Taiwan Journal of TESOL
Vol. 11.1, 31-74, 2014

CONSTRUCTING THE TAIWANESE COMPONET OF THE LOUVAIN INTERNATIONAL DATABASE OF SPOKEN ENGLISH INTERLANGUAGE (LINDSEI)

Lan-fen Huang

ABSTRACT

This paper reports the compilation of a corpus of Taiwanese students' spoken English, which is one of the sub-corpora of the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin, De Cock, & Granger, 2010). LINDSEI is one of the largest corpora of learner speech. The compilation process follows the design criteria of LINDSEI so as to ensure comparability across the sub-corpora. The participants, procedures for data collection and process of transcription are all recorded. Fifty third- or fourth-year English majors in Taiwan were given recorded interviews in English. Each interview was accompanied by a profile containing information about such learner variables as age, gender, mother tongue, country, English learning context, knowledge of other foreign languages, and amount of time spent in English-speaking countries and such interviewer variables as gender, mother tongue, knowledge of foreign languages and degree of familiarity with the interviewees. Data on another variable, the learners' English proficiency level based on the results of international standardised tests, was collected; this is not available in other sub-corpora of LINDSEI. The participants' proficiency was similarly distributed across B1 to C1 levels in the Common European Framework of Reference. The structure of the Taiwanese sub-corpus is discussed in comparison with eleven other published sub-corpora. The preliminary investigation, using corpus-linguistic approaches, reveals overall statistical information about the Taiwanese component and Version 1 of LINDSEI. The lexical analyses of the top 50 words and chunks show the characteristics of spoken English in the Taiwanese sub-corpus. The contributions and research potential of this newly-developed learner corpus are discussed, followed by an example of Contrastive Interlanguage Analysis of the most common chunk, *I think*, in the Taiwanese learners' speech. The release of this learner corpus is merely the first step. It is hoped that more corpus research will be done on Taiwanese learners, that corpora of other speech genres will be compiled and that research results will contribute to relevant areas in Applied Linguistics.

Key Words: LINDSEI, interlanguage, learner corpus, Taiwanese learners of English, *I think*

INTRODUCTION

Research on corpora has mostly focused on written English and contributed a great deal of corpus-based grammatical description and explanation. In contrast, relatively few studies have emerged of corpora of spoken languages, which call for a time-consuming and laborious transcription process. A similar trend is found in the investigation of learner corpora, which have been used to study the written language of learners from different mother tongue communities. However, relatively little research has been done on the interlanguage of spoken English. One of the few major accomplishments in the corpus studies of learners' spoken English is the compilation of the Louvain International Database of Spoken English Interlanguage (LINDSEI) Version 1 (Gilquin et al., 2010), which includes the spoken English produced by learners from eleven different first languages (L1s). The present paper first sets up the aims and briefly reviews the learner corpus research in Taiwan. Next, it introduces LINDSEI and reports the compilation process of the Taiwanese component. The structure of this sub-corpus is first compared to LINDSEI Version 1, and then given statistical and lexical analysis. Finally, its contributions and potential for future research are discussed.

Aims of the Research

This paper aims to report (a) the compilation of a sub-corpus of LINDSEI; and (b) the corpus-linguistic approaches to investigating this Taiwanese learner corpus. The participants were 50 third- and fourth-year university students majoring in English in Taiwan. The methods of data collection and transcription followed the requirements of LINDSEI in order to ensure comparability between the sub-corpora. Upon the completion of the corpus, corpus analytical methods were employed to conduct preliminary research, such as investigating basic corpus information, word frequencies and lexical chunks.

Learner Corpus in Taiwan

Corpus-based learner language has been studied for more than twenty years (see various papers in the edited volumes of Granger, 1998b; Granger, Gilquin, & Meunier, 2013). It has been widely acknowledged as a useful resource for such academic fields as Second Language Acquisition and English Language Teaching. A number of learner corpora have been made available (see the list prepared by the Centre for English Corpus Linguistics, 2013) and most of them are corpora of written English.

In Taiwan, to my knowledge, there are few learner corpora of written English: The Soochow Colber Student Corpus (Bernath, 1998), the Taiwanese Learner Corpus of English (Shih, 2000), the NCCU Foreign Language Learner Corpus (Chung, Wang, & Tseng, 2010) and the Taiwanese Learner Academic Writing Corpus (Chen, 2011). So far, only one Taiwanese learner corpus of spoken English has been compiled, consisting of speech by 15 students (Huang, 1991). The Taiwanese learner corpus of spoken English developed in this paper is likely to be the first and most complete learner corpus of English speech from Taiwan. In addition, it is a sub-corpus of the global collaborative project, LINDSEI, which makes it of great value. Its contributions and research potential are discussed in the later sections of this paper.

OVERVIEW OF LINDSEI

The LINDSEI project began in 1995 and in 2010 published its first version, which includes sub-corpora formed by eleven L1s: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish.[1] It involved 544 informal interviews and roughly one million tokens in total, with an average of 1,949 tokens in each one. About one third of the spoken data comes from the interviewers and two thirds from the learners (Gilquin et al., 2010).

In order to have comparable data across sub-corpora and to avoid the heterogeneity of interlanguage, the sub-corpora of LINDSEI must meet an established set of criteria. Each corpus consists of 50 to 53 informal interviews between a learner and an interviewer. All learners

[1] This Taiwanese sub-corpus was completed in late 2013. Another eight sub-corpora of different mother tongue backgrounds–Arabic (Saudi Arabia), Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian, and Turkish–are in progress. For more details, please see *LINDSEI Partners* (Gilquin, 2014) at http://www.uclouvain.be/en-307845.html (assessed on 25 January 2014).

are third- or fourth-year English-major students in countries where English is used as a foreign language and more than half the interviewers (64%) are native speakers (NSs) of English (Gilquin et al., 2010).

Each interview takes about 15 minutes to cover three tasks: set topics,[2] free discussion and picture description. The first task serves as a warm-up activity. One of three topics is chosen by the interviewee. This lasts five to six minutes, including some follow-up questions put by the interviewer. The second task, taking seven to eight minutes, consists of free discussion of general topics, such as life at university, hobbies, travel experience, what the student hopes to do after university, family, etc. The objective is not to stress and embarrass the interviewees with difficult questions but to get them to talk spontaneously. In the last few minutes, the interviewer asks the interviewee to look at a sequence of four pictures and tell the story that they illustrate. The student should not be given either the time or opportunity to make notes before describing the picture. It should be an improvised description.

All the interviews are orthographically transcribed and marked up according to the transcription guidelines (Gilquin, 2012) (see Appendix A). Each transcription is accompanied by a profile which contains information about such learner variables as age, gender, mother tongue, country, English learning context, knowledge of other foreign languages, and amount of time spent in English-speaking countries and such interviewer variables as gender, mother tongue, knowledge of foreign languages and degree of familiarity with the interviewees.

The eleven sub-corpora of LINDSEI offer a wide range of possibilities of research into Contrastive Interlanguage Analysis (CIA[3]). Comparisons can be made between different interlanguages as well as between any interlanguage and native speech in the Louvain Corpus of Native English Conversation (LOCNEC), which is compiled by De Cock

[2] The three set topics were: (a) *An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learnt from it.* (b) *A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.* (c) *A film/play you've seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad* (Gilquin et al., 2010, p. 8).

[3] The term, Contrastive Interlanguage Analysis (CIA) was coined by Granger (1996, 1998a).

(2004), using the same structure as LINDSEI. In addition, the written counterpart of LINDSEI, the International Corpus of Learner English (ICLE) (Granger, Dagneaux, Meunier, & Paquot, 2009) is a corpus of argumentative essays written by learners from sixteen L1 backgrounds. LINDSEI and ICLE share ten mother tongue backgrounds, which makes it possible to compare spoken and written interlanguages.

COMPILING THE TAIWANESE SUB-CORPUS OF SPOKEN ENGLISH

In this section, the compilation process of the Taiwanese sub-corpus of LINDSEI is reported in some detail, including the methods of recruiting participants, the conduct of informal interviews, and the transcription of audio files.

Recruitment of Participants

The participants were 50 third- or fourth-year undergraduate students majoring in English in the six universities in Taiwan,[4] which are listed in Table 1. These universities were included mainly because the contacts in the universities were willing to help in the recruitment of participants and the students in both the comprehensive and technical universities could be involved, which would allow representative data.

The participants were recruited through an advertisement on campus or at the invitation of their instructors. They were informed that the collected spoken data would be used for research purposes and had to give their permission by signing a learner profile questionnaire (Appendix B) on the day of the interview. The questionnaire used for the Taiwanese corpus was slightly adapted from that in LINDSEI by adding one question: *Have you ever taken an English proficiency test? If yes,*

[4] The LINDSEI team requires all contributors to a sub-corpus to submit 50 recordings and their accompanying profiles. In case of problems such as unintelligible sound quality or an incomplete learner profile for any of the contributors, 60 recordings were made in this case. In late 2013, 50 out of the 60 learners were sent to the LINDSEI team for further processing. Therefore, the data in the Taiwanese sub-corpus of LINDSEI reported in this paper may differ slightly from the final version included in the second version of LINDSEI.

please give the name of the test, your result and date of the test. Most of the learners gave their TOEIC scores, but some had IELTS, TOEFL, BULATS, GEPT and CSEPT grades.[5] Table 2 lists the distribution of the 50 learners' English proficiency in the four levels of the Common European Framework of Reference (CEFR). The learners' proficiency is mostly distributed across the B1 to C1 levels; therefore, it is best described as ranging from intermediate to advanced. The Taiwanese sub-corpus is similar to other sub-corpora in LINDSEI. Although information about the learners' proficiency in LINDSEI was not available, a tentative study, based on a random sample of five learners from each sub-corpus, indicates that 64% were rated as high-intermediate (or below) and 36% as advanced (Gilquin et al., 2010, pp. 10-11).

Table 1

Universities Participating in the Taiwanese Sub-corpus of LINDSEI

	University	**Number of participants (Percentage)**
1	Shih Chien University	6 (12%)
2	Wenzao Ursuline University of Languages	8 (16%)
3	National Cheng Kung University	13 (26%)
4	National Pingtung University of Education	10 (20%)
5	National Taiwan University of Science and Technology	7 (14%)
6	National Kaohsiung University of Applied Sciences	6 (12%)
	Total	50 (100%)

Four interviewers, one American, one British and two Taiwanese teachers of English, were involved in the data collection (see Table 3). Ideally, the interviewers should have been NSs of English, since it may

[5] The Test of English for International Communication (TOEIC), International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), and Business Language Testing Service (BULATS) are internationally recognised certificates. The General English Proficiency Test (GEPT) and College Student English Proficiency Test (CSEPT) are local tests developed in Taiwan.

be easier to develop natural communication when learners talk with someone who does not share the same L1. However, to fit in with the availability of the interviewers who were NSs, the learners and the compiler, 70% of the interviews were conducted by NSs and the remainder by the Taiwanese teachers of English. They were briefed beforehand on the way to conduct the interview and fully aware of the use of the transcripts and audio files for research purposes.

Table 2

The Distribution of the English Proficiency of the 50 Learners in the Four Levels of CEFR

Level	Number of participants (Percentage)
B1	13 (26%)
B2	17 (34%)
C1	19 (38%)
C2	1 (2%)
Total	50 (100%)

Table 3

The Interviewers' Gender and Mother Tongue

Interviewer	Gender	Mother tongue	Number of interviews (Percentage)
1	Male	British English	19 (38%)
2	Male	American English	16 (32%)
3	Male	Chinese	7 (14%)
4	Female	Chinese	8 (16%)
			50 (100%)

Procedures for Informal Interviews

On the day of the interview, the learners of English were asked to fill in a profile questionnaire (Appendix B), with the assistance of the compiler. This form included information about learner variables and was signed and dated to signify written consent to use the recorded interviews for research purposes. In order to make the best use of time without keeping the interviewers waiting, some learners filled in their questionnaires after the interviews. Either way, the learners were well aware of being recorded.

After filling in the questionnaires, the learners were given at least five minutes to prepare to talk on one of the three set topics. Then, the learners were invited to enter a classroom or meeting room where two small electronic recorders had been set up. The compiler left the room as soon as she had made sure that the recorders were working, because the students might have felt under pressure if two people had been listening to them.

As reported in the previous section, the whole informal interview took about 15 minutes. During this period, the interviewer tried his/her best to be friendly and to help students talk more by giving quick responses and specific questions, and the learners were given neither the time nor the opportunity to write notes. This interview aimed to collect spontaneous speech from the learners.

After the interviews, the learners were given a voucher for NT$200 (US$1 equals NT$30) to spend. The recordings and learner profiles were coded for the transcription process.

Process of Transcription

The 50 interviews were orthographically transcribed and marked up, following the guidelines provided by the LINDSEI project (Gilquin, 2012) (Appendix A), by two research assistants. The mark-up items include interview identification, speaker turns, overlapping speech, empty pauses, filled pauses and backchannelling, unclear passages, anonymisation, truncated words, foreign words and pronunciation, phonetic features, prosodic information, nonverbal vocal sounds, contextual comments, and task identification.

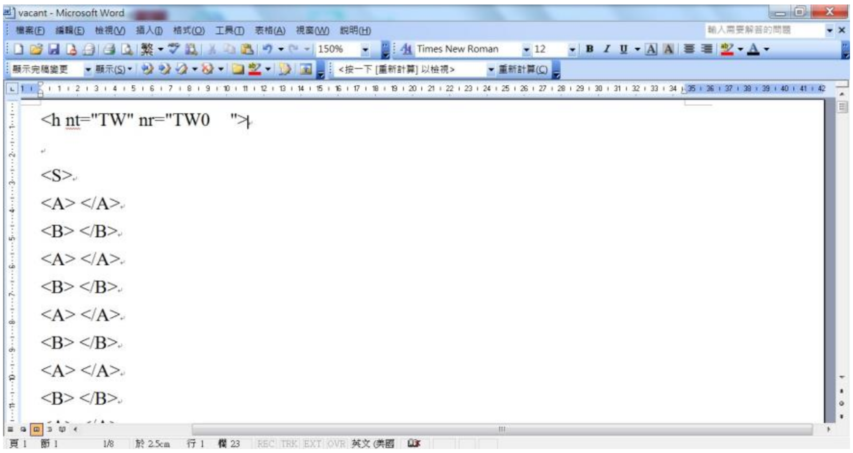
The transcription work for a 15-minute interview might take five to ten hours, depending on the transcribers' experience of transcribing. The two transcribers spent more time to begin with, when they were not yet very familiar with the transcription guidelines. All the transcripts were double-checked by the compiler. Each of them took about 30 to 60 minutes to finish.

The task of orthographic transcribing was less difficult. Few revisions were needed after the checking. Nevertheless, the mark-up process required more training. According to the two transcribers, among the twenty aspects of transcription in the guidelines, the marking-up of overlapping speech, empty pauses, and filled pauses and backchannelling was most difficult and time-consuming. In the process of double-checking, the compiler identified more discrepancies in these three items than elsewhere. This was probably because the transcribers had to play the recordings several times in order to locate appropriate places in both turns to annotate the tag <*overlap* />.[6] Without any facilitation from a timer, the duration of empty pauses was personally judged and classified in a three-tier system: one dot for a pause of less than one second, two dots for a pause of between one and three seconds and three dots for a pause of more than three seconds. The mark-up of filled pauses and backchannelling caused difficulty because, despite the varied use of them by the speakers, the transcribers had only six ways of marking and had to choose the most suitable: *(eh)* [brief], *(er)*, *(em)*, *(erm)*, *(mm)*, *(uhu)* and *(mhm)*. It was the compiler who ensured the consistency of transcription. In the cases that were not included in the guidelines (e.g. the vocal sound for hesitation or self-correction, which was transcribed as <*clicks tongue*>) the compiler consulted the LINDSEI project coordinator in Belgium.

In the process of transcription, two pieces of computer software were used, Microsoft Word and Windows Media Player. Figure 1 shows a template for transcribing in MS Word. (The transcriptions were converted to plain text after proofreading.) and Figure 2 is a screenshot of the transcribers' use of the template and Windows Media Player.

Another software programme, Audacity (2013 members of the Audacity development team, 2013) (Figure 3), was used to edit the sound recordings, in particular for deleting redundant time at the beginnings and ends of interviews. It also made it possible to manipulate the sound file, e.g. reducing its speed, playing it back several times, double-checking the length of empty pauses, etc.

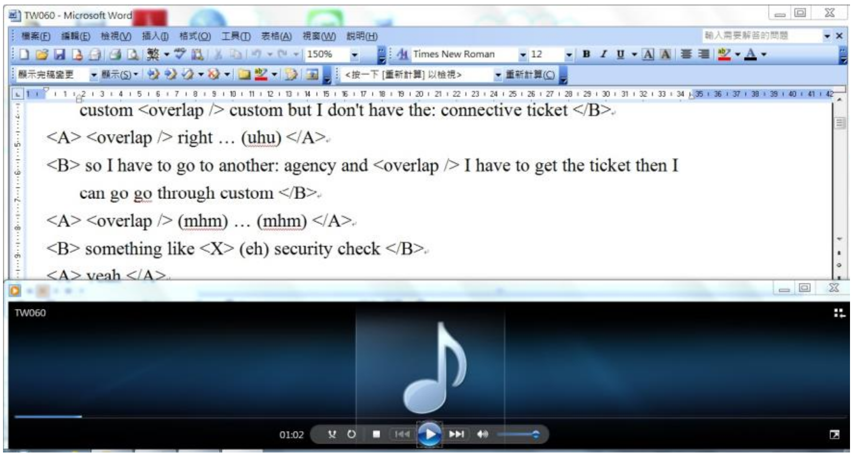
[6] The transcription guidelines for the LINDSEI project were made general in nature to accommodate all sub-corpora; therefore, the sub-corpora may not be used in the way that they were transcribed, being intended to serve in research enquiries of every kind. Three mark-up items, pointed out by the compilers of the LINDSEI German component, Brand and Kämmerer (2006), might be further processed by future researchers. Overlapping speech was not marked up at the exact syllable where it occurred but in front of the word. Similarly, syllable lengthening was indicated at the end of the word. Pauses were roughly indicated in the three-tier system: one dot for a short pause (< 1 second), two dots for a medium-length pause (1-3 seconds) and three dots for long pauses (> 3 seconds).



```
<h nt="TW" nr="TW0    ">

<S>
<A> </A>
<B> </B>
<A> </A>
<B> </B>
<A> </A>
<B> </B>
<A> </A>
<B> </B>
```

Figure 1. A screenshot of a template for transcribing speech in MS Word



```
    custom <overlap /> custom but I don't have the: connective ticket </B>
<A> <overlap /> right … (uhu) </A>
<B> so I have to go to another: agency and <overlap /> I have to get the ticket then I
    can go go through custom </B>
<A> <overlap /> (mhm) … (mhm) </A>
<B> something like <X> (eh) security check </B>
<A> yeah </A>
```

Figure 2. A screenshot of using a template with Windows Media Player

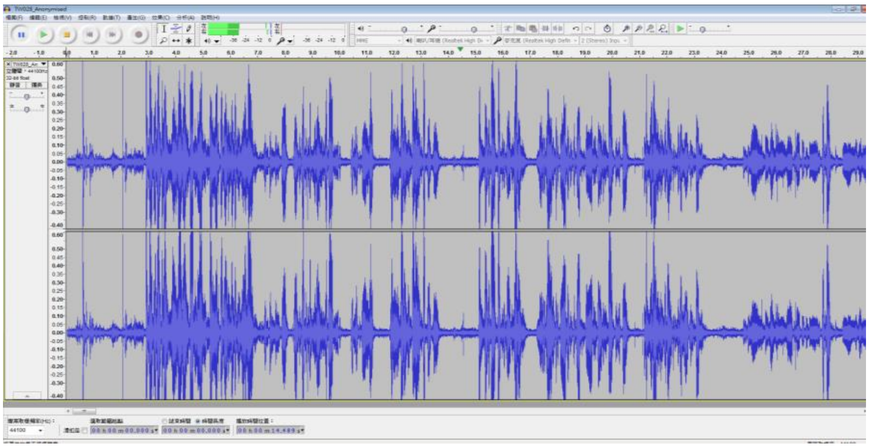


Figure 3. A screenshot of Audacity

STRUCTURE OF THE TAIWANESE SUB-CORPUS

As mentioned earlier, all the sub-corpora of LINDSEI must meet the same design criteria in order to elicit data comparable with those in other sub-corpora. This section presents the structure of the Taiwanese sub-corpus according to the variables in the profiles and discusses it with that of LINDSEI Version 1. Table 4 shows the structures of the Taiwanese sub-corpus and LINDSEI Version 1.[7]

The data in the Taiwanese sub-corpus are much more recent than those in LINDSEI Version 1. They were collected from November 2012 to June 2013, while those in LINDSEI Version 1 are from November 1995 to May 2005 (Gilquin et al., 2010). The Taiwanese sub-corpus comprises 50 interviews, while in LINDSEI Version 1 the average number of interviews is 50.4. The size of the Taiwanese sub-corpus is 110,280 tokens, which is close to the average size, 98,153 tokens, of the sub-corpora in LINDSEI Version 1. It is worth noting that the Taiwanese sub-corpus is larger than the other two national sub-corpora in Asia.

[7] The statistical information on the Taiwanese sub-corpus is generated by WordSmith Tools Version 6 (Scott, 2012). The average information of the eleven sub-corpora of LINDSEI Version 1 is provided in Gilquin et al. (2010).

Table 4

The Structures of the Taiwanese Sub-corpus and LINDSEI v.1

Corpus		**Taiwanese sub-corpus**	**11 sub-corpora of LINDSEI v.1 (on average)**
Recording dates		From 19 Nov 2012 to 3 Jun 2013	From 14 Nov 1995 to 9 May 2005
Composition of corpus	No. of interviews	50	50.4*
	No. of tokens (Turns A & B)	110,280	98,153
	No. of tokens (Turns B only)	69,577	72,013
	No. of tokens per task (Turns A & B)	Set topics: 36,905 (33%) Free discussion: 60,307 (55%) Picture description: 13,068 (12%)	Set topics: 40,244 (41%) Free discussion: 42,257 (43%) Picture description: 15,652 (16%)
	No. of tokens per task (Turns B only)	Set topics: 25,969 (37%) Free discussion: 35,450 (51%) Picture description: 8,158 (12%)	Set topics: 31,854 (44%) Free discussion: 28,626 (40%) Picture description: 11,533 (16%)
	Total duration	12 hours 54 minutes	11 hours 52 minutes
Interview	Average length (Turns A & B)	2,206	1,949
	Average length (Turns B only)	1,392	1,430
	Average duration	15 minutes 6 seconds	14 minutes 9 seconds
	Set topic	Country: 44% Experience: 34% Film/play: 22%	Country: 49% Experience: 23% Film/play: 28%

Table 4

The Structures of the Taiwanese Sub-corpus and LINDSEI v.1

(continued)

Learner	Average age	21.7	22.4
	Gender (percentage of female)	86%	79%
	Average no. of years of English at school	9.38	7.33
	Average no. of years of English at university	3.22	2.99
	Average no. of months in English-speaking countries	2.81	3.73
	English proficiency (in CEFR levels)	B1: 13 (26%) B2: 17 (34%) C1: 19 (38%) C2: 1 (2%)	N/A
Interviewer	Gender (percentage of female)	16%	71%
	Mother tongue (percentage of English NS)	70%	64%

* All sub-corpora in LINDSEI version 1 have 50 interviews respectively, except that the Chinese sub-corpus comprises 53 interviews and the Japanese sub-corpus 51 (Gilquin et al., 2010, p. 23).

There are 82.536 tokens in the Chinese sub-corpus and only 56,239 tokens in the Japanese sub-corpus. When the utterances by learners (Turns B in the corpora) are considered, the Taiwanese sub-corpus, makes up a total of 69,577 tokens, is slightly smaller than the average total of 72,012 in LINDSEI Version 1, but it is still larger than the total of tokens of the learners in the Chinese and Japanese sub-corpora, which amount to 63,542 and 37,126, respectively.

As noted in Overview of LINDSEI, each interview is made up of three tasks: a set topic, free discussion and picture description. In Table 4, it appears that, in the Taiwanese sub-corpus, free discussion represents more than half the corpus (55%), set topics account for one third and picture description produces the remaining 12%, while the set topic (41%) and free discussion (44%) are similarly represented in LINDSEI Version 1. However, in the latter, the breakdown of the figures varies significantly across sub-corpora. The distribution of the three tasks in the French, German, Greek and Japanese sub-corpora is similar to that in the Taiwanese sub-corpus.

In terms of the duration of interviews and length of utterances, the Taiwanese sub-corpus is similar to the other sub-corpora. The required time is 15 minutes per interview, but there are variations of time in all sub-corpora. The shortest interview lasts 11 minutes and 34 seconds and the longest lasts 19 minutes and 42 seconds. In terms of the choice of set topics, the first topic is most popular, accounting for almost half the interviews in all the sub-corpora.

The eligible participants were third- and fourth-year English majors and the average ages are 21.7 in the Taiwanese sub-corpus and 22.4 in LINDSEI Version 1. Across all sub-corpora, most of the learners are female. Before the learners in the Taiwanese sub-corpus enter university, they have studied English for 9.38 years, which is a longer period than the average, 7.33, in LINDSEI Version 1. In the past decade, the growing trend in Taiwan is for school children to begin learning English as early as possible. Therefore, it is not surprising that among the sub-corpora of LINDSEI, this figure is second to the Swedish sub-corpus (9.59 years). In addition to learners' time spent on English education in their home countries, the cumulative time they have spent in English-speaking countries is reported. Across the eleven sub-corpora, the average time varies remarkably between zero months in the Chinese and Greek sub-corpora and 13.78 months in the Swedish sub-corpus. While the Taiwanese sub-corpus has a relatively low average of 2.81 months, it includes a learner who spent seven years in Canada and 34 out of 50 (68%) who had never visited an English-speaking country. In addition to the above variables related to learners, their English proficiency levels are collected in the Taiwanese sub-corpus, which may be distinguished from other sub-corpora of LINDSEI.

The distribution of the interviewers' gender in this corpus contrasts with that in LINDSEI Version 1. While 16% of the interviewers in the Taiwanese sub-corpus were female, 71% in LINDSEI Version 1 were female. The percentages of the interviewers' mother tongue in the Taiwanese sub-corpus and LINDSEI Version 1 seem similar. However, the breakdown of the average figure shows that in the Greek, Japanese and Polish sub-corpora, there are no interviewers whose first language is

English (Gilquin et al., 2010, p. 37).

The learning context in Taiwan when the data were collected is summarised in Table 5. The aspects dealt with are the medium of instruction, the teaching focus, the availability of English-language media, and stays in English-speaking countries. The information provided describes the general situation as it was when the participants were in school. It does not necessarily reflect the situation since then.

QUANTITATIVE ANALYSIS OF THE TAIWANESE SUB-CORPUS

The Taiwanese sub-corpus of LINDSEI was investigated with the general corpus approaches by *WordSmith Tools 6* (Scott, 2012). The quantitative corpus investigation provides basic but overall information, which is used in this paper to examine whether the quantitative features of the Taiwanese sub-corpus are similar to those of LINDSEI Version 1.

Statistical Analysis

In Corpus Linguistics, *token* is used to refer to a single linguistic unit (in most cases, a word), and *type* means a distinct word. For instance, if the grammatical article *the* occurs 200 times in a corpus, it represents 200 tokens, but counts as only one type. It can be seen in Table 6 that the average number of tokens in the Taiwanese sub-corpus is 2,206, which has 464 types. As the design criteria are the same across each sub-corpus of LINDSEI, the average number of tokens in LINDSEI Version 1 is similar, with 1,949 tokens and 431 types. The average type/token ratio is also similar. It is 21.38 in the Taiwanese sub-corpus and 23.01 in LINDSEI Version 1. The type/token ratio (TTR) indicates the degree of lexical diversity. In a larger corpus, function words tend to be repeated; therefore, the larger the corpus is, the lower the TTR will be (Baker, Hardie, & McEnery, 2006, p. 162). This explains why the overall TTR in LINDSEI Version 1 is lower (1.51) than that in the Taiwanese sub-corpus (3.83), since the former is much larger than the latter.

Table 5

The Learning Context in Taiwan When the Taiwanese Sub-corpus was Collected

Medium of instruction	General classes were taught in Chinese (Mandarin) in schools of all levels. When the Taiwanese sub-corpus of LINDSEI was collected, the teaching of English to children in compulsory primary education began at different ages, depending on the local authorities (Chiu, 2007). In general, primary schools gave 1-2 hours a week of English classes, while junior and senior high schools gave 3-4 hours a week. In addition to learning English in compulsory education, most students in Taiwan started learning English at after-school learning centres (i.e. private cram schools). Teachers used a mixture of Chinese and English in classes. At university level, English classes for English majors were taught in both English and Chinese. Some courses by English native speakers were conducted in English only.
Teaching focus	The teaching focus tended to be more on form and accuracy than fluency and focused on reading and writing skills, as the curricula used to be exam-oriented.
Media	Newspapers and radio shows were mainly in Chinese. English-speaking TV programmes (mostly films and news) were subtitled in Chinese. This all meant that in everyday life students had little exposure to English through the national media. However, English newspapers, books and magazines are easily available in shops. English was also readily accessible through the Internet.
Stays in English-speaking countries	Staying or travelling in English-speaking countries was becoming very popular. Some universities in Taiwan offered exchange programmes for students to stay in an English-speaking country for a few months. Some study-abroad trips were also widely available in Taiwan. However, only a minority of the students actually stayed long in English-speaking countries. Almost one-third of the participants in the Taiwanese sub-corpus had never stayed in an English-speaking country when the LINDSEI interviews were conducted.
Other remarks about the status of English	English was and still is the most popular foreign language in Taiwan.

Table 6

The Statistical Information of Tokens and Types

Corpus		Tokens (running words)	Types (distinct words)	Type/token ratio (TTR)	Standardised TTR	STTR basis
LINDSEI Taiwanese sub-corpus	Average	2,206	464	21.38	27.92	1000
	Overall	110,280	4,225	3.83	27.97	1000
LINDSEI Version 1	Average	1,949	431	23.01	28.54	1000
	Overall	1,079,681	16,296	1.51	28.58	1000

To avoid skewing the ratios when comparing corpora of different sizes or texts of different lengths, the *WordList* in *WordSmith Tools 6* (Scott, 2013) is able to produce a standardised type/token ratio (STTR). In this case, the STTR is calculated on the basis of 1,000 words, which means that the first 1,000 words are calculated first and then the next 1,000 words, and so on. The STTRs are very close to each other, between 27.92 and 28.58. One of the uses of TTR is to measure lexical density. There are differing ways of calculating lexical density (Baker et al., 2006), but generally the lexical density is higher in a corpus (text) in written form (e.g. news writing in Biber, Conrad, & Leech, 2002) than in speech (e.g. conversation in Biber et al., 2002).

The statistical information in Table 6 presents the average and overall information on tokens, types, TTR and STTR. A more detailed table of the Taiwanese sub-corpus is shown in Appendix C. Future studies might compare the statistical information among individual learners or among learners of different proficiency levels.

Lexical Analysis

The statistical analysis in the previous section reports the overall and average information about the Taiwanese sub-corpus and LINDSEI Version 1. In the Taiwanese sub-corpus, however, the lexical analysis is based on the learners' language (Turns B), because learner corpus research centres on the linguistic features of interlanguage. The utterances by the interviewers (Turns A) and learners (Turns B) in the corpus were separated using *Windows PowerShell*.[8] All turns by learners were used to make a word frequency list, which was lemmatised with an English-language lemma list by Someya (1998).[9]

Table 7 lists the top 50 words in the learners' utterances. One of the characteristics of spoken English is the frequent use of first- and second-person pronouns (O'Keeffe, McCarthy, & Carter, 2007). This is also reflected in the Taiwanese sub-corpus. The pronoun forms, *I*, *you*, *my* and *we* are in the list of the top 50 words. As in O'Keeffe et al.'s (2007) study of the five-million-word Cambridge and Nottingham Corpus of Discourse in English (CANCODE), the top 50 words in the Taiwanese learners' spoken data included items of high frequency in conversations, such as *yeah*, *eh*, *er*, *mm*, *em*, and *oh*.

It seems that the learner language in the Taiwanese sub-corpus lacks the use of discourse markers, which is one of the main distinctive features in spoken English.[10] In Table 7, the words in italics, *and*, *so*, *like*, *but*, and *know* (combined with *you*), might be used as discourse markers. The common discourse markers, such as *well* (the 92nd item in the word list), and *I mean* (the 155th item in the word list), do not occur very frequently. These phenomena need to be further examined before more interpretations are offered.

Another common lexical analysis in Corpus Linguistics is of chunks, which are recurrent strings of words used together repeatedly. They are also called 'lexical bundles' (Biber, Finegan, Johansson, Conrad, & Leech, 1999) and 'clusters' (Scott, 2013). The *WordList* tool in *WordSmith 6* offers the function of automatically counting collocational patterns. Like the word list above, the learner language in the Taiwanese sub-corpus of LINDSEI was analysed. The cluster size in

[8] My thanks go to Mr Sheng Li, a PhD student at the University of Birmingham, UK, for his technical support with *Windows PowerShell* in September 2013.

[9] A lemma is the base form of a word. For example, the verb lemma WALK may cover all its inflections and/or spellings: *walk, walks, walked,* and *walking* (Baker et al., 2006, p. 104).

[10] Common features of spoken English include the following five categories: 1) deictic expressions, 2) situational ellipsis, 3) headers, tails and tags, 4) discourse markers and 5) polite and indirect language, vague language and approximation (Carter & McCarthy, 2006). The use of these five categories is common in spoken English but rare in written English.

the *WordList* tool was set between two- and five-words with a minimum of five occurrences.

Table 7

The Top 50 Words in the Learner Language in the Taiwanese Sub-corpus of LINDSEI

N	Word	Freq.	A percent of the running words (%)	Lemmas
1	**i**	3609	5.19	
2	the	3136	4.51	
3	***and***	2275	3.27	
4	to	2038	2.93	
5	be	1907	2.74	be[196] am[20] are[353] been[44] eing[15] is[955] m[27] was[263] were[34]
6	yeah	1561	2.24	
7	a	1261	1.81	a[1176] an[85]
8	eh	1258	1.81	
9	er	1035	1.49	
10	mm	985	1.42	
11	in	956	1.37	
12	that	941	1.35	that[900] those[41]
13	***so***	878	1.26	
14	have	823	1.18	have[673] d[28] had[46] has[62] having[9] ve[5]
15	think	816	1.17	think[748] thinking[13] thinks[17] thought[38]
16	they	774	1.11	
17	**you**	756	1.09	
18	***like***	748	1.08	like[740] liked[1] likes[7]
19	it's	746	1.07	
20	***but***	722	1.04	
21	of	655	0.94	
22	yes	642	0.92	
23	**my**	618	0.89	
24	because	605	0.87	
25	for	518	0.74	
26	not	496	0.71	

Table 7

The Top 50 Words in the Learner Language in the Taiwanese Sub-corpus of LINDSEI (continued)

27	it	495	0.71	
28	**we**	489	0.70	
29	em	449	0.65	
30	just	420	0.60	
31	she	419	0.60	
32	or	416	0.60	
33	go	409	0.59	go[266] goes[8] going[38] gone[2] went[95]
34	really	406	0.58	
35	he	398	0.57	
36	maybe	392	0.56	
37	don't	359	0.52	
38	will	344	0.49	
39	***know***	343	0.49	know[336] knew[3] known[1] knows[3]
40	there	326	0.47	
41	this	325	0.47	this[305] these[20]
42	her	324	0.47	
43	can	319	0.46	
44	me	318	0.46	
45	very	315	0.45	
46	with	302	0.43	
47	do	296	0.43	do[218] did[35] does[12] doing[17] done[14]
48	some	285	0.41	
49	oh	271	0.39	
50	one	271	0.39	one[269] ones[2]

Table 8

The Top 50 Chunks in the Learner Language in the Taiwanese Sub-corpus of LINDSEI

N	Word	Freq.	%	Texts	%
1	***i think***	600	0.86	50	100
2	and i	305	0.44	47	94
3	i i	264	0.38	44	88
4	in the	263	0.38	48	96
5	i have	248	0.36	44	88
6	i don't	241	0.35	48	96
7	yeah yeah	226	0.32	33	66
8	so i	209	0.30	45	90
9	a lot	164	0.24	37	74
10	want to	164	0.24	44	88
11	have to	157	0.23	33	66
12	and the	156	0.22	48	96
13	the the	154	0.22	38	76
14	because i	136	0.20	42	84
15	but i	136	0.20	45	90
16	***kind of***	135	0.19	30	60
17	and then	132	0.19	29	58
18	mm i	130	0.19	40	80
19	i was	126	0.18	34	68
20	i will	126	0.18	36	72
21	of the	126	0.18	34	68
22	think it's	123	0.18	39	78
23	yeah and	123	0.18	30	60
24	don't know	122	0.18	36	72
25	when i	122	0.18	36	72
26	er i	121	0.17	38	76
27	the woman	121	0.17	34	68
28	they are	117	0.17	36	72
29	go to	116	0.17	38	76
30	the i	114	0.16	32	64
31	to the	113	0.16	42	84
32	how to	111	0.16	34	68
33	i think it's	109	0.16	36	72
34	**i don't know**	108	0.16	34	68
35	like to	108	0.16	32	64
36	yeah i	108	0.16	30	60

Table 8

The Top 50 Chunks in the Learner Language in the Taiwanese Sub-Corpus of LINDSEI (continued)

37	lot of	106	0.15	29	58
38	***you know***	106	0.15	22	44
39	**a lot of**	105	0.15	29	58
40	mm mm	105	0.15	24	48
41	i can	104	0.15	33	66
42	have a	103	0.15	37	74
43	i want	99	0.14	36	72
44	eh i	98	0.14	35	70
45	yeah mm	96	0.14	30	60
46	i would	92	0.13	28	56
47	is a	90	0.13	39	78
48	i like	89	0.13	34	68
49	need to	89	0.13	26	52
50	the painter	86	0.12	27	54

The tool returned the result for the two- to five-word clusters that appear five or more times. Table 8 lists the top 50 chunks in the speech of Taiwanese learners. Most of them are grammatical groups, such as *in the*, *I have* and *I don't*. As noted by the software developer, the function of clustering produces words being used together which are not necessarily meaningful multi-word units (Scott, 2013, p. 395). As pointed out by O'Keeffe et al. (2007, p. 61), the chunks/clusters/bundles generated by corpus software might consist of (a) highly-frequent fragmentary word groups (e.g. *and I* and *in the* in Table 8), (b) syntactically incomplete but meaningful strings (e.g. *I have* and *kind of* in Table 8), and (c) semantically and pragmatically fixed expressions (e.g. *a lot of* in Table 8). If the top 50 chunks in Table 8 are viewed from the perspective of written English discourse, most of them lack any syntactic unity or semantic integrity, but in spoken English discourse, in which utterances are often not syntactically unified, they are apparently natural. The most frequent chunks in the Taiwanese learners' speech are similar to those in the five-million-word CANCODE corpus and the

North American spoken component of the CIC corpus,[11,12] investigated by O'Keeffe et al. (2007), in that these chunks represent the speaker-listener world of *I* and *you*.

The chunks in the speech of Taiwanese learners need to be further investigated. The five meaningful chunks are identified: *I think*, *kind of*, *I don't know*, *you know*, and *a lot of*. The most frequently used one is *I think*, which occurs 600 times in all 50 texts. It can be seen that there is a very sharp fall-off between the first one, *I think*, and the second, *and I*. In some of the previous studies of *I think* in the speech of Chinese learners, by Yang and Wei (2005) and Xu and Xu (2007), it is found that *I think* is one of the frequently-used chunks in Chinese learners' spoken English. In addition, further research can be done in conjunction with the LOCNEC corpus, a native speech counterpart of LINDSEI.

CONTRIBUTIONS OF THE TAIWANESE SUB-CORPUS OF LINDSEI

The establishment of the Taiwanese learner corpus of spoken English will make contributions in three ways: (a) by serving as a model for the compilation of corpora of spoken English in Taiwan; (b) by increasing the visibility of Taiwanese learners in international academia; and (c) by informing the teaching of spoken English to Taiwanese students. The last contribution will result from more studies using this corpus in the future. In this section, some possible research topics are proposed and an example of CIA of *I think* is given.

First, the Taiwanese learner corpus of spoken English will be the first publicly available learner corpus in Taiwan. It will serve as a model for the compilation of corpora. In Taiwan, where the development of corpus studies is still in its infancy, this learner corpus, in collaboration

[11] The CANCODE (Cambridge and Nottingham Corpus of discourse in English) corpus was jointly built by Cambridge University Press and Nottingham University and contains 5 million words of spoken English collected in Britain (O'Keeffe et al., 2007).

[12] The multi-billion-word ICI (Cambridge International Corpus) corpus (currently the Cambridge English Corpus) consists of corpora of written and spoken English from various sources, such as books, newspapers, advertising, letters, emails, websites, recordings of conversations, lectures, television, meetings, and radio speech as well as learner language (Cambridge University Press, 2014).

with the LINDSEI team in Belgium, provides research training for the compiler as well as the team members. The compiler benefits from interacting with international researchers in the field of Corpus Linguistics and from being involved in the process of transcribing, which is seen as an analytical tool (Swann, 2010). Both these advantages will help the compiler to exploit the potential of the collected data. The team members gain research experience and broaden their scope in the expectation that more corpus studies will be done in future.

Second, Taiwanese learners represent one group of Chinese speakers, as well as the Chinese sub-corpus compiled in mainland China, in the fields of corpus studies and interlanguage research. LINDSEI is currently the most comprehensive learner corpus project and includes international collaboration by twenty groups at the time of writing. Being one of the sub-corpora of LINDSEI, without doubt, increases the visibility of Taiwan in international academia and contributes to the research on spoken English. The spoken data collected in Taiwan will be shared with other research groups of L1s. This, compared with a self-designed learner corpus, enables researchers worldwide to conduct a wider range of investigations. Furthermore, the learner speech collected in Taiwan in 2012 and 2013 offers the most recent data of this kind, while those in the Chinese sub-corpus were compiled in 2001 (Gilquin et al., 2010). The information in the learner profiles of the Chinese sub-corpus shows that 48 out of 53 learners (90.6%) had received six years of English education at school before they began their first degree and none of the learners had ever stayed in an English-speaking country. By contrast, the learners in the Taiwanese sub-corpus had much greater exposure to English. They had on average nearly ten years of English learning before entering university and 21 out of 60 (35%) learners had stayed for an average of 6.8 months in countries where English is spoken.

Third, the usage patterns of Taiwanese learners can be identified, which will facilitate and improve the teaching of spoken English. The importance of corpus studies and applications has been stated in recent international conferences on Applied Linguistics held in Taiwan (e.g. the 18th International Symposium on English Teaching: Internet- and Corpus-based English Instruction (13-15 November 2009), the 2012 International Conference on Applied Linguistics and Language Teaching:

Technological and Traditional Teaching and Learning (19-21 April 2012), and the 2012 LTTC International Conference: The Making of a Translator (28-29 April 2012)). However, there has hitherto been no learner corpus of spoken English available for research purposes. It is worth noting that the Language Training and Testing Centre in Taiwan has undertaken to transcribe the speaking tests of the GEPT, which was developed in Taiwan, but it may take some time for the learner corpus to be published. In mainland China, some learner corpora have been made available, for example, the Spoken and Written English Corpus of Chinese Learners, Version 1.0 (Wen, Wang, & Liang, 2005) and version 2.0 (Wen, Liang, & Yen, 2008); and the Chinese Learner Spoken English Corpus (Yang & Wei, 2005). The data in these corpora were collected from speaking tests which involve retelling a story, describing a picture and discussing a topic. In the test-taking context, the learners' speech was restricted and unnatural. In contrast, the spoken English produced in the informal interviews for LINDSEI was relatively authentic. The learners were voluntary and the setting was outside the classroom and not exam-oriented.

Research Possibilities

The corpus of Taiwanese students' spoken English provides a range of possibilities for research. As mentioned earlier, the sub-corpora in LINDSEI have been employed in CIA, in which two types of comparison can be made: (a) between NS and learner languages (in this case, LOCNEC (De Cock, 2004) and the Taiwanese sub-corpus) and (b) between speakers of different mother tongues (the Taiwanese sub-corpus and any other sub-corpora of LINDSEI). In addition, there is a growing interest in quasi-longitudinal studies, i.e. comparing learners of the same L1 at different levels of proficiency. Information about learners' English proficiency levels is available (see Table 2) and reliable, because it is based on the results of international standardised tests of English proficiency. In both CIA and quasi-longitudinal studies, a number of investigations can be pursued, for example, into lexis, phraseology, organization of spoken discourse, and features of spoken English.

More specifically, some features of different genres of spoken English can be explored. Each sub-corpus of LINDSEI consists of data from three tasks: set topics, free discussion, and picture description (see

Table 4 for their distributions). The first two can be categorized as dialogic genre, whereas the last one is mainly monologic. For example, it will be possible to identify characteristics of productive fluency in these two genres and compare them in terms of such speech management strategies as repeats, filled pauses, and self-corrections.

As reported above, the use of *yeah*, *eh*, *er*, *mm*, *em*, and *oh*, found in native speech (O'Keeffe et al., 2007), is also identified in the Taiwanese learners' spoken data. The first type of CIA, comparing the LINDSEI-Taiwanese corpus with LOCNEC (De Cock, 2004), can be deployed to investigate these words. In a similar way, a quasi-longitudinal study of the use of these words by higher- and lower-level proficiency learners can be undertaken. Results of these kinds may shed light on the naturalness and spontaneity of spoken English and be applied to pedagogy.

Among the five features of spoken English (a) deictic expressions, (b) situational ellipsis, (c) headers, tails and tags, (d) discourse markers and (e) polite and indirect language, vague language and approximations (Carter & McCarthy, 2006), discourse markers have attracted much research attention (e.g. on Chinese learners: Fung & Carter, 2007; He & Xu, 2003; Huang, 2011; Liu, 2010;). The quantitative corpus studies have revealed the use of discourse markers by learners. Such research has been conducted across the eleven sub-corpora by Gilquin and Granger (2011, forthcoming). These researchers point out that using LINDSEI as an aggregate may conceal variations between learners of different L1s as well as between learners in a specific corpus. It seems that the L1 plays an important role for ESL learners.

In terms of practical applications, learner corpus research has certainly helped us to improve our understanding of learner language and to inform English Language Teaching. However, there is always more work to do. As De Cock (2010) notes in her call for more studies using spoken learner corpora in the classroom, the compilation of the Taiwanese sub-corpus of LINDSEI will certainly facilitate research on Chinese-speaking learners, which is one of the biggest groups to use English as a foreign language.

Example of Contrastive Interlanguage Analysis (CIA): *I think*

Learner corpora have been used in CIA and Computer-aided Error Analysis; learner corpus research has informed two fields: Second Language Acquisition and Language Teaching (Gilquin & Granger, forthcoming). In this section, one example of using the Taiwanese sub-corpus of LINDSEI is given: CIA in the case of *I think*.

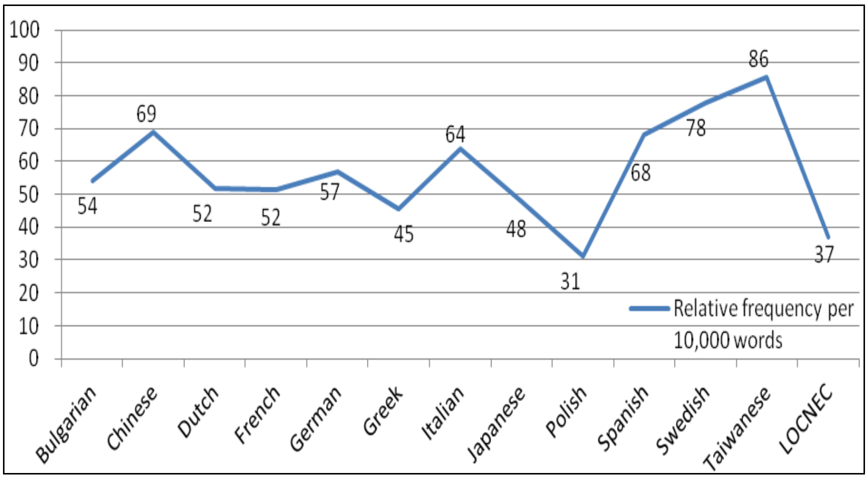


Figure 4. Relative frequencies of *I think* across the sub-corpora of LINDSEI and LOCNEC

On the principle that the LINDSEI sub-corpora are comparable to one another and to the native counterpart, LOCNEC, one possible investigation is the comparison of *I think* between two sub-corpora or across all sub-corpora and between the Taiwanese component and LOCNEC. *WordSmith Tools 6* (Scott, 2012) was used to produce the word counts and frequencies of the two-word chunk, *I think*, in the utterances of the interviewees (learners of English in LINDSEI and English speakers in LOCNEC). Figure 4 presents the relative frequencies of *I think* across the twelve sub-corpora of LINDSEI and LOCNEC. This frequency information is used as a point of entry into the data. It can be clearly seen that *I think* is much more frequently used in the Taiwanese sub-corpus than in the other corpora available for analysis.

In order to have a robust indicator of how significant the differences in frequency are between the two corpora and across all the corpora, statistical tests can be done to facilitate the interpreting of the data. The log-likelihood (LL) test is a common test of statistical significance in corpus studies.[13] When the relative frequency of *I think*, 86 times per 10,000 words, in the Taiwanese component of LINDSEI is compared with the 37 times per 10,000 words in its native counterpart LOCNEC, the LL score is +54.71,[14] which is much higher than the critical value (10.83) for the level of significance p<0.0001. This indicates that there is a statistically significant difference between the frequencies of *I think* in the LINDSEI-Taiwanese and LOCNEC and its overuse in the LINDSEI-Taiwanese relative to LOCNEC.

To compare the frequencies of *I think* in LINDSEI-Taiwanese and any other components of LINDSEI, the same statistical test can be employed. For instance, the Taiwanese and Chinese sub-corpora share the same first language, but the relative frequencies of *I think* (86 vs. 69 times per 10,000 words) in these two groups seem rather different. The LL score +0.09, below the critical value (3.83) for the level of significance p<0.05, indicates that the difference is not statistically significant.

From the above frequency comparison, some questions may be further explored; for example, how is *I think* used by Taiwanese learners and native speakers? Is it used for epistemic meanings or as a discourse marker? In native speech, the epistemic stance use of *I think* is most common (see studies such as Aijmer, 1997; Biber et al., 1999; Simon-Vandenbergen, 2000; Fortanet, 2004; O'Keeffe et al., 2007), but it can be interpreted differently, for example, as a hedge to express doubt in casual conversations and as an expression of opinion to show feelings of certainty and authority in political interviews (Simon-Vandenbergen,

[13] More discussion on statistical tests for corpus studies can be found in McEnery, Xiao and Tono (2006), Dunning (1993), Gries (2013, forthcoming).

[14] The LL calculator created by Paul Rayson (2011) of Lancaster University was used to perform the log-likelihood tests. The critical value of 15.13 for significance at the p<0.0001 level is applied in corpus studies. As suggested by Rayson, Damon and Brian (2004), setting the critical values in the LL test at a higher value for the significance level of 0.0001 can increase their reliability.

2000). Fortanet (2004) claims that in some cases of expressing opinion, *I think* seems to be associated with secondary functions, such as evaluation, vagueness and politeness.

If native usage is taken as the norm for teaching, how similar is learners' usage to that of native speakers, or how different? Do the Taiwanese learners have a strong preference for using *I think* over other modal expressions (e.g. *in my opinion*, *it seems to me*, *I would say*, *I believe*, *maybe* and *possibly*)? Do they also use other options as native speakers do? If not, some pedagogical intervention may be needed to raise their awareness of native usages of *I think* and other options. Yang and Wei's (2005) study of Chinese learners indicates that *I think* was over-used and the researchers claimed that in most cases *I think* was used as a 'conversational filler' (p. 40). Xu and Xu's (2007) investigation of discourse management chunks in Chinese learners' speech and native speech in ICE-GB reported that Chinese learners were unable to produce interpersonal chunks as varied as those of NSs. They also found that Chinese learners tended to literally translate chunks in Chinese and use first-person perspective language, such as *I think*, *in my opinion*, *I want to say*, *it's my turn* and *I don't agree*. The use of 'I-perspective' language was suggestive of self-centredness (Xu & Xu, 2007, p. 440). In the teaching of spoken English, the instruction of indirect language instead of *I think* might be of help for learners to improve interaction in certain contexts.

Moreover, it seems worthwhile to investigate whether or not the overuse of *I think* by the Taiwanese learners is due to L1 transfer. This can be done by comparing corpora of different L1s and by analysing a corpus of Chinese conversations (e.g. The Mandarin Topic-oriented Conversation Corpus (MTCC) and The Mandarin Conversational Dialogue Corpus (MCDC) (Institute of Linguistics, 2014) in order to identify and examine some Chinese equivalents of the usage of *I think* in English.

Whatever research questions are pursued, both the quantitative and qualitative analyses should be expanded to provide a more detailed description of the use of *I think* in learner language. This example of CIA shows the capacity of learner corpus research to shed light on the linguistic features typical of certain (groups of) learners.

CONCLUDING REMARKS

This paper reports the compilation of the Taiwanese sub-corpus of LINDSEI and preliminary investigation of this Taiwanese learner corpus. This learner corpus is of value in three aspects: (a) the procedure for the compilation of this spoken corpus is insightful for researchers who plan to carry out a similar project; (b) the spoken English corpus of Taiwanese learners will be published as a sub-corpus of LINDSEI (2nd edition), thereby increasing the visibility of Taiwan in academia and the possibilities of applying it; and (c) the research findings will serve as a reference for teaching English speaking.

As this Taiwanese learner corpus involves collaboration with international research teams, it certainly has a great deal of potential for future research. Its research possibilities are suggested by the example of investigating the two-word chunk, *I think*. However, the implications of learner corpus research are complex, open to interpretation from perspectives of Second Language Acquisition, Language Teaching, and English as a Lingua Franca, and not suggestive of easy, straightforward application. It requires much more preparation than ready-made materials. It is hoped that with the completion of this learner corpus and many others in the future, the use of learner corpora in Applied Linguistics will continue to increase.

REFERENCES

2013 members of the Audacity development team. (2013). *Audacity (Version 2.0.3)*. Ohio: GNU General Public License (GPL).

Aijmer, K. (1997). *I think* - an English modal particle. In T. Swan & O. J. Westvik (Eds.), *Modality in Germanic languages. Historical and comparative perspectives* (pp. 1-47). Berlin: Mouton de Gruyter.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Bernath, C. (1998). *Soochow Colber Student Corpus*. Soochow University, Taiwan.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Harlow, Essex: Person Education Limited.

Biber, D., Finegan, E., Johansson, S., Conrad, S., & Leech, G. (1999). *Longman grammar of spoken and written English*. Essex: Pearson Education Limited.

Brand, C., & Kämmerer, S. (2006). The Louvain international database of spoken English interlanguage (LINDSEI): Compiling the German component. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 127-140). Frankfurt: Peter Lang.

Cambridge University Press. (2014). *Cambridge English corpus*. Retrieved from http://www.cambridge.org/de/elt/catalogue/subject/item2701617/Cambridge-English-Corpus/?site_locale=de_DE

Carter, R., & McCarthy, M (2006). *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Centre for English Corpus Linguistics. (2013). *Learner corpora around the world*. Retrieved from http://www.uclouvain.be/en-cecl-lcworld.html

Chen, H. (2011). *A study on Taiwanese EFL learners' academic writing corpus*. Paper presented at the Learner Corpus Research 2011 "20 years of learner corpus research: looking back, moving ahead", Louvain-la-Neuve, Belgium.

Chiu, Y. W. (2007). The investigation of problems in English education in primary schools. *Training Information, 24*(4), 135-140.

Chung, S. F., Wang, S. Y., & Tseng, Y. W. (2010). The construction of the NCCU foreign language learner corpus. *Foreign Language Studies, 12*, 71-98.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, New Series*, *2*, 225-246.

De Cock, S. (2010). Spoken learner corpora and EFL teaching. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 123-137). London: Continuum.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Fortanet, I. (2004). *I think*: Opinion, uncertainty or politeness in academic spoken English? *RAEL: Revista Electronica de Linguica Aplicada, 3*, 63-84.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410-439.

Gilquin, G. (2012). *Transcription guidelines*. Retrieved from http://www.uclouvain.be/en-307849.html

Gilquin, G. (2014). LINDSEI Partners. Retrieved from http://www.uclouvain.be/en-307845.html

Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). *LINDSEI Louvain international database of spoken English interchange. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, G., & Granger, S. (2011). *The use of discourse markers in corpora of native and learner speech: From aggregate to individual data*. Paper presented at the Corpus Linguistics Conference 2011, Birmingham.

Gilquin, G., & Granger, S. (forthcoming). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies. Lund studies in English 88* (pp. 37-51). Lund: Lund University Press.

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: Longman.

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

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2013). *Twenty years of learner corpus research: Looking back, moving ahead* (Vol. Proceedings 1). Louvain-la-Neuve: Presses Universitaires de Louvain.

Gries, S. T. (2013). Statistical tests for the analysis of learner corpus data. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data*. Amsterdam & Philadelphia: John Benjamins.

Gries, S. T. (forthcoming). Quantitative designs and statistical techniques. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

He, A., & Xu, M. (2003). Small words in Chinese EFL learners' spoken English. *Foreign Language Teaching and Research, 35*(6), 446-453.

Huang, L. F. (2011). *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers* (Unpublished doctoral dissertation), University of Birmingham, UK. Retrieved from http://etheses.bham.ac.uk/2969/

Huang, L. Y. (1991). *Learner corpora and teaching spoken English in Taiwan universities*. Taipei: Chengchi University.

Institute of Linguistics, Academia Sinica. (2014). *Archives and linguistic representation of spoken Taiwan Mandarin*. Retrieved from http://mmc.sinica.edu.tw/mtcc_e.htm

Liu, B. (2010). *Discourse marker use by L1 Chinese EFL speakers* (Unpublished doctoral dissertation), University of Florida, US.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. Oxon: Routledge.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Rayson, P. (2014). *Log-likelihood calculator*. Retrieved from http://ucrel.lancs.ac.uk/llwizard.html

Rayson, P., Damon, B., & Brian, F. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Airon, & A. Dister (Eds.), *Le poids des mots: Proceedings of the 7th international conference on statistical analysis of textual data (JADT 2004) Louvain-la-Neuve, Belgium, March 10-12, 2004* (Vol. II, pp. 926 - 936). Louvain: Presses universitaires de Louvain.

Scott, M. (2012). *WordSmith tools* (Version 6). Liverpool: Lexical Analysis Software.

Scott, M. (2013). *WordSmith tools manual*. Liverpool: Lexical Analysis Software Ltd.

Shih, R. H. H. (2000). Compiling Taiwanese learner corpus of English. *Computational Linguistics and Chinese Language Processing, 5*(2), 89-102.

Simon-Vandenbergen, A. M. (2000). The functions of *I think* in political discourse. *International Journal of Applied Linguistics, 10*(1), 41-63.

Someya, Y. (1998). *e_lemma* (Ver. 2 for WordSmith 4). Retrieved from http://www.lexically.net/downloads/e_lemma.zip.

Swann, J. (2010). Transcribing spoken interaction. In S. Hunston & D. Oakey (Eds.), *Introducing applied linguistics: Concepts and skills* (pp. 163-176). Abingdon: Routledge.

Wen, Q. F., Liang, M. C., & Yen, X. Q. (2008). *Spoken and written English corpus of Chinese learners, Version 2.0*. Beijing: Foreign Language Teaching and Research Press.

Wen, Q. F., Wang, L. F., & Liang, M. C. (2005). *Spoken and written English corpus of Chinese learners, Version 1.0*. Beijing: Foreign Language Teaching and Research Press.

Xu, J. J., & Xu, Z. R. (2007). Discourse management chunks in Chinese college learners' English speech: A spoken corpus-based study. *Foreign Language Teaching and Research, 39*(6), 437-443.

Yang, H. C., & Wei, N. H. (2005). *Construction and data analysis of a Chinese learner spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.

ACKNOWLEDGMENTS

This work was supported by the National Science Council, Taiwan, under grant number NSC101-2410-H-158-012. Without this funding, the Taiwanese sub-corpus of LINDSEI would not have been possible. My gratitude goes to the LINDSEI team at the Centre for English Corpus Linguistics of the Université Catholique de Louvain, Belgium, in particular, the project leader, Prof Sylviane Granger and the coordinator, Dr Gaëtanelle Gilquin. The efforts of my project team members, Ms Hsiao-hui Lin, Ms Miranda Yu-ting Huang, Dr Jon Nichols, Mr Simon Kubelec, and Mr Chih-hao Hsueh are most appreciated. I would also like to thank Mr Alex Jou for his help with data collection and Mr Sheng Li for his technical support. Special thanks are due to my contacts in the six universities participating in this corpus and the Taiwanese learners who agreed to be interviewed and recorded.

CORRESPONDENCE

Lan-fen Huang, Language Centre, Shih Chien University, Taiwan
E-mail address: lanfen.huang@gmail.com

APPENDIX

Appendix A. LINDSEI Transcription Guidelines (Gilquin, 2012)

1. Interview identification

Each interview is preceded by a code of this type: <h nt="FR" nr="FR+*three-figure number*">

e.g. <h nt="FR" nr="FR004"> (4th interview with French mother tongue student)

Examples of country codes:

DUTCH	=	DU001
GERMAN	=	GE001
NORWEGIAN	=	NO001
SPANISH	=	SP001
SWEDISH	=	SW001

All interviews should end with the following tag (on a separate line): </h>

2. Speaker turns

Speaker turns are displayed in vertical format, i.e. one below the other. Whilst the letter "A" enclosed between angle brackets always signifies the interviewer's turn, the letter "B" between angle brackets indicates the interviewee's (learner's) turn. The end of each turn is indicated by either </A> or </B>.

e.g. <A> okay so which topic have you chosen </A>
<B> the film or play that I thought was particularly good or bad really </B>

3. Overlapping speech

The tag <overlap /> (with a space between "overlap" and the slash) is used to indicate the beginning of overlapping speech. It should be indicated in both turns. The end of overlapping speech is not indicated.

e.g. <B> yeah I went on a bus to London once and I'll never <overlap /> do it again </B>

<A> <overlap /> that's even worse </A>

4. Punctuation

No punctuation marks are used to indicate sentence or clause boundaries.

5. Empty pauses

Empty pauses are defined as a blank on the tape, i.e. no sound, or when someone is just breathing. The following three-tier system is used: one dot for a "short" pause (< 1 second), two dots for a "medium" pause (1-3 seconds) and three dots for "long" pauses (> 3 seconds).

e.g. <B> (erm) .. it’s a British film there aren't many of those these days </B>

6. Filled pauses and backchannelling

Filled pauses and backchannelling are marked as (eh) [brief], (er), (em), (erm), (mm), (uhu) and (mhm). No other fillers should be used.

e.g. <B> yeah . well Namur was warmer (er) it was (eh) a really little town </B>

7. Unclear passages

A three-tier system is used to indicate the length of unclear passages: <X> represents an unclear syllable or sound up to one word, <XX> represents two unclear words, and <XXX> represents more than two words.

e.g. <B> <X> they're just begging <XX> there's there's honestly he did a course .. for a few weeks </B>

If transcribers are not entirely sure of a word or word ending, they should indicate this by having the word directly followed by the symbol <?>.

e.g. <B> I went to see a<?> friend at university there and stayed </B>

Unclear names of towns or titles of films for example may be indicated as <name of city> or <title of film>.

e.g. <B> where else did we go (er) <name of city> it's in Bolivia </B>

8. Anonymisation

Data should be anonymised (names of famous people like singers or actors can be kept). Transcribers can use tags like <first name of interviewee>, <first name and full name of interviewer> or <name of professor> to replace names.

e.g. <A> I'm <first name of interviewer> . what's your name? </A>

9. Truncated words

Truncated words are immediately followed by an equals sign.

e.g. <B> it still resem= resembled the theatre </B>

10. Spelling and capitalisation

British spelling conventions should be followed. Capital letters are only kept when required by spelling conventions on certain specific words (proper names, I, Mrs, etc) – not at the beginning of turns.

11. Contracted forms

All standard contracted forms are retained as they are typical features of speech.

12. Non-standard forms

Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos*, *dunno, gonna, gotta*, *kinda*, *wanna* and *yeah*.

13. Acronyms

If acronyms are pronounced as sequences of letters, they are transcribed as a series of upper-case letters separated by spaces.

e.g. <B> yes not really I did sort of basic G C S E French and German </B>

If, on the other hand, acronyms are pronounced as words, they are transcribed as a series of upper-case letters not separated by spaces.

e.g. <A> (mhm) (er) you're doing a MAELT </A>

14. Dates and numbers

Figures have to be written out in words. This avoids the ambiguity of, for example, "1901", which could be spoken in a number of different ways.

e.g. <B> an awful lot of people complain and say well the grants were two thousand two hundred </B>

15. Foreign words and pronunciation

Foreign words are indicated by <foreign> (before the word) and </foreign> (after the word).

e.g. <B> we couldn't go with (er) knives and so on <foreign> enfin </foreign> we were (er) </B>

As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is also marked using the <foreign> tag.

e.g. <B> I didn't have the (erm) . <foreign> distinction </foreign> </B>

16. Phonetic features

(a) Syllable lengthening

A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with small words like *to*, *so* or *or*. Colons should not be inserted within words.

e.g. <B> that's something I'll I'll plan to: to learn </B>

(b) Articles

-when pronounced as [ei], the article *a* is transcribed as a[ei];

e.g. <B> and it's about (erm) . life in a[ei] (eh) public school in America I think </B>

-when pronounced as [i:], the article *the* is transcribed as the[i:].

e.g. <B> and the[i:] villa we were staying in was in one of the valleys </B>

17. Prosodic information: voice quality

If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it.

e.g. <B> <starts laughing> I don't have to assess it I only have to write it <stops laughing> </B>

18. Nonverbal vocal sounds

Nonverbal vocal sounds are enclosed between angle brackets.

e.g. <B> I hope so I've I've got some <coughs> friends out there </B>
e.g. <B> so I went back into Breda .. and sat down again <imitates the sound of a guitar> </B>

19. Contextual comments

Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example).

e.g. <A> no it's true it's nice to have your own bathroom </A>
<somebody enters the room>
<B> hi </B>

20. Tasks

The three tasks making up the interview (set topic, free discussion and picture description) should be separated from each other. This is done using the following tags: <S> (before the set topic), </S> (after the set topic), <F> (before the free discussion), </F> (after the free discussion), <P> (before the picture description), </P> (after the picture description). These tags should occupy a separate line and should not interrupt a turn.

e.g. <S>
<A> did you . manage to choose a topic </A>

Appendix B. Learner Profile

==

Text code: (to be filled in by the researcher)

==

Surname: **First name(s)**:
Age:
Male □ Female □

Nationality:
Country:
Native language:
Father's mother tongue:
Mother's mother tongue:
Language(s) spoken at home: (if more than one, please give the average % use of each)

Education:
Primary school - medium of instruction:
Secondary school - medium of instruction:

Current studies:
Current year of study:
Institution:
Medium of instruction:
English only □
Other language(s) (specify)_______________□
Both □

==

Years of English at school:
Years of English at university:

Stay in an English-speaking country:
Where?
When?
How long?

Appendix B. (continued)

Have you ever taken an English proficiency test? If yes:
Name of the test:
Result: Date:

==

Other foreign languages in decreasing order of proficiency:

==

I hereby give permission for my interview to be used for research purposes.

Date: .... Signature: ....

****

Section to be filled in by the interviewer
Interviewer: Male □ Female □
Native language:
Foreign languages (in decreasing order of proficiency):

Relation with learner: Familiar □ Vaguely familiar □
Unfamiliar □
(If possible, please be more specific, e.g. learner’s professor, TA, etc: ……...……………………..)

Appendix C. The Statistical Information of Tokens and Types in the Learner Language in the Taiwanese Sub-corpus of LINDSEI

N	text file	file size	tokens (running words) in text	tokens used for word list	types (distinct words)	type/token ratio (TTR)	standardised TTR	STTR basis
1	Overall	470375	69577	69577	3741	5.38	28.45	1000
2	TW_B001.txt	13721	1879	1879	448	23.84	30.40	1000
3	TW_B002.txt	12759	1827	1827	453	24.79	31.40	1000
4	TW_B003.txt	9232	1355	1355	372	27.45	29.60	1000
5	TW_B004.txt	11725	1696	1696	418	24.65	28.00	1000
6	TW_B005.txt	8477	1272	1272	332	26.10	27.60	1000
7	TW_B006.txt	8432	1363	1363	317	23.26	26.10	1000
8	TW_B007.txt	6648	1021	1021	314	30.75	30.80	1000
9	TW_B008.txt	9973	1753	1753	380	21.68	24.90	1000
10	TW_B009.txt	11408	1909	1909	418	21.90	27.50	1000
11	TW_B010.txt	11682	1862	1862	468	25.13	29.40	1000
12	TW_B011.txt	9702	1651	1651	370	22.41	26.10	1000
13	TW_B012.txt	7004	1099	1099	309	28.12	28.10	1000
14	TW_B013.txt	7697	1202	1202	298	24.79	24.90	1000
15	TW_B014.txt	8050	1239	1239	324	26.15	28.60	1000
16	TW_B015.txt	7706	1138	1138	319	28.03	29.00	1000
17	TW_B016.txt	9546	1241	1241	341	27.48	29.80	1000
18	TW_B017.txt	8831	1291	1291	327	25.33	26.50	1000
19	TW_B018.txt	10403	1452	1452	394	27.13	31.10	1000
20	TW_B019.txt	9409	1238	1238	318	25.69	27.50	1000
21	TW_B020.txt	7787	1047	1047	341	32.57	33.80	1000
22	TW_B021.txt	13500	1880	1880	441	23.46	29.40	1000
23	TW_B022.txt	9510	1398	1398	412	29.47	33.70	1000
24	TW_B023.txt	12323	1736	1736	433	24.94	30.40	1000
25	TW_B024.txt	9248	1553	1553	352	22.67	25.80	1000
26	TW_B025.txt	8172	1273	1273	306	24.04	26.30	1000
27	TW_B026.txt	7822	1149	1149	361	31.42	31.80	1000
28	TW_B027.txt	7505	1082	1082	312	28.84	29.10	1000
29	TW_B028.txt	8242	1341	1341	348	25.95	28.20	1000
30	TW_B029.txt	9172	1645	1645	382	23.22	28.10	1000

Appendix C. (continued)

N	text file	file size	tokens (running words) in text	tokens used for word list	types (distinct words)	type/token ratio (TTR)	standardised TTR	STTR basis
31	TW_B030.txt	10154	1575	1575	459	29.14	34.40	1000
32	TW_B031.txt	6936	1177	1177	278	23.62	24.40	1000
33	TW_B032.txt	6189	761	761	229	30.09		1000
34	TW_B033.txt	7810	1064	1064	314	29.51	30.30	1000
35	TW_B034.txt	11330	1639	1639	396	24.16	28.20	1000
36	TW_B035.txt	9424	1333	1333	353	26.48	29.30	1000
37	TW_B036.txt	8906	1267	1267	333	26.28	28.80	1000
38	TW_B037.txt	11180	1733	1733	381	21.98	23.90	1000
39	TW_B038.txt	6454	873	873	289	33.10		1000
40	TW_B039.txt	16675	2401	2401	488	20.32	28.00	1000
41	TW_B040.txt	10951	1397	1397	351	25.13	28.10	1000
42	TW_B041.txt	10065	1406	1406	377	26.81	29.30	1000
43	TW_B042.txt	6982	894	894	281	31.43		1000
44	TW_B043.txt	7206	1205	1205	278	23.07	24.20	1000
45	TW_B044.txt	8087	1147	1147	319	27.81	28.50	1000
46	TW_B045.txt	8314	1159	1159	303	26.14	27.70	1000
47	TW_B046.txt	7126	987	987	278	28.17		1000
48	TW_B047.txt	7990	1148	1148	269	23.43	24.40	1000
49	TW_B048.txt	10764	1569	1569	392	24.98	30.30	1000
50	TW_B049.txt	8837	1296	1296	330	25.46	28.30	1000
51	TW_B050.txt	13309	1954	1954	415	21.24	27.30	1000

「魯汶國際英語口語中介語語料庫」：
台灣英語學習者口語語料庫之建構

黃蘭棻
實踐大學

「魯汶國際英語口語中介語語料庫」(LINDSEI) (Gilquin 等 2010)為規模最大的英語學習者口語語料庫之一，目前共有二十個國際研究團隊參與。為確保各語料庫之間的可比性，台灣英語學習者口語語料庫依 LINDSEI 設計準則來建構。本文詳述語料庫建構流程—招募參與者、執行面談和謄寫音檔等。與其它子語料庫略為不同，台灣子語料庫收錄參與者的英語檢定成績，以歐洲語言共同參考架構(CEFR)為標準，程度大多介於 B1 和 C1 等級。本研究使用台灣子語料庫和 LINDSEI 第一版十一個子語料庫，進行量化語料分析、單詞分析和詞串分析。再以台灣子語料庫中頻率最多的詞串 *I think* 為例，初步量化比較中介語，並討論其研究潛力。台灣英語學習者口語語料庫透過國際合作，將提供國內外學者研究之用，並作為未來建構語料庫之參考。

關鍵詞：魯汶國際英語口語中介語語料庫、中介語、學習者語料庫、台灣英語學習者