English Learner Corpus: Global Perspectives with an Asian Focus

Chu-Ren Huang[1], Winnie Cheng[2], Hintat Cheung[3], Yasunari Harada[4], Huaqing Hong[5], Sophia Skoufaki[3], & Helen K.Y. Chen[1]
Department of Chinese and Bilingual Studies, the Hong Kong Polytechnic University[1]
Institute of Linguistics, Academia Sinica[1]
Department of English, the Hong Kong Polytechnic University[2]
National Taiwan University[3]
Waseda University[4]
National Institute of Education, Singapore[5]
Nanyang Technological University, Singapore[5]
churen.huang@inet.polyu.edu.hk, egwcheng@polyu.edu.hk, hintat@ntu.edu.tw, harada@waseda.jp, huaqing.hong@nie.edu.sg, sophiaskoufaki@ntu.edu.tw, helenkychen@gmail.com

Abstract

In the past decade, the use of corpora and computational tools in language education has played a crucial and critical role in innovations in language education. The current paper provides a survey of the current state of second language learner corpora in East Asia. Four types of learner corpus from four regions in East Asia are introduced, with further details on the design and compilation of each corpus, as well as extended research and applications based on each corpus. We also provide recommendations on best practices in implementing learner corpus data in language teaching and learning, and laying out a roadmap for future development and synergy among second language learner corpora with different first language speakers in East Asia.

Key Words: English learner corpus in East Asia Region, HFKSC, HKEC, VALIS, SCoRe, LTTC English Learner Corpus,

Introduction

The development of corpora and corpus-based approaches was probably the single most influential research innovation in the area of language teaching and learning in the past 50 years. The emergence of Learner Corpus, in addition, allows second language teaching and learning researchers to focus on issues specific to a community of learners, such as those who share the same mother tongue. Looking at Learner Corpus from a particular region has several important research implications. First, it facilitates sharing of resources and experience among learners with different but related linguistic and cultural backgrounds. Second, it allows direct study of the influence of first language on second language learning and helps to develop efficient learning and teaching strategies.

Survey of Current State of Learners Corpus in East Asia

A List of Learner Corpus: Learner Corpus around the World

Tono (2003) offers a list of learner corpora from around the world. In addition to the learner corpora from Europe and America, part of the list also introduces some learner corpora from East Asia region. Since then, there has been progressive development and compilation of additional learner corpus, for example, there are the Spoken and Written Corpus of Chinese Learners (SWECCL) and Bilingual Corpus of Chinese English Learners

(BICCEL) – a.k.a. Parallel Corpus of Chinese EFL Learners (PACCEL) of English majors, and Chinese Learner English Corpus (CLEC) and College Learners' Spoken English Corpus (COLSEC) of non-English majors from China. In Hong Kong, the Learner Corpus of English for Business Communication, Learner Corpus of Essays and Reports, and the Learner Journals have been compiled. Also, there have been other additions to the original list of learner corpora from other countries in East Asia region by Tono, such as the GEPT-Learner Corpus from Taiwan, the Thai English Learner Corpus (TELC) from Thailand, and the Corpus Archive of Learner English in Sabah/Sarawak (**CALES**) from Malaysia.

General Observations

Some general observations can be derived from the aforementioned list of learner corpora. First of all, most of these corpora focus on learners' output alone; second, most of the data used in the compilation of these corpora are derived from written data, rather than spoken data. Another shared feature of these corpora is that their data are mostly derived from timed output, such as examination with a fixed time frame. Moreover, almost all data are based on English as the foreign or second language. Finally, the data of each corpus is mostly collected from a particular country or region.

Further Research Questions

In addition to the above observations, some general research questions can be further proposed, for example, how do speakers who are involved in these projects transfer their native languages to the second languages? Also, if the data of the corpus are collected from a classroom environment, what are the instructors' roles in these data, and how about the textbooks, workbooks, and assessments? Another research question is, since some of these corpora focus on learners whose native language is Chinese, it will be of further interest to explore whether there are some common patterns and inherent variations of Chinese learners' English in Greater China (including China, Hong Kong, Taiwan, and Chinese speakers from other East Asia region). To extend this research even further, perhaps some common patterns and inherent variations among learners from East Asia region could be identified, as well as the implication of these patterns and variations in reflecting an Asian Pedagogy for teaching and learning English as a second language.

Learner Corpora from Four Regions

In this section, four different learner corpora from various regions in East Asia are introduced. These learner corpora include: the LTTC English learner corpus, the Singapore Corpus of Research in Education (SCoRE) project, the corpus work by Winnie Cheng, and the English Learners' Utterance Data Collection and Compilation at Waseda University by Yasunari Harada. In addition to introducing each corpus, details on the design and compilation of each corpus will also be given in what follows. Each of these corpora projects has provided not only rich information on the corpus design, but also further extended researches and application based on the data collected for each corpus.

LTTC English Learner Corpus

Introduction. The Language Teaching and Testing Center (LTTC) and the Graduate Institute of Linguistics (GIL) have been collaborating since 2007 on the construction of the

LTTC English Learner Corpus[1]. The corpus consists of language samples by Taiwanese learners of English who have sat the General English Proficiency Test (GEPT), a language proficiency examination administered by the LTTC. In the current, first phase of corpus construction, 2,000 written-production and 400 oral-production samples from the Intermediate GEPT examination have been processed.

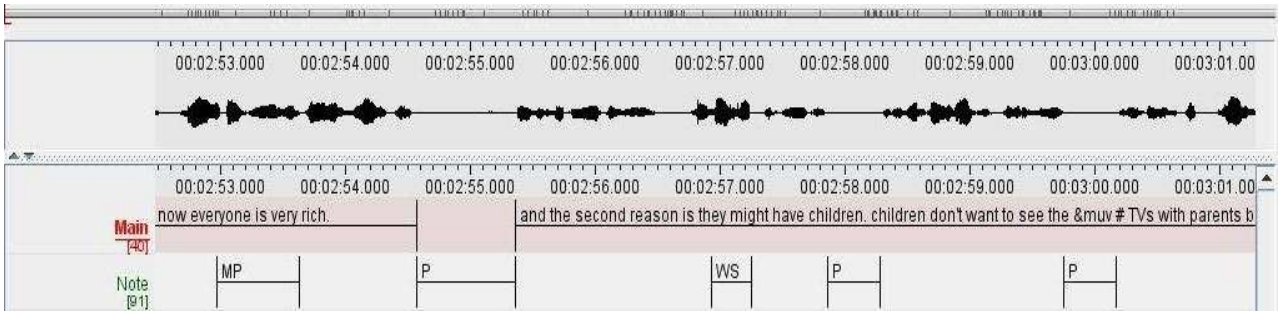
***Corpus design and compilation*.** As far as the content of the samples is concerned, the written samples which have been processed are short paragraphs on three topics. Two of these topics are questions about personal preferences (favourite food and idol) and the third asked test-takers to explain why many elementary-school children in Taiwan are nearsighted and to propose effective ways of preventing nearsightedness. The oral samples contain answers to ten questions and the description of a picture. Three sets of answers and image descriptions have been processed.

Metadata about the performance and characteristics of each test-taker are available for each sample. These metadata are the score that was assigned by test-graders to each sample, the broad region of Taiwan (e.g., North, East) where the test was taken, the age, gender, education level of the test-taker, his or her major if the test-taker was a college graduate, whether the test-taker was a student or not, and whether (s)he had lived in an English-speaking country, and if so, for how long.

The digitization of the written and oral samples relied on the principle of having each sample transcribed by two persons, comparing the two transcriptions through a computer program and then making revisions manually based on the differences between the transcripts. In terms of the written samples, the 2,000 hand-written samples were initially scanned and then each sample was typed in the computer independently by two student helpers. The two versions of each sample were compared by a computer program and revisions were done manually. The oral samples were initially recorded on cassette tapes, then digitized and finally student helpers were trained to transcribe them using the software ELAN (EUDICO Linguistic Annotator) (Hellwig, van Uytvanck, & Hulsbosch 2008) and add tags using the CHAT (CHILDES) format (MacWhinney 2008). Tags were added inside the body of the transcriptions for repetitions, self-corrections, incomprehensible sounds, lengthened vowels and other characteristics. Tags for filled (FP) and unfilled (P) pauses, mispronunciations (MS) and word stress (WS) errors were added in the tier below that of the main transcription. Figure 1 below shows the soundwave and the transcription of part of an ELAN file. The transcription appears in two tiers. The first tier contains the text and CHAT symbols and the second codes for unfilled pauses, mispronunciations, and word stress errors.

[1] This project is directed by Prof. Hintat Cheung of GIL. The project co-directors are Dr Zhao-Ming Gao from the Department of Foreign Languages and Literatures at NTU and Dr Siaw-Fong Chung from the Department of English at National Chengchi University. The project members are the postdoctoral research associate, Dr Sophia Skoufaki, the research assistant and administrator Ms Sumei Chen, and two PhD students, Ms Sally Chen and Ms Claire Chiyi Wu. Ms Shanju Lin was the research assistant and administrator in the academic year 2008-9.

Figure 1
Soundwave and Transcription Segment from an ELAN Oral File



Two student helpers transcribed each oral sample and these two versions were compared through a computer program. The samples were then revised manually.

The last stage in the processing of the sample files is part-of-speech tagging. So far, all the written samples and half of the oral samples have been part-of-speech tagged with the CLAWS4 tagger (Garside & Smith 1997). Following part-of-speech tagging, the aforementioned metadata were added to the samples through a software program so that each sample has the structure of a CHAT file. Figure 2 below shows a writing sample in CHAT format.

Figure 2
A Written Sample in CHAT Format

```
@UTF8
@Begin
@Languages:     en
@Participants:  TXT Writing Sample
@ID:    en|lttc|TXT|||||||
@No.:   1002
@Age:   18
@Gender:        Female
@Region:        Central Taiwan
@Education:     Senior Hight School
@Major: None
@Time of living in English-speaking countries:  Never or Blank
@Student or not:        student
@Testing year:  2008
@Test Level:    intermediate
@Test Paper No.:        IW-0801
@Composition Score:     2.5
*TXT:   The problem of nearsightedness is serious in Taiwan.
%POS:   The_AT problem_NN1 of_IO nearsightedness_NN1 is_VBZ serious_JJ in_II Taiwan_NP1 ._.
*TXT:   Most elementary school's students have nearsight.
%POS:   Most_DAT elementary_JJ school_NN1 's_GE students_NN2 have_VHO nearsight_NN1 ._.
*TXT:   Each family may have a computer, a television, PSP or Wii.
%POS:   Each_DD1 family_NN1 may_VM have_VHI a_AT1 computer_NN1 ,_, a_AT1 television_NN1 ,_, PSP_NP1 or_CC Wii_NP1 ._.
*TXT:   These are elementary school's students ' favorite.
%POS:   These_DD2 are_VBR elementary_JJ school_NN1 's_GE students_NN2 '_GE favorite_NN1 ._.
*TXT:   Students like using leisure hours on them after school.
%POS:   Students_NN2 like_II using_VVG leisure_NN1 hours_NNT2 on_II them_PPHO2 after_II school_NN1 ._.
*TXT:   Chatting with classmate, playing games and watching television are fun.
%POS:   Chatting_VVG with_IW classmate_NN1 ,_, playing_VVG games_NN2 and_CC watching_VVG television_NN1 are_VBR fun_JJ ._.
*TXT:   Students feel relax when they do these things.
%POS:   Students_NN2 feel_VVO relax_VVO when_RRQ they_PPHS2 do_VDO these_DD2 things_NN2 ._.
*TXT:   These are why many elementary school's students have nearsight.
%POS:   These_DD2 are_VBR why_RRQ many_DA2 elementary_JJ school_NN1 's_GE students_NN2 have_VHO nearsight_NN1 ._.
@End
```

In these files the metadata appear as the header of each file, followed by the writing sample. Every other line in the sample has the same content as the line above it but it also includes CLAWS part-of-speech tags.

Studies based on corpus data. A wide range of research topics are being conducted to aid corpus construction. These investigations will help to evaluate what kind of samples should be added to the corpus, the tags which will be added to the samples, and the format in which they will be presented. In what follows, each of these projects will be summarized.

With the project 'Automatic essay scoring: using Support Vector Machine', Dr Gao aims to identify which features of the GEPT writing samples have the greatest impact on scoring. A software program was created to locate the feature set in the writing samples which achieves the closest match with the examination paper scores. Results indicate that predicting pass or fail is more difficult than predicting the score of a writing sample through this approach and that the samples written on one specific topic were more difficult to score automatically than the samples on the other topics. In future research, this automatic scoring software program will be modified through the inclusion of additional features so that its performance within and across essay topics will be improved. Moreover, in order to better assess its efficacy, the program will be applied to writing samples of different proficiency levels.

In the project 'Corpus-based Lexical Semantic Research in SLA,' Dr Chung and her associates examine topic variation (Chung & Wu 2009), confusable near-synonyms (Hsieh & Chung 2009) and the misuse of the unaccusative verb 'happen' (Wang & Chung 2009).

More specifically, in testing the effect of topic variation on writing performance, Dr Chung and Ms Wu examine how a) topic variation in writing can be measured and b) a learner corpus can help in discovering the relationship between topic variation and writing performance. These research questions were inspired by the claims that topic familiarity (Kellogg 1994, 1999) and skill in memorizing and retrieving relevant linguistic and encyclopedic knowledge from memory (Hayes 1996) affect significantly the written performance of language learners. The data for this study are the writing samples that have been processed for the corpus. Data were analyzed to examine whether passage lengths, type and token counts, and/or vocabulary difficulty vary as a result of topic variation. The finding that vocabulary difficulty differs among topics supports the claim that topic variation can be measured using quantitative methodology. Since vocabulary difficulty is shown to significantly differ among topics, its relationship with score bands is also examined. The results of this analysis indicate that topics are in close relationship with the scoring results and capable writers usually have greater ability to write in various topics in comparison with poor writers.

The research project on confusable near-synonyms 'make' and 'do' indicated that learners tended to confuse the two by producing instances such as '*do choice' (probably as a result of *zuo4 xuan3ze2* from Mandarin) and '*make harm' (also a translation from *zao4cheng2 shang1hai4*). In this project, researchers compared the learner data with UKWac (cf. Bailey & Thompson 2006), a collection of web-based materials with more than two billion words, available through the Sketch Engine (Kilgarriff & Tugwell 2001). Considering the varied topics covered in a general corpus, the UKWac returned many instances of 'do+from' and 'made+from'. The study found that most of the uses of 'do+from' are followed by abstract nouns and those of 'made+from' by concrete nouns. However, since the three topics in the collected GEPT samples do not yield many of these instances, a large-scale comparison between the learner data and a general corpus was not possible. This could be a direction of comparison in future work.

In the study examining the unaccusative verb 'happen', it was found that among all the instances (23) of the bare form 'happen,' 52% were used incorrectly as in 'Why this situation

happen in Taiwan?'(1148.txt). On the other hand, among all instances (28) of the *-ed* form 'happened', 75% were used correctly by the learners as in 'How did this happened? Maybe it's happened because of the TVs and…' (0158.txt). The study discusses results in the context of Mandarin L1 transfer in second language acquisition.

The project 'Error tagging system' by Dr Skoufaki aims to a) devise an error tagging system which will lead to detailed error-search results, b) tag a large variety of errors in the corpus, and c) tag a more extensive number of discourse errors than those tagged in other learner corpora. A review of the error tagging systems in well-known learner corpora has led to the formation of a working error tagging system where error tags have a hierarchical structure and are separate, rather than complex, in order to conduct flexible error searches. In particular, the working error tagging system borrows its architecture from the FreeText error tagging system (Granger 2003), where the general structure of an error tag is '<linguistic level><linguistic sublevel>'. The content of the tags is a combination of FreeText and CLEC tags (Gui & Yang 2002) plus some additional ones which were added during the trial error tagging of 30 writing samples. Statistical analyses have revealed the most common errors and significant correlations between errors and grade bands. Trial-tagging more texts and measuring intra- and inter-tagger agreement will be the next step in the development of the error tagging system.

With respect to the tagging of discourse errors, coherence errors cannot be identified reliably (Mann & Thompson 1988). Therefore, a bottom-up method was used to identify coherence errors. Rhetorical Structure Theory (RST) (ibid.) was considered appropriate for various reasons (Skoufaki forthcoming). In RST, coherence relation tags link units of analysis and show their relevant discourse functions from the author's perspective. One of the constraints on the formation of these diagrams is that they and their constituent sub-diagrams should have the form of a schema, that is, the abstract representation of coherence relation diagrams (Mann & Thompson 1988). The hypothesis which supports the use of RST diagrams for coherence-error detection is that coherence errors will be indicated by diagrams which violate the aforementioned constraint.

RST diagrams of 45 writing samples equally distributed across topics and score bands were constructed with the RST Annotation Tool software [2] and 'abnormalities' in the diagrams were categorized according to the coherence errors they indicate. Results show that 'abnormalities' in RST diagrams can indeed indicate coherence errors. However, some coherence errors do not have any consequences for RST diagrams, so the tagger's intuition is necessary for their identification. Moreover, the tagger's intuition is also needed whenever an abnormality in the diagram points to more than one possible coherence errors.

The project 'The relationship between formulaic language and pauses' by Dr Skoufaki tests quantitatively Fillmore's claim that the ability to talk with a few pauses is largely due to the use of formulaic expressions (Fillmore 1979/2000), while applying it to L2 speech and refining it in view of relevant experimental findings. Previous research has indicated that fluent L2 learners make fewer intra-clausal pauses but not fewer inter-clausal pauses than non-fluent learners (Lennon 1990, Mizera 2006) and that they make briefer intra-clausal, but not inter-clausal pauses than non-fluent learners (Mizera 2006). The research hypothesis is that the frequency of fixed multi-word expressions and lexical collocations in Taiwanese low-intermediate ESL learners' speech samples will correlate negatively with the frequency and/or mean duration of intra-clausal, not inter-clausal, pauses. The data are answers by 30 subjects to three questions from the oral part of the corpus. The hypothesis is supported by experimental data. However, pauses may have not been consistently measured because of varying acoustic properties of the phonemes before and after the pauses. This issue and the

[2] This software was downloaded from http://www.isi.edu/licensed-sw/RSTTool/index.html.

psycholinguistic underpinnings of the significant correlations will be examined through further research.

In the project 'How Second-language Speakers Realize Pitch Accents in English Utterances', Ms Sally Chen examines a) the patterns through which second-language learners realize pitch accents in English utterances and b) whether different task formats give rise to different pattern varieties. This project is motivated by the dearth of research in suprasegmental features in L2 speech and by previous research by Mennen in L1 Dutch and Greek (Mennen 1998) and L2 Greek produced by Dutch learners (Mennen 2004, 2006). The first study indicates that Dutch has later pitch peaks while Greek has earlier ones. The last two studies indicate that even advanced Dutch learners of Greek produce later peaks in their L2 Greek due to L1 transfer. The data for this study are recordings of two passages as read aloud by a learner who had received a grade of three, the median grade of the test, on a five-point scale. A group of ten native English speakers were recruited to serve as controls. They were given the same test materials and their readings were recorded under a test scenario replicating that of the L2 learners. Preliminary results show that in general, the L2 learners address more tones in their production.

Finally, in her PhD research project with working title 'The manifestation of discourse cohesion and organization in L2 learners' spoken and written texts', Ms Wu examines differences in the use of cohesive devices and discourse organization between the spoken and written production of GEPT test-takers. With respect to discourse cohesion, past research has shown that cohesive devices are used in a more complex way in writing rather than in speech (Chafe 1985), that writing is characterized by higher lexical density (e.g., Stromqvist et al. 2002) and discourse components are connected through dependency ties (Redeker 1984). In terms of discourse organization, since it has been accepted that language of higher linguistic complexity is much more frequently found in written texts than in spoken ones, this project aims to explore exactly how different modalities give rise to different linguistic patterns. Ms Wu is currently focusing on causal conjunction in the writing samples of the corpus.

***Direction for future research*.** As mentioned above, in the near future the projects on automatic essay scoring, error tagging, formulaic expressions and pauses, pitch realization, and discourse cohesion will be continued and extended when necessary. The research project on lexical semantics will be extended through a new sub-project, the investigation of the uses of the near-synonyms 'watch', 'see', and 'look' in the writing samples of the corpus. Moreover, two new projects will be undertaken. Metaphoric expressions, similes and idioms in the writing samples will be examined by Dr Chung for their conventionality and relationship with specific topics and errors. The contribution of vocabulary diversity and difficulty to the essay scoring of the intermediate GEPT writing samples and whether it remains stable across topics will be examined by Dr Skoufaki and Dr Chung.

The aforementioned research projects have indicated not only avenues for future research but also ways in which the LTTC English Learner Corpus can be extended and utilized. Most projects have indicated a need for more writing data and some for writing data from learners at a higher level of English proficiency. Therefore, the corpus is currently being extended through the addition of 2,000 more writing samples from the Intermediate GEPT examinations and 1,000 such samples from the High-Intermediate examinations. In addition, the variety of the research projects conducted suggests that the corpus could be of use to the general public and to other researchers. Consequently, the research team plans to provide various levels of access through a web-based version of the corpus without compromising the database. Research is currently being conducted in this direction.

The Singapore Corpus of Research in Education (SCoRE) project

***Introduction*.** The SCoRE project is one of our endeavours of compiling a multimodal corpus database of education discourse of Singapore schools. Since 2003, funded by **the Ministry of Education, Singapore**, through **the Centre for Research in Pedagogy and Practice** (CRPP), **National Institute of Education** (NIE), **Nanyang Technological University** (NTU), Singapore, the project team has amassed a large collection of data on classroom interactions, teaching materials and students' assignments in Singapore primary and secondary schools from its various research projects. The aim of this project is to compile the data collected by several research panels of **CRPP** to a multimodal, multilevel annotated corpus database - **the Singapore Corpus of Research in Education** (SCoRE).

The main goal of the SCoRE project is to provide a large general purpose resource for education researchers. In specific, we attempt to achieve the following objectives in this corpus database project:

- to build a multimodal, multilevel annotated classroom discourse corpus database;
- to develop a sophisticated query engine for education researchers to do corpus-based studies on classroom discourse;
- to provide empirical modelling of classroom interaction patterns;
- to develop a bundle of automated discourse annotation and query tools;
- to inspire and facilitate corpus-based investigation of linguistic variation within or across class sessions and disciplines.

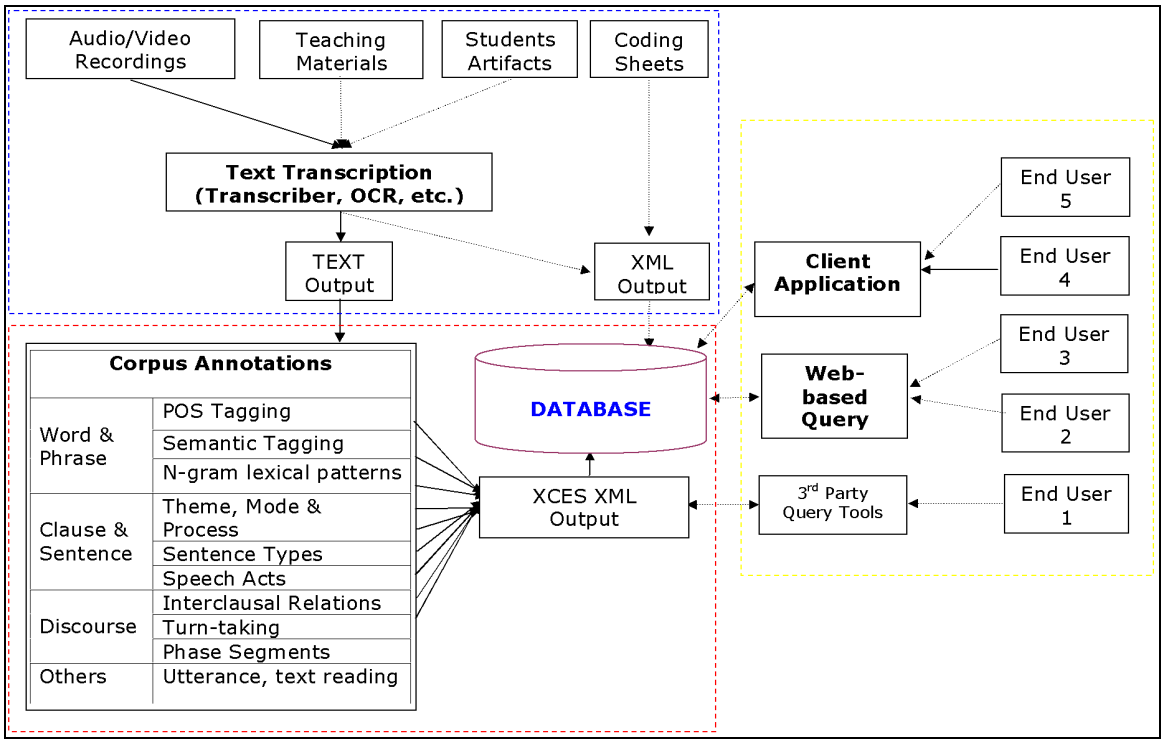
In addition to the collection of transcripts of classroom recordings, teaching materials and students artefacts, the project also attempts to annotate a number of corpus data at different linguistic and discourse levels. The proposed deliverables include a speech subcorpus, a lexical subcorpus, and several multilevel annotated subcorpora at different development stages. Eventually all these subcorpora will be indexed and incorporated into a large corpus database, which will be provided with sophisticated query tools for both online and offline queries.

It is held that building such a computerised corpus database of classroom discourse and attendant query tools can significantly benefit education researchers, linguists, teacher training practitioners, and curriculum designers with a corpus-based empirical approach and proof.

Components of the corpus and compilation of the corpus database. The corpus consists of teaching materials, classroom interactions, and students' artifacts. The current release data are mainly transcripts of classroom audio/video recordings taped in more than 350 primary and secondary schools in Singapore. Presently, there are 277 transcripts processed and annotated in the SCoRE corpus; another 248 recordings of lessons have been transcribed and now are under cleaning and annotation. Based on the statistics of the 277 lessons, on average there are 5000 words in one hour recording, thus the size of the classroom component of the SCoRE corpus will be about 5 million words.

Other than the classroom discourse, SCoRE project is also collecting the relevant teaching materials (e.g., text books, handouts, etc.) and students' artifacts in the corresponding classes. In this way, SCoRE corpus will eventually have approximately 50 million words in total. Basically, the construction of the corpus database of education discourse consists of three phases of tasks, which are presented briefly in the figure below.

Figure 3
The Construction of the Corpus Database of Education Discourse



- Data collection and Manipulation

The raw data are collected from various sources, which include audio/video recordings of classroom lessons, teaching materials (textbooks, PowerPoint slides, OHT transparency, etc.), students' artefacts and classroom coders' coding sheets. The audio/video recordings are transcribed by a team of well-trained, full-time and part-time transcribers. Cleaned up and prepared for annotation, these transcripts are converted to some target formats with the conversion tools that we have developed in house. To protect privacy, we also index personal and institutional names to anonymise them before the data is released to the public.

- Feature selection

As a general purpose education research corpus, it is ideal to annotate as many linguistic and paralinguistic features as possible, for different research goals. These features, ranging from morphosyntatic to discourse, will be the critical part of the corpus database compilation.

The design of the corpus has been negotiated between different interest groups in CRPP, regarding questions such as what annotation schemes to adopt, and what standards or guidelines to adhere to. However, the linguistic and paralinguistic features are interwoven from different perspectives of linguistic theories or schools. The features that we will be annotating are listed in the table below.

Table 1
Linguistic Features to Annotate

LEVELS	TAGGING	EXPLANATION	TOOLS USED
Token/Word/ Phrase	POS & Semantic Tagging	Automatic tagging, with manual post-check unknown words	Wmatrix (English) Autotags (Chinese)
	N-gram Lexical Patterns		
Clause/ Sentence	Theme, Mood & Process	SFG & Dialogue act	MMAX2
	Sentence Types, Nominalizations	SFG and Speech-Act Annotation	MMAX2
	Speech Acts		MMAX2
Discourse	Interclausal Relations	SFG	MMAX2
	Turn-taking, IRFs	Dialogue analysis	MMAX2
	Phase/episode TRS	Dialogue analysis	MMAX2
Others	media, localized language, teaching/learning strategies, code-mixing/switching, etc	Some done while transcribing	Transcriber/ MMAX2

- Annotation tools and methods

Many of those features from Table 1 above cannot be automatically annotated with handy computer tools, yet semi-automatic processing or editing tools are helpful. In regard to the features to annotate, we have surveyed the existing tools in the literature and selected a few suitable tools (Sim, Hong & Kazi, 2005). Basically, the third-party annotation tools we customized to process our data include CLAWS POS tagger, USAS semantic tagger, ICTCLAS, Ngram Statistics Package, MMAX tool, and SYSTEMICS.

- Database building

As the corpus consists speech and text data of multiple levels of annotations, a sophisticated database should be employed to facilitate any extensive queries of these heavily-annotated data. Thus, the corpus will be stored in a relational database, which is proven to be reliable and fast (Wittenburg et al, 2005; Davies, 2005). Users can then access the database by means of a web interface, client applications making use of our APIs or 3rd party query tools. A web-based query is expected to be the main method of query for users to access the database. Technologically savvy users could also access the corpus using web services interface to retrieve relevant data and process the data accordingly. The use of 3rd party tool serves to satisfy some researchers' interests in exploitation of the data offline.

- Web-based Query

A web-based query interface to SCoRE database is an ideal approach to make this corpus available to education researchers and other potential users, because with such an approach:

- the database resides on the server side, and remains transparent to the users;
- users do not need to download and install any client program, and just need to login into the designated webpage and query the database.

We will basically provide two types of web-based query for the users: a "lazy query" with a wide range of selectable fixed parameters; and an advanced query with wildcards or regular expressions.

Educational applications of the database. With the aforementioned deliverables, the potential educational applications of this project can be carried out from the following several perspectives:

- A multimodal classroom interaction corpus:
  Digital audio/video recordings of the classes make it possible to conduct in-depth classroom research on knowledge acquisition and pedagogy.

- A multilingual and multidiscipline corpus:
  The corpus sampling of SCoRE project covers several curriculum subjects (Language, Literature, Science, Social Studies, History, and Mathematics) mediated in four official languages (English, Chinese, Malay and Tamil) in Singaporean primary and secondary schools. In this way, it is helpful for comparison studies across languages and disciplines.

- Empirical model of classroom practice:
  There are many practices that are believed to be particularly useful in classrooms, but whose effectiveness has not been empirically demonstrated. Many of these can be investigated in the searchable database of classroom interactions.

- Resource for teacher training or professional development:
  With the abundance of sample classroom interactions collected and annotated, teachers and teacher trainees can gain new perspectives about task design and input that leads to effective teaching practices and better understanding of how learners may grasp one level of meaning rather than others.

- Critical classroom discourse analysis:
  With the availability of authentic materials collected from real classroom interactions, researchers interested in conversation analysis can get access to meticulously recorded, carefully transcribed and linguistically annotated samples of natural data.

Conclusion. The aims set in the SCoRE project are by all counts quite ambitious, yet the construction of such an education discourse database can bridge the gap in educational research, which has been criticised for lacking an evidence-based corpus of knowledge to justify the effects and effectiveness of teaching practices in classroom. We believe that our effort and experience can also benefit a larger research community from various perspectives.

HKFSC and HKEC at the Hong Kong Polytechnic University

***Introduction*.** Winnie Cheng from Hong Kong has been involved in corpus work since mid-1990's when the conversational corpus, i.e. the first sub-corpus of the Hong Kong Corpus of Spoken English (HKCSE) (the other three sub-corpora being academic, business and public), was compiled and analysed (Cheng & Warren 1999a). Since then, Cheng has been actively involved in the teaching of corpus linguistics in a number of subjects at undergraduate and postgraduate levels and the use of corpora in language teaching. Her corpus-related research ranges from compilation of specialised corpora and corpus analysis, particularly discourse intonation and phraseological patterns. Many of the corpus research findings have been and are being used in teaching a range of topics in the study of linguistics, namely Discourse Analysis, Conversation Analysis, Pragmatics, Lexical Studies, and Intercultural Communication, not to mention Corpus Linguistics. Moreover, corpora and

corpus evidence have been used in teaching, primarily for the writing of teaching materials (Cheng & Warren 2000)

Corpus compilation. In corpus compilation, apart from the Hong Kong Corpus of Spoken English (HKCSE), Cheng has recently been building professional-specific corpora which are open to public access on the website of the Research Centre for Professional Communication in English (RCPCE) (http://www.engl.polyu.edu.hk/RCPCE/). The corpora are Hong Kong Corpus of Policy Addresses, Hong Kong Financial Services Corpus (HKFSC) (about 7 million words), and Hong Kong Engineering Corpus (HKEC) (about 6 million words).

Studies based on corpus data. In corpus research, a large-scale project was to prosodically transcribe and analyse 50% (about 1 million words) of the HKCSE (Cheng, Greaves & Warren 2005). The most recent work is Cheng, Greaves & Warren (2008), which is the first book to apply David Brazil's (1997) Discourse Intonation systems (prominence, tone, key and termination) to the study of authentic, naturally-occurring spoken discourses (HKCSE). It describes the four Discourse Intonation systems in terms of how the system works and how they are manifested in the corpus, both across the sub-corpora and also across speakers in the corpus. The book is accompanied with a CD containing the prosodically transcribed corpus together with iConc which is the software designed and written by Chris Greaves specifically to interrogate HKCSE (prosodic). The book has raised and discussed issues that are of importance in Conversation Analysis, Corpus Linguistics, Discourse Analysis, Discourse Intonation, Pragmatics, and Intercultural Communication.

Other studies have analysed various discourse intonation systems and genres in the 1-million-word HKCSE (prosodic), including the intonation of declarative-mood questions (Cheng & Warren 2002), the intonation of Q&A sessions in public discourses (Cheng 2004b), in service encounters in a five-star hotel in Hong Kong (Cheng 2004a), the use of rise and rise-fall tones in the HKCSE (Cheng & Warren 2005), comparing the intonation of speakers in TV quiz shows in Britain and Hong Kong (Cheng & Warren 2006b), a study of pitch concord (Cheng 2008c), and the discourse intonation patterns of word associations in public discourses (Cheng & Warren 2008).

Some other interesting investigations into the HKCSE have compared a range of conversational behaviours and strategies between Hong Kong Chinese and English-speaking Westerners in intercultural conversations, including constructing and negotiating ideologies of race and ethnicity (Cheng 1999), the study of humour (Cheng 2003b), preference organisation, discourse topic development, simultaneous talk and discourse information structure (Cheng 2003a), a study of collocational and intonational patterns in public speeches (Cheng 2004c), discourse patterns (Cheng 2007b), co-constructing prejudiced talk (Cheng 2008a), and disagreement and indirectness (Cheng & Tsui forthcoming).

Based on observations of corpus linguistics in the 1980's, Sinclair (1991) expounds the theory of 'units of meaning' which states that "in all cases so far examined, each meaning can be associated with a distinct formal patterning … There is ultimately no distinction between form and meaning" (Sinclair 1991: 496). The meaning of language is primarily expressed by linguistic units called 'the lexical item' (Sinclair 1996) which is a unit in the lexical structure to be selected independently and which then selects lexical or grammatical patterns for its expression. Cheng (2006, 2009a) examined the extended meanings of lexical cohesion in a corpus of SARS (Severe Acute Respiratory Syndrome) spoken discourse. Major findings are that the item 'SARS' does not display any strong collocates and colligates with the definite article and prepositions. The semantic preference is typically to do with the impact of SARS on Hong Kong and those directly affected by it. The semantic preference of 'the impact of SARS' is found to fuse with the semantic prosody of a negative sense of

'embattled/besieged'. However, in the post-SARS texts, a different semantic prosody, 'positive assessment', is observed.

As mentioned above, corpora evidence has been used in teaching and the writing of teaching materials (Cheng & Warren 2000). For instance, research into the grammatical and pragmatic characteristics of spoken English has found that the ability to do inexplicitness, indirectness and vagueness is a key component in the repertoire of all competent discoursers, particularly in conversations (Cheng 2007c, Cheng & Warren 1999b, 2001c, 2003). Other studies have examined the pragmatic functions of actually (Cheng & Warren 2001a), tag questions (Cheng & Warren 2001b) and the interactional strategy of interruption (Cheng 2007a) in the conversational sub-corpus of the HKCSE.

Corpus data have also enabled identification of patterning that differs from traditional models of the English language. A number of speech act studies comparing English presented in ELT textbooks in Hong Kong and English used in natural communicative situations outside of the classroom have found that textbook accounts of language use are often decontextualised and lack an empirical basis. For example, in their studies of the speech acts of compliment and compliment responses (Cheng 2003a) and thanking (Cheng 2009c), disagreement (Cheng & Warren 2005), giving an opinion (Cheng & Warren 2006a), checking understandings (Cheng & Warren 2007), the authors suggest that English textbook writers need to incorporate a wider range of and more accurate forms into their materials in order to better reflect the realities of actual language use.

Corpora and corpus evidence have also been used in the design of language learning tasks and activities. Concordance-based activities, for example, are designed to familiarise learners with various types of investigations and to stimulate the development of appropriate learning strategies through practice. These language learning approaches concur with the contemporary task-based language learning approach (Willis and Willis 2007) which emphasises the development of tasks and activities to engage learners in using the language. These are adopted to exploit the pedagogic context by focusing on both the authenticity of the source text and "its authenticity by the learner, which arises out of the involvement of the learner with the material, via the task" (Mishan 2004: 219). Cheng (2008b), building on Greaves & Warren (2007), outlines the steps and specifics of the activities to raise learners' awareness of the prevalence and importance of phraseology, and to develop in learners the computational and analytical skills needed to conduct an initial study of the phraseological profile of a text. The steps are described as follows:

1. Learners work with two texts: Policy Addresses given by the Chief Executive of Hong Kong in October 2006 and October 2005.
2. Compile a list of the ten most frequent words in each text.
3. Compile a list of the twenty most frequent phrases in each text.
4. Monitor and record the frequency with which the most frequent words and phrases found in 2005 Policy Address occur in the 2006 Policy Address and vice versa.
5. Discuss the findings from the two texts.
6. Throughout the analysis of the two policy addresses, the differing lengths of the two Policy Addresses are noted, and so direct comparison of frequencies need to take this into account. (Cheng 2008b: 26)

One of the latest developments in corpus research is the analysis of potential phraseogical patterns specific to corpora, with the use of the highly innovative ConcGram software (Greaves 2009), which is able to fully and automatically uncover all of the phraseology in a text or a corpus of texts, irrespective of variation (Cheng et al. 2009). ConcGram can identify all the potential configurations of between two and five words in any corpus, based on a

window of any size, to include the associated words even if they occur in different positions relative to one another (i.e. positional variation) and even when one or more words occur in between the associated words (i.e. constituency variation) (Cheng, Greaves & Warren 2006). Analysis of concgrams have been described in recent studies (Cheng et al. 2009); Cheng (2009b) highlights the importance of introducing phraseology in EFL curricula, and how corpus tools may be implemented in CALL environments to help students gain knowledge of phraseological items.

English Learners' Utterance Data Collection and Compilation at Waseda University

Introduction. In this subsection we introduce VALIS[3], an English learners' spontaneous utterance data collection and compilation project headed by Yasunari Harada. The utterance data in this project have been collected from undergraduate students who are for the most part native speakers of Japanese, at School of Law, Waseda University. One important aspect of this project is that the process of data collection is embedded in general English classes for first-year students and focuses on relatively spontaneous utterances in face-to-face oral communication. By incorporating both video and audio data, the goal of the current project is to enrich the learner data, as currently spoken learner corpora are hard to find. Another important feature in this project is that the utterance data are linked with proficiency and learning history of students who uttered each segment. Moreover, as the data collection in this project is embedded in students' in-class activities aimed at improving performance at a specific English task, the data accumulated would eventually witness longitudinal improvements of students' performance and proficiency, while the data collection activities in turn help students integrate the four basic skills of listening, speaking, reading and writing into integrated communicative interaction. Thus it turns out to be one of the pedagogically effective means for helping those students learn to communicate in spoken and written English.

Background. Before contemplating on the process of data collection and compilation of this project, the principal researcher had employed the *Versant English Test*[4], or *PhonePass* at the time, for evaluating listening and speaking proficiency of students in his English classes, starting in the school year of 2000-2001. The current version of this test is a 12-minute automated spoken English test for non-native adult speakers delivered over the phone or via the Internet. The rest presents six different tasks, including reading printed sentences aloud, repeating sentences presented aurally, answering short questions, building sentences, retelling short stories, and answering open questions. Each test-taker is assigned a test paper with a unique Test Identification Number, and they call a local Test Delivery System (TDS), which will deliver and administer the test and collect the responses by the test-taker. The responses will then be sent via the Internet to a speech recognition and scoring system that will analyze each call and assign scores. The scores of test-takers are available immediately after the analysis via the web for both test administrators and test-takers with proper identification numbers. The test results are reliable, in the sense that scores do not easily improve.

Most Japanese students who have taken this test, however, reflect that this test is very difficult to answer, and their responses to the open question task tend to be missing (silent) or scanty, as they are not used to immediately responding to questions aurally presented in English. This is also due to the factor that the emphasis during their high school English classes is more on studying textbooks and taking written exams, also partly because Japanese

[3] VALIS is for {Vast | Versatile | Visual} {Accumulative | Autonomous | Acoustic} {Learner | Learning | Language} {Information | Interaction} {Storage | System}. See Harada *et. al.*(2007), Kawamura *et. al.*(2008) and Harada *et. al.*(2008) for details.

[4] See Bernstein & De Jong (2001) and Balogh *et. al.* (2005).

students are not used to squarely answering to direct questions from teachers or their peers in any language. In order to better prepare those students for the test, the principal researcher initiated the following oral response practices in his classes.

Oral response practices. In order to help students better prepared for taking this automated spoken English test, especially for the section of open questions which are not scored automatically, and hopefully help them make progress in their spoken proficiency of English, the principal researcher introduced oral response practices in his classes. In each practice, three students form a group, with one plays the role as the questioner, another the time-keeper, and the last one the respondent. For each class, ten questions pertaining to one particular topic are prepared in advance and printed on a business-card size piece of paper. The questioner picks up one of those ten question cards and reads the question aloud to the respondent twice. The respondent has ten seconds to think and formulate the answer and 45 seconds to speak whatever comes to their minds. The time-keeper prompts the respondent by saying "Start!" ten seconds after the question is read, and cueing "Stop!" 45 seconds later After the response is given, the questioner and the time-keeper each gives a score to the response based on a rubric provided to the students and writes the score on a peer-review sheet for the respondent. Then, the three students rotate their respective roles and go on to the next question. Usually, in a session of 90-minute class, 20 to 25 minutes are devoted to this activity, and students are assigned to write a 400-word essay on the topic, then peer-review the essays written a week earlier. After incorporating this activity in his class, the principal researcher was surprised by two simple facts of life he had not suspected earlier: (i) students like this practice very much and (ii) human relationships in his classes improve drastically through this activity as this helps students know more about each other.

Spoken utterance data collection. The English classes in which those data collection activities are conducted convene in computer cluster rooms, so that many of the students' activities are digitally recorded at various stages. For instance, students are expected to read a picture book, a chapter book, a graded reader, or a paperback novel of their choices, from several hundred volumes the principal researcher brings to the classroom for the session every week. Students are assigned to report on the title of the book, number of pages read, and time spent for reading the materials in an Excel file. Students either write an essay or revise the one written a week earlier. In addition, they have to submit an initial version written in class, a revised version completed as homework, and a final version revised after peer-review and peer-evaluation in class, all in Word files. Students also engage in small-group presentations on what they saw on the web news sites or what they discussed during the oral interaction practices. These are presented in PowerPoint files and are collected later on. All electronic files are easy to collect and to analyze, but students' face-to-face oral interactions are ephemeral and much harder to store and analyze. In the fiscal 2004-2005, the principal researcher designed and built a digital audio recording device for utterance data collection, supported by Waseda University Grant for Special Research Projects 2004A-033 and experimented with the system in the next fiscal year of 2005-2006 with Waseda University Grant for Special Research Projects 2005B-022. His research project was awarded Grant-in-Aid for Scientific Research (B) 18320093 from the Japanese Ministry of Education, Culture, Sports, Science and Technology during the three fiscal yeas of 2006-2009, which enabled use of sufficient video cameras with wireless microphones in class and computer equipment for annotating the collected audio data. Currently, the project is funded by Grant-in-Aid for Scientific Research (B) 21320109 from the Japan Society for the Promotion of Science during the five fiscal yeas of 2009-2014, to continue the data collection and compilation efforts for another five-year period.

Audio and video recording devices. The numbers of students in the English classes where the data collections take place are currently around 24 and at most 36, partly because of the

curriculum design and partly because of facility constraints, namely the number of personal computers per classroom. As the students are organized into groups of three (or less) during the oral interaction practices, a maximum of 12 tracks had to be recorded simultaneously. No substantial overlap is expected to occur between the end of the question and the beginning of the response so one microphone per group seemed to suffice in order to pick up the question read twice and the response spontaneously given. Some other practical considerations were also taken into account when making decisions regarding the audio-recording device, such as:

- recording quality: linear PCM with highest sampling rate / highest bit rate possible
- storage and post-processing of sound files to be handled on Windows machines
- equipment to be carried, deployed and used in any classroom

Eventually, we decided on the following basic configuration of our digital audio recorder:

- Alesis ADAT HD24 XR: 24-Track Hard Disk Recorder
- Alesis MultiMix 12R: 8ch microphone fader (amplifier/mixer)
- Sony ECM-360: electret-condenser microphone
- microphone cables
- a portable container on wheels for the equipment

Along with this audio recording device operated by the teacher, the current data collection utilizes 13 sets of one Sony DCR-SR100, a video camera with 30GB internal hard-drive, plus one Sony HCM-HW1, wireless Bluetooth microphone, to be used by each group of students[5] (with one backup set). The internal hard drives of those video cameras are recognized as external drives when connected to Windows machines via USB 2.0 cables and the time stamps of those files keep track of when the segment was recorded.[6] As long as all the video cameras' internal clocks are synchronized, it is easy to tell when a given file was shot from the time stamp of the file. In addition, the video cameras come with 5.1 channel surround sound recording systems, with the center channel assigned to the Bluetooth wireless microphone when attached. With a reasonably decent stereo playback system, you can tell which direction a given voice is coming from, which may be of help in identifying the speaker of a given piece of utterance.

Annotation tools and computational environments. The utterances collected in this project are either mainly in English or mainly in Japanese, although we also find portions where the languages switch as students encounter difficulties. For the part of audio data where the language is supposed to be in English, a given segment may represent either a question prepared and printed by the principal researcher and read aloud by a student during oral interaction practices, followed by a relatively spontaneous and unprepared response by another student to whom the questions is just read. On the other hand, the recorded audio and video data also cover students' intra- and inter-group interactions presented in their native Japanese language. The intra-group interactions often are reflected from students' coordination or confirmation of roles during the oral interaction practices, support and help when a student encounters some difficulty in reading or answering a question, or their chatting that could be related or unrelated to the topics of questions in practice. Inter-group interactions may be reflected in the instructions given by teacher and/or TA, questions from students to the teacher and/or TA, and interactions among students in groups in their native language.

[5] The time-keeper is now also in charge of videotaping using the camera provided.

[6] Unfortunately, the file system of Alesis ADAT HD24 XR is not compatible with Windows and the time stamps on Windows files created as files are transferred from the recording device to a Windows hard-drive reflects when the files are converted rather than when the audio recording originally took place, so that we need to manually keep track of when a given track was recorded, etc.

As the workforce for post-recording processing are, for the most part, only available from undergraduate and graduate students in social science and humanities departments, also with limited or no experience in annotation tools or Unix operating system, it was necessary to search for annotation tools that run on Windows machines. After some experimentation with other tools, we tentatively decided to employ TableTrans, which is based on WaveSurfer, with some additional functionality, available from LDC AGTK. [7] As we tried to expand the number of transcription annotators, we encountered several serious problems with TableTrans. One problem is that it is not designed for graphic user interface, which almost all students in the social science and humanities department grew up with. Documentation and other information for this system are difficult to locate from the web, and compatibility with Japanese character encoding and fonts does not work well. We were unable to figure out how to "undo" some simple operations and an error on keyboard operation may do unrecoverable damages to the annotation. For those reasons, a new candidate for transcription annotation will be required in dealing with:

- how to install and use TableTrans
- how to handle the original audio data file and the output annotation file
- what character encoding to use
- how to consult annotation guidelines, and other principles and rules online
- how to keep work logs

For those reasons, we immediately found that the use of TableTrans for direct transcription annotation was not a scalable approach in increasing the number of new annotators.

Browser-interface for query and annotation. Currently, we are using a web-browser interface for searching for a particular audio segment or response, with specific information such as the question being answered, the date of the utterance, the identification tag and/or the proficiency levels of the student who produced the response, etc. The same web-browser interface is used to assign transcription annotation work to student annotators, which enabled us to hire new annotators and ask them to start working on their assignments without waiting for them to get used to a completely unfamiliar software tool such as TableTrans. The annotators can just click on icons for the sound file and use audio-playback software to listen to the segment and type in the transcription annotation into the web pages. A master annotator has to work with TableTrans to segment the recorded tracks into meaningful units, but the scalability of annotation work has greatly improved with the new system.

Annotators and administrators can search for particular sound segments for efficient work flow. The following figure 4.1 and 4.2 are screenshots of the browser-interface tool we use currently.

[7] Many researchers in conversation analysis in Japan use MultiTrans, which is also based on WaveSurfer with some additional functionality. This is mainly because it fits with their purposes but partly because there is a patch for the Japanese language encoding easily available, thus there are enough documentation in Japanese readily available on the web.

Figure 4.1
Data Listing for Annotators

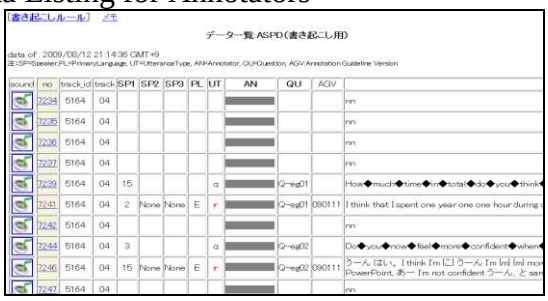


Figure 4.2
Data Listing for Administrator

While annotators can only see listings of the audio segments assigned to them and cannot see actual student names or listen to the segments where the speakers are identifying themselves, administrators can see a listing of all audio segments. For administrators, the same utterance segment may be transcribed by multiple annotators, so that one record in Figure 4.2 represents one transcription rather than one audio segment.

Figure 5.1
Annotation Data Input

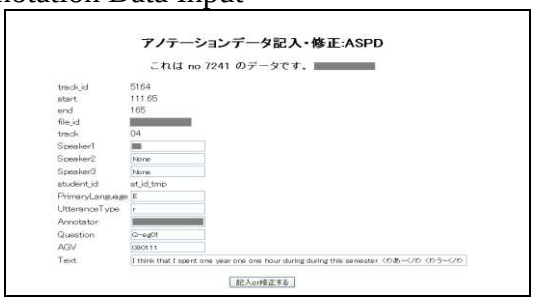
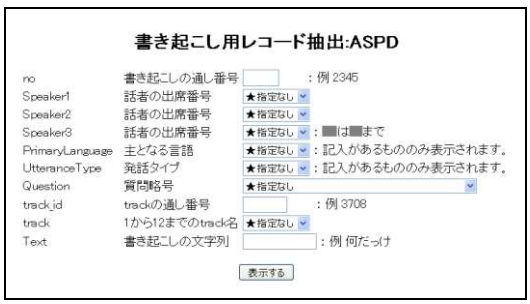


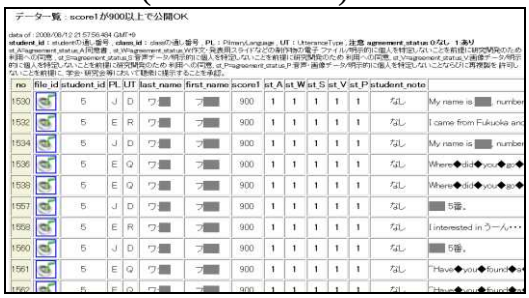
Figure 5.2
Data Query for Annotators



Annotators can click on the icons for the sound file they are going to work on and type in their transcriptions in the text field in Figure 5.1 and submit the data. They can also search from the audio segments assigned.

Figure 6.1
Administrator Query Page

Figure 6.2
Search Result (not actual data)



Principal researchers can perform search using tags such as the speaker's class, identification number (for the data collection purpose), question, date of the data collection, and other profiles of the student when a separate database containing student profile information is linked for this query purpose. Figure 6.1 is a sample query of this kind and Figure 6.2 is a sample listing to a query, in which responses by students whose TOEIC score is equal or greater than 900 are listed (which is not a real data).

Concluding remarks. One important additional aspect of the current data collection project is that it aims to digitally archive longitudinal improvements in students' English proficiency, including their oral fluency in face-to-face interactions. Along with the audio recording and data analysis mentioned above, we collect video recordings in the practice sessions, as audio recordings do not fully capture what is happening in face-to-face communication or possible changes in the learner, including features like eye contact and gaze, etc. Although it is still difficult to annotate students' attitudes and facial expressions, what teachers of English needs to address in their language teaching is not only knowledge, skills, and proficiency of the target language in their students, but also students' attitudes and feelings toward the target language. Another advantage of incorporating video data is images are much easier for the teachers to browse and scan than audio data alone.

In order to observe longitudinal improvements in students' activities, the data collection of the present project is not limited to the audio recording of oral response practices. Other records of students' performance, such as their spontaneous utterances, questions, responses and formal or informal in-class presentations are included as well. In addition, for students' profiles, their TOEIC and Versant English Test scores, also their personal histories of learning English are all being recorded. Finally, additional written works from students, like

essays (in Word files), slides (in PowerPoint files), and extensive readings (in Excel files) are all been kept as part of the database.

As yet another finding after the fact, through the data collection embedded in classroom activities the principal researcher learned two additional simple facts of life: (i) college students today do not mind cameras and microphones, and in fact love to be videotaped, and (ii) those intrusive data collection devices function as scaffolding for them to learn to speak in a foreign language, which happens to be English in this case. Students show changes in their attitudes toward learning and communicating in English for the better through those experiences.

Best Practice Recommendations

Based on the experiences of building four different types of learner corpora, we can extract some useful recommendations for L2 corpus construction. Before contemplating on the design of a proposed new learner corpus, it is important to define the goals and objectives of the project and decide the source of the data and how the data will be collected. While a new project may often be motivated by new ideas and objectives, it is also important for the corpus designer to study existing corpus, available recording devices, annotation tools and their documentations, database management systems and their interface designs, and tools for transcription and other annotation, with a view to the nature and expertise of the workforce that would be readily available for the project. These factors all crucially constrain what can actually be performed in real life situations and would serve as guidance to the construction of the new corpus. Published books and journal papers are reliable source of information, but these tend to report only successful attempts and new findings. On the other hand, through personal communication with researchers in the field, we may learn as much or even more from carefully planned but failed attempts and envisioned projects that did not materialize for financial reasons. Audio quality of speech recordings is quite important in case of spoken data collection while ease of operation of recording devices and file management systems dictate the number of data streams that can be collected at one time in a given location. At the moment, multi-track digital recording devices are getting cheaper and Bluetooth wireless microphones are becoming more readily available.

Another practical recommendation for researchers who are interested in building learner corpus is that they should be prepared to secure enough funding for various aspects of the contemplated project at various stages. While researchers should take into consideration details of equipments required for data collection, storage and transcription, and other annotation work, it is similarly important for them to consider funding for dedicated and temporary human resources needed for data collection, data management, as well as transcription and other annotation efforts. One should also consider multi-regional cooperation in terms of different funding restrictions and constraints: different funding agencies may enforce different policies for research funding so that it may be easier to procure equipment with budget from one source while human resources may be more easily supplied from some other source. For a better result of longitudinal research, researchers should also pay attention to how they could secure dedicated facility for recording. Finally, researchers should attend to securing subjects for data collection and clear rights to carry out human research.

Lastly, we should be aware of various real-life trade-off relationships among various factors that are involved in data collection and data sharing in case of learner corpora. In an optimal work flow of data collection conducted in an ideal world, data solicitation procedures should be standardized and human intervention should be minimized, while audio recordings

should take place in a sound-proof or at least sound-insulated recording rooms. On the other hand, if researchers are interested in student-student interactions during classroom activities, peer prompts and solicitation should be incorporated and recording would take place in noisy classroom environments. Quality of audio recordings and signal-to-noise ratio may be compromised, but there is no way the research can conduct this kind of data collection in a recording room. Personal computer interface can standardize the solicitation procedure but students would have to learn to speak to computers rather than to humans. Ideally speaking, learner data should be made publicly available with rich and comparable annotation on English proficiency levels of the speakers, learner history, and transcription by different kinds of annotators through web browser interface. However, the more information one reveals for one piece of audio segment, there is a higher possibility of identifying its speaker. Also, for researchers, audio and video data are precious source of research but students would not necessarily agree to make them publicly available for anyone. Thus, a careful non-disclosure agreement scheme and a database management system that can control restricted access to the data are indispensable.

Conclusion

In this paper, we introduced 4 different learner corpora from Hong Kong, Japan, Singapore and Taiwan. Based on discussion on these projects, we also made some best practice recommendations for future researchers and teacher to follow. Our survey paper captures the state-of-the-art developments of this exciting new direction in a very vibrant research region. Similar research has already been undertaken by colleagues in other Asian countries. In addition, innovations in corpus-based language technology, as well as the popular emergency of social network sites and tools on the web have already introduced new possibilities in the construction of learner corpus. The topic of the impact of emerging technology and new media should be a fertile ground for future research.

Acknowledgment

We would like to thank Professor Tien-En Ko, Jessica Wu, and Rachel Wu for their meticulous and tireless organization and generous support that made the panel possible. We would also like to thank Professor Yukio Tono and Dr. Henriëtte Hendriks for contributing as commentators at the panel. Although they are not co-authors of this paper, their comments played an important role in shaping ideas in this survey paper. Any remaining errors are, of course, ours.

Selected Bibliography for Learner Corpus

Bailey, S., & Thompson, D. (2006). UKWAC: Building the UK's First Public Web Archive. *D-Lib Magazine*, 12 (1). Downloaded on 13 January 2009 from http://www.dlib.org/dlib/january06/thompson/01thompson.html

Balogh, J., Barbier, I., Bernstein, J., Suzuki, M., & Harada, Y. (2005). A Common Framework for Developing Automated Spoken Language Tests in Multiple Languages. *JART Journal*, 1(1), 67-79.

Bernstein, J. & De Jong, J. H.A.L. (2001). An Experiment in Predicting Proficiency within the Common Europe Framework Level Descriptors. In Y.N. Leung et al. (Eds.), S*elected Papers from the Tenth International Symposium on English Teaching* (pp. 8-14). Taipei, ROC: The Crane Publishing.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. (ETS TOEFL Monograph Services MS-25). Princeton, NJ: Educational Testing Service.

Bird, S. (2001). *Linguistic annotation*. Available on-line from http://www.ldc.upenn.edu/annotation/

Brazil, D. (1997). *The communicative role of Intonation in English*. Cambridge: Cambridge University Press.

Cazden, C. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. R. Olson, N. Torrance, and A. Hildyard (Eds.), *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing* (pp. 105-122). Cambridge: Cambridge University Press.

Cheng, W. (1999). Constructing and negotiating ideologies of race and ethnicity in intercultural conversations. *Anglistica*. *English and the Other*, 3(1), 15-31.

Cheng, W. (2003a). *Intercultural conversation*. Amsterdam/Philadelphia: John Benjamins.

Cheng, W. (2003b). Humour in intercultural conversation. *Semiotica*, 146(1/4), 287-306.

Cheng, W. (2004a). // → did you TOOK // ↗ from the miniBAR //: What is the practical relevance of a corpus-driven language study to practitioners in Hong Kong's hotel industry? In U. Connor and T. A. Upton (Eds.), *Discourse in the professions* (pp. 141-166). Amsterdam: John Benjamins.

Cheng, W. (2004b). 'well thank you David for that question': The intonation, pragmatics and structure of Q&A sessions in public discourses. *The Journal of Asia TEFL*, 1(2), 109-133.

Cheng, W. (2004c). // → FRIENDS // ↘↗ LAdies and GENtlemen //: Some preliminary findings from a corpus of spoken public discourses in Hong Kong. In U. Connor and T.A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 35-50). Amsterdam/New York: Rodopi.

Cheng, W. (2006). Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse. In J. Flowerdew and M. Mahlberg (Eds.), *Special Issue of International Journal of Corpus Linguistics: Corpus Linguistics and Lexical Cohesion*, 11(3), 325-344.

Cheng, W. (2007a). 'Sorry to interrupt, but …': Pedagogical implications of a spoken corpus. In Campoy, M. C. and Luzon, M.J. (Eds.), *Spoken corpora in applied linguistics* (pp. 199-216). Bern: Peter Lang.

Cheng, W. (2007b). Discourse patterns in intercultural conversations. In B. Kraft and R. Geluykens (Eds.), *Cross cultural pragmatics and interlanguage English* (pp. 221-242). Muenchen: Lincom Europa.

Cheng, W. (2007c). The use of vague language across spoken genres in an intercultural corpus. In J. Cutting (Ed.), *Vague language explored* (pp. 161-181). London: Palgrave Macmillan.

Cheng, W. (2008a). Co-constructing prejudiced talk: Ethnic stereotyping in intercultural communication between Hong Kong Chinese and English-speaking westerners. In A. M. Y. Lin (Ed.), *Problematizing identity: Everyday struggles in language, culture and education* (pp. 171-191). New York: Lawrence Erlbaum.

Cheng, W. (2008b). Concgramming: A corpus-driven approach to learning the phraseology of discipline-specific texts. *CORELL: Computer Resources for Language Learning*, 1, 22-35.

Cheng, W. (2008c). A corpus study of pitch concord. In a cura di M. Pettorino, A. Giannini, M. Vallone, and R. Savy (Ed.) *La comunicazione parlata* (*Spoken Communication*) (pp. 58-73). Naples: Liguori Editore.

Cheng, W. (2009a). Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse. In J. Flowerdew and M. Mahlberg (Eds.), *lexical cohesion and corpus linguistics* (pp. 65-83). Amsterdam/Philadelphia: John Benjamins.

Cheng, W. (2009b). *Income/interest/net*: Using internal criteria to determine the aboutness of a text. In Aijmer, K. (Ed.), *Corpora and language teaching* (pp. 157-177). Amsterdam/Philadelphia: John Benjamins.

Cheng, W. (2009c). 'thank you': How do conversationalists in Hong Kong express gratitude? In B. Fraser and K. Turner (Eds.), *Language in life, and a life in language: Jacob Mey – A Frestshrift* (pp. 39-48).UK: Emerald.

Cheng, W., & Tsui, A. (forthcoming). 'ahh ((laugh)) well there is no comparison between the two I think': How do Hong Kong Chinese and native speakers of English disagree with each other. *Journal of Pragmatics*. (12-MAY-2009, 10.1016/j.pragma.2009.04.003, http://dx.doi.org/10.1016/j.pragma.2009.04.003)

Cheng, W., & Warren, M. (1999a). Facilitating a description of intercultural conversations: The Hong Kong Corpus of Conversational English. *International Computer Archive of Modern English (ICAME) Journal*, 20, 5-20.

Cheng, W., & Warren, M. (1999b). Inexplicitness: What is it and should we be teaching it? *Applied Linguistics*, 20(3), 293-315.

Cheng, W., & Warren, M. (2000). The Hong Kong Corpus of Spoken English: Language learning through language description. In L. Burnard and T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 133-144). Frankfurt am Main: Peter Lang.

Cheng, W., & Warren, M. (2001a). The functions of *actually* in a corpus of intercultural conversations. *International Journal of Corpus Linguistics*, 6(2), 257-280.

Cheng, W., & Warren, M. (2001b). "She knows more about Hong Kong than you do isn't it": Tags in Hong Kong conversational English. *Journal of Pragmatics*, 33(9), 1419-1439.

Cheng, W., & Warren, M. (2001c). The use of vague language in intercultural conversations in Hong Kong. *English World-Wide*, 22(1), 81-104.

Cheng, W., & Warren, M. (2002). The intonation of declarative-mood questions in a corpus of Hong Kong English: // ↘↗ beef ball // → you like //. *Teanga: Journal of the Irish Association of Applied Linguistics. Special Issue of Corpora, Varieties and the Language Classroom*, 21, 151-165.

Cheng, W., & Warren, M. (2003). Indirectness, inexplicitness and vagueness made clearer. *Pragmatics*, 13(3/4), 381-400.

Cheng, W., & Warren, M. (2005). // ↗ CAN i help you //: The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English. *International Journal of Corpus Linguistics*, 10(1), 85-107.

Cheng, W., & Warren, M. (2005). // → well I have a DIFferent // ↘ THINking you know //: A corpus-driven study of disagreement in Hong Kong business discourse. In F.

Bargiela-Chiappini, F. and M. Gotti (Eds.), *Asian business discourse(s)* (pp. 241-270). Frankfurt am main: Peter Lang.

Cheng, W., & Warren, M. (2006a). I would say be very careful of…: opine markers in an intercultural Business corpus of spoken English. J. Bamford and M. Bondi (Eds.), *Managing interaction in professional discourse. Intercultural and interdiscoursal perspectives* (pp. 46-58). Rome: Officina Edizioni.

Cheng, W., & Warren, M. (2006b). "// → you need to be RUTHless //: Entertaining cross-cultural differences. *Language & Intercultural Communication*, 6(1), 1-22.

Cheng, W., & Warren, M. (2007). Checking understandings in an intercultural corpus of spoken English. In O'Keefe and S. Walsh (Eds.), *Corpus-based studies of language awareness. Special issue of language awareness*, 16(3), 190-207.

Cheng, W., & Warren, M. (2008). // → ONE country two SYStems //: The discourse intonation patterns of word associations. In A. Ädel and R. Reppen (eds.) *Corpora and discourse: The challenges of different settings* (pp. 135-153). Amsterdam: John Benjamins.

Cheng, W., Greaves, C., & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *International Computer Archive of Modern English (ICAME) Journal*, 29, 47-68.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, 11(4), 411-433.

Cheng, W., Greaves, C., & Warren, M. (2008). *A corpus-driven study of discourse intonation*. Amsterdam/Philadelphia: John Benjamins.

Cheng, W., Greaves, C., Sinclair, J. McH, & Warren, M. (2009). Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics*, 30(2), 236-252.

Cheng, W., Warren, M., & Xu X. F. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. *System*, 31(2), 173-186.

Christie, F. (2002). *Classroom discourse analysis: A functional perspective*. London: Continuum International Publishing Group Ltd.

Chung, S.-F. & Wu, C.-Y. (2009, March). *Effects of topic familiarity on writing performance: A Study based on GEPT Intermediate Test Materials*. Presented at The 2009 LTTC International Conference on English Language Teaching and Testing.

National Taiwan University, Taiwan. March 6-7. [Full paper appeared in conference CD Rom].

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2): 213-238.

Cullen, R. (2001). The use of lesson transcripts for developing teachers' classroom language. *System*, 29, 27-43.

Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, Vol 10(3), 307-334.

Diederich, J., Pedersen, M., & Pedersen C. (2005). NIEtro: Shallow topic detection and tracking for the analysis of classroom interactions. *Proceedings of the Conference of Redesigning Pedagogy: Research, Policy, Practice*. Singapore, May 30 - June 1, 2005.

Edwards, A. D., & D. P. G. Westgate. (1994). *Investigating classroom talk* (2nd Ed). Lewes: Falmer.

Fillmore, C.J. (1979). On fluency. In D.Kempler and W.S.Y. Wang (Eds.) *Individual differences in language ability and language behavior* (pp. 85–102). New York: Academic Press. Reprinted in H. Riggenbach (Ed.) (2000). *Perspectives on fluency* (pp. 43-60). Ann Arbor: University of Michigan Press.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102-121). Longman, London.

Gordon, E., Campbell, L., Hay, J., Maclagan M., Sudbury, A., & Trudgill, P. (2004). *New Zealand English: Its origins and evolution*. Cambridge: Cambridge University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20(3), 465-480.

Greaves, C. (2009). *ConcGram 1.0. A phraseological search engine.* Amsterdam: Benjamins.

Greaves, C. and Warren, M. (2007). Concgramming: A Computer-driven Approach to phraseology of English. *ReCALL Journal*, 17(3), 287-306.

Gui, S.C. & Yang, H.Z. (2002). *Chinese learner English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Halliday, M.A.K. & Christian M.I.M. Matthiessen. (2004). *An introduction to functional grammar* (3rd revised edition). London: Hodder Arnold.

Han, X., & Cong, T. H. (2003). Teachers' language use in Hong Kong Mandarin classroom. In Cong, T. H. (ed.), *New perspectives on teaching Chinese.* Beijing: Beijing University Press.

Harada, Y., Maebo, K., Kawamura, M., Suzuki M., Suzuki, Y., Kusumoto, N., & Maeno, J. (2008). Toward Construction of a Corpus of English Learners' Utterances Annotated with Speaker Proficiency Profiles: Data Collection and Sample Annotation. In Tokunaga, T. and Ortega, A. (Eds.): *LKP 2008, Lecture Notes in Artificial Intelligence (LNAI) 4938*, (pp. 171-178). Springer-Verlag Berlin Heidelberg.

Harada, Y., Maebo, K., & Kawamura. (2007). VALIS 2.0: Transcription of What was (not) Uttered. *IPSJ SIG Technical Reports* 2007-CE-90 (1), Vol. 2007, No. 69, (pp. 1-8). Information Processing Society of Japan.

Harada, Y., Maebo, K., Kawamura, M., Maeno, J., Kusumoto, N., Suzuki, Y., & Suzuki M. (2007). VALIS: {Vast | Versatile} {Accumulative | Autonomous | Academic} {Learner | Learning | Language} {Information | Interaction} {Storage | System}. *IPSJ SIG Technical Reports* 2007-CE-88 (24), Vol. 2007, No. 12, (pp. 169-176). Information Processing Society of Japan.

Hargreaves, D. (1996). *Teaching as a research based profession: Possibilities and prospects*. London: Teacher Training Agency.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1–27). Mahwah, NJ: Erlbaum.

Hellwig, B. & Van Uytvanck, D., & Hulsbosch, M. (2008). *EUDICO Linguistic Annotator (ELAN) version 3.6 manual.* Downloaded on 13 January 2009 from http://www.lat-mpi.eu/tools/elan/

Hong, H. (2005). SCoRE: A multimodal corpus database of education discourse in Singapore schools. *Proceedings of the Corpus Linguistics Conference* Series Vol. 1, No. 1 (ISSN 1747-9398). Birmingham, UK, July 14-17, 2005.

Hsieh, Y.-C. & Chung, S.-F. (2009). *'Do' and 'Make': A corpus-based study*. To be presented at the American Association for Corpus Linguistics Conference. University of Alberta**,** Canada. October 8-11 2009.

Huddlestone, R. & Pullum, G. (2002). *The Cambridge grammar of the English language*, Cambridge: Cambridge University Press.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning, *English Language Research Journal*, 4, 27-45.

Kawamura, M., Harada, Y., Maebo, K., Kusumoto, & Maeno, J. (2008). VALIS: Distributed Network Support for Annotation and Constrained Sharing of Utterance Data Obtained

from Japanese College Learners of English. *IPSJ SIG Technical Reports* 2008-CE-93 (22), Vol. 2008, No. 13, (pp. 155-162). Information Processing Society of Japan.

Kellogg, R. T. (1994). *The psychology of writing*. New York: Oxford University Press.

Kellogg, R. T. (1999). Components of working memory in text production. In G. Rijlaarsdam & E. Esperet (Series Eds.) & M. Torrance & G. C. Jeffery (Vol. Eds.), *Studies in writing: Vol. 3. The cognitive demands of writing: Processing capacity and working memory in text production* (pp. 43–61). Amsterdam: Amsterdam University Press.

Kilgarriff, A. & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of the ACL Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse, 32-38.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.

MacWhinney, B. (2008). *The CHILDES project. Tools for analyzing talk – Electronic version. Part 1: The CHAT transcription format.* Downloaded on 13 January 2009 from http://childes.psy.cmu.edu/manuals/chat.pdf

Mann, W. C. & Thompson, S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), 243-281.

Markee, N. (2000). *Conversation analysis*. London: Lawrence Earlbaum Associates.

Mennen, I. (1998). Second language acquisition of intonation: the case of peak alignment. In M.C. Gruber, D. Higgins, K. Olson & T. Wysocki (Eds.), *Chicago Linguistic Society 34, vol. II: The panels*, 41 (1), 327–341.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics*, 32, 543-563.

Mennen, I. (2006). Phonetic and phonological influences in non-native intonation: an overview for language teachers. *QMUC Speech Science Research Centre Working Paper*, 9, 1-18.

Mishan, F. (2004). Authenticating corpora for language learning: A problem and its resolution, *ELT Journal*, 58(3), 219-227.

Mitchell, R. (2000). Applied linguistics and evidence-based classroom practice: The case of foreign language grammar pedagogy. *Applied Linguistics*, 21(3), 281-303.

Mitchell, R., and Lee, J H W. (2003). Sameness and difference in dlassroom learning cultures: Interpretations of communicative pedagogy in the UK and Korea. *Language Teaching Research*, 7(1), 35-63.

Mizera, G.J. (2006). *Working memory and L2 oral fluency*. Unpublished PhD dissertation. University of Pittsburg, PA.

Müller, C. & Strube, M. (2003). Multi-level annotation in MMAX. In P*roceedings of the 4th SIGdial Workshop on Discourse and Dialogue,* Sapporo, Japan, 4-5 July 2003, 98-107.

Müller, C. & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (Eds.), *Corpus technology and language pedagogy: New ressources, new tools, new methods*. Frankfurt: Peter Lang, pp.197-214. (English Corpus Linguistics, Vol.3).

Rayson, P. (2001). Wmatrix: a web-based corpus processing environment. Software demonstration presented at *ICAME 2001* conference, Université catholique de Louvain, Belgium. May 16-20, 2001.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Rayson, P. (2005). *Wmatrix: a web-based corpus processing environmen*t, Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/**.**

Redeker, G. (1984). On differences between Spoken and Written Language. *Discourse Processes,* 7, 43-55.

Sim, T. J., Hong, H., & Kazi, S. A. (2005). A survey of the transcription, annotation, and query tools for the development of a classroom discourse corpus. In *Proceedings of the Conference of Redesigning Pedagogy: Research, Policy, Practice*. Singapore, May 30 - June 1, 2005.

Sinclair, J. McH. (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.

Skoufaki, S. (forthcoming). Devising a discourse error tagging system for an English learner corpus. In Wible, D. and Li, M.-Y. (Eds.), *Second language reading and writing: investigations into Chinese and English.* Taoyuan, Taiwan: National Central University Press.

Stromqvist, S., Johansson, V., Kritz, S., Ragnarsdottis, H., Aisenman, R., & Ravid, D. (2002). Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language and Literacy,* 5(1), 45-68.

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer, P. Rayson , A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16* (pp. 800-809). UCREL, Lancaster University.

Trappes-Lomax, H., & Ferguson, G. (2002). *Language in language teacher education*. Amsterdam and Philadelphia: John Benjamins.

Wang, L.-C. & Chung, S.-F. (2009). 'What is Happened?': A Corpus-based Analysis of L2 English. To appear in The *Proceedings of the 2009 International Asian Conference on Education* (ACE). Osaka, Japan. October 24-25 2009.

Willis, D. and Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

Wittenburg, P., Broeder, D., Piepenbrock, R., & Veer, K. (2005). *Databases for linguistic purposes: a case study of being always too early and too late*. Available on-line from http://emeld.org/workshop/2004/Wittenburg-paper.html.