行政院國家科學委員會專題研究計畫　成果報告

學術英文學習者語料庫之建立與分析

計畫類別：個別型計畫
計畫編號：NSC92-2411-H-009-011-
執行期間：92 年 08 月 01 日至 94 年 01 月 31 日
執行單位：國立交通大學語言教學與研究中心

計畫主持人：郭志華

計畫參與人員：郭志華

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 4 月 26 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　　　　學術英文學習者語料庫之建立與分析　　　　※
※　　　EAP Learner Corpus: Development and Analysis　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫
計畫編號：NSC92－2411－H－009－011
執行期間：92 年 8 月 1 日至 94 年 1 月 31 日

計畫主持人：郭志華

執行單位：國立交通大學語言教學與研究中心

中華民國九十四年四月三十日

行政院國家科學委員會專題研究計畫成果報告

學術英文學習者語料庫之建立與分析

EAP Learner Corpus: Development and Analysis

計畫編號：NSC92-2411-H-009-011
執行期限：92 年 8 月 1 日至 94 年 1 月 31 日

主持人：郭志華　國立交通大學語言教學與研究中心

一、　中文摘要

學術英文(English for Academic Purposes, EAP)對非母語人士一直是困難重重，尤其學術論文之寫作，即使英文程度良好的專業人士也常常無法寫得和母語人士神似(native-like)。研究指出，這種非母語性質有時並非是錯誤，而是是否使用某些字彙及句構，或使用之多寡所造成，因此語言學者亟欲瞭解這種非母語性質的來源。學習者語料庫之建立與分析提供了探討這個問題的一個實證的研究方法。

隨著電腦科技的進步，語料庫語言學(corpus linguistics)的研究近年來愈來愈受重視。學習者電腦語料庫(computer learner corpus)之建立與分析結合了語料庫語言學與第二語言習得(Second Language Acquisition, SLA)之研究方法與理論，以蒐集之大量學習者之語料，進行電腦分析，探討學習者語言使用之特色、模式、錯誤、或非母語性質(non-nativeness)，且可進而與母語人士語料比較，深入瞭解學習者之中階語(interlanguage)。

本計畫因此建立了一個學術英文學習者語料庫以探討分析我國學習者在學術英文寫作上之語用特色、模式、及弱點。我們也建構一個對應的英語為母語人士的學術英文語料庫，用以比較分析。兩個語料庫都包含人文(應用語言學)和科學(積體電路設計)兩個領域已發表之期刊或會議論文。我們分別分析了兩個語料庫中論文的幾個語用方面，包括語料統計分析(如全部及平均字數，全部用字與不同用字比例、高頻率用字、功能字與內容字等)、第一人稱代名詞、助動詞、搭配字、及一些特別用字如 *given*，也探討學術英語中常用的句構如句尾副詞性分詞片語及被動句型。然後我們比較兩個語料庫的分析結果。

分析結果顯示，高級程度的學術英語學習者和英語為母語人士在用字文法及一些我們探討的學術英語特色上仍有細微地方之差異。例如我們發現我國學術論文寫作者論文的整體用字密度較低，而同一字的出現頻率較高，低頻率字佔的比率較低，所用的搭配字較少。同時比較高頻率字及摘要字的搭配動詞也顯示出我國論文作者較少用或不會用的字。在句構上則有特殊句型如 *given* 的特別用法及 *using* 的句尾副詞性分詞片語，這些顯示了我國學術英文學習者用字及句型上尚待補強的地方。

另一方面，分析也顯示學習者和母語人士在用字和句型上相似的地方，這些則表示我們的作者已經擁有的英文或學術英文的知識。

要幫助我國學術論文學習者有效寫作如母語人士般的論文，我們不只需要錯誤分析，還需要有進一步更細微的語用分析。

關鍵詞：學術英語、語料庫語言學、學習者語料庫、中階語

Abstract

English for Academic Purposes (EAP), particularly writing research articles, has been problematic to non-native learners. Even advanced professional non-native writers are not able to write native-like research articles. Studies have indicated that this problem is not necessarily caused by errors, but rather by the use, or lack of use, of some words or structures.

With advances in computer technology, corpus linguistics has gained great momentum recently. Research on computer learner corpora can combine research methodologies and theories of corpus linguistics and Second Language Acquisition

(SLA) to investigate the features, patterns, errors, or non-nativeness of learner language from both qualitative and quantitative perspectives. With the help of linguistic analysis software, we can also compare learners' language use with that of native speakers so that a better picture of learners' interlanguage can be drawn.

This study, therefore, constructs an EAP learner corpus to investigate the characteristics and language use patterns of Chinese EAP learners. A corresponding native-speaker corpus is also constructed for comparison. Each corpus contains RAs in two fields: IC design and applied linguistics, representing engineering and humanities research respectively. We explore various aspects of language use in both corpora including general profile (such as total and average word forms, type-token ratio, high-frequency word list, and content/function words ratio), first-person pronouns, modals, collocation, and special academic vocabulary such as *given*. Also, the common structural patterns in academic writing such as final adverbial participial phrases and passive constructions are examined. Comparisons between native and learner, and engineering and humanities are made, respectively.

Results from analyses reveal that advanced EAP learners show subtle but distinct differences from native EAP writers in terms of not only general lexico-grammatical writing ability but a number of academic linguistic features. For example, based on some parameters of our inquiry, that is, word profile, lexical density, vocabulary span, and collocation, it is found that the learner texts show a lower lexical density but higher token-type ratio (i.e., the recurrence rate of words), lower percentage of low-frequency words, and fewer collocates of words. Qualitative analysis such as comparison of the top 200 high-frequency words or collocations of important summative nouns in two corpora further reveals the words that EAP learners do not use or use rarely. Structurally, sophisticated structural patterns such as various pre-posed "given" patterns or final adverbial participial clauses occur much more frequently in native texts than in our advanced learner texts.

On the other hand, the learner texts show some similar lexico-grammatical usages which are characteristic of academic writing. This may reflect the aspects of academic writing that our advanced EAP learners already acquire.

To write more effective and native-like research papers, our advanced EAP learners need more than errors analysis.

Keywords: EAP, corpus linguistics, learner corpus, interlanguage

二、緣由與目的（**Introduction**）

Corpus linguistics refers to linguistic research, primarily quantitative, based on the collection and analysis of natural language data. It emerged in the early sixties when a number of linguistic researchers began to question the validity of intuition-based linguistic analysis and description. With advances in computer technology, computer corpus has attracted a growing interest and contributed to many fields of language-related research such as frequency analysis of lexico-grammatical features of text, collocation analysis, register or language variation analysis, lexicography, and even machine translation. As Leech (1992) indicated, corpus linguistics provides a new way of thinking about language, which is challenging some of our most deeply-rooted ideas about language.

Until very recently, however, rare attempts have been made to use computer corpus in SLA research, particularly learner language (Granger 1998). Current SLA research is mainly based on introspective data and language use data of the elicited type (Ellis 1994). The compilation of computer learner corpus enables SLA researchers to access and assess (1) empirical data, (2) not only errors but also more sophisticated linguistic behavior such as avoidance and overuse (Granger 1998), (3) learning difficulties, (4) what Granger calls "Contrastive Interlanguage Analysis"

(1998: 12), and (5) specialized text such as EST, or genre such as research articles (Flowerdew 2002).

Methodologically, research on computer learner corpora combines corpus linguistics and Second Language Acquisition (SLA) to investigate the features, patterns, errors, or non-nativeness of learner language from both qualitative and quantitative perspectives. Studies have investigated the design criteria of learner corpus or the construction of computer analysis system (Biber 1990; Thomas & Short 1996; Granger 1998; Pienemann 1992; Jagtman & Bogaerts 1994). On the other hand, a number of studies have already used learner corpora to explore the interlanguage of specific groups of learners (Milton & Freeman 1996; Milton & Hyland 1997; Lorenz 1998; Upton & Connor 2001).

In the 80s and 90s, corpus-based research has focused on the construction of large-scale, general-purpose corpora, such as the well-known British National Corpus (BNC). In the past few years, nevertheless, the use of small-scale, specialized genre-based corpora has gradually been recognized (Flowerdew 1996); most of them are compiled for the specific purpose of EAP research. These corpora are mainly collections of academic writing samples of native speakers and used to inform EAP pedagogy of "standard" or "native" models of academic writing (Flowerdew 2002).

From a perspective of SLA, nevertheless, we need to learn how non-native EAP writers deviate from native norms of EAP, and what specific language use patterns characterize their interlanguge. We can also compare their writing samples with those of native writers to find out the sources of their "non-nativeness" as well as their possible learning difficulties. Only a couple of studies have concentrated in this area. (Granger 1993; Milton 1998, 1999; Upton & Connor 2001). Results from such studies have revealed interesting interlanguage contrasts for EAP pedagogy.

This study, therefore, constructs an EAP learner corpus to investigate the characteristics and language use patterns of Chinese EAP writers. A corresponding native-speaker corpus is also constructed for comparison. In addition, as EAP learners are often advanced learners who are already equipped with a certain amount of lexico-grammatical and genre knowledge and able to produce grammatical sentences but whose writing works are still regarded "non-native," we focus our investigation on the possible sources of non-nativeness, such as under-use or over-use of certain words or structures, rather than language use errors. The learner corpus, consequently, consists of published RAs by Chinese writers. Three major goals of the study are as follows:

1. To compile a genre-based, special-purpose EAP learner corpus as well as a corresponding NS corpus;
2. To investigate the interlanguage of advanced Chinese EAP writers in terms of vocabulary span, lexical density, collocation, genre-specific grammatical structure, etc.;
3. To identify possible sources of "non-nativeness" of these writers on the basis of a comparison between learner corpus data and NS corpus data.

With respect to research methodology, we first collected published RAs by both native and non-native writers in two fields: IC design (*IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*) and applied linguistics (*English for Specific Purposes*), representing engineering and humanities research respectively. We then explored various aspects of language use in both corpora (learner corpus and NS corpus) including general profile (such as total and average word forms, type-token ratio, high-frequency word list, and content/function words ratio), first-person pronouns, modals, collocation, and special academic vocabulary such as *given*. Also, the common structural patterns in academic writing such as final adverbial participial clauses and passive constructions were examined. Comparisons between native and learner, and engineering and humanities were made, respectively.

三、結果與討論（Results and Discussions）

We developed and analyzed an EAP learner corpus and compared it with a corresponding NS corpus to characterize the language use patterns of Chinese EAP learners. Computer software was used to conduct frequency analysis and concordance. Results from our analysis are reported below.

Corpus Composition and General Text Statistics

Table 1 shows the composition of the two corpora. The total running words of them are close, 173807 and 176995, respectively. However, the learner corpus consists of 40 RAs while the NS corpus consists of only 20 RAs. The average length of RAs differs greatly.

Table 1 Composition of the Two Corpora

Corpora/ composition	Learner Corpus	NS Corpus	Total
Running words	173807	176995	350802
No. of RAs	40	20	60
Ave. length	4345	8850	

Table 2 demonstrate general text statistics of the two corpora, revealing differences of the two in type /token ratio (i.e., lexical density) and token/type ratio (average frequency of word) . The higher lexical density implies that native writers generally use more different words in RAs than learners. This suggests that Chinese learners may have a smaller vocabulary repertoire. On the other hand, the higher token/type ratio of the learner corpus suggests that each word occur more times.

Table 2 Text Statistics of the Two Corpora

Corpora/ statistics	Learner Corpus	NS Corpus	Total
Tokens*	173807	176995	350802
Types*	9767	11029	10736
Type-token	$562/10^4$	$623/10^4$	
Tokentype	17.8	16.0	

*Token refers to running words, or word-forms.

*Type refers to different word-forms.

Frequency Analysis

Table 3 below shows the results from a frequency analysis. The frequencies of the 100 most frequent words are counted and the percentages of text these words constitute are calculated. Studies have hypothesized that the high-frequency words (overused) would have consistently higher percentages in learner corpus than in NS corpus (Ringbom 1996). Our data, however, do not support this hypothesis. The main difference seems still lying in the size of vocabulary, which is reflected in the percentage of low-frequency words (40.18% vs. 43.45%).

Table 3 Word Frequency Profile (The 100 most frequent words) of the Two Corpora

Corpora/ frequency rankings	Learner Corpus (% of text)	NS Corpus (% of text)
1-10	23.05	24.41
1-20	28.73	30.64
1-30	32.12	33.96
1-40	34.84	36.30
1-50	37.13	38.35
1-60	39.23	40.20
1-70	41.11	41.79
1-80	42.80	43.21
1-90	44.39	44.49
1-100	45.87	45.68
(freq. of 1)	40.18	43.45

Analysis of the top 200 high-frequency words gives a consistent result. (As a result of the limitation of space, it is impossible to provide the list here.) 124 of the 200 high-frequency words (62%) are common, and 76 words (38%) are different. Among the different words, we find such words as *according, could, question, remaining, shows, order* in learner corpus, while *following, may, would, problem, present, sequence, found, given* are in NS corpus.(Some of these words aroused our interest and were further explored in terms of their collocations; this will be reported in the later part of this report.) The function words/content words ratios of the two corpora are very close (95/105 in

learner corpus and 98/102 in NS corpus).

Linguistic Features

Table 4 below shows a summary of the results from analyses of three linguistic features in the genre of RA. (Detailed data such as further breakdown of various modals are omitted here to save space.) We can observe that the occurrences of passive constructions in the two corpora do not differ (in terms of quantity). On the other hand, interestingly, NS writers use much more first-person pronouns than Chinese writers. This suggests that Chinese EAP writers may learn and follow the teaching of many EST/EAP style manuals, that is, avoid personal involvement and hence avoid the use of first-person pronouns in formal academic writing, while NS may think it appropriate to use first-person pronouns to express their commitment and emphasize their contribution in some parts of RA (though they use as many passives as Chinese writers). NS EAP writers also use more modals than Chinese writers. Modality is one of the semantic-grammatical features of language. It is mainly concerned with the opinion and attitude of the speaker/writer. RA writers tend to use modals to express various degrees of certainty, probability, expectation, and tentativeness. They also use modals to qualify their research results in order to show modesty. Chinese EAP writers may under-use modals as a result of their lack of knowledge of special functions of some modals such as *would* and *must*, and over-use some modals they are more familiar with such as *can* and *could.*

Table 4 The Use of Passive, Modals, and First-Person Pronouns

Corpora/ Features	Learner Corpus (occurrences)	NS Corpus (occurrences)
Passive	3149	3104
Modals	1545	1779
First-Person Pronouns	926	1584

As indicated earlier, it is found that *given* appears in the list of top 200 high-frequency words of the NS corpus, but not in that of the learner corpus. We are interested in looking into the various usages of this word by both NS and Chinese writers in RAs. Table 5 below shows the use of "given" in the two corpora. The data show very sharp contrast between the two corpora. It is obvious that Chinese writers under-use *given* as a result of unfamiliarity with some special usages of this word which do not occur at all in the learner corpus.

Table 5 The Use of "given"

given	pattern	Learner Corpus (IC)	NS Corpus (IC)
Pre-posed	"a given" + n.	5	60
	"any given" + n.	0	6
	"the given" + n.	0	30
	subtotal	5	96
Post-posed	n. + "given"	0	11
	"given" as prep.	39	30
Total		44	137

One of the author's previous studies (Kuo 1998) has identified a grammatical structure which characterizes RAs with a number of discourse functions -- the adverbial participial clause. The pre-posed participial clause is easier to interpret as temporally preceding its main clause and occurs often in general English text; however, it has limited rhetorical functions, mainly providing a contingency or a cause to what is described in the main clause. The post-posed adverbial participial clause can perform a much wider range of rhetorical functions, including describing a subsequent event, providing a reason or result, giving an accompanying explanation or purpose, and indicating means or condition; The structure occurs frequently in NS RAs, as indicated earlier. Since the previous study revealed that the participial clause of *using* is particularly prevailing in RAs since most RAs need to indicate the means of research, which may be methodology, equipment, materials, etc. In this study, we, therefore, examine the occurrences of *using* at pre-posed and post-posed position, and its alternative prepositional phrase *by using*. As shown in Table 6, NS writers use *using* as pre-posed or post-posed adverbial participial clause much more frequently than Chinese writers. The difference is especially distinctive in the

post-posed part. In contrast, Chinese writers use *by using* much more frequently than NS writers. These results suggest that Chinese RA writers may be unfamiliar with the rhetorical functions of this grammatical structure.

Table 6 The Use of "*using*"

using	Learner Corpus (IC)	NS Corpus (IC)
Pre-posed	6	21
Post-posed	29	145
by using	53	18
Total	88	184

Collocation: Summative Nouns

To further explore possible differences in the use of collocation, we choose three summative nouns which appear in the top 200 high-frequency word lists of both corpora: *approach*, *model* and *system*. Summative nouns are important as they function to introduce or present the major research concept, product, or method. Then we search all verb collocates of each of these summative nouns in either corpus. Table 7 provides the collocates of *approach* we find in each of the two corpora. We can observe that NS writers can use more verb collocates with summative noun *approach.* This again shows how Chinese writers differ from NS RA writers in active vocabulary.

Table 7 Collocates of *approach*

Corpora	Learner Corpus (IC)	NS Corpus (IC)
Verb Collocates of *approach*	propose adopt present use describe extend show establish take demand apply employ be based (on) consider provide incorporate	follow be based (on) consider develop use outline present propose lack describe report disfavor explore pursue evaluate apply rely (on) involve adopt employ advance take advocate derive compare choose try
Total	16	27

四、計畫成果自評（Self-evaluation）

This project compiles and analyzes an EAP learner corpus. It further compares it with a corresponding NS corpus. Specifically, text statistics, frequency analysis, linguistic features, and collocation of summative nouns are explored. The results show several aspects of RAs in which advanced Chinese EAP learners deviate from NS writers. For example, the lower lexical density implies that Chinese writers may have a smaller vocabulary repertoire than NS writers. A consistent result is shown in the percentage that low-frequency words constitute in text, particularly words that occur only once in the

corpus. Low-frequency words in the learner corpus constitute a smaller percentage than in the NS corpus. Analysis also reveals that NS writers use much more first-person pronouns and more modals than Chinese writers. With respect to the specific usages of words common in EAP, it is found that Chinese EAP writers are not familiar with a number of phraseological patterns of "given." In addition, NS EAP writers use the post-posed adverbial participial clauses, such as "using" to indicate means of research, more often than Chinese EAP writers. These results provide significant implications for EAP writing pedagogy as well as materials development for NNS EAP learners.

On the other hand, the study is limited in a couple of aspects as a result of the size of both corpora and the fields we choose. We are not very sure about the generalizability of the results to other fields.

However, we are convinced that computer learner corpus research is opening a new horizon for both linguistic descriptions and SLA research.

五、參考文獻 (References)

Biber, D. (1990). Methodological Issues regarding Corpus-Based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, 5, 257-269.

Ellis, R. (1994). *The Study of Second Language Acquisition.* Oxford University Press, Oxford.

Flowerdew, J. (1996). Concordancing in Language Learning. In Pennington, M. (ed.) *The Power of CALL*. Houston, TX: Athelstan, 97-113.

Flowerdew, L. (2002). Corpus-Based Analyses in EAP. In Flowerdew, J. (ed.) *Academic Discourse.* Pearson Education limited, Harlow, 95-114.

Granger, S. (1993). International Corpus of Learner English. In Aarts, j., de Haan, P., & Oostdijk, N. (eds.) *English Language Corpora: Design, Analysis and Exploitation.* Rodopi, Amsterdam, 57-71.

Granger, S. (1998). *Learner English on Computer*. Longman, London and New York.

Jagtman, M. & Bogaerts, T. (1994). COMOLA: A Computer System for the Analysis of Interlanguage Data. *Second Language Research*, 10, 1, 49-83.

Leech, G. (1992). Corpora and Theories of Linguistic Performance. In Svartvik, J. (ed.) *Directions in Corpus Linguistics.* Mouton de Gruyter, Berlin, 105-122.

Lorenz, G. (1998). Overstatement in Advanced Learners' Writing: Stylistic Aspects of Adjective Intensification. In Granger, S. (ed.) *Learner English on Computer*. Longman, London and New York, 53-66.

Milton, J. & Freeman, R. (1996). Lexical Variation in the Writing of Chinese Learners of English. In Flowerdew, L. & Tong, K. K. (eds.) *Entering Text.* The Hong Kong University of Science and Technology, Hong Kong, 127-143.

Milton, J. & Hyland, K. (1997). Qualification and Certainty in L1 and L2 Students' Writing. *Journal of Second Language Writing*, 6, 2, 183-205.

Milton, J. (1998). Exploiting L1 and Interlanguage Corpora in the Design of an Electronic Language Learning and Production Environment. In Granger, S. (ed.) *Learner English on Computer*. Longman, London and New York, 186-198.

Milton, J. (1999). Lexical Thickets and Electronic Gateways: Making Texts Accessible by Novice Writers. In Candlin, C. & Hyland, K. (eds.) *Writing: Texts,*

Processes and Practices. Longman, London and New York, 221-243.

Pienemann, M. (1992). COALA – A Computational System for Interlanguage Analysis. *Second Language Research,* 8, 1, 59-92.

Ringbom, H. (1998). Vocabulary Frequencies in Advanced Learner English: A Cross-Linguistic Approach. In Granger, S. (ed.) *Learner English on Computer.* Longman, London and New York, 41-52.

Thomas, J. & Short, M. (eds.) (1996). *Using Corpora for Language Research.* Longman, London and New York.

Upton, T. A. & Connor, U. (2001). Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre. *English for Specific Purposes*, 20, 4, 313-329.