第了 • He tripped up on the essay part of the exam.
(5.91)(g,1392) 考試得優等成績 • get a good result in the exam
(5.91)(e,1892) 英文考試 an examination in English = an English ~

8. Output of the query "我今天心情很低落"

Input：我今天心情很低落

我-1.51 今天-5.33 心情-6.48 很-3.26 低落-8.36
(分數)(檔案出處)(中文句子)(英文句子)
(15.3)(f,808) ((我)我今天情緒低落 [高昂] I feel down [up] today.
(14.84)(h,2404) 心情鬱悶,懊喪,情緒低落 get the bump
(11.35)(f,1485) 我覺得很好((指心情或健康)) I feel first-class .
(10.2)(f,1496) 今天早上我覺得精神很好 • I feel first-rate this morning.
(10.2)(t,2159) 今天我很忙 T today I am very busy.
(9.74)(e,1934) 他心情很愉快 • He was in excellent spirits.
(9.74)(m,1506) 覺得心情很好 • feel like a million (dollars)
(8.59)(w,661) 今天海上的浪很大 • The sea is very wavy today.
(8.59)(f,1914) 今天很熱 • It is flipping hot today.
(8.59)(h,123) 今天天氣很好 • It's not half fine,today.
(8.59)(a,3238) 今天過得真快 [令人覺得很短暫] • Today was a short day.
(8.59)(h,2129) 今天很熱 • It is hot today.
(8.59)(e,1942) 除了這裏 [南方] 以外,今天各地的天氣都很好 except here [in the South].
(8.59)(c,2809) 今天很冷 It is cold today.
(8.59)(n,699) 今天很暖和 • It is nice and warm today.
(8.59)(p,1125) 今天很冷 • It's perishing today.
(8.59)(d,2337) 今天的天氣真的很溫和 • It is distinctly warm today.
(8.36)(m,2219) 士氣高 [消沉,低落] • M morale is high [low,falling].
(8.36)(c,1550) 那消息使士兵們士氣低落 • The news chill ed the morale of the soldiers.
(8.36)(d,3350) 這幾天他似乎情緒十分低落 • His spirits seem to be droop ing these days.
(8.36)(a,838) 逢有韻律的,不押韻的,(後有韻律的品質很落的) • I
(8.09)(g,505) 這首曲子使我心情動盪 • That tune get s me.
(8.09)(g,2879) 我沒有氣力 [心情]做這件事 • I can't rise to it.
(8.09)(m,2156) 我沒有心情 [沒心情]工作 • I was in the [no] mood for work.
(8.09)(u,366) 我了解你的心情 I understand how you feel.
(6.94)(v,806) 我今天早上發不出聲音 • I have no voice this morning.
(6.94)(w,1221) 你造就了今天的我 • You have made me what I am today.

9

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

附件三

報告人姓名	高照明	服務機構 及職稱	國立台灣大學外國語文學系 助理教授
會議　時間 　　地點	6 月 1 日—2 日 香港理工大學	本會核定 補助文號	NSC 89-2411-H-002-074 90 臺會綜二字 0021987 號
會議 名稱	（中文）2001 年國際資訊科技與多媒體英語學習會議 （英文）ITMELT2001		
發表 論文 題目	（中文）語料庫句型檢索系統 （英文）A Pattern Retrieval Program for Exploring Corpus-Based Phraseology		

表 Y04

報告內容應包括下列各項：

一、參加會議經過

此次會議的主辦單位是香港理工大學 Hong Kong Polytechnic University, 共 29 篇論文發表，參與者將約 3 百人，會議的規模相當龐大。除了來自香港各大學與國外的學者外大都是中小學教師。因此有幾個場次是 penal session 或 tutorial。

二、與會心得

有 2 篇論文值得注意。Meena Singhal 博士所發表的 CALL for Reading in English: an Interactive Web Program for College-Level ESL Students 以網路互動式的教學方法訓練以英語為第二語言的學生閱讀的各種策略，進而增進閱讀與寫作的能力。,Birgid Winkler 博士所發表的 Working with an English Learners' Dictionary on CD-ROM 以實際觀察與訪談紀錄探索學生使用電子辭典的方式，從而瞭解一般英語學習者的需求。

三、考察參觀活動(無是項活動者省略)

四、建議

1. 從 Meena Singhal 與 Birgid Winkler 及其他學者的論文中，我們可以發現電腦輔助語言教學不只是運用現有資訊科技，包括人工智慧，第二語言習得理論，測試理論等其它相關理論與實證研究是電腦輔助語言教學成功的重要基礎。
2. 鑑於國內對於尖端電腦輔助語言教學理論與實際仍然遠落後歐美甚至香港，國科會與教育部有必要推動此一領域之研究。

五、攜回資料名稱及內容

Bruce Morrison, David Gardner, Ken Keobke, and Mary Spratt (2001) ELT Perspectives on IT & Multimedia. Selected Papers from the ITMELT 2001 Conference 1st & 2nd, 2001. Published by the English Language Cetre, The Hong Kong Polytechnic University.

六、其他

國內雖有類似會議但水準有待提升。

表 Y04

A Pattern Retrieval Program for Exploring Corpus-Based Phraselogy

GAO Zhao-Ming
Department of Foreign Languages and Literatures
National Taiwan University
zmgao@ccms.ntu.edu.tw
http://ccms.ntu.edu.tw/~zmgao/pattern.htm

Data-Driven Language Learning

- Tim Johns (1991) "language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data."
- Vance Stevens(1993) "With DDL, the learner's own discovery of grammar based on evidence from authentic language use becomes central to the learning process."

The Importance of Concordancers in Data-Driven Language Learning

- Gregory Hadley (forthcoming) "Data-driven learning is very much a "research-then-theory" method of studying grammar. Language learners start with a question, and then come to their conclusions after analyzing the corpora with a concordancer program."

Vance Stevens(1993)

- "… concordancers are the type of software that most closely approaches fulfilling the potential of computers in language learning. In a sense, they are working approximations of expert systems. They bring cognitive and analytic skills in students to bear on the manipulation of comprehensive databases for the purpose of solving real-language problems."

Key Function of Concordancers

- A concordancer allows a user to explore authentic language use by inspecting the context of an input keyword in the corpora.

Major Inadequacy of Traditional Concordancers

- They often display hundreds or even thousands of lines of unorganized data. Looking for a particular usage is just like looking for a needle in a hay.

Some of the Features of an Advanced Concordancer

- 1. It should support queries involving syntactic patterns.
- 2. It should facilitate the retrieval of collocations.

What is a pattern?

- Hunston and Francis (2000", p. 37)
- "The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning.

The Role of Patterns in Lexicography, Language Learning, and Linguistics

- 1. Subcategorization: Chomsky (1957, 1965)
- Verb Patterns:
- 2. Hornby (1975) *Guide to Patterns and Usage in English*
- *Oxford Advanced Learners' Dictionary*
- *3. Collins Cobuild Grammar Patterns*
- Hunston and Francis (2000) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*

The Role of Patterns in Lexicography, Language Learning, and Linguistics

- 4. Collocations: Benson et al. (1997) *The BBI Dictionary of English Word Combinations*
- 5. Levin (1993) *English Verb Classes and Alternations*
- 6. Construction Grammar: Fillmore, Kay, and Goldberg

What can be found using our pattern retrieval program?

- 1. Verb Patterns
- 2. Idioms
- 3. Collocations

2

English Pattern Retrieval Program

Input a pattern consisting of keywords and the following syntactic categories.
[N V ADJ ADV O NP VP PP INF S *] e.g. input V responsibility to find a verb followed by the word responsibility.

Or Select from the following patterns.

VNPS

English Pattern Retrieval Program

Input a pattern consisting of keywords and the following syntactic categories.
[N V ADJ ADV O NP VP PP INF S *] e.g. input V responsibility to find a verb followed by the word responsibility.

Or Select from the following patterns.

VNPS

English Pattern Retrieval Program

Input a pattern consisting of keywords and the following syntactic categories.
[N V ADJ ADV O NP VP PP INF S *] e.g. input V responsibility to find a verb followed by the word responsibility.

Or Select from the following patterns.

VNPS

How is the program implemented?

- 1. Corpus Collection
- 2. Identify Sentence Boundary
- 3. Preprocess the corpus with Brill's Part-of-Speech Tagger
- 4. Use Abney's Partial Parser
- 5. Index words and syntactic structures

The output of Brill's tagger

- Such/JJ an/DT act/NN is/VBZ abhorrent/JJ to/TO my/PRP$ sense/NN of/IN right/NN ./.

Output of Abney's Parser

Future Research

- 1. Increase the corpus size.
- 2. Enhance the accuracy of structure match using other taggers and parsers.
- 3. Allow more flexible pattern search.
- 4. Optimalize the system.
- 5. Investigate the relationship of meaning and patterns.