Building a Singapore Learner Corpus of English Writing for Pedagogy

Zhang Ruihua, Guo Libo and Hong Huaqing

Abstract

This paper documents the development of a Singapore learner corpus of English writing for pedagogy, which has been constructed at Nanyang Technological University, Singapore. This corpus comprises sample English artefacts produced by students at 3 levels, i.e. Primary 6 (Year 6), Secondary 4 (Year 10) and Junior College 2 (Year 12). It is built to capture and compare learners' developmental features in terms of vocabulary, grammar and discoursal devices at different learning stages and therefore theorize on the nature of English writing development of learners in Singapore. The texts are tagged with meta information of learners' school level, gender, ethnic group and grade. Issues of corpus design, e.g. representativeness in sampling, are also addressed. Finally, pedagogical implications and potential applications of the project are presented.

Keywords: Corpus development, sample artefact, text type, POS tagging, semantic tagging, discoursal annotation.

Introduction

The implementation of an English-dominant bilingual policy in Singapore over the past decades has contributed to the establishment of the English language as lingua franca in the country. Singapore English as a localized variety has triggered some corpus construction research and four major Singapore English corpora have been constructed, i.e. the Singapore Component of the International Corpus of English (shortened as ICE-SIN), the NIE Corpus of Spoken Singapore English (Deterding & Low, 2001), the Grammar of Spoken Singapore English Corpus (Lim, 2004) and the Singapore Corpus of Research in Education (SCoRE) (Hong, 2005), a corpus of classroom discourse collected from primary and secondary schools in Singapore. However, none of the above-mentioned corpora is learner corpus and all of them comprise spoken data except the written component of the ICE-SIN. In other words, so far there has not yet been a corpus focusing on learner English writing in Singapore. For learner corpora, there are many English learner corpora available worldwide, such as the Cambridge Learner Corpus (CLC), the International Corpus of Learner English (ICLE), the Longman Learners' Corpus (LLC), and the Hong Kong University of Science and Technology (HKUST) Corpus. Nevertheless, most of these corpora were constructed out of the language data produced by learners of English as a foreign language. Many of the widely accessible corpora were created as tools for linguistic research with no pedagogical goals in design. As a result, their content and design did not necessarily meet pedagogical needs (Braun, 2007). This situation has been changing recently and learner corpus researchers are becoming progressively aware of the importance of

Zhang Ruihua, Dr., Nanyang Technological University, Singapore,ruihuaz@gmail.com
Guo Libo, Assit. Prof., Nanyang Technological University, Singapore,libo.guo@nie.edu.sg
Hong Huaqing, Dr., Nanyang Technological University, Singapore

second language acquisition (SLA) theory and SLA researchers are beginning to acknowledge the potential value of learner corpora (Granger,2009).

This paper documents the development of a Singapore learner corpus of English writing for pedagogy, which has been constructed at Nanyang Technological University, Singapore. This corpus is designed as a representative corpus of Singaporean learner English writing, comprising sample English artefacts produced by students at 3 levels, i.e. Primary 6 (Year 6), Secondary 4 (Year 10) and Junior College 2 (Year 12). The corpus consists of approximately 3 million words and the included texts are tagged with such meta information as age, gender, ethnic group and grade. The purposes of building such a learner corpus are the following:

- to generate linguistic profiles for Primary 6 (P6), Secondary 4 (Sec4), and Junior College 2 (JC2) levels;
- to analyze these profiles, and ascertain whether students' English learning at a particular stage has met the requirements of the Syllabus and if not in what ways, to provide a firm, linguistic ground for subsequent assessment of the status of English teaching and learning and for policy making;
- to enable the classroom teacher to make informed decisions about his or her student's writing and the design of classroom materials;
- to enable curriculum designers to set informed targets for writing development across stages of development; and
- to contribute to the theorization on the nature of English writing development (Christie & Derewianka, 2008).

This paper consists of five sections. This introduction section gives the background, rationale and purposes of building such a learner corpus. Section 2 addresses the design of the corpus, followed by the presentation of corpus development in detail. Section 4 discusses the potential applications of the corpus. Then Section 5 concludes the paper by stating that although it takes time and efforts, in the long run, it is worth building such a corpus and this effort will greatly benefit education researchers, English teachers and students in Singapore.

Corpus Design

Sampling Principles

A total of 17 above average schools were involved in this corpus project. Ten primary schools, 4 secondary schools, and 3 junior collegeswere selected and invited to support the corpus construction. Among these participating schools, 6 schools are located in the western part of Singapore, 6 in the north/central and 5 in the east. There is no school located in the south of Singapore, so no school from the south was included in the project.

The main criteria for student sampling included gender, ethnicity, and the ability of English writing. The sampling of students was based on classes to facilitate the administration. The above average classes were selected from each school. In the school, students are not streamed based on the English language but on the overall

performance, so some of the pupils from the selected top classes might only have middle ability in English. The reason for sampling informants from the high and middle ability groupswas that this corpus was intended to reflect the best possible standard students could achieve at a certain level. As the corpus was intended to reflect the English writing development of Singaporean learners, the students from each level were supposed to be from groups of the same level of ability. So for primary and secondary students, we chose to exclude the low ability groups. However, the above-mentioned sampling principles did not apply to junior college students because JC students are usually grouped on the basis of specialty rather than their performance or grade. In addition, the admission into junior colleges demonstrated that their English proficiency was at least adequate. Therefore, it was the JC teachers who chose the classes for the project based on their experience. We selected only general schools rather than schools specifically run for Chinese, Malay or Indian students, so the final composition of the participating students roughly corresponds to the proportion of each major ethnic group in total population of Singapore, that is, 84% for Chinese students, 8% for Malay students, 8% for Indian and others. The ratio of male informants to female ones is52:48. The total number of informants for this project is 2294, with 1117 P6 students, 426 Sec4 students and 751 JC2 students.

For each informant, 4 pieces of artefacts were collected. For primary pupils, among these 4 pieces, 2 were formal class writing assigned by English teachers: one narrative composition and one piece of situational writing; and the other two were free writing, such as journals or reflections. For Sec4 students, 2 pieces of formal narrative/expository writing and 2 journals/reflections were collected. For JC students, 2 general papers and 2 journals/reflections were requested, but two of the JCs provided 2 pieces of General Paper and 2 pieces of Application Question answer. The third JC provided 2 pieces of General paper and 2 pieces of free writing, namely, 2 reflections. The final corpus comprises 7369 texts of 8 text types, totalling 2,911,279 tokens. Its distribution of text types across school levels is shown in Table 1.

Representativeness

The participating students come from 17 schools which are located in different regions in the country. The informants comprise students from various ethnic groups and three different levels. The corpus includes 7369 scripts of various text types produced by the students at the three levels, such as narrative, situational writing, expository and argumentative for formal writing, and journals and reflections for free writing. All of these constitute a representative sample of English writing development in the Singapore educational system.

Subcorpora

The corpus contains English artefacts from learners of different levels, namely, P6, Sec4, JC2, and different text types, i.e. narrative, expository, argumentative for formal writing, and journals and reflections for free writing, so the corpus can be readily partitioned into different sub-sets by defining the selection category(ies). For example, the artefacts from JC2 may constitute a subcorpus and the texts of all narrative writing

may form another. In the same vein, the texts from all girls can be put in one subcorpus and the texts from all Chinese students in another. Linguistic features can be compared and contrasted across these subcorpora. In particular, the English writing development can be tracked by investigating the subcorpora of three levels.

Table 1. Distribution of text types across school levels

Level	No of students	Formal writing	Free writing	Size (tokens)
Primary school	1117	Continuous writing: 1671 texts; 608,246 tokens Situational writing: 1370 texts; 171,178 tokens	Journal: 589 texts; 132,461 tokens Reflection: 387 texts; 76,732 tokens	988,617
Secondary school	426	Narrative: 357 texts; 202,887 tokens Expository: 322 texts; 166,879 tokens Situational writing: 159 texts; 57,462 tokens	Journal: 29 texts; 7,640 tokens Reflection: 444 texts; 131,994 tokens	566,862
Junior college	751	General paper: 1044 texts; 963,627 tokens Application question: 771 texts; 284,093 tokens	Reflection: 226 texts; 108,080 tokens	1,355,800
Total	2294	5694 texts; 2,454,372 tokens	1675 texts; 456,907 tokens	2,911,279

Corpus development

Concerns and problems with regard to the development of the corpus are reported in this section. Basically, the compilation of this corpus consists of 4 phases of tasks, namely, collecting scripts, digitising scripts and adding header information, and part of speech (POS) and semantic tagging by using Wmatrix (Rayson, 2012).

Collecting Scripts

Collecting students' scripts was crucial in the whole process of corpus construction. Before the scripts were collected, the consent from the parent/guardian of the student and the assent from the principal/head of department were obtained. After the original scripts were collected, they were photocopied for data entry and scanned as PDF files for later reference. The original ones were returned to the schools afterwards.

Digitising Scripts & Adding Header Information

A group of research assistants were asked to digitise the scripts. Each piece was saved as a separate plain text file, with its meta information coded in its file name. A partial coding scheme for non-confidential information is shown in Table 2:

Table 2. Partial coding scheme

Text type	**School level**	**Gender**	**Ethnic group**
N1—Narrative 1	p—primary school	m--male	c—Chinese
N2—Narrative 2	s—secondary	f--	m—Malay
S1—Situational writing 1	school	female	i—India
S2—Situational writing 2	j—junior college		e—Eurosian
E1---Expository 1			o—Others
E2---Expository 2			
GP1—General paper 1			
GP2—General paper 2			
AQ1—Application question 1			
AQ2—Application question 2			
J1—Journal 1			
J2—Journal 2			
R1—Reflection 1			
R2—Reflection 2			

The data entry basically follows these instructions:

- Includes only the student's original work (not correction/revised work);
- Ignores the teacher's corrections and comments;
- Computerized with absolute fidelity to the original, including punctuation;
- No spelling mistakes or errors are corrected.
- Each piece of work is saved in a separate file with a unique name; work responding to the same question paper from the same class is saved in a separate folder with a unique name.

File names comprise the codes for text type, school, class, student ID and the marks for that particular piece of work, if applicable. For example, a file containing a piece of narrative writing (N1) with 12 marks from a Chinese girl (student ID: 01; Class number: 1) of XX Primary School is named as N1-p9fc101-12 (case-sensitive). Folder names contain the codes for text type, school and class. For example, a folder containing Journal 1 from Class 1 of XX Primary School is named as J1-p9-1 (case-sensitive). In the meanwhile, the header information, i.e. the file name and the title of the text, was added to the text. The header information was put in pointed brackets at the beginning of the file, as presented in Figure 1.

<N1-p9fc101-20>
<title> Endangered Species </title>
"Ahh!" I streched, "The breeze is nice here!" I was on a school field trip in Malaysia I sprinted out all the way by myself and went to the famous flea market. The place was packed like sardines. I managed to squeeze in. I walked slowly. There were high heels, sneakers, slippers. I thought "Wow! I did not realise that Malaysia had such nice shoes. Next time, I must bring my family here next time! Too bad, Ms shayla does not allow us to buy things" People were all pushing and squeezing.

Figure 1. An example of header information

For secondary students, an English writing prompt like ‘force’ might lead to a narrative composition or expository essay, so the coding of formal writing text type depended on the judgment of the research assistant who keyed in the particular script. They were required to indicate it clearly in the error report if they were not sure about the text type of the script. The data was checked later based on the error report and mistakes in coding were corrected.

POS Tagging

A web-based environment Wmatrix was used to annotate this learner corpus. This web browser provides a simple interface and remote access to the corpus annotation tools and all processing is carried out on theremote web server. For part of speech tagging, the hybrid tagger CLAWS (Garside & Smith, 1997), which has been continually developed since the early 1980s and“assigns a part-ofspeechtag to every word in running text with about97% accuracy” (Rayson, 2012: p1), is used in Wmatrix. CLAWS is a robust tool, which has been tested over a large amount of data, such as the British National Corpus (Leech et al, 1994b). Another reason for choosing it as the POS tagging tool for this learner corpus was because the Wmatrix semantic analysis system, which will be discussed in the next section, only accepts as input text which has been POS tagged using CLAWS.

After all the texts were computerised, checked and cleaned up, they were joined on the basis of various categories, such as school level, text type, by using the Wordsmith tools. The files containing joined texts were uploaded to the Wmatrix server

for POS tagging and the tagged results (CLAWS7) were downloaded separately. A sample output is shown in Figure 2. For example, in "I_PPIS1 suggested_VVD cheekily_RR" in line 1, "I", "suggested" and "cheekily" are running words in the text, and "PPIS1", "VVD" and "RR" are the POS tags assigned to them respectively: "PPIS1" for 1st person sing. subjective personal pronoun (I), "VVD" for past tense of lexical verb, and "RR" for general adverb (for the full CLAWS7 tagset go to http://ucrel.lancs.ac.uk/claws7tags.html).

Semantic Tagging

A semantic tagger (Rayson& Wilson, 1996) SEMTAG is used in Wmatrix to tag texts semantically by assigning a semantic label to each token in the text to indicate its semantic category. This tool is the only available tool for semantic annotation to

I_PPIS1 suggested_VVD cheekily_RR ._.
My_APPGE younger_JJR brother_NN1 ,_, James_NP1 ,_, glanced_VVD towards_II
my_APPGE direction_NN1 and_CC gave_VVD ma_NN1 mischievous_JJ smirk_NN1 ._.
Before_CS I_PPIS1 knew_VVD it_PPH1 ,_, the_AT race_NN1 had_VHD already_RR
begun_VVN ._.
Both_DB2 of_IO us_PPIO2 raced_VVD down_RP the_AT escalators_NN2 like_II
banshees_NN2 gone_VVN berserk_NN1 ,_, laughing_VVG and_CC shouting_VVG at_II
the_AT top_NN1 of_IO our_APPGE voices_NN2 like_II lunatics_NN2 ._.
Passers-by_NN2 shot_VVD dirty_RR looks_VVZ at_II us_PPIO2 but_CCB we_PPIS2
could_VM not_XX care_VVI less_RRR ._.
When_CS we_PPIS2 almost_RR reached_VVD the_AT end_NN1 of_IO the_AT
escalator_NN1 ,_, I_PPIS1 realised_VVD that_CST I_PPIS1 was_VBDZ far_RR
ahead_II21 of_II22 him_PPHO1 ._.

Figure 2. An example of POS tagged text

English and achieves the success rate of about 92% (Rayson & Wilson, 1996). The tagset (for the full semantic tagset go to http://ucrel.lancs.ac.uk/usas/) includes 21 major categorieswhich are further expanded into 232 subcategory labels. Table 3 presents the 21 categories at the top level of the hierarchy:

Table 3. 21 major semantic categories

A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Figure 3.A sample semantic tagset

A13.1	Degree: Non-specific
A13.2	Degree: Maximizers
A13.3	Degree: Boosters
A13.4	Degree: Approximators
A13.5	Degree: Compromisers
A13.6	Degree: Diminishers
A13.7	Degree: Minimizers
E4	Happy/sad
E4.1	Happy/sad: Happy
E4.2	Happy/sad: Contentment
E5	Fear/bravery/shock
E6	Worry, concern, confident

A sample semantic tagset is presented in Figure 3. A13.1-7 stand for 7 subcategories of degree adverbs respectively: non-specific, maximizers, boosters, approximators, compromisers, diminishers and minimizers. E 4.1-2 stand for the semantic fields of “Happy/sad: Happy” and “Happy/sad: Contentment”. E5 stands for the semantic field of “Fear/bravery/shock” and E6 for “Worry/concern/confident’. SEMTAG has a lexicon of over 36,000 single wordsand an idiom list of over 15,000 entries. A single semantic tag is assigned to idioms, such as *all in all, have a screw*

loose. SEMTAG uses sevenmajor techniques to disambiguate senses of words and assigns a contextually appropriate semantic tag to the target word: pos tags, generallikelihoodranking for single-word and template tags, overlapping template resolution, domain of discourse, text-based disambiguation, contextual rules and local probabilistic disambiguation (Rayson, 2008).

A semantically tagged excerpt by the SEMTAG is shown in Figure 4:

Right_Z4 before_Z5 my_Z8 eyes_B1 ,_PUNC I_Z8mf saw_X3.4 my_Z8 very_A13.3
own_A9+ brother_S4m fall_M1 backwards_X9.1- ._PUNC
In_Z5 an_Z5 attempt_X8+ to_Z5 get_A9+[i10.2.1 back_M1[i10.2.2 up_S8+[i12.2.2
,_PUNC he_Z8m reached_M1[i13.2.1 out_M1[i13.2.2 for_Z5 the_Z5 handrails_H2
of_Z5 the_Z5 escalator_O2 ._PUNC
Instead_A6.1- ,_PUNC two_N1 of_Z5 his_Z8m fingers_B1 got_A2.1+ stuck_A1.7+
in_N4[i14.2.1 between_N4[i14.2.2 the_Z5 handrails_H2 ._PUNC
He_Z8m fell_M1 on_E6-[i15.3.1 his_E6-[i15.3.2 back_E6-[i15.3.3 and_Z5 hit_E3-
is_A3+ head_B1 against_Z5 the_Z5 jagged_O4.4 edges_O2 of_Z5 the_Z5 escalator_O2 ._PUNC
Istantaneously_Z99 ,_PUNC a_Z5 pool_W3/M4 of_Z5 blood_B1 formed_T2+ around_Z5
him_Z8m and_Z5 his_Z8m white_O4.3 top_M6 turned_M2 bloody_Z4 red_O4.3 ._PUNC
Millions_N1 of_Z5 thoughts_X4.1 raced_S7.3+ through_Z5 my_Z8 mind_X1 ._PUNC
Was_A3+ my_Z8 dearest_E2+++ brother_S4m going_T1.1.3[i16.2.1 to_T1.1.3[i16.2.2
die_L1- ?_PUNC
Why_A2.2 did_Z5 I_Z8mf even_A13.1 suggest_Q2.2 this_M6 challenge_A12- ?_PUNC
I_Z8mf was_Z5 immobilised_M8 with_Z5 **fear_E5-** and_Z5 a_Z5 chill_O4.6- went_M1[i17.2.1 down_M1[i17.2.2 my_Z8 spine_B1 ._PUNC

Figure 4. A semantically tagged excerpt

From the last line but two in the above text, it can be seen that in the sequence "with_Z5 fear_E5- and_Z5 a_Z5 chill", *fear* was tagged as "E5" which stands for "fear/bravery/shock". "The semantic annotation is designed to apply to open-class or 'content' words. Words belonging to closed classes (such as prepositions, conjunctions, and pronouns), as well as proper nouns, are marked by a tag with an initial Z" (Rayson, 2008: p 66), as shown in the sequence "with_Z5". Subcorpora with semantic tagging can be used to compare and contrast the use of words from a particular semantic

category. For example, we can examine the use of emotion words in the three subcorpora of school levels and track the development of their uses.

Discoursal Annotation

The UAM CorpusTool (O'Donnell, 2013) can be used to annotate discoursal features. So far the Singapore Learner Corpus has been partially annotated with the discoursal features of hedging/boosting and grammatical metaphor. For annotation using the UAM CorpusTool, first of all, a project[2] was created and by 'Adding Layer' the kind of analysis, i.e. grammatical metaphor or hedging/boosting, was specified. Then the plain text files to be annotated were uploaded and incorporated into the project. The crucial step was to define a scheme to annotate the desired features (see O'Donnell, 2008 for the details about how to edit a scheme). Figure 5 shows the complex annotation scheme for grammatical metaphor:

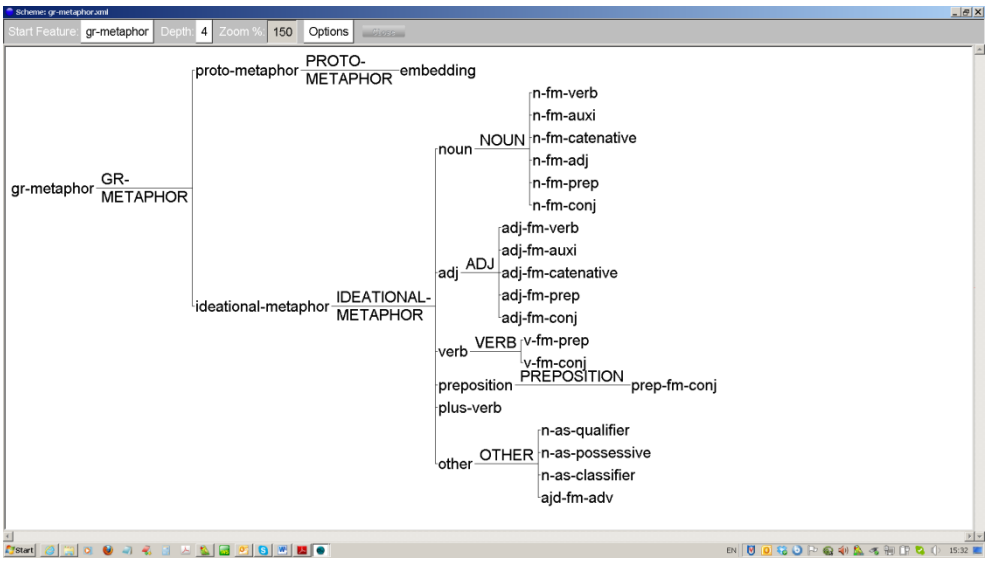

Figure 5. Annotation scheme for grammatical metaphor

The UAM CorpusTool makes use of the hierarchy representation from Systemic Functional Linguistics. The annotation scheme can grow to contain more choices, depending on what analyses you would like and how complex you like it to be. Figure 6 is a screenshot of the project window for annotating grammatical metaphors.

[2] A project in the UAM CorpusTool refers to a piece of annotation work done to a set of texts.

The middle box has several choices to be made for the annotation: noun-from-verb, noun-from-auxiliary, noun-from-catenative, noun-from-adjective, noun-from-preposition and noun-from-conjunction. Double clicking on one of the choices will

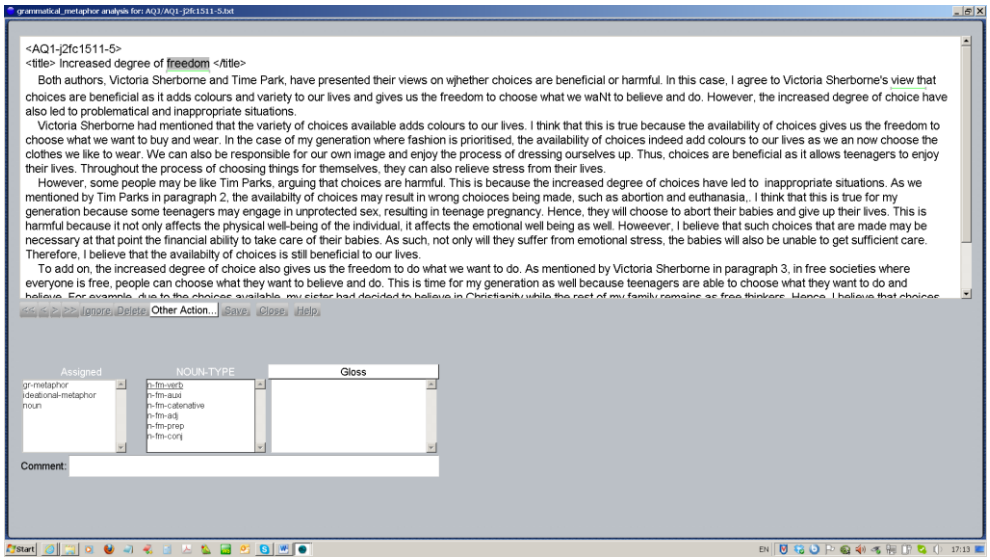

Figure 6. A project window for annotating grammatical metaphor

move the choice to the Assigned box, where you may have noticed the hierarchy of the choices: grammatical metaphor--ideational metaphor--noun. This tool will automatically produce the statistics for the annotated features. By using this tool, many discoursal features can be annotated and counted to serve various research purposes.

Corpus Application

According to Römer (2008), there are two types of pedagogical corpus applications: indirect and direct applications. The indirect application involves using a learner corpus to identify what is particularly difficult for a certain group of learners and to put special emphasis on these points inthe design of instructional materials. In other words, the indirect application involves deriving insights about second language acquisition from learner corpus analyses and drawing pedagogic implications from these insights. The direct application, on the other hand, is characterized by using learner corpora or data from learner corpora directly in the classroom for data-driven language learning (e.g., Mukherjee & Rohrbach, 2006; Seidlhofer, 2002). In practice, learner corpora are more frequently linked with the indirect applications as the corpora directly used in classrooms for data-driven language learning are more likely to be corpora of native speakers, which are usually set as the norms of language learning.

This corpus is intended for pedagogic purposes. Constructing such a learner corpus of English writing can not only fill the gap of learner corpus of writing in Singapore but also create a platform for investigating Singaporean learners' developmental features in their English learning. This corpus can generate linguistic profiles to capture and compare learners' developmental features in terms of vocabulary, grammar and discoursal devices at different learning stages and therefore theorize on the nature of writing development of learners (Christie & Derewianka, 2008). Research findings derived from this corpus can provide insight into how much learning and what kind of learning takes place at different stages, illuminate problematic areas of learners, reveal age, gender and ethnic differences in learners' English writing, and offer insightful implications for English language teaching, learning, classroom materials design and curriculum development.

Potentially, this corpus can be employed to identify the linguistic features that characterize learners from different stages of English learning and their writing development in multilingual Singapore. The analysis may focus on but will not be restricted to the following features:

Lexical features:	Lexical diversity/richness&density
	Collocational patterns
	Lexical innovation
	Spelling errors
Grammatical features:	Morphological features related to tense and number
	Tense-related features
	Article-related features
	Preposition-related features
	Agreement-related features
	Modals-related features
	Order-related features
Discoursal features:	Metaphorization-related features (Guo & Hong, 2009)
	Use of cohesive devices
	Use of rhetoric devices
	Metadiscourse features
	Development of different genre conventions, e.g. how to write a narrative essay, an exposition, etc.

As mentioned earlier, this corpus has only been partially annotated with discoursal features of grammatical metaphor and hedging/boosting; to analyse other discoursal features, the corpus needs to be further annotated with the UAM CorpusTool. This kind of annotation will be conducted at the next phase.

Concluding Remarks

In this paper, we have outlined the significance, corpus design, development and potential applications of the Singapore Learner Corpus of English Writing. The construction of such a corpus has filled a gap of learner corpus in Singapore and opened up many possibilities for future research on second language writing development.

Although it is highly labour-intensive and time-consuming in terms of data entry, in the long run, this effort is worthwhile and will considerably benefit the researchers, English teachers and students in Singapore and beyond. Currently, due to some constraints by the funding organizations, it is only open to the project team. However, it will be available to the public after it has acquired the permission of the authorities concerned in the future.

References

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL: the Journal of EUROCALL*, 19(3), 307-328. doi:10.1017/S0958344007000535

Christie, F., & Derewianka, B. (2008). *School discourse: Learning to write across the years of schooling*. London, England: Continuum.

Deterding, D. (2007). *Singapore English*. Edinburgh, Scotland: Edinburgh University Press.

Deterding, D., & Low, E. L. (2001).The NIE Corpus of Spoken Singapore English (NIECSSE). *SAAL Quarterly*, 56, 2-5.

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). London: Longman.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam, Netherlands/Philadelphia, Pennsylvania: John Benjamins.

Guo, L., & Hong, H. (2009). Metaphorization in Singaporean student writing: A corpus-based analysis. In R. Silver, C. Goh, & L. Alsagoff (Eds.), *Language acquisition and development in new English contexts* (pp. 112-131). London, England: Continuum.

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

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: The tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING 94, pp. 622–628). Kyoto, Japan.

Lim, L. (Ed.). (2004). *Singapore English: A grammatical description*. Amsterdam, Netherlands/Philadelphia, Pennsylvania: John Benjamins.

Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (Eds.), *Planning, gluing and painting corpora: Inside the applied corpus linguist's workshop* (pp. 205-232). Frankfurt, France: Peter Lang.

O'Donnell, M.(2013).*The UAM Corpus Tool*. Retrieved from http://www.wagsoft.com/CorpusTool/.

O'Donnell, M. (2008). The UAM CorpusTool: Software for corpus annotation and exploration. *Proceedings of the XXVI Congreso de AESLA*, Almeria, Spain, 3-5

April 2008.

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

Rayson, P. (2012). *Wmatrix: a web-based corpus processing environment*.Computing Department, Lancaster University.Retrieved from http://ucrel.lancs.ac.uk/wmatrix/.

Rayson, P., &Wilson, A. (1996). The ACAMRIT semantictagging system: progress report. In L. J. Evett, & T. G. Rose (Eds.), *Language Engineering for Document Analysis and Recognition, LEDAR, AISB96 Workshop Proceedings* (pp. 13-20). Brighton, England.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1, pp. 112-131). Berlin, Germany: Walter de Gruyter.

Scott, M. (2012). *WordSmith Tools* version 6. Liverpool, England: Lexical AnalysisSoftware.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson.(Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213-234).Amsterdam, Netherlands/Philadelphia, Pennsylvania: John Benjamins.

Acknowledgements

We would like to express our gratitude to all the participants for their participation in this survey and the class English teachers for their assistance in administering the survey. We are also indebted to the Singaporean Ministry of Education and Office of Education Research of the National Institute of Education for the funds for the project entitled 'Building a Singapore Learner Corpus of English Writing for Pedagogy' (Project No. OER 21/10 GLB). The research team consists of Dr Guo Libo, PI; Dr Hong Huaqing, Co-PI; Dr Zhang Ruihua, Research Fellow; Mohammad Noor Mohamed Hussein, Research Assistant (July 2011-June 2012); Cham Charmaine, Project Manager. We also wish to acknowledge the early work done by Dr Gong Wengao.

Singapur İngilizce Yazımı Öğrenci Derleminin Oluşturulması

Özet

Bu çalışmada Nanyang Teknoloji Üniversitesi'nde oluşturulan Singapur İngilizce Yazımı Öğrenci Derlemi'nin oluşturulması anlatılmaktadır. Derlem 3 ayrı seviyedeki (İlkokul-6.sınıf, Ortaokul-7. sınıf, Lise-12. sınıf) öğrencilerin ürettiği İngilizce yazılardan oluşmaktadır. Öğrencilerin değişik evrelerdeki gelişimsel özellikleri kelime bilgisi, dilbilgisi ve söylem araçları bakımından incelenmekte ve böylece Singapur'daki öğrencilerin İngilizce yazımının gelişimi ortaya konmaktadır. Metinler öğrencinin okul seviyesi, cinsiyeti, mensup olduğu etnik grup ve sınıfına göre etiketlenmiştir. Derlem tasarımıyla ilgili olarak örneklemi temsil etme gücü gibi konulara da değinilmiştir. Son olarak, pedagojik çıkarımlar ve projenin olası uygulamalarına yer verilmiştir.

Anahtar sözcükler: Derlem geliştirme, örnek eser, POS etiketleme, semantik etiketleme, söylem açımlama.