外國語文研究第十二期
2010 年 06 月頁 71~98

國立政治大學外語學習者語料庫之建構

鍾曉芳　王淑儀　曾郁雯*

中文摘要

近年來學習者語料庫的建置與研究愈來愈受到重視，然多數學習者語料庫仍是以英語為主，較少著重在其他外語學習語料庫之建構。國立政治大學外國語文學院乃於其新建計畫中，提出了以英語、日語、韓語、法語、俄語及阿拉伯語等六種語言之學習者為主的學習者語料庫計畫，參與計畫之教授於課程中所蒐集的學生寫作文本為本計畫主要語料。本文將詳述本學習者語料庫之建置過程，包含其建置理念、未來展望及相關應用。透過學習者語料庫，相關研究人員和語言教學老師們能更了解台灣外語學習者在學習過程語言使用之特性與困難，亦期盼達到教學成效與學術研究成果之提升。

關鍵詞：學習者語料庫、語料庫、語言學、英語教學、語料庫介面、外語教學

* 國立政治大學英國語文學系
98.10.30 到稿　98.12.27 通過刊登

The Construction of The NCCU Foreign Language Learner Corpus

Siaw-Fong Chung, Shu-Yi Wang, Yu-Wen Tseng*

Abstract

Greater interest is being shown in learner corpora in recent years. Many learner corpora exist for English but only a few for other languages. The NCCU Foreign Language Learner Corpus is a newly-created learner corpus including texts in six languages – English, Japanese, Korean, French, Russian and Arabic. This corpus is composed of learners' assignments in various forms written in the different languages collected by participating professors of this project. This corpus thus provides details of the linguistic features of Taiwanese students in their process of learning different foreign languages. This paper outlines the details in the creation of this corpus, including its rationales, future prospects and possible applications of this corpus. The corpus will be beneficial to both researchers and language teachers who intend to investigate Taiwanese learners' production of a specific foreign language.

Key Words: learner corpus, corpora, foreign language learning, interface, language teaching

* Department of English, National Chengchi University.

The Construction of The NCCU Foreign Language Learner Corpus

Siaw-Fong Chung, Shu-Yi Wang, Yu-Wen Tseng

1. Introduction

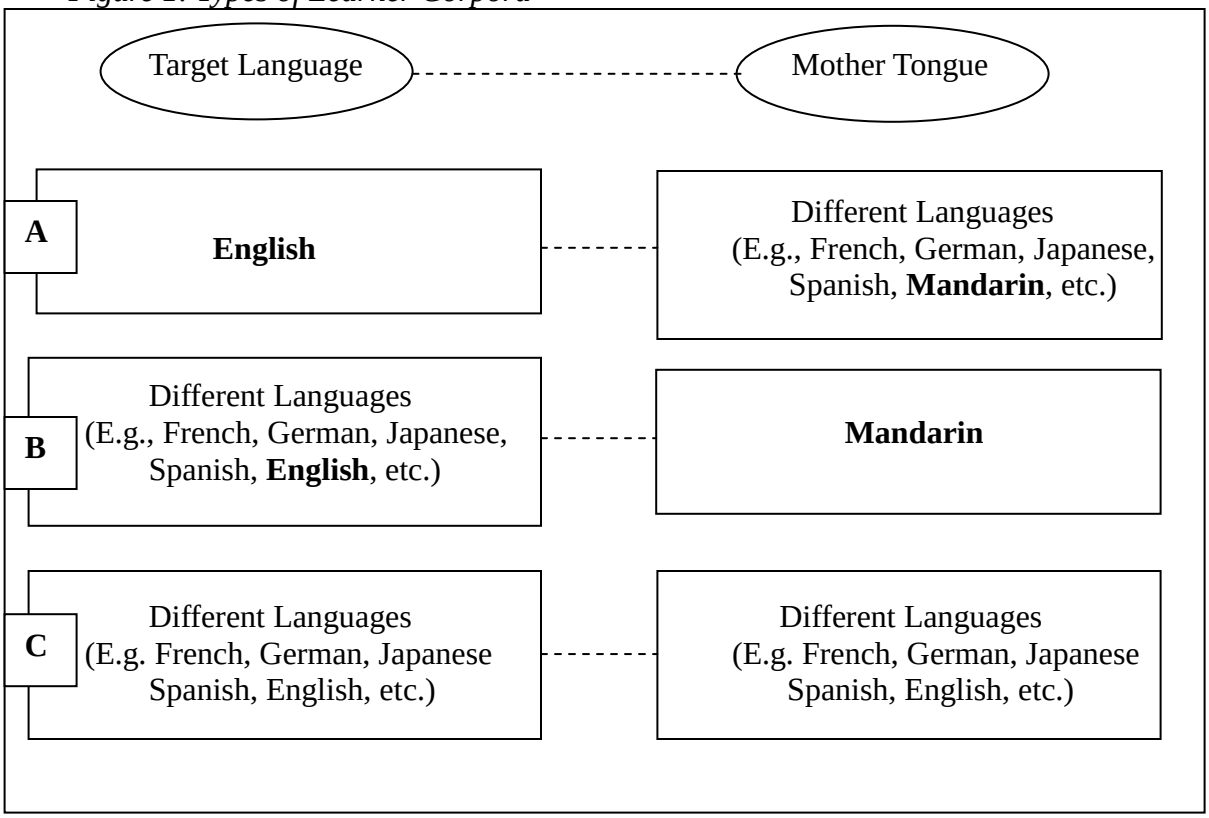
Learner corpora usually refer to a collection of written and/or spoken texts produced by foreign or second language learners. These types of corpora document data verbatim from learners' production of a target language in which specific features such as errors or non-standard characteristics in the learners' language are considered as interlanguage (Selinker, 1972; Corder, 1981) between the mother tongue and the target language. The most often used methodology in analyzing a learner corpus is contrastive interlanguage analysis (CIA) (cf. Gilquin, 2001; Granger, 1996), a method in which features in a learner corpus are checked with those in a reference corpus which is based on native speaker data. When comparing content of the two corpora, the existence of certain features or the lack of them will be considered as specific characteristic of the learners' learning process.

To date, many learner corpora of English have been created and these corpora include English data by foreign or second language learners of various backgrounds. The International Corpus of Learner English (ICLE) (Granger, *et, al*., 2009) is an established learner corpus documenting learners of different mother tongue backgrounds in Europe. ICLE also collected data written by Chinese studying in the Europe. As for learner corpora based on texts produced by Chinese learners, the Spoken and Written English Corpus of Chinese Learners or SWECCL (Wen, Wang, & Liang, 2005 & 2007) from China is a collection of test materials based on the English produced by Chinese learners of English in China. A recent Taiwan-based learner corpus of English has been collected by the Language Testing and Training Center (LTTC) based on texts produced by examinees taking the General English Proficiency Test (GEPT) (cf. Cheung, Chung & Skoufaki, 2010). These corpora are all based on texts produced by learners of English. There are few learner corpora that are of texts produced by learners of other foreign languages and there are even fewer which comprise of a collection of foreign languages within one same corpus. CPATEI (Spanish-English Learners Written Parallel Corpus) (Lu & Lu, 2009) is a project in Taiwan which collects learner data in Spanish produced by Taiwanese learners.[1] The same project also collected data from texts in Japanese, German and Chinese written by Taiwanese learners. Another project is the project of International Corpus of Crosslinguistic Interlanguage (ICCI).[2] The ICCI project aims both at collecting data from learners of English as well as from learners of different foreign languages in countries such as Austria, China (Hong Kong), Israel, Poland, Singapore, Spain and from Taiwan. The Taiwan data in the ICCI project come mainly from students studying foreign languages at the LTTC. These projects have a similar aim – to collect data from learners of various mother tongue backgrounds who are

[1] http://corpora.flld.ncku.edu.tw/
[2] http://cblle.tufs.ac.jp/llc/icci/

learning different target languages. The following Figure 1 summarizes the three main types of learner corpora.

Figure 1: Types of Learner Corpora



In Figure 1 above, most of the existing learner corpora fall under type 'A' with English as the target language produced by learners from various language backgrounds. Type 'B' is a different kind of learner corpus because only one type of learners is targeted at – learners whose mother tongue is Mandarin Chinese. In type 'A,' learners whose mother tongue is Mandarin constitute part of the many types of learners' language backgrounds. As for type 'B,' learners who speak Mandarin Chinese as their mother tongue constitute the only type of language background while the targeted languages are many, including English which, in contrast, is the only targeted language in type 'A'. In type 'C', language data are produced from learners of different language backgrounds who are learning different target languages.

In this paper, we will detail the construction of a learner corpus based on learners at National Chengchi University (NCCU) who are learning different languages, i.e., type 'B' in Figure 1. This newly created learner corpus is called the NCCU Foreign Language Learner Corpus (hereafter NCCU Learner Corpus), which is funded by the College of Foreign Languages and Literature in NCCU under the NCCU Top-Universities Program. The main objective of this learner corpus project is to establish a corpus of different

languages by collecting NCCU learners' written texts in both soft- and hardcopies. In terms of data collection, the College of Foreign Languages and Literature in NCCU is privileged in the sense that it includes language courses taught in twenty-three different languages. Therefore, in terms of learning environment, NCCU provides a good resource of data collection based on Taiwanese learners of various foreign languages. Since learners of various target languages can be found in NCCU, a learner corpus built from these languages will benefit research in the fields of second language teaching and language pedagogy.

The above are some of the motivations which explain the rationale behind the establishment of a foreign language learner corpus in NCCU. The overall aim is to enhance the quality of language education and to boost research using local based data. As of the second semester of the academic year of 2008, there were sixteen participating professors in this project and they are experts in the following languages: English, French, Japanese, Korean, Russian, and Arabic. At this stage, only written assignments have been collected for these languages. Spoken data will only be considered at a later stage in the development of the learner corpus.

In this paper, we introduce the features of the NCCU Learner Corpus and at the same time, we provide documentation of how this corpus came into shape. We review some learner corpora and discuss the steps necessary to create our learner corpus, all of which are crucial information for the construction of a learner corpus. In addition, we also provide future prospects of this learner corpus and discuss the applications of the corpus. In the section below, we first review two of the learner corpora that we have mentioned previously – the ICLE and the SWECCL.

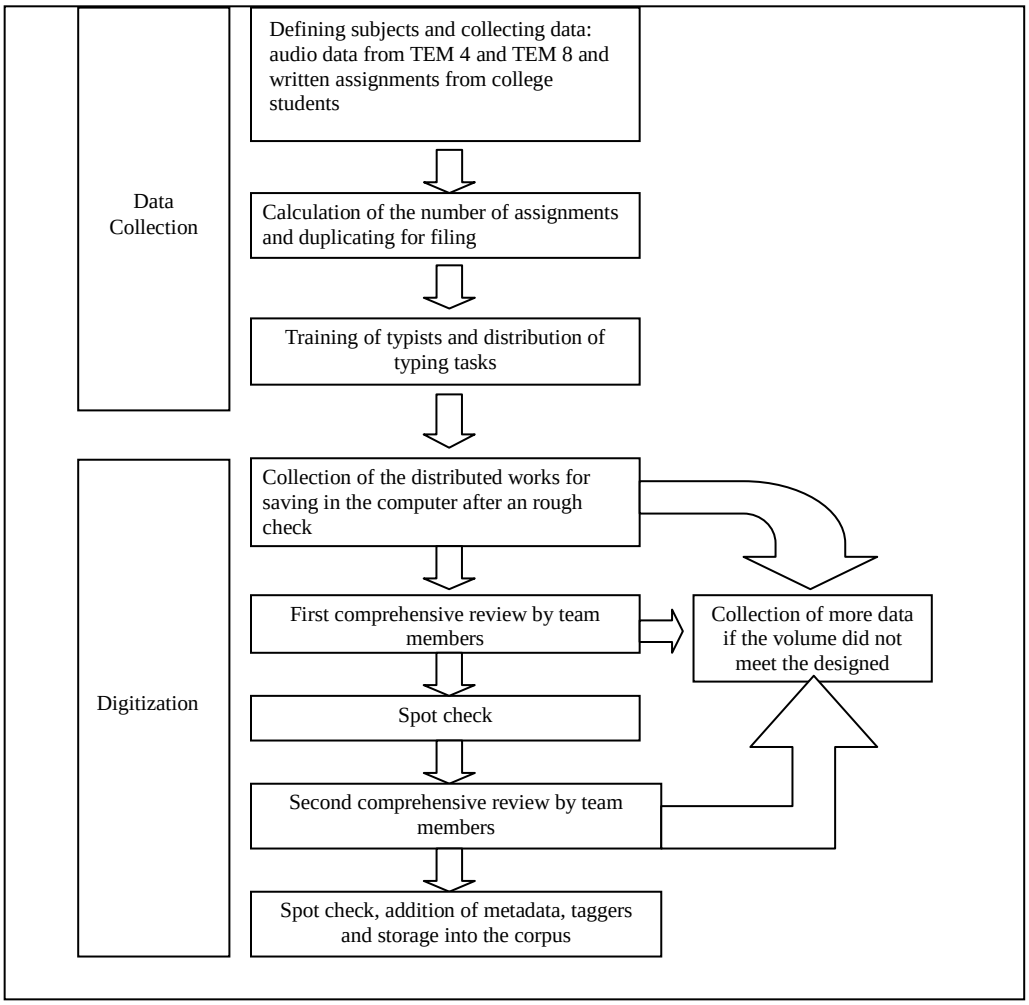
2. Learner Corpora in Use: ICLE and SWECCL.

Learner corpora in English are seen in various forms. SWECCL 1.0 and 2.0 (Wen, *et, al*, 2005 & 2007), two versions of the Spoken (SECCL) and Written (WECCL) English Corpus of Chinese Learners created in China, were launched from 1996 to 2007. The team of the SWECCL project collected recorded audio files for the SECCL from Test for English Majors (TEM) and English learners' writings in China for the WECCL.

The steps involved in data collection to establishment of the SWECCL corpus can be summarized by the authors of this work in Figure 2 below.

First, the project team decided to collect data from the TEM and writing assignments from college students. After collecting all data, the team calculated the volume of data and made duplicates for filing. When the data were all classified, the team started the typing work including pre-training and assigning works to typists. The typists submitted the digitalized data for electronic storage. The team then conducted a comprehensive review of the digitalized data. To ensure that all the data was valid, spot checks were conducted after two different comprehensive reviews and then the metadata and taggers were added for storage into the corpus.

Figure 2. Flow Chart of the Establishment of the SWECCL



Each of the two versions of SWECCL contains around 2,000,000 tokens, respectively and all data were tagged. The SWECCL used CLAWS4 (Garside, 1987; Garside. & Smith, 1997), a grammatical tagging system established by Geoffrey Leech, Roger Garside and Michael Bryant at Lancaster University in the United Kingdom, as its parts-of-speech (POS) tagging system and the corpus was also lemmatized and error-tagged. As for the SECCL (spoken), features of grammatical errors, mispronunciation, disfluency, self repetition and pause fillers were also tagged. Example (1) below provides a tagged sample of spoken errors in SECCL (Wen, Wang, & Liang, 2005: 27-29).

(1) *Grammatical error: has <had>*
Mispronunciation: need <leed>
Disfluency (long pause) …
Self repetition
Pause fillers: um, yeah…

Since it is difficult to define all of the spoken features of learners of English, the SWECCL project categorized the only the five main features that learners would most probably produce. In (1), since the past tense was misused while the correct usage should be *has*, the error is tagged in brackets. When a speaker pronounced *need* as *leed*, the mispronunciation is tagged. The long pause, self repetition and pause fillers are common features in the language produced by second language learners and these features are all tagged.

As for the WECCL (written), a tagged sample is provided in (2).

(2) *Spelling error: environment <sp-enviroment>*
Grammatical error: works <gr-work>
Missing error: and <mis-and>

In (2), the team categorized three main errors, namely spelling error, grammatical error, and omission error. The type of error is marked in brackets along with the original data.

SWECCL 2.0 enriched the sources of its corpus source by adding recorded files from TEM 8, a test for English majors at level 8, with more diversity in the writing topics. In addition, the subjects in WECCL 1.0 were all English major students while WECCL 2.0 also included work from learners of different majors.

The International Corpus of Learner English (ICLE) (Granger,*et, al*., 2009) was another learner corpus constructed for a project led by Sylviane Granger at the Université catholique de Louvain. Starting in the 1990s, the project team collected writings of English learners with different mother tongues. The first version was released in 2002 and the second in 2009. Version 1.0 and 2.0 differ in the size of the corpus and learners of different mother tongues involved. In addition, a built-in concordancer is available in the latest version that enables the users to analyze the data in the CD-ROM.

The collaborators in the ICLE project followed a rigorous process of data collection. The learners were required to be young adults who were university undergraduates with advanced language proficiency, and they had to be learners of English as a foreign language. The data are mainly in the form of argumentative academic writing with around 200,000 words in each sub-corpus. Table 1 below shows the distribution of essay types and their word counts in each sub-corpus.

Table 1: Distribution of Learner Data from ICLEv2 (Granger, Dagneaux, Meunier, & Paquot, 2009: 25-38)

National sub-corpus	Number of essays			Corpus size (words)
	Argumentative	Literary	Others	
Bulgarian	302	0	0	200,194
Chinese	982	0	0	490,617
Czech	197	46	0	201,687
Dutch	252	11	0	234,723
Finnish	357	33	0	274,628
French	295	52	0	226,922
German	422	15	0	229,698
Italian	133	61	198	224,222
Japanese	366	0	0	198,241
Norwegian	312	4	1	211,725
Polish	361	3	1	233,920
Russian	275	1	0	229,584
Spanish	199	52	0	198,131
Swedish	302	53	0	200,033
Turkish	280	0	0	199,532
Tswana	519	0	0	199,173
TOTAL	5554	331	200	3,753,030

From this table we can see that the ICLE corpus is comprised of learner data from learners of sixteen different countries whose written productions of English were collected.

The data collected were lemmatized and part-of-speech tagged with CLAWS4. In addition, in each text file, the essay code, deleted quotes and references, and illegible words have been added as Figure 3 to 5 illustrate.

Figure 3. The Snapshot of the Code for Deleted Quotes <>*

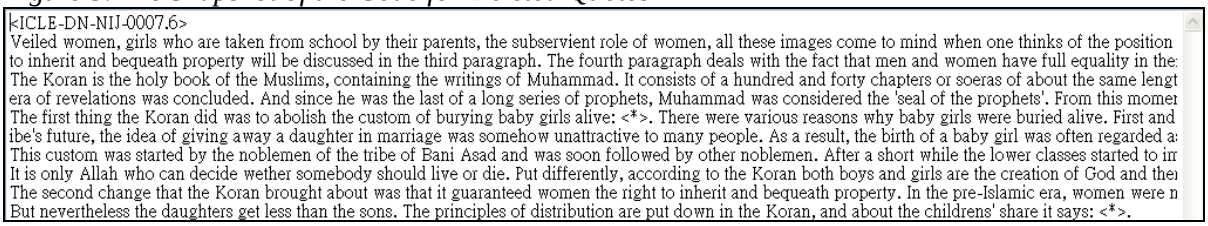


Figure 4. Snapshot of the Code for Deleted Bibliographic References <R>

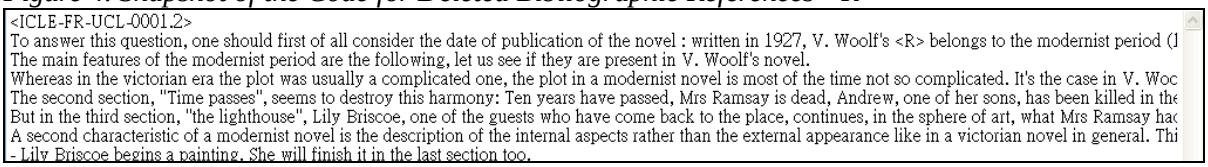
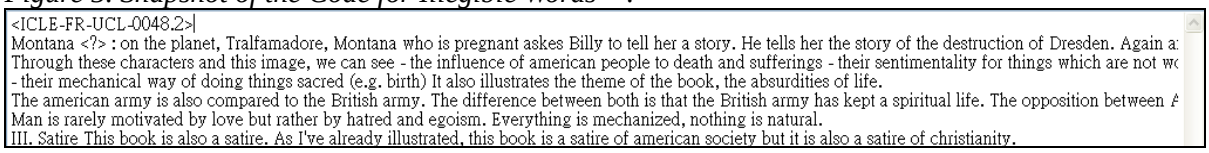


Figure 5. Snapshot of the Code for Illegible Words <?>



The ICLE project also designed a learner profile questionnaire that allowed for analysis of a variety of task and learner variables which were later used as the basis for the search criteria in the concordancer tool, as illustrated in Figure 6 to follow. For the task variables, the medium in this corpus is currently writing, and the field is all general English. Other metadata were also included especially for describing the nature of different texts. These metadata also served as filters while a query is made via the corpus interface. The learner variables contain six variables (age, gender, mother tongue, region, knowledge of other foreign languages and their time spent in staying in an English speaking country) and two fuzzy variables (learning context and proficiency), which were sorted when users input a query into the sub-corpus. However, the way ICLE presents the metadata differs from SWECCL as the information is stored in a database linked to the text files.

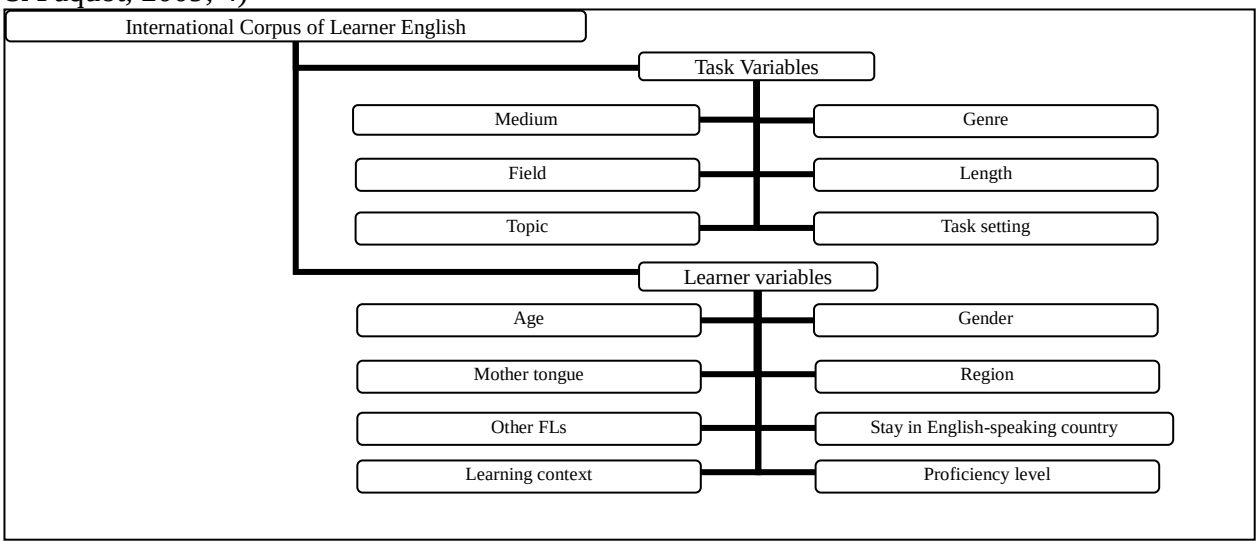
The metadata in the SWECCL are the control variables, such as writing topics, writing time, modes or student types. Examples of the metadata are provided in (3) below.

(3) (a) SECCL 1.0
<SPOKEN><TEM4><GRADE2><YEAR99><GROUP67><TASKTYPE1><SEXF><RANK05>

(b) SECCL 2.0
<SPOKEN><TEM4><GRADE2><YEAR03><GROUP130><TASKTYPE 1 2 3><SEX T1=F, T2=F, T3=F, F><RANK=3>

From (3), one can see that the genre type, type of test, students' grade, college entrance year, group, task type, students' gender and rank in a group were all the filter while sorting search results in the corpus.

Figure 6. Task and Learner Variables Included in ICLEv2 (Granger, Dagneaux, Meunier, & Paquot, 2009, 4)



In addition to the above mentioned corpora, there are also other learner corpora in use nowadays, listed in Tables 2 and 3 to follow.

Most of these corpora are learner corpora of English (Table 2). The English learner corpora are on the increase from time to time, as more and more researchers are becoming interested in creating learner corpora. Since English data can be found relatively easier, the expansion in the number of English learner corpora has been quicker than the other languages. In the following Table 3, there are details of learner corpora of other languages which have also been created recently.

Table 2. Learner Corpora of English in Use

Learner Corpora of English					
Corpora	**Year of Creation**	**Corpus Size (Tokens)**	**Data Types**	**Learners' Language Background**	**Annotation**
ICCI (The International Corpus of Crosslinguistic Interlanguage)	2008	Under construction		Japanese, and learners of foreign languages from Austria, China (Hong Kong), Israel, Poland, Singapore, Spain and Taiwan	Unknown
LTTC (The Language Training and Testing Center)	2008	262,178	Written	Mandarin	POS tagged (to date)
CLEC (Chinese Learner English Corpus)	2002	1,070,602	Written	Mandarin	POS tagged/ lemmatized
SWECCL 2.0 (Spoken and Written English Corpus of Chinese Learners)	2002	2,000,000	Spoken /Written	Mandarin	POS tagged /spoken feature /error tagged/ lemmatized
MELD	2000	100,000	Written	Various	Partially

(The Montclair Electronic Language Learner Database) (Fitzpatrick & Seemiller, 2001)					tagged by annotators without a predetermined error list
ISLE (Italian and German Spoken Learners' English) (Atwell, Howarth, & Souter, 2003)	2000	11,484 utterances	Spoken	German and Italian	Spoken features annotated
TLCE (Taiwanese Learner Corpus of English) (Shih, 2001)	1999	730,000	Written	Mandarin	POS tagged / lemmatized
PELCRA (Polish and English Language Corpora for Research and Applications)	1997	500,000	Written	Polish	Parallel and comparable corpora between English and Polish (POS tagged by CLAWS)
SWECCL 1.0 (Spoken and Written English Corpus of Chinese Learners)	1996	2,000,000	Spoken/ Written	Mandarin	POS tagged /spoken feature tagged/error tagged
HKUST (Hong Kong University of Science and Technology)	1991	25,000,000	Written	Cantonese	POS tagged and partial error tagged
ICLE (International Corpus of Learner English)	1990	3,753,030	Written	Various	With headers/ not tagged
JEFLL (Japanese EFL Learner)	unknown	700,000	Written	Japanese	Unknown

From Table 2, we can see an overview of the development of English learner corpora in use around the world as well as the main features of each corpus. In the following Table 3, four learner corpora for various foreign languages are shown.

Table 3. Learner Corpora of Other Languages in Use

Learner Corpora of Other Languages					
Corpora	**Year of creation**	**Corpus Size (Tokens)**	**Data Types**	**Learners' Language Background**	**Annotations**
NCCU Foreign Language Learner corpus (Various Languages)	2008	Under construction	Written	Mandarin	To be carried out in the second phase of the project
CATE (Spanish, Japanese, German and Chinese) (Lu & Pai, 2007)	2005	Unknown	Written	Mandarin	Lemmatized and POS tagged
USP Multilingual Leaner Corpus (Tagnin, 2006)	2002	200,000	Written	Brazilian	Unknown

The above has discussed two types of existing learner corpora currently in use. In the following section, we provide the steps being used to build the NCCU Foreign Language Learner Corpus at the first stage.

3. Steps in Creating the NCCU Foreign Language Learner Corpus

As Granger (2002) mentioned, a learner corpus will require strict design by considering factors such as learner age, language proficiency, mother tongue background, etc., thus the NCCU Foreign Language Learner Corpus defines different sets of data based on years of learning at university, age, tasks, topics, modes of tasks, etc. Detailed information about its creation is introduced in the following sections.

3.1 Collection and Digitization of Assignments

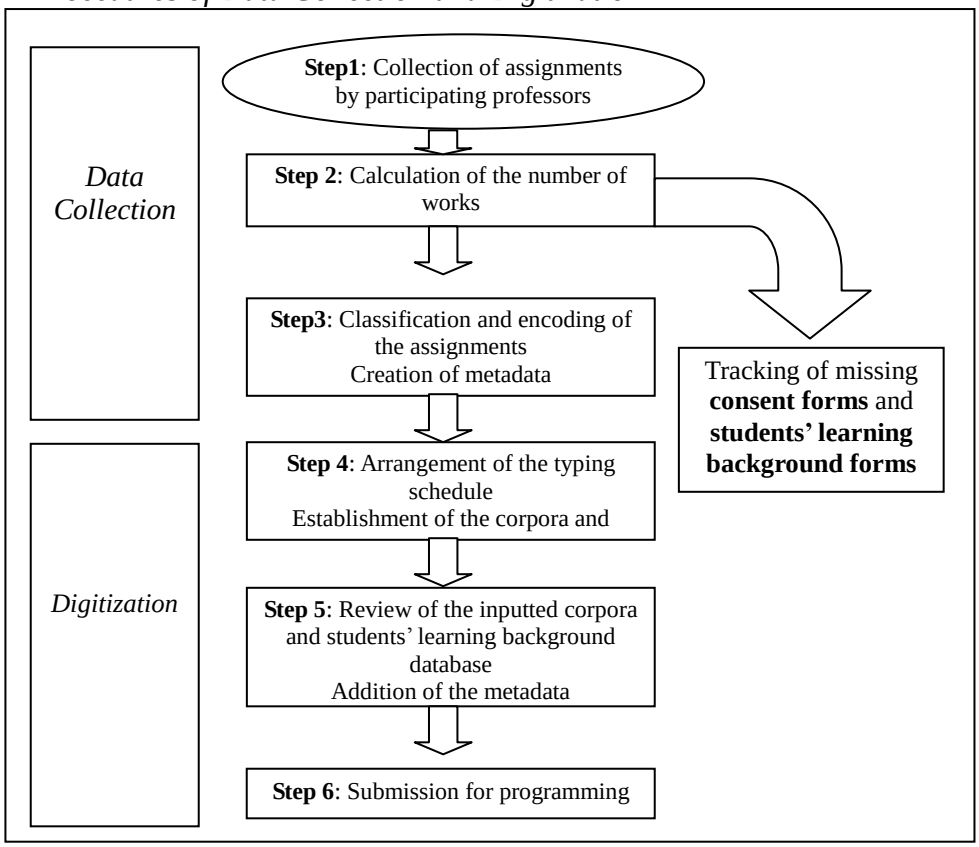
Since the College of Foreign Languages and Literature in NCCU is composed of a wide range of foreign language departments including English, Arabic, Japanese, Korean, Turkish, French, Slavic languages and other European language programs, it is an advantage for us to build up a multilingual learner corpus. Sixteen professors from different language department/program/language center and one professor from computer science are participating in this project. So far, we have collected and uploaded data from English, Arabic, Japanese, Korean, French and Russian (to date). The steps taken for collecting and digitizing data are given in Figure 7 and are described as following.

Step 1: Collection of assignments by participating professors

Subjects of NCCU foreign language learner corpus were learners taught by the sixteen participating professors. These students had agreed to add their assignments as part of our learner corpus project. They did so by signing a consent form handed out by professors who taught them. Students were informed and given an opportunity to accept or reject signing the consent form. Details of the language background of the learners were also collected by requesting them to fill in a standardized language background

form. Students' identity was kept anonymous when their assignments were displayed in the learner corpus.

Figure 7. Procedures of Data Collection and Digitization



The modes in which the students' assignments could come in included homework, in-class exercises, or exam tasks. Participating professors provided hard- and/or softcopies of the assignments, which were centrally collected by the project assistants.

Step 2: Calculation of the number of texts and review of the required attachments

After receiving the assignments from the participating professors, the project team members calculated, double-checked and re-organized the assignments and recorded them in statistical form. Each time a set of assignments was submitted, the professors provided a description form indicating the nature of the assignments (the metadata here). Thus, the team members also needed to assure that all of the required forms were submitted. If any form was found missing, the team members would immediately ask the subjects involved to re-submit the forms. A collected assignment would be invalid without a complete set of consent, language background and assignment descriptors, and thus not included in the learner corpus. Therefore, every step was taken to ensure that a standardized procedure was followed.

Step 3: Classification and encoding of assignments and creation of metadata

The next step involved classifying and encoding assignments. This was done by languages, by sets of assignments and by subjects. Each assignment was encoded as the following in line with the classification procedures. First, a language code was provided, followed by the assignment set number. Then, the student code was added. The final code would be, for example, E001001 in which the 'E' stands for the language of English; the first '001' stands for the first set of the assignments provided by the professor and the following '001' stands for the first subject (the code is underlined in this paper in order to show the two '001's). After coding the assignment, the next step was to code the metadata descriptions enclosed with the assignments. The example in (4) shows one set of metadata. Each piece of assignment was coded as (4) below. Here is an example of the metadata:

(4) *#Subject: E001002*
#Language: English
#Course: College English
#Grade: freshman
#Collected_date: 2009/07/07
#Topic: Writing narration
#Mode: blogs
#Type: descriptive
#Type: narrative
#Purpose: homework
#Group/individual: individual
#Version: first

In (4) above, one can see the metadata of the subject, language, course, grade, collection date, topic, mode, type of assignment, purpose, group/individual and version used for defining a text. Each of these variables may be used as a filter for sorting queries on the website.

Step 4: Arrangement of typing schedule and establishment of the corpus interface

At least two independent typists were recruited for each language– the first typed the handwritten assignments; the second served as a checker. Since double checking was necessary for assuring the typed data were the same as the collected data, these two typists double-checked each other's typed assignments. Before the start of every typing session, which took place in a similar office in the college building, the team members would give clear instructions about the typing procedures, the format in which the work was to be saved and the coding logic. All typists were also requested to sign a consent form to preserve the confidentiality of the corpora data. All typing took place during office hours and none of the assignments were taken home by the typists. All data were typed as .txt files and students' information were recorded in excel format (see Figures 8 and 9).

Figure 8. Saving the Format and Finished Text

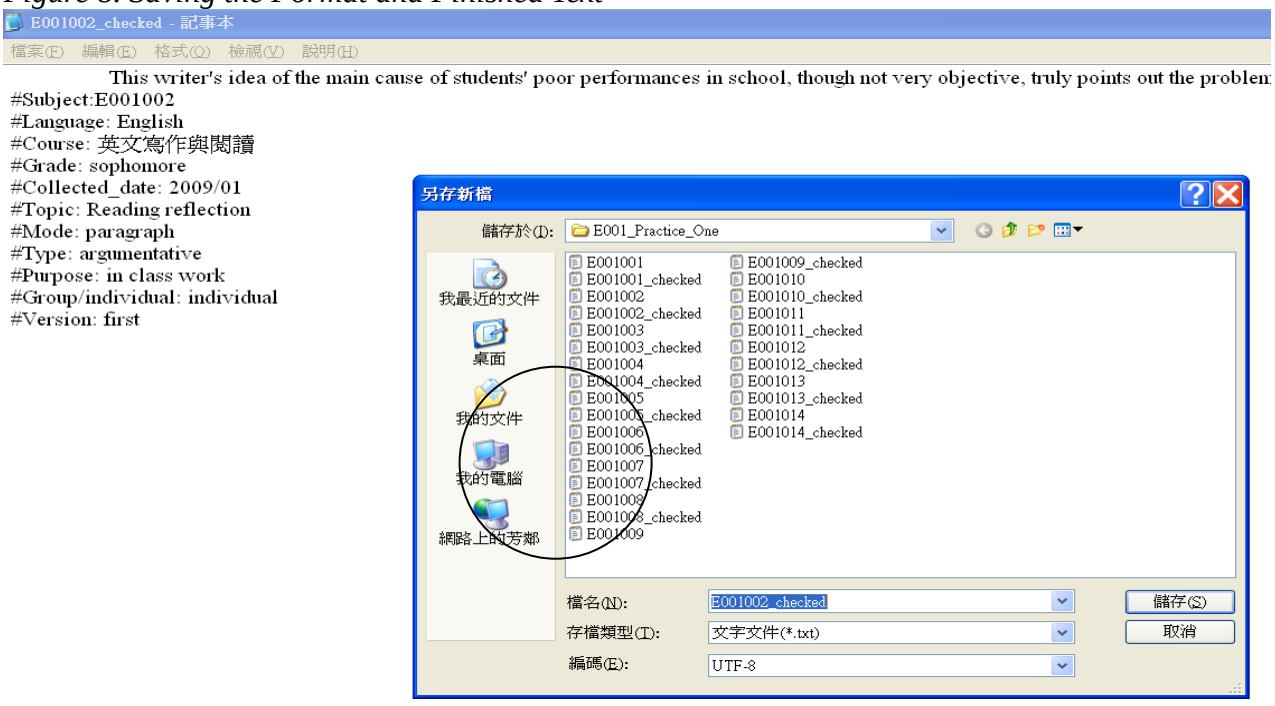
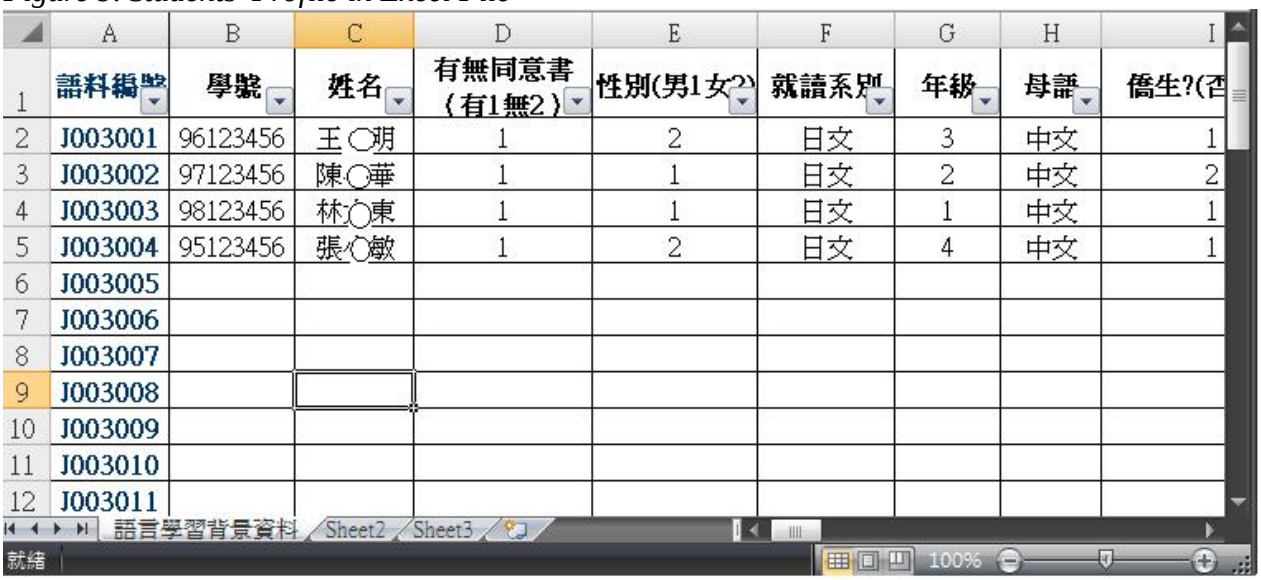


Figure 9. Students' Profile in Excel File



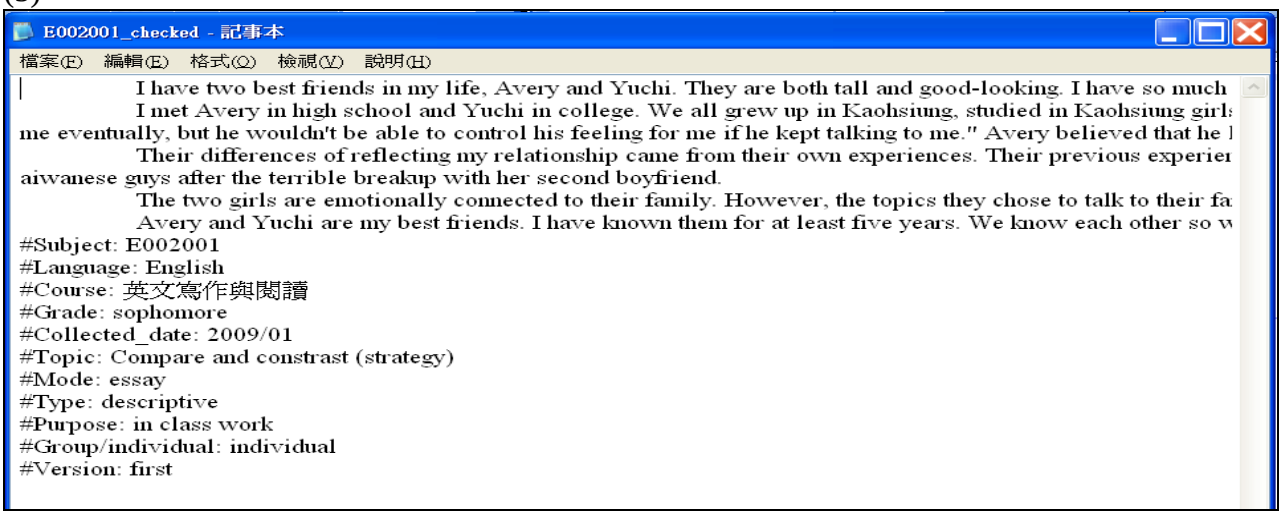
	A	B	C	D	E	F	G	H	I
1	語料編號	學號	姓名	有無同意書（有1無2）	性別(男1女2)	就讀系別	年級	母語	僑生?(否
2	J003001	96123456	王○明	1	2	日文	3	中文	1
3	J003002	97123456	陳○華	1	1	日文	2	中文	2
4	J003003	98123456	林○東	1	1	日文	1	中文	1
5	J003004	95123456	張○敏	1	2	日文	4	中文	1
6	J003005								
7	J003006								
8	J003007								
9	J003008								
10	J003009								
11	J003010								
12	J003011								

From Figure 8, one can see a completed text containing a text body with metadata, which has been checked. All checked files were saved as new files in the UTF-8 format. Figure 9 is a sample of a students' profile. The language learning background of the students was recorded in Excel. Mapping to each subject number (column B), all related information including register number, name, gender, department, grade, mother tongue and other information were saved as the variables in the system.

Step 5: Review of typed assignments and students' learning background database and addition of metadata

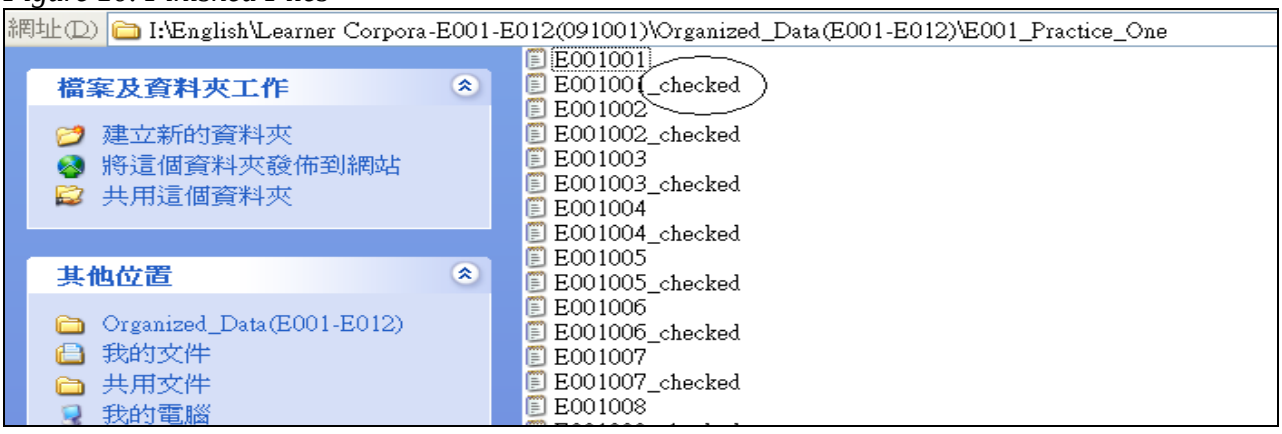
When a typist finished a piece of assignment, a second typist reviewed all the typed data in terms of its content, format and the completeness of the students' background information. After the data had been confirmed by the second typist, the metadata for each piece of assignment was attached to the bottom of the data. The following writing sample in (5) shows the appearance of a completed text.

(5)



After that, the typed texts were saved as independent files with marked "checked," shown in Figure 10 below.

Figure 10. Finished Files



The electronic copies were re-formatted by our typists so that they followed the same coding convention as that of hardcopies.

Step 6: Submission for programming

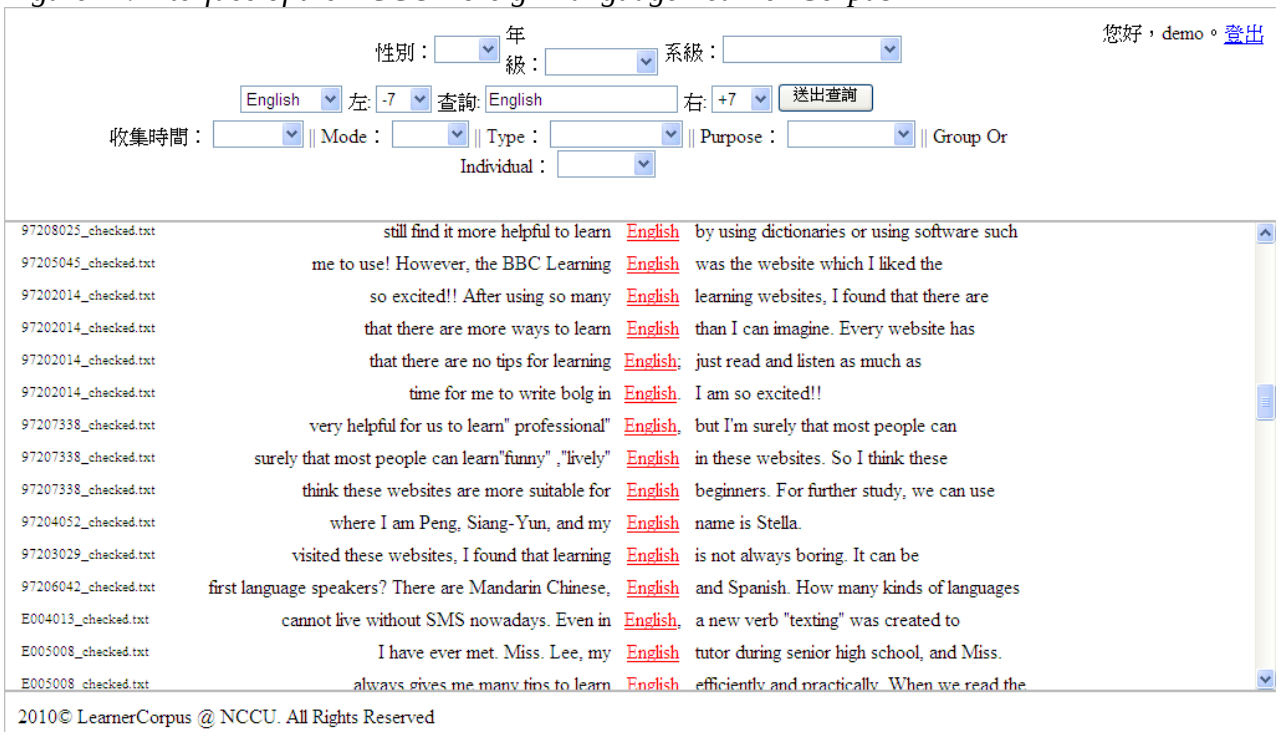
When the typing and attachment of metadata were all accomplished, the assignment sets would then be submitted to the project programmer to be uploaded onto the learner corpus database. The next section will focus on the creation of the interface.

3.2 Creation of the Interface of the NCCU Foreign Language Learner Corpus

The NCCU Learner Corpus is now available at http://learnercorpus.nccu.edu.tw. The interface has already been created and the content is currently available for research purposes.

Currently-collected data from students' written assignments have been archived, and basic search functions, such as vocabulary search, are available in use now. Furthermore, detailed data descriptions, such as collection dates, types of assignments and nature of assignments, etc., can be found. Most importantly, the metadata of each item of data searched, such as gender and age, are clearly presented in the interface. The NCCU Learner Corpus website is presented as follows (Figure 11).

Figure 11. Interface of the NCCU Foreign Language Learner Corpus



Since our learner corpus is multi-lingual, the data encodings of the different languages are very different; most are not ASCII-based characters. Although Ross (1984) suggested data handling methods for each non-Roman character set, it is still hard to process corpora texts in different languages and display multi-lingual results in the same interface. In order to simplify data encoding from the different languages, we used UTF-8 for the internal encoding to store our corpora data because UTF-8 can be used to handle

most of the world's writing systems. Thus, a program was written to convert all non-UTF-8 encoded data into the UTF-8 coding system. We also extracted all metadata to enable search functions according to the entries in the metadata, as shown in the top of Figure 11.

Figure 12. Scope of the NCCU Foreign Language Learner Corpus Web System

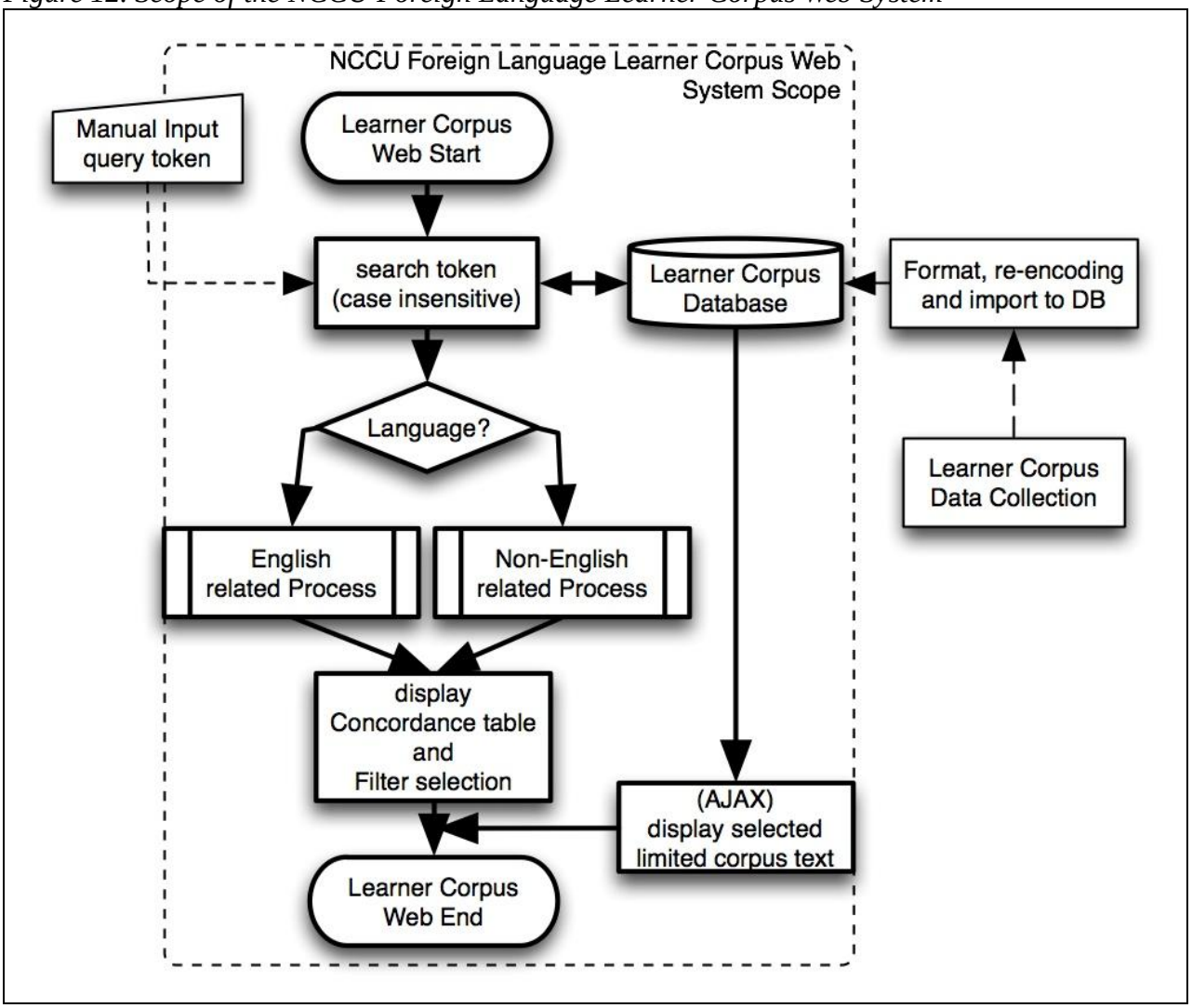


Figure 12 is the flow chart presenting the process when users enter a query at the interface. When a user searchers for a token, the first step is to select the language. Once the language is defined, the system will decide which process should be operated, and the interface will display the concordance table and the filter selection. Users can use the second level of filters to narrow down the range of texts.

For the interface, two levels of filter functions are shown. The first level can be designated before a query takes place, i.e., at the top row in Figure 11 above (enlarged in Figure 13 below). The categories one can first delimit are gender, year, and affiliation. The second row in Figure 13 shows the types of languages (there are five at this stage)

and the designated window size.

Figure 13. Querying Interface with Basic Filters and Designated Window Size

As for the second level of filtering, metadata such as collection date, mode, type, purpose and nature of assignments are processed, as shown in the third row in Figure 14 below. These five filtering categories can only perform the logic "AND" function if two or more categories are selected.

Figure 14. Second Level of Filters

As for the logic of the programs, in order to process different languages, two separate systems were created – one is based on roman-character languages (English and French) and the other on non-roman-character languages (Japanese, Korean and Arabic).[3]As can be seen in Figure 14 above, there is an option for displayed left/right window size. The nature of the two separate systems used for running the different languages will have an impact on the display of window size, since each language has to be defined differently according to window size. At the current stage, a word in the roman-character languages is judged using a space as a delimiter. Non-roman-character languages, however, are only counted by the amount of memory occupied in the program.[4]

In Figure 14 above, when a search query is made, an exact match of the window size for presenting will be produced. All matching data are displayed in the formed of aligned concordance lines. We also applied the AJAX (Asynchronous JavaScript And XML) function in which users can obtain more data of the left or right of the context by clicking on 'more previous context' (更多前文) or 'more after context' (更多後文).

In the current simple version of the interface, punctuation problems occur in all languages and they are inconsistent, mainly due to the variations in the coding for of the different languages. In the case of the roman-character languages, since their delimiter is a space, the query result will show punctuation adhering to query words. For example, as we query *computer*, ***CALL (****computer* also appears as part of the result, as shown in

[3] Learner data from Russian had not been uploaded yet at the time when this paper was being written. Huge difficulties were encountered when processing these different languages. This paper reports work to the current stage so far.

[4] Thus far, this is not an ideal way to display the information but once segmentation takes place, the size of the displayed window will be greatly improved.

Figure 15 below. This is because the acronym ***CALL***, the left bracket '(' after it, and the query word *computer* were written as a single word by the learners. As all data from learners were kept verbatim, the missing space may cause this problem, although it can also be taken as a special feature of the learners' language.

Figure 15. Example of Punctuation Problems

As for the non-roman-character languages, we used '\b' flag as a regular expression to identify an exact-match word. The window size of languages such as Japanese and Arabic is based on the amount of memory occupation in the program on the left and right of the query word. Figure 16 below is an example of querying '留学' in Japanese where the left window size is -8 and the right window size is +3.

Figure 16. Selecting Different Window Size for Different Sides (Japanese)

In addition, we have also added a wild-card function (*) for all of the languages. For example, when querying '얼*나' in Korean, the system will show the results in Figure 17 below.

Figure17. Wild-Card Function (in Korean)

In addition to the displays of English, Japanese and Korean shown above, the interface also enables queries for French and Arabic at the present phase, As demonstrated in Figures 18 and 19 below:

Figure 18. Queries for French

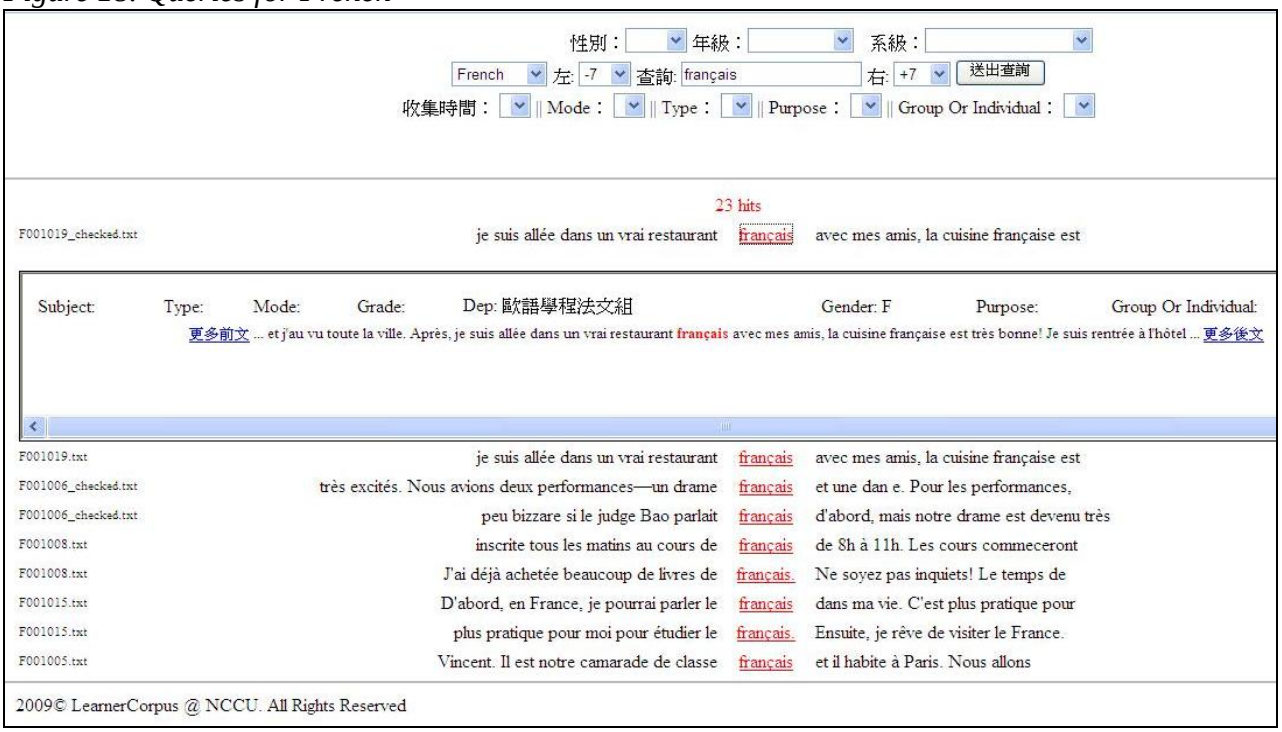


Figure 19. Queries for Arabic

4. Future Improvements of the NCCU Foreign Language Learner Corpus

In order to create a learner corpus with annotations, we suggest to implement the following improvements – POS and error Tagging.[5]

4.1 Lemmatization and Parts-of-Speech (POS) Tagging

Lemmatization refers to the conversion of all derived forms to their root form, such as *see* for *seeing*, *saw* and *seen*. The lemmatized forms will be tagged along with POS. According to Hunston (2002), tagging functions as a label to allocate POS to each word in a corpus. For instance, in the SWECCL, the tagging methodology is facilitated by CLAWS4. CLAWS4 was initially used for tagging the British National Corpus with 100 million words and the task was completed in 1994. It has since been utilized for tagging other English corpora. CLAWS4 contains ten grammatical categories with 137 taggers which can be used to label words. In the case of tagging in the SWECCL, the accuracy can reach 94.5%. A sample of tagged text is shown in (6) below for the SWECCL.

(6) Education <NN1> is <VBZ> a <AAT2> lifelong <JJ> processs <NN1> . <.>
NNI: singular common noun
VBZ: be verb "is"
AAT2: singular article
JJ: general preposition
.: period

Even though CLAWS4 was not originally designed for use with a second language learner corpus, the POS tagging can help researchers to conduct studies on language learning and teaching by observation of the grammatical features used by learners. For our learner corpus, we will follow the tagsets in CLAWS4 for English. As for the other languages in our learner corpus, POS tagging will be carried out based on the features of each language. At this stage, we are still looking into tools suitable tools for use in tagging these languages. The Hidden Markov Model-based POS tagger for Arabic is claimed to achieve a state-of-the art performance of 97% (Shamsi & Guessoum, 2006). For French data, the TreeTagger might be possible for its part-of-speech tagging[6]. No (2007) has proposed a KWGInterpreter for Korean POS tagging. For Russian, RussianPOSTagger is available for tagging Russian texts[7]. For Japanese data, the Japanese tagger ChaSen might be suitable for tagging POS[8]. These tools still require evaluation for use at a later stage when the actual tagging process is conducted. We have still to evaluate the use of these tools in our design of the NCCU tagging process.

4.2 Error Tagging

As for error tagging, the errors produced by learners can also be highlighted in the system. The following example in (7) shows an example with its error tagged from the

[5] The tagging of the different languages requires specific tools for the respective languages. This is one of the obstacles we need to overcome at this stage.
[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[7] http://rupostagger.sourceforge.net/
[8] http://chasen-legacy.sourceforge.jp/

WECCL (written).

(7) Examples of written errors are provided as following:
Spelling error: environment <sp-environment>
Grammatical error: works <gr-work>
Missing error: and <mis-and>

Manual tagging is required for all languages included in this project, since a fully automatic tagging system for errors is hard to accomplish. The most difficult part of manual tagging is the lack of consensus on a similar error rated differently by different raters. Granger (2003) pointed out that elaborated guidelines for tagging should be utilized with detailed principles for handling error categories. In general, two taxonomies for error coding have been commonly agreed upon in previous work, including linguistic category classification and a target modification taxonomy (Tono, 2003). The former refers to linguistic features such as lexis and tense, and the latter refers to features that differ from the form used by native speakers, such as omission and the change of order (Díaz-Negrillo & Fernández-domínguez, 2006). In the following section, we show several applications of the NCCU Learner Corpus.

5. Applications of the NCCU Foreign Language Learner Corpus

Learner corpora have been adopted in the analysis of various aspects of linguistic analysis, including the lexical analysis of words, collocations and colligations, as well as the analysis of syntactic structures. They provide an authentic resource for analyzing and observing learners' language, which might have implications for second language (L2) acquisition.

Contrastive interlanguage analysis (CIA) compares the language used by native speakers and that produced by language learners (Granger, 1996). As Granger noted, in the ICLE project, the function of the corpus collected is devoted to the CIA analysis according to two types of comparison are usually made- the comparison of native language (the reference in which corpus) and interlanguage (or non-native varieties), as well as the comparison between (or among) interlanguages. Due to the differences in the type of learner corpora (the ICLE is type A while the NCCU learner corpus is type B, cf. Figure 1 previously), comparisons will possibly be conducted cross-linguistically to investigate how learners might perform in different languages.

The data in learner corpora are often contrasted with that of the native speaker corpora by centering on a linguistic feature to examine whether that feature is used more frequently (or overuse) or less frequently (underuse) than native speaker corpora. For example, Liu and Shaw (2001) evaluated EFL learners' knowledge of the verb *make*, which appears at a high frequency and has various meanings, by comparing the results of a learner corpus and a native speaker corpus. They questioned learners' qualitative knowledge of vocabulary instead of gauging the quantity of words learners know. The result showed that learners' knowledge of a word is different from that of the native speakers'. In another study, Chen (2006) analyzed her self-collected corpus comprised of

papers written by Taiwanese MA TESOL students by using ten journal articles from two TESOL journals as the reference corpus. She explored the learners' use of conjunctive adverbials and found that the connectors were overused and sometimes misused at the word-level. Palacios-Martínez and Martínez-Insua (2006) examined Spanish learners' use of the existential *there* by analyzing two learner corpora in comparison to two native speaker corpora. They found that the uses of *there* differ in frequency, structural complexity, polarity and pragmatic value. In Gilquin, Granger, and Paquot's (2007) evaluation of learners' EAP writing, they compared learner corpus data with the native speaker data and found a number of problems that learners might encounter in academic writing. The learner corpora, moreover, can be utilized for materials design and corpus-informed tools for learning.

In addition to research of written learner corpora, study of the results produced by spoken corpora also help researchers identify problems and features of learner language. Shirato and Stapleton (2007), for example, examined a spoken learner corpus of Japanese learners of English, proving that the learner corpus is a useful tool for revealing how learner language differs from the native speakers'.

With raw data, using the learner corpus, researchers can also investigate and contrast the raw frequencies of words or collocations. Though less sophisticated, Granger (1996: 45) still confirmed this as a "very fruitful undertaking", in regard to Granger, Meunier, and Tyson's (1994) research on learner lexicon, which reveal learners' overuse of *but* and under use of *and*. The application of concordancing further provides evidence of how learners use a word in context and how it differs from the usage of native speakers.

When the corpus is parsed and tagged, research focusing on word categories and syntactic structures can be conducted. By tagging the errors in L2 corpora, studies of learners' errors under the framework of computer-aided error analysis can be conducted. The ICLE has developed an error tagging system which utilizes purpose-built menu-driven error editors. The Standard Speaking Test (SST) speech corpus has also adopted a machine learning technique to detect learners' errors automatically (Izumi, Uchimoto & Isahara, 2000). Research can thus be conducted to investigate interlanguage errors of specific linguistic features; for example, the connector usage in essays written by EFL learners of English (cf. Granger & Tyson, 1996).

Using the NCCU Learner Corpus, Chung and Tseng (2009) carried out a preliminary analysis of the preposition *to*, focusing on the collocations and senses of *to* used by language learners. The result showed that learners' misuse of this preposition only occurs in lower frequency words, indicating that learners learn the *to*-collocates in chunks. The errors were further analyzed, and possible reasons for the errors committed were proposed. Moreover, comparisons to other languages are also possible as the NCCU Learner Corpus features a multilingual learner corpus. Through analyzing learners' language use, researchers and teachers can both benefit from uncovering the features of learner language and from revealing difficulties learners encounter, which would provide both a comprehensive understanding of learners' knowledge about the language and at the same time provide some implications for language teaching.

6. Conclusion

More and more attention has been given to learner corpora in corpus building these years. This paper introduces a newly created learner corpus called the NCCU Foreign Language Learner Corpus. The construction of this corpus was facilitated by the numerous varieties of language courses available at NCCU.

At this stage, the NCCU Learner Corpus has been uploaded to an online interface, and some basic search functions have already been included. Additional training courses will be given on a continual basis to all members. The major difficulty faced by this project is the lack of professional programmers in languages other than English. Thus far, our project has accomplished the first stage of data collection, although the data are presented as raw data at the moment. These data are ready to be used for analyses despite the absence of annotations, which are expected to be added in the second phase of this project. In addition to keeping a comprehensive record of students' learning processes and teachers' pedagogical materials, the ultimate objective of this project is to encourage language educators to make further innovations in the pedagogical approaches, to investigate the possible reasons for learners' language errors, and to carry out research into linguistic and educational issues as well as to provide a better understanding for language learning. It is certain that both language instructors and learners will benefit immensely by this project.

Through using corpora, teachers can also investigate how students use certain vocabulary items in writing and discover how these items have been used incorrectly. This may thus prompt teachers to make advances in research.

Furthermore, by working on this project together, teachers can observe how features of different languages may influence language learning among students who learn more than two languages at the same time. This is one of the characteristics of contrastive interlanguage analysis, in which cross-referencing is carried out for different languages. At the college level, this project not only serves to unite language education and linguists but also to encourage the exchange of teaching philosophies by teachers of different languages. Based on the abovementioned advantages, this paper has outlined the need and necessities in creating a foreign language learner corpus based on Taiwan contexts.

Acknowledgements

We would like to thank the funding from the NCCU Top-Universities Program, and Professor Nai-ming Yu, the Dean of the College of Foreign Languages and Literature and Professor Hsueh-ying Judy Yu,Vice-Dean of the College of Foreign Languages and Literature for their support of this project. We would also like to thank all of the participating professors–Professors Wen-lang Soo, Yoshida Taeko and Su-chin Wang from the Department of Japanese, Professors Yun-sen Sung, Gui-ying Peng, Hsiang-lin Yeh and Pei-chi Chang from the Department of Slavic, Professors Feng-lan Luo from the Department of Arabic, Professors Chieh-tsung Chang and Byeong-seon Park from the Department of Korean, Professors Yao-chueh Juan, Katarzyna Stachura and Thomas Haquette from the Program in European Languages (French), Professor Simon Smith from the Foreign Language Center, and Professor Chao-lin Lui from the Department of Computer Science for contributing in different ways to this project. This paper documents the background to the developments of the learner corpus, details how the NCCU Foreign Language Learner Corpus is built, which is under the supervision of the first author. Further research based on individual languages as future research outcomes of this project will be highly encouraged. Finally, we would also like to thank the research assistants Tzu-yu Liu, F.-Y. August Chao, Liang-Chun Jonathan Wang, Yi-Chen Joy Hsieh and Chun-Hung Chen for their continuous support.

References

Atwell, Eric, Peter Howarth, and Clive Souter. (2003). The ISLE Corpus: Italian and German spoken learners' English. *ICAME Journal*, 27, 5-18.

Chen, Cheryl Wei-Yu. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, 11(1), 113-130.

Chung, Siaw-Fong and Yu-Wen Tseng. (2009). Learning Prepositions: A Corpus-based Study in Taiwan EFL Contexts. Poster presented at the *Third International Conference Grammar and Corpora*. Mannheim, Germany. September 22-24.

Cheung, Hintat, Siaw-Fong Chung and Sophia Skoufaki. (2010). Indexing Second Language Vocabulary in the Intermediate GEPT. In the *Proceedings of the Twelfth Academic Forum on English Language Testing in Asia (Language Testing in Asia: Continuity, Innovation and Synergy).* The Language Training and Testing Center, Taiwan. March, 5-6. pp.118-136.

Corder, Stephen Pit. (1981). *Error analysis and interlanguage*. London; New York: Oxford University Press.

Díaz-Negrillo, Ana and Jesùs Fernández-domínguez. (2006). Error tagging systems for learner corpora. *RESLA*, 19, 83-102.

Fitzpatrick, Eileen and Steve Seegmiller. (2001). The Montclair electronic language learner database. In George Antoniou & Dorothy Deremer (Eds.), *Proceedings of the International Conference on Computing and Information Technologies* (pp.

369-375). World Scientific.
Garside, Roger. (1987). The CLAWS Word-tagging System. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.
Garside, Roger and Nicholas Smith. (1997). A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech, & Anthony McEnery (Eds.), *Corpus Annotation: Linguistic information from computer text corpora* (pp. 102-121). New York: Addison Wesley Longman.
Gilquin, Gaëtanelle. (2001). The Integrated Contrastive Model. Spicing up your data. *Languages in Contrast*, 3(1), 95–123.
Gilquin, Gaëtanelle, Sylviane Granger, and Magali Paquot. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6, 319-335.
Granger, Sylviane. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg & Mats Johansson (Eds.) *Languages in Contrast. Text-based cross-linguistic studies. Lund Studies in English 88* (pp. 37-51). Lund: Lund University Press.
Granger, Sylviane. (2002). A bird's-eye view of computer learner corpus research. In Sylviane. Granger, Joseph Hung & Stephanie Petch-Tyson (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-33). Amsterdam & Philadelphia: Benjamins.
Granger, Sylviane. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3), 465-480.
Granger, Sylviane, FannyMeunier and Stephanie Tyson. (1994). New insights into the learner lexicon: a preliminary report from The International Corpus of Learner English. In L. Flowerdew & A.K.K. Tong, (Eds.), *Entering Text* (pp. 102-113). The Hong Kong University of Science and Technology.
Granger, Sylviane and Stephanie Tyson. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15, 19-29.
Granger, Sylviane, Esrelle Dagneaux, Fanny Meunier and Magali Paquot. (2009). *International corpus of learner English.* Louvain: UCL.
Hunston, Susan. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.
Izumi, Emi, Kiyotaka Uchimoto and Hitoshi Isahara. (2000). SST speech corpus of Japanese learners of English and automatic detection of learners' errors. *ICAME Journal*, 24, 31-48.
Liu, Eric. T. K. and Philip. M Shaw. (2001). Investigating learner vocabulary: A possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *IRAL, 39*, 171-194.
Lu, Hui-Chuan and Pai, Fang-I. (2007). *Construcción de corpus de aprendices taiwaneses de español(CATE).* XXV International Aesla Conference. Murcia, Spain.
Lu, Hui-Chuan and Lu, Hsueh-Lo. (2009). The Construction of Spanish-English Learners' Written Parallel Corpus (CPATEI). *Foreign Language Studies*, 9(1), 135-156.

No, Yongkyoon. (2007). KWGInterpreter: a lemmatizing POS tagger for the Korean language. *Proceedings of the 2007 Joint Conference of LAK, MLSK, and KSLI*, pp. 86-94, Linguistic Association of Korea.

Palacios-Martínez, Ignacio and Ana Martínez-Insua. (2006). Connecting linguistic description and language teaching: Native and learner use of existential there. *International Journal of Applied Linguistics*, 16(2), 213-231.

Ross, Hugh McGregor. (1984). *Handling non-Roman character sets with computers*. In Proceedings of *Translation and Communication: Translating and the Computer*, London, pp.135-146.

Selinker, Larry. (1972). Interlanguage. *International Review of Applied Linguistics*, 10, 201-231.

Shamsi, Fatma Al and Ahmed Guessoum. (2006). A Hidden Markov Model-based POS tagger for Arabic. *8es Journées internationales d'Analyse statistique des Données Textuelles*.

Shih, Rebecca Hsue-Hueh. (2000). Compiling Taiwanese learner corpus of English. *Computational Linguistics and Chinese Language Processing*, 5(2), 87-100.

Shirato, Junko and Paul Stapleton. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, 11(4), 393-412.

Tagnin, Stella. E.O. (2006). A multilingual learner corpus in Brazil. In Andrew Wilson, Dawn Archer, & Paul Rayson, (Eds.), *Corpus Linguistics Around the World* (pp. 195-201). Amsterdam: Rodopi.

Tono, Yukio. (2001). The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. *Proceedings of ASIALEX2001*, 257-262, Korea.

Tono, Yukio. (2003). Learner corpora: Design, development and applications. *Proceeding of the Corpus Linguists 2003 Conferenc*. 800-809, Lancaster. United Kingdom.

Wen, Qiu-Fang, Li-Fei Wang and Mao-Cheng Liang. (2005). *Spoken and Written English Corpus of Chinese Learners (1.0).* Beijing: Foreign Language Teaching and Research Press

Wen, Qiu-Fang, Li-Fei Wang and Mao-Cheng Liang. (2007). *Spoken and Written English Corpus of Chinese Learners (2.0).* Beijing: Foreign Language Teaching and Research Press

Yergeau, Francios. (1998). UTF-8, a transformation format of ISO 10646. Retrieved October 30, 2009, from The Internet Engineering Task Force (IETF) website: http://www.ietf.org/rfc/rfc2279.txt