Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 87-100

Compiling Taiwanese Learner Corpus of English

Rebecca Hsue-Hueh Shih *

Abstract

This paper presents the mechanisms of and criteria for compiling a new learner corpus of English, the quantitative characteristics of the corpus and a practical example of its pedagogical application. The Taiwanese Learner Corpus of English (TLCE), probably the largest annotated learner corpus of English in Taiwan so far, contains 2105 pieces of English writing (around 730,000 words) from Taiwanese college students majoring in English. It is a useful resource for scholars in Second Language Acquisition (SLA) and English Language Teaching (ELT) areas who wish to find out how people in Taiwan learn English and how to help them learn better. The quantitative information shown in the work reflects the characteristics of learner English in terms of part-of-speech distribution, lexical density, and trigram distribution. The usefulness of the corpus is demonstrated by a means of corpus-based investigation of learners' lack of adverbial collocation knowledge.

Keywords: learner corpus, Taiwanese Learner Corpus of English (TLCE), Second Language Acquisition (SLA), English as Foreign Language (EFL), quantitative analysis, lexical density, collocation.

1. Introduction

A computer corpus is a body of computerized written text or transcribed speech. Computer corpora are useful for a wide variety of research purposes, in fields such as lexicography, natural language processing, and all varieties of linguistics. The first computer corpus made its appearance in the early 1960s when two scholars at Brown University compiled a one-million-word corpus, known as *the Brown Corpus* [Francis & Kucera, 1964]. It contains a wide range of American English texts with grammatical annotation. For decades, this pioneering work was an important source for linguistic scholars who wished to perform quantitative as well as qualitative that is crucial for a broad coverage system. Third, a static WSD model is unlikely to be robust and portable, since it is very difficult to build a single model relevant to a wide

* Department of Foreign Languages and Literature, National Sun Yat-sen University, Kaohsiung, Taiwan
E-mail: hsuehueh@mail.nsysu.edu.tw

analysis of language structure and use [Francis & Kucera, 1982]. In the early 1970s, an equivalent British collection, *the Lancaster-Oslo-Bergn* (LOB) Corpus, was designed and compiled to facilitate comparative studies. Quantitative information on the distribution of various linguistic features in these two corpora became available [Johansson & Norheim, 1988; Nakamura, 1993]. The two corpora and other subsequently compiled corpora are similar in structure and size, and are considered to be first generation corpora.

With the fast development in technology needed for text capture, storage and analysis, the scale of computer corpora has increased considerably, and a corpus of one million words seems to be inadequate for large scale studies on lexis. In the early 1980s, the publisher Collins and Birmingham University compiled the first mega-size corpus, the Cobuild Corpus, for the production of a new English dictionary. The scale of the corpus reached 13-million words by the time the dictionary was published in 1987 [Collins, 1987]. In preparation for a new generation of language reference publications, the corpus was transformed into *the Bank of English* in 1991 and has been growing larger in size ever since. Another well-known mega-corpus, *the British National Corpus*, was compiled between 1991 and 1994 by a consortium of academics and publishing houses. This corpus consists of 100 million words of part-of-speech tagged contemporary written and spoken British English. Access to the corpus was originally restricted within Europe, and it was not until very recently that the corpus was made accessible worldwide. Due to the need for comparative studies of different English varieties as in the first generation, *the International Corpus of English* compilation project was launched in the 1990s [Greenbaum, 1996] to gather written and spoken forms of national varieties of English throughout the world. The project aims to collect up to 20 subcorpora, each containing one million words of English used in countries where English is the first language, and in countries such as India and Singapore where English is an additional office language. The corpus will enable researchers to use each national subcorpus independently for descriptive research and also to undertake comparative studies.

For nearly fifty years, machine-readable language corpora have greatly benefited people in both linguistics and publishing houses. Linguistic scholars have been able to better understand language structure and use with the aid of quantitative data. Publishers have produced new pedagogical tools that reflect the real use of language. However, it was not until the 1990s that scholars in the EFL and SLA sectors began to recognize the theoretical as well as practical potential of corpora and to believe that with the aid of quantitative information, computer learner corpora can form an authoritative basis for obtaining further insights into the interlanguage systems of language learners. Publishing houses also realize the vital role that learner corpora play on designing EFL tools, which can be improved "with the NS (native speaker) data giving information about what is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners" [Granger, 1998a] However, it is difficult to create learner corpora on the huge scale of native corpora mainly because each collection is usually confined to classroom language.

In 1993 *the International Corpus of Learner English* (ICLE) was launched [Granger, 1993] through academic collaboration worldwide. At present, the corpus contains 14 different national varieties, some of which are subdivided regionally, and each subcorpus contains 200,000 words. A great deal of comparative research has been done based on the ICLE, providing statistics-based interpretation of the learners' lexicon, grammar, and discourse [Granger, 1998b]. Another learner corpus, and probably the largest corpus of single group learners so far, is *the Hong Kong University of Science and Technology Learner Corpus* [Milton & Tong, 1991], which consists of five million words of written English from Cantonese learners. This corpus is intended to be used for the development of English teaching materials in Hong Kong. SLA scholars in Japan soon followed the trend, and several learner corpus projects were launched, such as *the JEFLL corpus* of around 200,000 words from Japanese EFL learners' written data, *the SST Corpus* of 1 million spoken words of learners, and *the CEJL Corpus* of junior high school to university students. In China, *Chinese Middle School Students' Written English* and *Chinese Middle School Students' Spoken English* are two learner corpora forming *the Corpus of Middle School English Education* that was compiled at South China Normal University beginning in 1998. Apart from academic circles, publishing houses such as Longman and Cambridge University Press have also compiled their own learner corpora for the development of their own language related publications.

While many countries around the world have been creating their own learner corpora, little work has been done in Taiwan. *The Soochow Colber Student Corpus* [Bernath, 1998], which was compiled between 1984 and 1995 at Soochow University, can be viewed as a pioneering Taiwanese corpus of learner English. It contains around 227,000 words of written text from junior and senior students of Soochow University and National Taiwan University. No other corpus of comparable size was compiled until 1999 when a one-million-word learner corpus project, *the Taiwanese Learner Corpus of English*, was launched at Sun Yat-sen University. This corpus is a collection of written data from college students majoring in English at the university. The data has been annotated for various linguistic features using the TOSCA-ICLE tagger/lemmatizer [Aarts, Barkema, & Oostdijk, 1997], assigning to each word its lemma and a tag of its morphological, syntactic and semantic information. With the permission of the compiler of the Soochow Colber Student Corpus to incorporate 85% of its contents, consisting of written data from students majoring in English, the scale of the TLCE has increased from its original 530,000 to 730,000 words. The corpus continues to grow in size. Currently, the TLCE is probably by far the largest annotated learner corpus of English in Taiwan. In the following sections, a complete description of the TLCE will be given, including its purpose, design criteria, method of data capture and documentation, corpus structure and grammatical annotations. The quantitative characteristics of the TLCE as well as its pedagogical application will be depicted and illustrated at the end of the paper.

2. Compilation of the TLCE

2.1 Purpose

The history of the computer learner corpus is less than a decade old, but it has been widely considered as "a useful resource for anyone wanting to find out how people learn languages and how they can be helped to learn them better" [Leech, 1998]. Learner output is indeed hard data that SLA scholars can utilize to depict learners' interlanguage systems. The TLCE has been compiled in the hope that it will become a useful resource for SLA scholars who want to understand the internal learning process of Taiwanese learners of English, and in the hope that with corpus-based research findings, EFL teachers will be able to tailor their teaching to students' needs.

2.2 Corpus Design Criteria

It is important to have clear design criteria when compiling a learner corpus because of the heterogeneous nature of learners and learning situations. Clear criteria help make it possible to interpret research results correctly and help justify the results of comparative studies on different corpora.

Table 1 shows the design criteria of the TLCE. The subjects who have contributed data to the corpus are students majoring in English at the three universities, ranging from freshmen to seniors (aged 19 to 22). Their English proficiency varies from intermediate to advanced levels. The TLCE includes written production of two different genres, namely, informal writings and essay writings. Informal writings consist of daily or weekly journals, which the learners are encouraged to keep during their writing courses, and essay writings are the compositions they are asked to submit regularly for their courses. The types of compositions are mainly descriptive, narrative, expository and argumentative.

Table 1. *TLCE Design Criteria*

attributes	
age	19-22
level	Intermediate to advanced
Mother tongue	Chinese
Learning context	EFL
medium	Written text
genre	journals and compositions

2.3 Data capture and documentation

The data of the TLCE are in three forms: electronic files, printouts and handwritten texts. More than half of the collection has been submitted through e-mail, which is the easiest way of gathering data for the corpus. E-mail or Microsoft Word files are converted into text files. Another source of data, learners' printouts, have been scanned and transformed into a machine readable format. Post-editing of the scanned data is

necessary to remove scanning errors. The most time-consuming task is the collection of handwritten texts; all the data have to be keyboarded. As the issue of spelling errors is not a concern in the project and errors would hinder part-of-speech tagging in the subsequent annotation work, all the data in the corpus are spellchecked.

The documentation of each piece of writing is needed for researchers to create their own subcorpora according to selection based on pre-defined attributes, and to carry out different comparison studies. For this reason, details about attributes are recorded as an SGML file header for each text. The information includes the university where the learner is studying, the academic year in which the text is collected, the school year (proficiency level) of the learner, and the genre of the text. For instance, the header

<#nsysu-891-f-DES>

indicates that the text is a descriptive composition written by a freshman at Sun Yat-sen University in the first semester of the 1989 academic year.

2.4 Corpus structure

As stated in Section 2.2, journals and compositions are the two genres of writing collected in the corpus. Journals are informal writings from students, recording what concerns them the most during a day or a week. The journals are sent to their teachers through e-mail systems. Compositions are the essay writings based mainly on different writing strategies: description, narration, exposition and argumentation. The first two are often taught in the first year at universities, whereas the expository and argumentative types are practiced in the second and the third years. Table 2 illustrates the structure of the corpus, including the total numbers of texts and words, and the percentage of the corpus each genre represents.

Table 2. *The Structure of the TLCE*

Text Types	**Total number of texts**	**Total number of words**	**Proportion (%)**
journal	823	213091	29.4
composition			
Description/narration (first year)	435	134363	18.5
Exposition/argumentation (second/third years)	738	333734	46.1
others	109	43156	6.0

As indicated in the table, the ratio of journals to compositions in the corpus stands at around 3 to 7. Expository and argumentative types of writings are most numerous, making up more than 46% of the whole corpus. Data classified as *others* came originally from *the Soochow Colber Student Corpus* with type labels that did not fit into the TLCE categories. For instance, they are labeled as autobiographical writings, letters, imaginative writings or creative writings.

2.5 Grammatical Annotation

Computer corpora are either raw corpora or annotated corpora. Raw corpora simply contain plain text, whereas annotated corpora have extra encoded features obtained through part-of-speech tagging or syntactic parsing. Part-of-speech tagging is a process of attaching a category and probably other attributes to each word, whereas syntactic parsing provides the structural analysis of each sentence. The former is usually done automatically by rule-based, probabilistic or mixed taggers, and the average tagging accuracy is about 95%; the latter can be done by automatic full/partial parsers outputting one or more syntactic structures for a sentence.

The text in the TLCE is currently part-of-speech tagged using the TOSCA-ICLE tagger [Aarts et al., 1997]. TOSCA-ICLE is a stochastic tagger, supplemented with a rule-based component, which tries to correct observed systematic errors of the statistical components. Each word is given its lemma, and a part-of-speech tag, which consists of a major wordclass label, followed by attributes for subclasses and for its morphological information. There are 17 major word classes in the tag set (see Appendix A) and a total of 270 different attribute combinations.

3. Quantitative Analysis

A major advantage of the corpus approach lies in the usefulness for conducting quantitative analysis. The quantitative features of a corpus provide a basic but global view of the characteristics of the learners' writings. The following findings depict the characteristics of the TLCE as a learner corpus.

3.1 Part-of-speech Distribution

Figure 1 shows the part-of-speech distribution of the corpus. The graph only indicates those parts of speech individually making up at least 5% of the total corpus. As can be seen, Nouns (N) and verbs (VB) exist in similar proportions in the corpus. Pronouns (PRON) are third, followed by prepositions (PREP), adverbs (ADV), adjectives (ADJ), articles (ART) and conjunctions (CONJUNC). Note that the words in nominal form (N or PRON) make up nearly one third of the whole corpus.

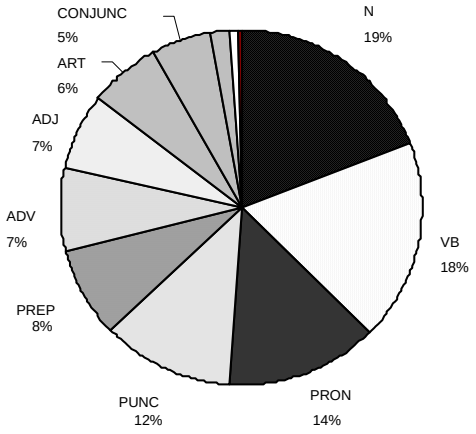

Figure 1. *POS Distribution*

3.2 Type/Token Ratio (Lexical Density)

For open classes, N, VB, ADV, and ADJ, it is desirable to know their type/token ratios. The type-token ratio, also called the lexical density, is often used as a measure of the lexical complexity of a text. Here, it is used as the measure of the word versatility of an open class. It is the ratio of different words to the total number of words in the class and is calculated by the formula

$$Lexical\_Density = \frac{number\_of\_separate\_words(type)}{total\_number\_of\_words(token)} * 100 .$$

Although N and VB have similar distributions as shown in Figure 1,their lexical densities show great discrepancy. As can be seen in Figure 2, the lexical density of N is four times higher than that of VB. This phenomenon is also found in the pair consisting of ADJ and ADV, where ADJ has a much higher density value than ADV. In other words, although the frequency counts of VB and ADV in the learner corpus are similar to those of N and ADJ, respectively, the variety of actual words used in the categories of VB and ADV is much more limited than in the N and ADJ categories.

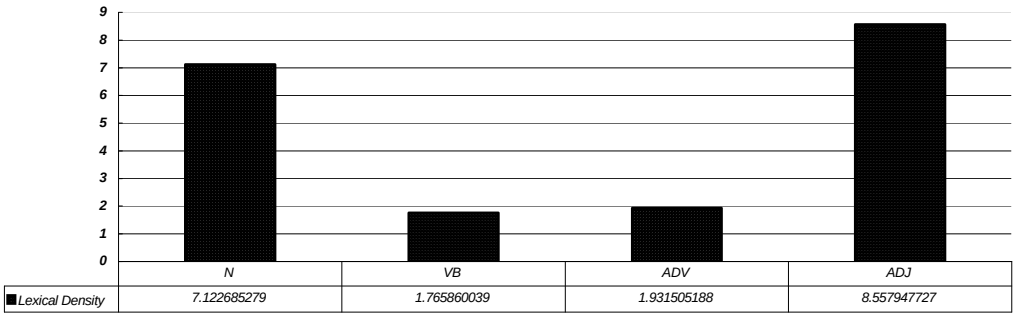

Figure 2. *Lexical Density*

3.3 Part-of-speech trigrams

A POS trigram is a pattern of three adjacent POSs. It reveals to a certain extent the habitual use of syntactic structures by language learners. The corpus has a total of 777,096 trigrams from 2202 different patterns. Hence, the type-token ratio of POS trigrams is as low as 2.8. Table 3 shows the distribution of the front rank trigram patterns according to frequency of use. As can be seen, the first 50 patterns make up a large proportion of use in the distribution diagram. In fact, it is calculated that the top 220 ranking patterns make up 82% of the trigrams. In other words, learners use only 10% of the POS trigram patterns in 80% of their writings. These figures demonstrate the serious lack of structural variations in learners' writings.

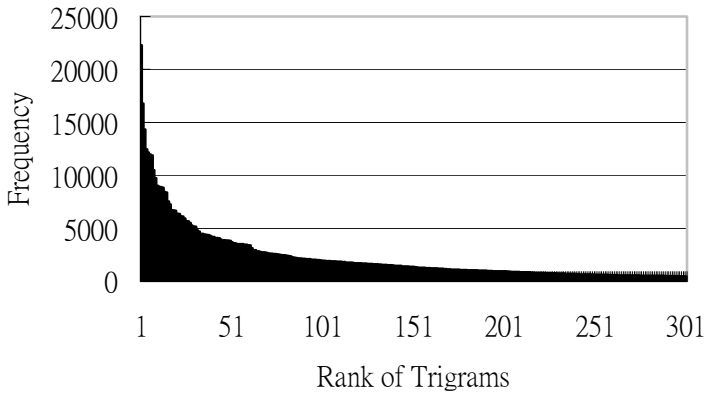

Figure 3. *Trigram Distribution*

4. Pedagogical Application

The main purpose of compiling the TLCE is to provide Taiwanese researchers in the SLA and EFL areas with a large quantity of authentic learner data, which can be used to conduct qualitative analysis based on quantitative information. With the availability of this useful resource, they can utilize advanced corpus analysis tools to systematically uncover the features of non-nativeness existing in learner English. The findings will enable EFL teachers to focus on areas where remedial work is needed. In this section, a pedagogical application of the TLCE is demonstrated through an investigation of learners' lack of adverbial collocation knowledge from both overuse and underuse perspectives. A series of experiments were carried out based on a contrastive approach, comparing learner English (from the TLCE) with native English (from a one-million-word subset of the BNC).

4.1 Top 10 adverbs in the BNC and the TLCE

A frequency list of adverbs with the "-ly" suffix was obtained from each corpus, and their top 10 adverbs were taken into consideration. The left column of Table 3 shows the top 10 adverbs used by the learners, and the right column shows those used by the native speakers. The bracketed number following an adverb indicates the adverb's rank in the other corpus.

Table 3. *Top 10 adverbs in the two corpora.*

	Top NTLCE(learner)	BNC (native)
1	really(1)	really(1)
2	finally(12)	probably(23)
3	usually(4)	*particularly(96)*
4	especially(10)	usually(3)
5	suddenly(l8)	actually(8)
6	easily(I3)	early(22)
7	recently(20)	certainly(27)
8	actually(5)	nearly(32)
9	*deeply(60)*	simply(51)
10	quickly(14)	especially(4)

As can be seen in the table, four of the top 10 adverbs in the TLCE appear in the BNC list, namely, *really*, *usually*, *especially* and *actually*. The rest fall into BNC's top 20 group except for *deeply*, whose counterpart is ranked 60. This implies that *deeply* is very overused by Taiwanese learners. By contrast, *particularly*, with the third highest rank in the BNC list, is the one least used by the learners. Sections 4.2 and 4.3 provide a closer examination of these two phenomena, respectively, based on the contexts in which they appear.

4.2 Overuse phenomenon

This experiment examined the context in which *deeply* appears in the TLCE. The adverb can be used to intensify adjectives or verbs. According to the estimation of Mutual Information, the top 10 adjectives or verbs that highly collocate with *deeply* those listed in the left column of Table 4.

The middle and right columns show the adverbs (including *deeply*) which are used by the learners and native speakers, respectively, to intensify words. They are listed in the descending order of their joint frequencies with the corresponding words. As can be seen, *deeply* seems to be chosen most often when learners wish to use an adverb to modify these words, whereas in the BNC, the native speakers use other synonyms (words in bold type) more frequently than *deeply* to intensify the same set of words. *Extremely distressed*, *strongly/greatly influenced*, *greatly impresse*d, *strongly/greatly attracted*, *firmly convinced* and *extremely confused* are collocations that do not exist in the TLCE. This finding suggests that instead of the

monotonous use of *deeply*, Taiwanese learners should be made aware of native speakers' strong preference for the above collocations.

Table 4. *Adverb Alternatives*

Intensified words	**Adverbs in TLCE**	**Adverbs (Synonyms) in BNC**
Distressed	Deeply	extremely, deeply, …
Influenced	deeply, directly, rapidly	strongly, greatly, deeply,…
Moved	deeply, really, suddenly, …	deeply,…
impressed	deeply, especially, really	greatly, deeply,…
attracted	deeply, really, fully	strongly, greatly, deeply,…
convinced	deeply, obsessively	firmly, deeply,…
touched	really, deeply	deeply,…
concerned	deeply, obsessively	deeply,…
confused	deeply, really	extremely, deeply,…
interested	really, deeply, keenly	deeply,…

4.3 Underuse phenomenon

To understand the learners' use of *particularly*, its concordancing lists from the corpora were investigated. There are only 4 instances of the adverb in the TLCE, whereas in the BNC, there are 217 examples. Following is the complete TLCE list and a selected sample of the BNC list:

TLCE concordancing list

First of all, the government, <particularly> the Ministry of Administration,
self-defense, the teachers, <particularly> the elementary school teachers,
is still applied universally , <particularly> in cram schools for high schools
take your words seriously, <particularly> in foreign countries. They might

BNC concordancing list (selected)

ncy food aid in 1990. 'We're <particularly> concerned about the situation in
may have been linked with a <particularly> violent six-week strike by rail
n French international thinking <particularly> over France's role as the motor
ront and other radical groups, <particularly> among the rapidly expanding
he past six months, and many, <particularly> the US, are expected to argue
st and provide grants for artists, <particularly> students, in the region. Thr
strial and social development, <particularly> after Renault was nationalised in
hey still have a very useful role, <particularly> when it is the function of t
the landscape study shown here. 1 <particularly> liked the rounds for their v
hot poker. These colours work <particularly> well in late summer and early

As can be seen in the TLCE concordancing list, there are only two different functions of *particularly* in the learners' writings: it is used to modify either a noun phrase or a preposition phrase. However, there are more functions of the adverb in the native speakers' writings. Apart from noun and prepositional phrases, the native speakers also use it to intensify clauses, verb phrases, adjectives and even

adverbs. Table 5 shows the percentage of each of the grammatical functions used in each corpus. The findings are two fold. First, the learners seem to possess limited knowledge of *particularly*'s grammatical behaviours. Only two out of the six functions are actually found in the TLCE. Second, the learners are not clear about the possible uses of *particularly*. Its collocation with adjectives makes up 42% of the BNC examples, the highest among all, but yet it is not used in this way by the learners at all. The above findings suggest that learners should be informed of the grammatical function of the adverb during the learning process.

Table 5. *Distribution of Grammatical Collocations of "particularly"*

Grammatical collocation	**BNC(%)**	**TLCE(%)**
ADJECTIVE	42	-
PREPOSITION PHRASE	28	50
NOUN PHRASE	15	50
CLAUSE	7	-
VERB	6	-
ADVERB	2	-

5. Summary and Outlook

This is the first large-scale tagged Taiwanese learner corpus of English. Preliminary results show several interesting characteristics of the learner corpus in terms of its part-of-speech distribution, the lexical density of its main categories, and the distribution of its trigram structures. An example of pedagogical application has been used to illustrate the usefulness of the corpus. These efforts have been made in the hope that scholars in language education and research will benefit from this pioneer learner corpus, which will be made available soon on website with software tailored to facilitate corpus analysis.

Acknowledgements

Financial support from the National Science Council of the Republic of China of this work under contract No. NSC 89-2411-H-110-024 is gratefully acknowledged. I would also like to express my gratitude to Colman Bernath, the compiler of the Soochow Colber Student Corpus, for allowing his corpus to be incorporated into the TLCE. I am also greatly indebted to the following colleagues for helping me collect data: Dr. Shu-ing Shyu, Dr.Ching-yuan Tsai, Dr. Shu-li Chang, Dr. Shu-Fang Lai, Dr.Yuan-jung Cheng, Hue-jen Wen, Alex K.T. Chung, Chu-jen Loh, Dr. Ting-yao Luo at Sun Yat-sen University and Dr. Zhao-ming Gao at National Taiwan University.

References

Aarts, J., Barkema, H., & Oostdijk, N. 1997. *The TOSCA-ICLE Tagset* . Nijmegen: University of Nijmegen, The Netherlands.

Bernath, C. 1998. *Soochow Colber Student Corpus.* Available: ftp://ftp.scu.edu/tw/scu/english/colber.

Collins. 1987. *COBUILD English Language Dictionary*. London and Glasgow: Collins.

Francis, W., & Kucera, H. 1982. *Frequency Analysis of English Usage:Lexicon and Grammar*. Boston: Houghton Mifflin.

Francis, W. N., & Kucera, H. 1964. *Manual of Information to accompany ' a Standard Sample of Resent-Day Edited American English, for Use with Digital Computers'* Department of Linguistics, Brown University.

Granger, S. 1993. The International Corpus of Learner English. In J. Aarts, P. d. Haan, & N. Oostdijk (Eds.), *English language Corpora: Design, Analysis and Exploitation.* pp. 57-69 Armsterdam: Rodopi.

Granger, S. 1998a. The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer*. pp. 3-18 London and New York: Longman.

Granger, S. (Ed.). 1998b. *Learner English on computer*. London and New York: Longman.

Greenbaum, S. (Ed.). 1996. *Comparing English Worldwise: The Interational Corpus of English*. Oxford: Clarendon Press.

Johansson, S., & Norheim, E. H. 1988. The subjunctive in British and American English. *ICAME Journal* (12), 27-36.

Leech, G. 1998. Preface. In S. Granger (Ed.), *Learner English on Computer* . New York: Longman.

Milton, J., & Tong, K. (Eds.). 1991. *Text Analysis in Computer Assisted Language Learning*. Hong Kong: Hong Kong University of Science and Technology.

Nakamura, J. 1993. Quantitative comparison of modals in the Brown and LOB corpora. *ICAME* (17), 29-48.

Appendix A: part of speech set of TOSCA Tagger

Label	Major word class
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	Proforin
PRON	Pronoun
PRTCL	Particle
PUNC	Punctuation
VB	Verb