● Research Report ●

Introducing the SILS Learners' Corpus: A Tool for Writing Curriculum Development

Victoria MUEHLEISEN

Abstract

This report describes the SILS Learners' Corpus, a large electronic database of essays written in English by students at the School of International Arts at Waseda University. The corpus is being created to better understand the state of students' writing as they enter SILS and as it develops through the course of their first few semesters. The corpus will be immediately useful for the SILS language program developers in creating course material for the writing classes, but the wide range of linguistic and education backgrounds of SILS students and the large numbers of essays included means that the corpus will also be a valuable resource for researchers investigating a variety of topics in second language acquisition. The report includes a description of the data collection process and the computer program that has been created for managing the corpus.

1. Introduction: The Challenge of the SILS Writing Curriculum

Compared to the typical undergraduate school in Japan, the student body of School of International Liberal Studies is exceptionally diverse. Students come from all over the world, and while all of them have sufficient skills in English to begin to study in an English-medium academic program, many of them have had little or no training in academic essay and research writing in English, and some have not even done much writing in their native language. The students entering in 2006, for example, range from Japanese public high school graduates who have only ever written sentences or short paragraphs in their classes of English as a Foreign Language, to native and near-native English speakers who attended junior-high and high school in English speaking countries where they have done all their course work through the medium of English. In between, we find non-native English speakers who have done intensive English work in preparatory schools or short study-abroad stays or who have studied in international English-medium schools in Europe, Asia, or Africa.

Regardless of their past writing experience, however, all of these students need to learn the basics of university-level academic writing, and they need to do so quickly. The required writing program at SILS, taught through Waseda University International (WUI), is designed to bring students up to the necessary standard. The program contains three levels, with the Advanced level being equivalent to the first-year composition classes at many American universities and covering the basics of argumentative essay and research paper writing. However, the majority of students who enter SILS are not yet ready for that level and are placed into the Basic or Intermediate levels. The challenge for the writing curriculum is to take all the students up to the Advanced level within three semesters, regardless of their starting point, and in order to do so, the program and materials developers at WUI need to identify the weaknesses of the Basic and Intermediate students and find ways to deal with them. The SILS Learners' Corpus is being developed to help provide the data with which to do that.

In the corpus project, we are collecting assignments written by SILS students for their required writing classes and compiling them into a large electronic database to be used for

research. With funding from the Support Program for Contemporary Educational Needs（現代的教育ニーズ取組支援プログラム）grant provided by the Japanese Ministry of Education, Culture, Sports, Science and Technology, we have purchased equipment, hired a computer programmer to create a database program for the project, and hired graduate students to assist in the data collection and entry. This research report describes the SILS Learners' Corpus project and some of the ways in which it might be used to develop the writing curriculum for our students.

2. What is a corpus, and what can it be used for?

In Hunston's (2002: 2) concise definition, a corpus is "a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study." In the days before computers, corpus texts were usually collected on index cards or slips of paper, but these days, the term *corpus* almost always refers to a set of electronically collected and stored texts.

Corpora are now being used for a wide range of linguistic studies concerning everything from lexicography to syntax to stylistics. Researchers in first and second language acquisition have also begun to assemble and use corpora of learners' texts in order to construct models of the learners' linguistic knowledge and ability. While language teachers have probably always inductively built up models of typical students' linguistic knowledge as they read, correct, and respond to their students' work, the use of corpora allows them to analyze this at a deeper level. Writing teachers, for example, usually have some ideas of the kinds of problems and mistakes their students are likely to make, but these intuitions are not very detailed. Just from looking at individual student papers, it can be difficult to see how prevalent a particular problem is, and which particular groups of students share them.

To take a simple example, one which many SILS teachers have probably noticed, many of our students use the word *however* incorrectly, as if it were a coordinating conjunction. The example below, taken from a SILS student paper, illustrates this.

> *For example, people who did not study or go to abroad know that the American nation is generally cheerful, however they cannot know what they talk in their daily life or feel atmosphere of New York City.*

In order to decide how to deal with the problem of *however* in the writing curriculum, we need to know several things. How common is this problem really? Do most students misuse *however*, or is it a problem of just a few students? Do these students consistently misuse *however,* or do they also use it correctly? Can we assume that students placed into the Advanced level class are unlikely to make this error, so that we can focus on teaching the use of *however* in just the Basic and Intermediate level classes? Is this error only found with non-native English speakers, or is it also a problem for native English speakers? The answers to these questions cannot be found by looking at the papers from just one class or one semester. This, of course, is where the SILS Learner Corpus comes in. We can use the corpus to find and examine all the uses of *however*, both correct and incorrect, in a large set of student papers to look for patterns.

A learner corpus can be used for more than just finding examples of students' mistakes. Some corpus-based studies of learner language have examined non-native students' overuse or underuse of particular vocabulary items and structures in comparison to native students; for example, Virtanen (1998) looked at the overuse of direct questions by Swedish and Finnish learners of English, Lorenz (1998) studied the overuse of adverbial intensifiers such as *absolutely* and *extremely* in English essays of German students, and Nesselhauf (2004) examined the underuse of verbs such as *take* and *have* in constructions such as *take notice*. With our learner corpus, we will be able to look for patterns of overuse or underuse characteristic of SILS students and adjust the writing curriculum in response.

Researchers in second language acquisition have also been using learner corpora to compare the writing from students of different linguistic

backgrounds to try to determine the influence of first language on second language acquisition. For example, Hinkel's (2002) work focuses on differences in linguistic and rhetorical features such as the use of pronouns and hedges in the English writing of students whose native languages include Chinese, Japanese, Korean, Arabic, Indonesian and Spanish. With the wide range of backgrounds of SILS students, all completing the same assignments as part of the same curriculum, the SILS corpus should prove to be a valuable resource for extending this kind of research.

Finally, learner corpora have been used to study the development of second language skills over time. One ambitious project of this type is the Japanese EFL Learner (JEFLL) corpus (http://leo.meikai.ac.jp/~tono/jefll.html), created by Yukio Tono at Meikai University. It includes English essays written by students in Japanese schools from junior-high through university level. While our SILS corpus only includes college students, for students who are initially placed into the Basic level class when entering SILS, we will have three semesters of developmental data, as the students progress through the Basic, Intermediate and Advanced levels. This data will have special value due to the fact that it can be indexed by individual students, so that we can see the development not only of groups of students but of individual writers as well. In the future, we may be able to expand data collection to include essays from some students at later stages of their SILS career, for example, after returning from study abroad, to make the corpus more useful for the study of language development.

3. The SILS Learners' Corpus

The SILS Learners' Corpus is intended to "capture" the nature of SILS student writing, so, ideally, we would like to collect all the written assignments submitted by all SILS students for all of their classes. Practically, however, that is not possible. For one thing, we need students' permission and cooperation in order to collect their work for the corpus, and while most students have so far agreed, we can never expect that all will do so. Another more serious difficulty is that SILS students submit written assignments in many different formats and under different conditions – most essays and research papers are submitted on paper, not electronically, and many are handed in at the end of the semester when it might be difficult to obtain permission for their inclusion in the corpus. For this reason, we have begun by focusing our collection efforts on the required writing classes taught by Waseda University International.

3.1 Background Data Collection

At the beginning of each semester, the WUI writing teachers explain the corpus project and hand out forms to collect permission and background information from the students. Those who agree to participate sign and return the forms, and their essays are then collected for the corpus when they are handed in for class. The system has been set up to be as painless as possible for both the teachers and students – the assignments are collected when the students hand them electronically through the on-demand system used for all the WUI writing courses, so the students and teachers do not have to anything more after the students have given permission. Even after that, however, students retain the right to ask that any particular essays be kept out of the corpus, by simply writing "Do not include this in the corpus," at the top of the essay. All of the teachers and the vast majority of students in the WUI classes in the Spring semester of 2006 agreed to participate in the project.

The permission form is intended for the protection of both the students and the researchers, and is modeled on the form used by the Uppsala Student English Project (USE), a learner corpus compiled at Uppsala University in Sweden. (See Axelsson 2000 for details on USE). It explains why the corpus is being created, how the data will be preserved and the students' privacy protected, and who will have access to the data and for what purposes.

In designing the background information form, we looked at the one used by the USE project as well as the guidelines suggested by the International Corpus of Learner English (ICLE), a project coordinated by the University of Louvain (http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm). It includes questions on the native language(s) of the student and his or her parents, the languages spoken at home and in the

communities where the student has lived, the languages used as the medium of instruction in the schools that student has attended, as well as the length of study and place of study of English as a foreign language. Students are also asked for their gender, age, and most recent TOEFL score (although some decline to provide the last two items). For the immediate purpose of evaluating the effectiveness of the SILS writing curriculum, most of this information is probably unnecessary – the most important variable for that is the level (Basic, Intermediate, or Advanced) at which the student has been placed, something which we already know. However, we are asking for all the background information in order to make the corpus data useful to researchers studying a wide range of issues related to second language acquisition. Using the corpus, for example, a researcher would be able to compare the essays of students who speak Chinese as a first language to those of students who speak Japanese as a first language, or to compare the essays of students who went to English-medium high schools with those of students who did not.

3.2 The Database Program

To store and process the background data and essays, we have designed an original database program. Figure 1 below shows the screen in which the background information is entered. Note that nothing can be entered unless the permission box has been checked, indicating that the person entering the data has confirmed that there is a properly signed permission form on file. When the background information is first put into the database, the students' actual names and school ID numbers are included so that we can match the background information with the essays, but at the same time, a corpus ID number is automatically assigned by the program to each student. Later, when the students complete the Advanced level course and their essays are no longer being collected, their names and student numbers will be removed from the database, leaving only the corpus ID number to link the background data to the essays, and thus preserving the students' privacy.

The essays are entered into the same database program. Figure 2 below shows some of the additional data which is entered together with the essays themselves, including the details of the particular class (year, semester, level and teacher) and a description of the assignment. This wealth of detail sets the SILS project apart from most other learner corpus projects, which typically mix together assignments written for different classes and teachers or in response to different assignment tasks. In her survey of ten learner corpora projects, for example, Pravec (2002) found that only five included any information about the topic of the essays, and none of these included detailed assignment descriptions. The SILS corpus also includes both first and second drafts of the same essays, and is set up to include teachers' feedback on the first draft (although few teachers

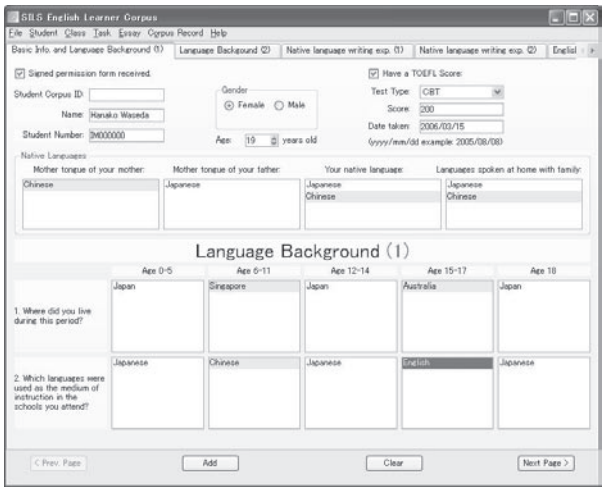

Fig. 1 Student data entry screen of the SILS Learners' Corpus database program.

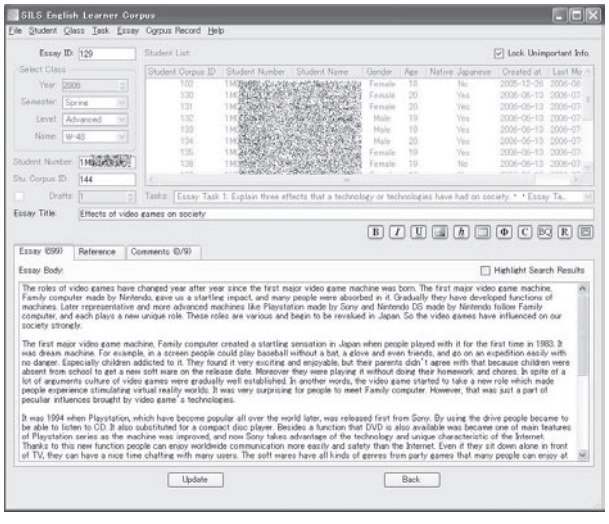

Fig. 2 The essay entry screen of the SILS Learners' Corpus database program.

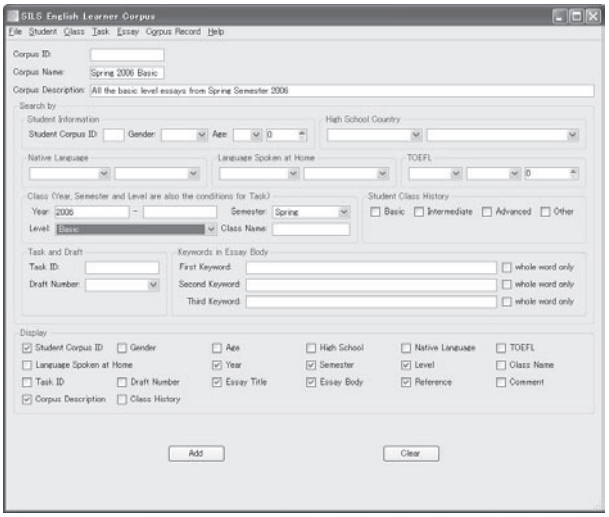

Fig. 3 The corpus generation function of the SILS Learners' Corpus database program.

have contributed data so far.) None of the learner corpora known to this author include second drafts or teachers' comments, so the SILS corpus should prove to be a valuable resource for studying how students revise their essays and respond to teachers' feedback.

Finally, the database program has a corpus generation function, which is used to tailor the output to create a corpus made up of only the essays fitting a particular criterion. Figure 3 below shows the settings which will generate a corpus of all the first drafts of essays written by basic level students in the spring semester of 2006. Note that it has been set up so that students' names and school ID numbers can never appear in the output from the program. This is an additional safeguard on students' privacy.

3．3 Format of the Corpus Data

As the essays are entered, they are converted from their original format, usually Microsoft Word documents, to XML text format in order to make the data usable for corpus processing software. (See Ide 2002 for a brief introduction to XML in corpora design). In doing so, some layout and formatting data, such as information about fonts and margins, is lost but the students' words and basic formatting information, such as paragraph breaks and bold or italic font, are preserved. When the corpus is generated, codes are also added for the variables which have been selected.

Transforming the essays and compiling them into a corpus makes them suitable as input to corpus-processing software which can be used to look for patterns in the texts. Figure 4 below shows data from the SILS corpus being used with the concordancing program AntConc created by Laurence Anthony at the School of Science and Engineering at Waseda University. A corpus called "advanced. txt", made up of the first drafts of essays in the database written by Advanced level students in the spring semester of 2006, has been loaded into the program, and *however* (spelled with a lowercase "h", to pick only the cases in which it has not been used at the start of a sentence) has been specified as the search term. The concordancing program brings together all 96 instances of the use of *however.*

From just a quick glance as these concordance lines, we can see that in the most of the cases, *however* has been used correctly, and we can pick out some examples in which it is not. In a few cases, we need to view more context in order to understand how *however* is being used. Using the File View of AntConc, we can expand the context of the concordance line, as shown in Figure 5. In this view, the XML tags and which mark the start and end of paragraphs, can also be seen.

The SILS corpus uses only a few XML tags, such as those to mark paragraph breaks and italic font, but many of the learner corpora being compiled around the world have additional annotation. Two of the most common types of annotation are part-of-speech (POS) tagging and

Fig. 4 The Concordance function of the AntConc program.

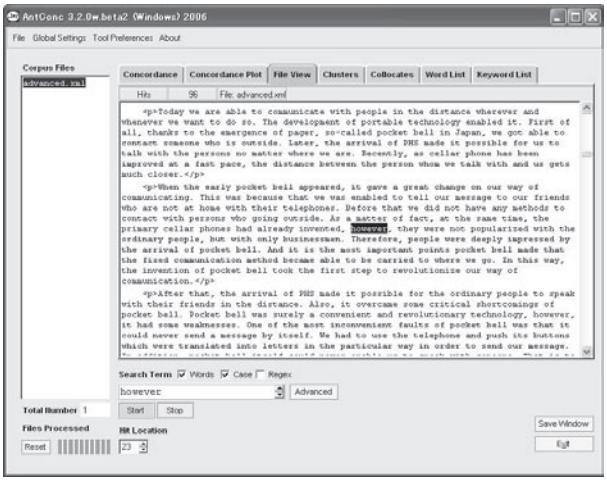

Fig. 5 The File View function of the AntConc program.

error tagging. In part-of-speech tagging, each word in the corpus is identified and marked for syntactic category; the tags are added into the text of the corpus itself, with one POS tag for each word. POS tagging allows researchers to automatically distinguish between word forms with the same spelling (the verb *run* versus and the noun *run*, for example), or to easily search for particular syntactic constructions such as passives. For most large corpora, POS tagging is done by an automatic tagging program which uses dictionaries and probability logarithms to choose the most likely part-of-speech for a particular word. (One well-known program is CLAWS-Constituent-Likelihood Automatic Word Tagging System, developed at Lancaster University). POS taggers, however, have been developed to handle texts written by native speakers and are not as accurate in processing texts by non-native writers since these writers often make mistakes in grammar. To be most useful, then, the automatic POS tagging of a learner corpus needs to be checked afterwards by hand and corrected where necessary, a labor-intensive, and therefore expensive, task. This type of automatic tagging followed by manual checking has, in fact, been done for many of the commercially sponsored large learner corpora projects around the world, but for few of the publicly-funded ones. (See Tono 2003 for an overview of this.) If time and money are available, we will consider tagging all of the essays in the database for POS, but it may be more practical for individual researchers to tag the subcorpora they are using for a particular project.

Error tagging involves the identification of student errors in the texts. It would obviously be very useful for researchers to be able to find, for example, cases in which a preposition is used incorrectly or in which an article is missing. There are also some generalizations about word usage which might be missed if spelling errors are not identified. In a non-error tagged corpus, for example, it would necessary to search both for "surprise" and for its commonly misspelled variant "suprise" in order to be sure of getting an accurate picture of the word's use. However, as Tono (2002: 804) explains, most kinds of errors cannot be identified automatically; while programs have been developed for a few simple kinds of errors, such as missing articles, the detection of most kinds of errors requires human judgment. A computer program may be able to tell that a sentence is grammatically ill-formed, but it cannot tell why it is strange. To correctly tag an error, it is necessary to judge what the writer was intending to say. The need for human intervention increases the cost of error tagging, and it also introduces problems of validity and consistency, both from one human tagger to another and from one corpus to another. There have been some attempts to develop error editing software and generic error tagsets (see, for example, Dagneaux, Denness and Granger 1998 for the error editor developed for ICLE, and Tono *et al.* 2001 for a generic tagset and editor) but at this point, because of the time and money that would be required, there are no plans to error tag the entire SILS Learners' Corpus. Individual researchers, of course, may wish to tag specific parts of the corpus for particular projects, and if they do so, we will ask that the tagged data be kept in the SILS corpus to be made available to others.

4. Plans for using the SILS Learners' Corpus

So far, the efforts of those involved in the SILS Learners' Corpus project have focused on designing the database and developing a procedure for collecting and entering the students' essays. The database design and debugging process took almost an entire year, and as we have started to enter large amounts of data, we continue to discover small glitches in the program that need to

be fixed. But by we now have entered most of the essays from the Spring 2006 semester classes and are preparing to enter those for the Fall classes. Our corpus will soon contain several thousand essays, so we are ready to move to the next phase of the project － actually using the corpus.

We plan to start with some practical investigations related to the textbooks and teaching materials used in the WU classes, choosing some specific vocabulary items and grammatical points which are taught in class to see if and how the students are using what they are taught in their essays. For example, we plan to look at how students are using subordinating conjunctions such as *however* and *therefore*, and see if there is any improvement in their accuracy in the essays they write after they have studied these specific points in class. We hope to be able to make useful suggestions for the ordering of material in the writing courses and for items to be included as new materials are developed.

The SILS Learners' Corpus will also be available to other researchers at Waseda. In the next few years, we hope to build up an exceptional resource useful for anyone interested in the areas of second language acquisition and language pedagogy.

References

Anthony, Laurence. 2006. *AntConc* (version 3.2.0w.beta). Waseda University. URL: http://www.antlab.sci.waseda.ac.jp

Axelsson, Margareta Westergren. 2000. USE - The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal* 24:155-7.

Dagneaux, Estelle, Sharon Denness and Sylvianne Granger. 1998 Computer-aided error analysis. *System* 26/2: 163-174.

Granger, Sylviane (ed.). 1998. *Learner english on computer.* New York: Longman.

Granger, Sylviane, Estelle Dagneaux, and Fanny Meunier (eds.). 2002. *International Corpus of Learner English. Version 1.1.* Université catholique de Louvain: Centre for English Corpus Linguistics.

Hinkel, Eli. 2002. *Second language writers' text: Linguistic and rhetorical features*. Mahwah, New Jersey: Lawrence Erlbaum.

Hunston, Susan. 2002. *Corpora for applied linguistics*. Cambridge: Cambridge University Press.

Ide, Nancy 2000. The XML framework and its implications for corpus access and use. *Proceedings of Data Architectures and Software Support for Large Corpora.* Paris: European Language Resources Association, 28-32.

Lorenz, Gunter. 1998. Overstatement in advanced learners' writing: stylistic aspects of adjective intensificiation. In Granger, 1998.

Nesselhauf, Nadia. 2004. How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In Guy Auston, Silvia Bernardini, and Dominic Stewart. *Corpora and language learners.* John Benjamins.

Pravec, Norma A. 2002. Survey of learner corpora. *ICAME Journal* 26: 81-114.

Tono, Yukio. 2003. Learner corpora: Design, development, and applications. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) *Proceedings of Corpus Linguistics 2003*. Lancaster University.

Tono, Yukio, Tomoko Kaneko, Hitoshi Isahara, Toyomi Saiga and Emi Izumi. 2001. The standard speaking test (SST) corpus: A 1 million-word spoken corpus of spoken of Japanese learners of English and its implications for L2 lexicography. In Sangsup Lee (ed.) *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary,* pp. 257-262. The 2nd Asialex International Conference. Yonsei University, Korea, 8-10 August 2001.

Virtanen, Tuija. Direct questions in argumentative student writing. In Granger 1998.