Survey of learner corpora

Norma A. Pravec
Montclair State University

1 Introduction

A learner corpus is a computerized textual database of the language produced by foreign language learners (Leech 1998). Because of its capability to store and process language, the computer provides the means to investigate learner language in a way that was not possible previously. As pointed out by Leech, a database of learners' language that is large and that has been carefully assembled should prove to be a very useful resource to anyone who wants to find out how languages are learned and how to help make the learning process even better. For a corpus to be most useful, however, various types of annotation, such as part-of-speech tagging (POS), error tagging, semantic tagging, discoursal tagging, or parsing, can be added so that meaningful linguistic patterns can be extracted.

The main purpose in compiling a learner corpus is to gather objective data that can aid in describing learner language (Granger 1998). Generally, learner corpora are important because they provide a deviation from the standard, ie the language of the native speakers of a particular language. Through the investigation of authentic natural language data, researchers can focus on theoretical and/or pedagogical issues while educators can concentrate on the needs of learners. For example, by examining a computerized error corpus, researchers have access not only to learner errors, but also to learners' total interlanguage[1] (Granger 1998:6). From a pedagogical perspective, research into learner corpora has led to the creation of EFL tools, such as the *Electronic Language Learning and Production Environment* tool driven by the Hong Kong University of Science and Technology (HKUST) learner corpus. This EFL tool is an electronic pedagogical system that was designed to assist learners in improving the lexical, grammatical, and discoursal aspects of their reading and writing in English (Granger 1998: 187).

Learner corpora are predominantly found in Europe and Asia. In addition, there is one learner corpus that is being compiled in the US. Despite the wealth of corpora that now exist or are being assembled, however, there has been no systematic comparison of the attributes of these corpora to suggest to the researcher, the educator, or the learner which corpus would be the most useful to answer a particular question. This survey offers specific information about those corpora that have been compiled mainly with the written text of learners of English as a second or foreign language (ESL or EFL)[2]. It presents a systematic comparison of the main features of these corpora along with extensive details about each corpus including the size of the corpus, the purpose of the corpus, the proficiency level of the learners, and the availability of learner background information. In addition, this survey provides the organization of the corpora, such as whether the data are stored in flat files or relational files, and the kind of markup language that is used. Moreover, annotation, ie the type of tagging that is done, and whether the annotation is done manually or automatically by using tagging tools, as well as the various tools that are employed to access the corpora, are presented.

This study aims to provide detailed information about various learner corpora that will be accessible not only to researchers, but also to ESL or EFL educators and learners. By having this knowledge, a researcher who is interested in exploring a particular linguistic aspect of the learner's written language will be able to isolate the corpus that is appropriate to the type of research to be conducted. Educators will be able to find a corpus that can provide examples directly related to a particular lesson being taught, and learners will be able to identify the corpus that can provide assistance in learning English more effectively.

2 Basic information about the corpora

The currently existing corpora are presented in Table 2.1:

Name of Corpus	**Type of Corpus**	**Location of Corpus**	**Language Background**
CLC	Commercial	England	Various
HKUST	Academic	University of Science & Technology, Hong Kong	Cantonese
ICLE	Academic	University of Louvain-La-Neuve, Belgium	Various
JEFLL	Academic	Meikai University, Japan	Japanese

JPU	Academic	University of Pecs, Hungary	Hungarian
LLC	Commercial	England	Various
MELD	Academic	Montclair State University, USA	Various
PELCRA	Academic	University of Lodz, Poland	Polish
TSLC	Academic	Hong Kong University, Hong Kong	Cantonese
USE	Academic	Uppsala University, Sweden	Swedish

2.1 Purpose of the Corpus

ICLE

One of the most important learner corpora is ICLE (International Corpus of Learner English), which provides a computer collection of essays written by advanced learners of EFL, ie university students of English in their third or fourth year of study, from different native language backgrounds. ICLE, the result of collaboration between a large number of universities internationally, was launched in 1990 by Sylviane Granger and is currently being coordinated by her at the University of Louvain-la-Neuve in Belgium. It is important because it was the first learner corpus created in an academic setting.

For Granger, the main purpose in compiling a learner corpus is the gathering of objective data for the description of learner language, which is considered imperative for valid theory and research (Granger 1998). The primary goal of ICLE is the investigation of the interlanguage of the foreign language learner. In particular, ICLE was established to provide an empirical resource for large-scale comparative studies of the interlanguage of advanced EFL learners with significantly different native language backgrounds. The research goals of ICLE are two-fold. The first aim is to collect dependable evidence on learners’ errors and to compare them cross-linguistically in order to determine whether they are universal or language specific. In addition, the comparison is carried out to determine to what extent they are affected by factors in the learner’s cultural or educational background. The second objective of ICLE is to investigate aspects of ‘foreign-soundingness’ in non-native essays which are usually revealed by the overuse or underuse of words or structures with respect to the target language norm. This investigation is done by means of a comparison between individual L2 sub-corpora and native English corpora, such as the International

Corpus of English, the Lancaster – Oslo/Bergen corpus, and the Louvain Corpus of Native English Essays.

Besides these research goals, as director of the Université Catholique de Louvain Centre for English Corpus Linguistics, a specialist research center with three core areas of research activity (French-English contrastive linguistics, translation studies, and computer learner corpora), Granger is concerned with ensuring that ICLE is used to benefit learners. To that end, the Centre for English Corpus Linguistics is active in encouraging research into the potential applications of learner corpora to pedagogical materials and learning aids. For example, the Centre has created an error tagging tool, called the *Error Editor*, which enables researchers to tag errors in learner texts and to subsequently put together comprehensive lists of typical learner errors. It is the center's hope that this research will facilitate the development of computer-aided language learning programs, and spelling, grammar, and style checkers that are more appropriate for learners of English than those currently available (http://www.fltr.ucl/ac.be/FLTR/GERM/ETAN/CECL/introduction.html).

At the present time, ICLE is comprised of the texts of students from 14 different native language backgrounds. The language backgrounds of the subjects whose essays constitute ICLE include French, German, Dutch, Spanish, Swedish, Finnish, Polish, Czech, Bulgarian, Russian, Italian, Hebrew, Japanese and Chinese. ICLE is divided into individual sub-corpora of collected data that are representative of each of these language backgrounds. In addition, ICLE continues to expand as new varieties are added on a regular basis (Granger 1998: 10). New additions include a Brazilian Portuguese sub-corpus, for which information can be accessed at http://www.bricle.f2s.com, as well as Lithuanian, Norwegian, and South African (Setswana) sub-corpora (http://www.fltr.ucl.ac.be.fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm).

JPU

JPU (Janus Pannonius University), established by Jozsef Horvath, is the first to employ a large database of Hungarian learner English. The corpus, which is discussed in great detail in his online dissertation (http://www.geocities.com/writing_site/thesis/), contains the essays of advanced level university students that were collected from 1992 to 1998. The JPU corpus includes five sub-corpora: a Russian Retraining sub-corpus, an Electives sub-corpus, a Language Practice sub-corpus, and the two most sizable sub-corpora, a Postgraduate sub-corpus, and the Writing and Research Skills sub-corpus.

Horvath's primary interest is the analysis of learner English for language education purposes. However, the corpus was also developed for descriptive and analytic purposes. Horvath regards the current status of JPU to be satisfactory for linguistic and language educational study (Horvath 1999: Section 4.1.1). Students' scripts can be analyzed for different purposes: by keeping a record of students' performance, longitudinal studies are made possible; the collection can be submitted to theoretically and practically relevant analysis; linguistic and pedagogical information can be extracted from the corpus; the corpus can be exploited for language education; and researchers can compare and contrast the individual learner corpora that comprise JPU, or they can compare and contrast the individual JPU learner corpora with L1 collections (Horvath 1999: Section 4.1.1).

PELCRA

The PELCRA project, a cooperative effort by the Department of English Language at Lodz University in Poland and the Department of Linguistics and Modern English Language at Lancaster University in England begun in 1997, is aimed at the collection of different kinds of corpus material including the Polish Learner English Corpus. This corpus Polish Learner English Corpus portion of the PELCRA project is a compilation of learner data from a range of learner styles at different proficiency levels, from beginning learners to post-advanced learners (teenagers and adults), that consists of the written work of students (http://www.uni.lodz.pl/pelcra/corpora.htm).

The PELCRA learner corpus contains data that were collected from the exam essays of Polish learners of English at the Institute of English Studies in Lodz and two teacher-training colleges affiliated with the University of Lodz. The texts were gathered over a period of three consecutive years, from 1998 to 2000. The students whose essays are included in the corpus were at different levels of proficiency, ranging from Cambridge First Certificate to Cambridge Proficiency Exams (English exams taken by students all over the world). Each year, all students of the Institute submitted their year-end exam essays. This means that in the database there are up to three essays written by the same student, but supposedly the student was at a higher proficiency level each time an essay was submitted. To delineate the proficiency levels, the essays are organized according to the year in which they were written, eg first-year students, fourth-year students. Because their resources of staff and money are limited, however, the developers of the PELCRA learner corpus decided to concentrate on the advanced level first in order to build up a collection of usable data more quickly. Although the current data come from exam essays, indications are that there will not be an exclusive focus on exam tasks in the future.

The PELCRA learner corpus was established for research, pedagogy, and the development of appropriate learner materials. Areas of investigation include word order differences between the Polish and English languages with a particular emphasis on pre-modification and post-modification, questions of definiteness and determiners, prepositions and collocations, the use of general for specific items in the lexis in addition to a large field of errors of avoidance (http://www.uni.lodz.pl/pelcra/corpora.htm). Agnieszka Lenko, the primary PELCRA learner corpus researcher, is mainly interested in researching lexical errors.

USE

USE (Uppsala Student English), maintained at Uppsala University in Sweden, is a collection of texts written primarily by Swedish university students who are advanced learners of English with a high level of proficiency. Piloted in 1998, and started in 1999, USE was developed mainly for language research and for pedagogical purposes, and to some extent it can also be used for course evaluation. Another purpose of the corpus is to serve as an instrument to diagnose the language difficulties of Uppsala students at different levels (http://www.hit.uib.no/icame/ij24/use.pdf).

HKUST

Established by John Milton, HKUST (Hong Kong University of Science & Technology), is comprised of the texts of Chinese students of English (mainly Cantonese speakers) at the advanced high school level, just prior to entering tertiary institutions, and enrolled in university. In addition to providing an opportunity to analyze the writing of Chinese learners, this collection of scripts furnishes data about student performance that are critical for the development of new pedagogical instruments. It is believed that knowledge of the students' linguistic and pedagogical needs assists in the creation of pedagogical aids, such as electronic composition and grammar tutorials, which target the needs of these students. In this way, students will have some of the realistic and practical advice they would otherwise have only from a human authority (Milton & Chowdhury 1994: 129–130). The scripts can also be fundamental to the design of syllabi and materials production. In addition, through the analysis of the variations between the writing produced by students under examination conditions and in out-of-class assignments, student performance can be better understood so as to enable them to write more effectively in either situation.

TSLC

Also in Hong Kong, TSLC (TELEC Secondary Learner Corpus), which was developed by Quentin Allan and begun in 1994, contains text written by Chinese learners of English, whose mother tongue is Cantonese, at the secondary school level. TSLC is a resource that is being developed by a team of teacher educators and materials writers at the Teachers of English Language Education Centre (TELEC), a teacher education facility housed in the Department of Curriculum Studies at Hong Kong University. The three main functions of TELEC are to run workshops periodically for in-service teachers; to conduct research into English language teaching; and to administer *TeleNex*, a computer network that is designed to provide professional support to secondary level English teachers in Hong Kong, many of whom do not have specialized knowledge of the English language.

The *TeleNex* network is the most visible aspect of TELEC's roles and it is available over the internet, without charge, to all secondary level English teachers in Hong Kong. *TeleNex* is composed of two hypertext databases, *TeleGram* and *TeleTeach*, as well as a range of theme-based conference corners. *TeleGram*, a database specifically designed for the Hong Kong teaching environment, is a pedagogical grammar that provides information about English grammar and usage. *TeleTeach*, on the other hand, is a database of graded teaching materials which have been designed so that they can be printed out for use in the classroom. TSLC was used in the development of the three components of the *TeleNex* network.

The main purpose of TSLC is to provide the data for the TELEC staff to carry out linguistic analysis into the particular areas of English in which secondary students in Hong Kong experience difficulty. These analyses are then used to supply information for the approach that is taken to compile the *Students' problems* files in *TeleGram*. The *TeleGram* files deal primarily with problems of production that include morphology, lexis, syntax, punctuation, ellipsis, style, and register, in addition to collocation and coherence at both sentence and discourse level.

Moreover, the corpus is used, together with various modern English corpora, to conduct investigations into the writing of Hong Kong secondary students. These examinations have revealed informative patterns of incorrect usage, such as overuse, underuse, and lexical, collocational, or syntactic errors, as well as correct usage. All of these corpora have proven to be invaluable to the TELEC staff when they need to answer teachers' questions about various aspects of grammar and usage through the *TeleNex* conference corners (Allan forthcoming).

JEFLL

Based in Japan, JEFLL (Japanese English as a Foreign Language Learner) was launched by Yukio Tono in 1996. This interlanguage corpus of English learners was created in order to fully understand the process of L2 acquisition in that particular EFL context. Although JEFLL is concerned with the interlanguage of Japanese EFL learners, it is significantly different from ICLE in that it has been constructed with not only the texts of advanced learners, but also with the essays of beginning and intermediate learners (http://www.lb.u-tokai.ac.jp/tono/jefll.html). In fact, texts are taken from students at the junior high, high school, and university level. Since JEFFL does not focus on one learner level, it is developmental in nature, and this makes it possible for a researcher to analyze interlanguage development on certain lexical, grammatical, or semantic features. Tono has also used his corpus to carry out interlanguage error studies (Tono 1998, 1999, 2001).

CLC

CLC (Cambridge Learner Corpus) is one of the two commercial learner corpora discussed in this survey. The commercial corpora stand out from the other corpora because they were established to assist English Language Teaching/Training (ELT) publishers in compiling ELT dictionaries and other ELT resources, such as ELT course books. These learner corpora help publishers create tools that address the specific needs of the target user, the ELT student. ELT students' requirements include having access to full information about grammar, sociolinguistic information about register that is reliable, and information concerning spoken English that is not necessarily obvious to them (Gillard & Gadsby 1998:159). Although publishers of ELT dictionaries have traditionally used native-speaker English to obtain information about the current usage of the language, these corpora have been compiled with learner English in order to analyze learners' errors and to use the analysis in compiling ELT dictionaries and other ELT resources (Gillard & Gadsby 1998: 159–160).

Part of the Cambridge International Corpus, a very large collection of English texts which are stored in a computerized database, CLC is comprised of exam scripts written by students from various native language backgrounds who take English exams around the world. Authors, editors, and lexicographers use CLC when they are working on books for Cambridge University Press. They can search CLC to find examples of how learners use English. For example, they can find out which words, patterns, and grammatical structures are used successfully by learners. But even more valuable than this, they can find out which areas of English cause the majority of problems for learners. This in turn helps them present the appropriate information in Cambridge dictionaries and ELT course books (http://uk.cambridge.org/elt/reference/clc.htm).

In addition, the data from CLC is used to answer questions about the way that students learn at different proficiency levels. CLC is also used to ensure that the assessment of students' exams is done consistently from country to country and from year to year (http://uk.cambridge.org/elt/reference/clc.htm). Furthermore, CLC will be used as a resource and a model for Natural Language Processing software in a project that is underway to grade candidates' essays. This software will be used to compare the language in a new essay with those already graded in the corpus. Thus, it can provide an appropriate grading of the new essay (http://www.cambridge-efl.org/rs_notes/0001/rs_notes1_6.cfm).

LLC

Like CLC, LLC (Longman Learners' Corpus) also makes up part of a larger group of databases that consists of many millions of words, ie the Longman Corpus Network, which provides in-depth knowledge about words, usage, language trends, and grammatical patterns in English (http://www.longman.com/dictionaries/corpus/lccont.html). LLC is made up of essays and exam scripts, which have been sent in by students and teachers throughout the world. These essays and exam scripts are representative of various native language backgrounds and various proficiency levels. LLC also seeks to assist learners of English. For instance, in compiling dictionaries for learners of English, it is the lexicographer's job to predict what a student wants to know, and then to explain it in such a way that the student is able to understand it. A dictionary can only be useful to the student if it includes the word or phrase the student wants, ie if the student is able to find it and is able to understand the information when it is located. At each of these points, LLC can provide assistance to the lexicographer and ultimately to the learner of English (Gillard & Gadsby 1998: 160).

LLC has been very useful in compiling the *Longman Essential Activator*, a new 'production dictionary', which guides the learner to exactly the right word needed for a particular context (http://www.longman.com/dictionaries/which_dict/essact.html). It is intended for intermediate-level students. It is semantically organized like the *Longman Language Activator* which was assembled by using the earliest information in LLC and published in 1993 for the advanced learner.

The purpose of the *Longman Essential Activator* is to help students progress from the small number of simple words learned in the early stages of English study to being able to express themselves naturally and accurately through the learning of a wider range of words and phrases. Errors that are common and easily correctable are included in a specific 'help box' with several objectives, which include clearly showing a learner error, making it clear that it is an error rather than an example of correct usage, and providing the learner with the correct way of expressing the idea that triggers the error (Gillard & Gadsby 1998: 163–164).

MELD

MELD (Montclair Electronic Language Database), established and maintained by Eileen Fitzpatrick and Milton S. Seegmiller at Montclair State University in the US, is the only collection of English texts written by university students at an advanced level of proficiency from a variety of native language backgrounds in an ESL context. The main purpose of MELD is to provide a database for research into second language acquisition. Through the classification and annotation of the corpus for part of speech information and errors in learners' written productions, usage information can be easily retrieved and researched.

2.2 Corpus size

Acquiring the data to assemble a learner corpus is not an easy task. Indeed, the compilation of a learner corpus is a meticulous process involving a significant amount of time and effort. Yet, for learner language to be adequately represented in a corpus, consideration for the size of a learner corpus is important. Otherwise, the sample size may cause the investigation into learner language to be insufficient, or at the very least, to be more difficult. Table 2.2 gives an overview of sizes of the corpora dealt with in this survey:

Name of Corpus	**Size of Corpus**	**Additional Information**
CLC	> 10,000,000 words	N/A
HKUST	> 25,000,000 words	N/A
ICLE	> 2,000,000 words	Each sub-corpus consists of 200,000 words
JEFLL	> 500,000 words	Target size of corpus set at 1,000,000 words
JPU	> 400,000 words	Target size of corpus set at 500,000 words
LLC	~ 10,000,000 words	N/A
MELD	~ 50,000 words	Another 50,000 words have been collected, but exist without any markup as of yet

PELCRA	500,000 words	Another 1,500,000 words exist as transcripts, but without any markup as of yet
TSLC	> 3,000,000 words	N/A
USE	~ 1,000,000 words	N/A

It is evident from this table that there is no uniformity in the size of the corpora. Each corpus has been designed to address the needs of those involved in its establishment. In most instances, data continues to be collected in order to increase the size of each corpus.

2.3 Register of text, task setting, topic choice, essay length

2.3.1 Register of text

The register of the texts, ie the writing style, for the academic learner corpora presented in this review is academic writing. While the register of the texts comprising LLC is academic writing, the register of the texts included in CLC consists of academic writing as well as writing that covers general English and business English.

2.3.2 Task setting

The task setting refers to whether the task was timed or untimed, whether the task was part of an exam or not, or whether EFL tools were or were not used by learners (Granger 1998: 8).

The English texts used in the CLC corpus, for example, are responses to exam questions from the University of Cambridge Local Examination Syndicate English exams, which are taken by students around the world. Although not explicitly mentioned in the CLC website, the Cambridge examination website (http://www.cambridge-efl.org/exam/general/bg_fce.htm) indicates that the task setting is timed writing (lasting approximately one and one and a half hours) due to the fact that various examinations are involved. So far, scripts have been selected from the Upper Main Suite examinations, which are designed for academic purposes.

The texts used in the LLC consist of essays and exam scripts. This indicates that the task setting includes both timed and untimed writing.

The Hong Kong *Use of English Examination*, taken each year by students leaving secondary school and subsequently used as the placement instrument for tertiary English-language programs, provides a portion of the scripts used for

HKUST (Milton & Chowdhury 1994: 127). Aside from this timed writing portion of the corpus, a representative corpus of untimed writing has also been developed. This allows analyses for variations that can demonstrate the difference between the writing students produce under examination conditions and in out-of-class assignments (Milton & Chowdhury 1994: 128).

The task setting of the texts that comprise TSLC includes both timed writing involving exam scripts and untimed writing involving compositions written in class. At the present time, the corpus contains student writing that is representative of the following text types: personal letters, formal/business letters, letters to the editor, newspaper or magazine editorials, feature articles, speeches, oral reports, and free compositions. Within these text types, the genres that are covered are the following: narratives, recounts, descriptions, explanations, and arguments.

The task setting for JEFLL is timed writing involving free compositions written during 20-minute sessions without the use of a dictionary. The texts that comprise the corpus consist of descriptive and argumentative essays.

ICLE is comprised of both timed and untimed essays. In addition, they may or may not have been part of an exam, and they may or may not have involved the use of EFL tools. Although the topics are varied, the content is similar in that the topics are all non-technical and argumentative, rather than narrative, for example. The corpus also contains a small proportion of literature exam papers (Granger 1998:9–10).

Although not explicitly stated, it can be assumed that the task setting for JPU is untimed writing, because the two primary types of texts that make up the corpus are essays and research papers. In this corpus, an essay is defined as any submission to a university course that is non-fiction and for which the method of gathering data is not strictly specified. Within this group, there are sub-divisions: personal reflective essays, narrative-based and descriptive writing, and a combination of essay and research paper for a content course. In this third type of text, the writer typically makes use of reference materials, but the presentation of the ideas does not adhere to a standard research pattern, nor is the writer obligated to follow the academic standards of a dedicated research paper (Horvath 1999: Section 4.1.6).

For PELCRA, since the texts are taken from exam essays, it can safely be assumed that the task setting is timed writing. The essays are for the most part argumentative, but there are also some narrative and descriptive essays in the corpus.

The task setting of the texts comprising USE is untimed writing. In fact, students are encouraged to revise their writing before submission. The corpus consists primarily of argumentative and reflective essays. Texts are taken from essays written by students participating in English courses. In addition, essays written for literature and culture courses are included in the corpus. Information about these essays can be found at http://hem.passagen.se/ylvaberg/useinfo1.htm as well as at http://www.hit.uib.no/icame/ij24/use.pdf.

In the ESL corpus MELD, the task setting is untimed writing. The range of text types in the corpus includes argumentative essays, cause and effect analyses, and the comparing and contrasting of different topics.

2.3.3 Topic choice

Topic choice is another important feature to consider because it affects lexical choice (Granger 1998: 8). However, topic choice for these corpora was difficult to ascertain. In fact, actual topic choices were uncovered for only five of the corpora.

For example, in the *UE* section of HKUST, students were given the choice of writing on four expository topics (Milton & Chowdhury 1994: 128). Topics such as current affairs are included in these texts. The essays in ICLE encompass a variety of topics, such as cultural insight, censorship, crime and punishment, morality and leadership, the European dream, and man and nature, among many others (http://www.englund.lu.se/research/corpus/corpus/swicle.html). Students participating in JPU write about various topics. These include writing about a personal learning experience outside school that might help people discover an ability that is transferable to other fields (Horvath 1999: Section 3.3.3.1.7), and discussing aspects of Hungarian newspaper articles that were published on the day the student was born (Horvath 1999: Section 4.3.9). For USE, each student writes up to five essays representing five different topics; for instance, in the students' first essay, entitled 'English, My English', students are asked to describe their relationship to the English language and assess their strengths and weaknesses in the four skills, ie writing, speaking, listening, and reading. Additional information about USE topics can be found at http://www.hit.uib.no/icame/ij24/use.pdf. Student authors whose essays comprise MELD write about various topics including distance learning vs learning in a traditional setting. Finally, although the specific topics were not ascertained for JEFLL, the essays cover six different topics.

2.3.4 Essay length

Essay length was obtained for seven of the corpora. The essays in MELD average 500 words. The average length of an essay in ICLE is approximately 500 to 1000 words. The essays in PELCRA range between 300 and 1000 words. The length of the essays used to create USE is between 800 and 1000 words, plus or minus 200 words. In HKUST, the essays are approximately 1000 words in length. On the other hand, the essays in TSLC average between 300 and 500 words, while the essays in JEFLL range between 20 and 150 words depending on the proficiency level of the learner.

2.4 Availability of learner background information

The presence of subjects' background information in a learner corpus is important and the linking of this information to the scripts in the corpus is equally important. This information is important because it provides a researcher with the means to focus on texts that match some particular predefined attributes. In this way, the researcher can create, if desired, a customized sub-corpus for the purposes of investigation. For example, a wide range of comparisons can be performed on the data, such as female vs male learners, intermediate vs advanced learners, or Japanese learners vs Polish learners (Granger 1998: 12). In most instances, learner background information has been made part of these corpora and, where possible, specific information about each corpus has been included in this survey.

In MELD, data relevant to the essays has been collected on the student writers. Through the use of a learner profile questionnaire, student authors provide such items as age, gender, and language and educational background information that are linked to the data.

CLC contains very comprehensive data with respect to the examinees themselves. This information is usually collected from the Candidate Information Sheets that are completed by all candidates taking the examinations, and from the scores given to them for the components of the examination as well as for the examination as a whole. However, it is important to note that no information is stored about individually named examinees, because their names are removed when the scripts are keyed in. As a result, all information is anonymous. (http://www.cambridge-efl.org/rs_notes/0001/rs-notes1_6.cfm).

In LLC, each script is coded for the student's nationality, level, text type (essay, letter, exam script, etc), target variety (British or American English), and for the country of residence. This corpus has been designed to provide balanced and representative coverage for each of these categories (Gillard &Gadsby 1998: 160).

Learner background information has also been included in ICLE. For instance, age, sex, mother tongue background, knowledge of other foreign languages, and the amount and/or type of practical experience in the English language are incorporated into the corpus. As with MELD, this information is recorded via a learner profile questionnaire that is completed by all learners (Granger 1998: 10).

Learner profiles exist for all the students involved in PELCRA. In fact, specific information regarding learner variables include age, sex, learning experience, and languages spoken by the learners, as well as visits made by these learners to English-speaking countries.

Extra-linguistic learner background information is also available in USE, even if all of it has not yet been coded. Background data for each student, which is provided through a special questionnaire, is coded in a separate database. After the author's name has been removed from all of the essays, they are then converted to a uniform format (http://hem.passagen.se/ylvaberg/useinfo1.htm). Included in these learner variables are age, sex, first language, educational background, and the amount of time spent in English-speaking environments (http://www.hit.uib.no/icame/ij24/use.pdf).

Background information on the students involved in JPU has been saved on computer disk and has been made part of the corpus, but details concerning specific learner variables could not be ascertained.

For JEFLL, the header information in the essays contains all the learner-related variables, but specific information about the variables was not uncovered. Additionally, extratextual information, such as grade, school type, the topics for the compositions, etc, is also available.

Detailed background information about students can be extracted from HKUST. Each paper was numbered with a key corresponding to a database entry.

Learner background information exists for TSLC, but this information has been archived and is only available to the TELEC staff.

2.5 Accessibility of corpora to researchers

As has been made evident thus far, a great deal of information about learner language is present in each of the corpora discussed in this review. Having access to a resource such as a learner corpus is important for researchers, educators, and/or learners. Although the use of most of these corpora is restricted to the researchers, educators, and students affiliated with each individual corpus, by providing portions of the corpora either online, through CD-ROMs, or by other means, these corpora have been, or will be, made publicly available.

CLC, for example, is only for the in-house use of authors and writers working for Cambridge University Press and for members of the staff at the University of Cambridge Local Examination Syndicate. On the other hand, LLC is available for academic research, and at this time, around ten million words can be supplied. Alternatively, subsets of the larger corpus, either by language level or by students' mother tongue background, can also be provided (http://www.longman.com/dictionaries/research/resapp.html).

ICLE is only available for linguistic research and cannot be used for commercial purposes. At this time, most of the corpus can only be used by researchers at those institutions involved in the establishment of the individual corpora that make up the corpus. However, the Czech sub-corpus is currently available online at http://kvt.ujep.cz/~flaskaj/icle/iclecorp.htm. In addition, an example of the first essay in the PICLE (Polish) sub-corpus is provided in raw form at http://main.amu.edu/pl/~przemka/rawsmpl.html, while the tagged version of the same essay can be viewed at http://main.amu.edu.pl/~przemka/tasmpl.html. Moreover, a CD-ROM of ICLE will be available for research purposes in May 2002. It will contain over two million words of EFL writing representing eleven mother tongue backgrounds, that is, Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. The CD-ROM will include a search interface that will enable researchers to select data on the basis of learner variables such as mother tongue background, age, sex, and/or task variables such as text type, timed/untimed setting, or the use of reference tools. It will also contain a detailed description of the different corpora and a text by each national coordinator describing the status of English in the learners' countries of origin. The CD-ROM will be advertised at http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm. Granger can also be contacted for information about it at granger@lige.ucl.ac.be. Future versions of the CD-ROM will include other national sub-corpora.

A portion of JPU is currently accessible on the internet through the following website: http://www.geocities.com/jpu_corpus. Specifically, essays written by the female students in the Postgraduate sub-corpus were made available online to researchers as of October 2001. Additional sub-sections of the corpus are listed on the website, but are not yet accessible. In fact, Horvath is currently in the process of making the full corpus available on the internet.

PELCRA does not yet exist as a finished product. At this time, the corpus is not available to outside researchers. It may, however, be made available at some point in 2002. While it is not currently available publicly, samples of student essays are accessible at http://www.uni.lodz.pl/pelcra/samples.htm.

USE is not complete yet, so not all of the parts have been converted and/or been made anonymous. What to do about access has not been decided as of yet. Nonetheless, processed components can be made available for non-profit research, and if there is such interest, USE can be contacted through their website. In addition, if there is an interest in an exchange, USE would be happy to hear from the researcher.

HKUST is available to outside researchers who are interested in collaboration. Milton should be contacted at lcjohn@ust.hk.

TSLC is only available for use by teacher educators, materials writers, and students at TELEC. *TeleNex* has been created for the benefit of English teachers and students in Hong Kong, and although access to *TeleNex* is available only to registered English teachers in Hong Kong, a sampler of files from the two databases, *TeleGram* and *TeleTeach*, as well as sample messages from the conference corners, are accessible at http://www.TeleNex.hku.hk. TELEC is currently considering plans for access to outside researchers. All inquiries should be addressed to the project manager, whose address is available at the aforementioned website.

Currently, JEFLL is being used privately. However, Tono expects to make it available in the public domain, for both research and commercial purposes, through the internet in approximately two years.

MELD is currently designing web pages to make the data publicly available. In addition, after the database has been completed, it can easily be re-tooled to accommodate other languages besides English, thus providing a valuable resource for the foreign language departments at Montclair State University.

3 Tagging of the corpora

There are many types of annotation that can be done on a corpus: POS tagging, semantic tagging, discoursal tagging, error tagging and parsing.

CLC is unique in that two and a half million words, ie one quarter of the corpus, have been encoded for learner errors. This tagging feature allows the search for particular types of errors for which many examples can always be found. It also provides the means to see which words or structures produce the most errors in learner English. Furthermore, sub-corpora, such as the responses to a particular examination, can be set up to further refine these searches. In addition, which errors are typical of particular language groups can be identified (http://uk.cambridge.org/elt/reference/clc.htm; http://www.cambridge-efl.org/rs_notes/0001/rs_notes1_6.cfm). The method used to tag the corpus, ie manually or automatically, was not ascertained, however.

LLC is not POS tagged. However, part of the corpus has been manually tagged for error, although this portion is only for internal use at Longman Dictionaries. LLC has been used as the basis for the *Longman Dictionary of Common Errors*, which provides an insight into the common errors made by students in their written and spoken work. It provides clear, user-friendly techniques to avoid and correct these errors. Also based on the LLC are the production dictionaries that have been mentioned previously, ie the *Longman Language Activator* for students at an advanced level, and the *Longman Essential Activator* for intermediate level students.

MELD includes annotation by POS information. In addition, the data is tagged for learner error. POS tagging is being done automatically; error tagging is being done manually at this time.

Developed by the TOSCA Research Group for Corpus Linguistics of the University of Nijmegen (The Netherlands), the TOSCA-ICLE Tagging Unit was made for the tagging and parsing of the sub-corpora of ICLE. Included in its tagset are 17 major word-classes. In addition, there are a total of 220 different tags representing features for sub-classes and additional semantic, syntactic, and morphological information (http://lands.let.kun.nl/TSpublic/tosca/icle.html). Versions of the tagger/parser/lemmatizer[3] are now generally available for MS-DOS.

Moreover, this corpus has been error tagged through an error coding system that the researchers created. The system, developed at Louvain-la-Neuve, involves a number of steps, including the manual correction of learner errors, and the assignment and insertion of an error tag to each error, which is supported by a specially designed editing tool called the *Error Editor.* The *Error Editor* tagging system is hierarchical in that the error tags consist of one major category code and a series of sub-codes. There are seven major category codes, ie Formal, Grammatical, LeXico-grammatical, Lexical, Register, Word redundant/ word missing/word order, and Style. These codes are then followed by one or more sub-codes. Catalogs of typical learner errors are provided within this system. Once the process of error tagging has been completed, by using standard text retrieval software tools, the error tagged texts can be searched according to error code which can then be analyzed. This process allows for error counts, the retrieval of specific error types, and the ability to view errors in context (Granger 1998). Tono points out that although the *Error Editor* is available for research purposes, the same things can be done with other Extensible Markup Language tag-inserting editors if one knows the procedure. It is important to note that, while the data can be POS-tagged using the TOSCA tagger and error-tagged using the *Error Editor*, the data included in the CD-ROM will not be tagged.

JPU is a semi-annotated collection of texts. While learner background information and other information are tagged, eg course, course year, and genre, the corpus does not include word class or grammatical tags. Without this kind of annotation, a disadvantage is produced: in its present form, the corpus cannot provide for automatic-processing and information output that are completely reliable. However, Horvath also sees this lack of tagging components as an advantage. In his view, the linguist researcher must rely partially on intuition, based on pedagogical practice and observation, and partly on linguistic evidence to overcome the lack of annotation in order to interpret the data (Horvath 1999: Section 4.1.3). The method used to annotate learner background and other information was not ascertained, however.

Some of the data in the PELCRA learner corpus has been manually POS tagged. It was anticipated that, in time, all of the data would be automatically tagged using CLAWS POS tagging software for English text, with the standard 'C7' tagset (http://www.comp.lancs.ac.uk/computing/research/ucrel/). The C7 tagset, which is extensive, can be viewed at http://www.comp.lancs.ac.uk/ucrel/claws7tags.html. Since the CLAWS software was trained on native speaker English, there is concern that it may not be appropriate for use in tagging the data in PELCRA. Therefore, although an automatic POS tagger will eventually be used on the data, no final decision has been made at this time as to the choice of tagger. The sample essays mentioned previously were tagged using CLAWS with the C7 tagset, however, and examples of the POS tagging that has been done on them can be viewed at http://www.uni.lodz.pl/pelcra/samples.htm.

In addition, an error tagset is currently being developed for PELCRA. Moreover, Lenko has done some manual tagging of errors for her own purposes, but this was on small samples of approximately 140 essays.

At the present time, only part of USE has been POS tagged. The tool that is used for tagging the database is the Brill tagger which was trained on the written component of the British National Corpus Sampler, that is, on one million words. The Brill tagger, a POS tagger that is based on transformation rules rather than statistical methods, yields results comparable to those of statistical methods. The tagging of the Sampler has been manually post-edited and should, as a result, have a very high accuracy rate, which is one of the reasons it was chosen. No regular error tagging has been done on the corpus even though the erroneous, or other, usage of some features has been investigated. However, if funding is obtained, there is a possibility that it will also be tagged for error.

Researchers working with HKUST are of the opinion that a considerable amount of research can be conducted on this interlanguage corpus in its untagged form. However, having an annotated corpus from which to extract data relating to the particular patterns of this interlanguage would be very useful. Toward that objective, a random sample of about one percent of the corpus had been manually tagged for error and POS as of 2001. An attempt at large-scale tagging of all lexical expressions in the corpus has been the objective of the researchers involved in the corpus. It is their hope, through this annotation, to address such issues as whether the frequency of error in relation to non-error represents actual difficulty for the students, how the variables of writing circumstances affect learners' writing, and the degree to which error probability can be measured in order to produce automatic tagging algorithms (Milton & Chowdhury 1994: 128–129). Some progress toward this end has been achieved (Milton 2000).

The researchers working with HKUST created their own tagging mechanism and tagset, which included categories of error and non-error, to test how effectively an automatic tagger can detect errors. The choice of tags for labeling the interlanguage was based on patterns that arose within the texts. The establishment of the error taxonomy, ie classification, makes it possible to query the database automatically, and to then produce comprehensive lists containing specific error types (Granger 1998: 15). It is the researchers' intention to improve the speed and consistency of word-class tagging by using CLAWS.

TSLC has not been tagged. Although POS and error tagging are being explored, funding constraints have put this on hold for the time being.

Various types of tagging have been performed on JEFLL. They include POS tagging, error tagging, semantic tagging, and parsing. CLAWS is used for POS tagging and SEMTAG, created by the University Centre for Computer Corpus Research on Language at Lancaster University, is utilized for semantic tagging. Error tagging, on the other hand, is done manually. The Apple Pie Parser, a probabilistic syntactic parser developed by Satoshi Sekine at New York University with postediting features, is used for parsing the corpus. This parser can run on UNIX as well as Windows (http://cs.nyu.edu/cs/projects/proteus/app/). In addition, the *Tag Editor* (version 1.2) and the *Error Editor* (version 1.0) were also employed. These are tools that were developed for the Standard Speaking Test corpus project, which is a one million word corpus of spoken English by Japanese learners. Tono is one of the coordinating members of this project. Through the tagging that has been done on JEFLL, detailed lexical and semantic analyses can be performed.

4 Organization of the corpora

4.1 Types of databases

A database can generally be regarded as any large collection of information, such as research notes in a word-processed document, a collection of files that contain the text of a novel, or bibliographic records (http://www.kcl.ac.uk/

humanities/cch/year1/dbms/intro/tabular.html). Much more descriptive than the simple term ‘database’, however, are the terms ‘flat file database’ or ‘relational database’. With regard to the organization of a learner corpus, whether it is a flat file or a relational database must be considered.

A flat file database is one in which all of the data is included in a single table. It requires that a field be created for each piece of data, such as age, gender, and language background, to be tracked in a tabular format of columns and rows. It is particularly well-suited to handle data that occur naturally in small chunks, such as author, register, topic, etc, something text-analysis software is not well-adapted to handle. The most common type of flat-file software is the spreadsheet, which can handle text as well as numbers (http://www.kcl.ac.uk/humanities/cch/year1/dbms/intro/tabular.html).

While flat-file databases are especially good for maintaining records on a single subject, a relational database is better when there is a need to view and work with data from several files. A relational database is the most powerful technique currently available for managing complex kinds of datasets. Like a flat-file program, a relational database begins with tabular data, but the major restrictions of flat-file database management programs are avoided. In the relational database design, the data are divided into separate tables that are linked together by common fields. In this way, when a user makes a query, the component tables, which are related by software, allow only the parts of the data that reveal a pattern or answer a particular question to be brought together and put in an order that is helpful to the user (http://www.kcl.ac.uk/humanities/cch/year1/dbms/intro/tabular.html).

4.2 Mark-up languages

Another important feature of a learner corpus, ie the markup language, must also be considered. The markup language is concerned with the encoding of a corpus. The encoding, referred to as annotation or tagging, added to the texts that comprise a corpus, is a metalanguage that is generally done in some form of markup language (Horvath 1999: Section 2.3.1). Two commonly used markup languages in the corpora surveyed in this review are XML and SGML.

The Extensible Markup Language (XML) is the universal format for presenting structured documents and data on the World Wide Web. The functionality of the Web is improved through XML’s design because it provides more flexible and adaptable information identification. It is called extensible because it is not a fixed format like HTML (hyper-text markup language), which is a single, pre-defined markup language. As a metalanguage, XML allows the design of customized markup languages for a limitless number of different types of documents. This is made possible because it is written in Standard Generalized Markup Language (SGML), the international standard metalanguage for defining descriptions of the structure for different types of electronic documents. SGML is quite large, powerful, and complex as a metalanguage. A lightweight, abbreviated version of SGML, XML retains enough of its functionality to make it useful, but the optional features that make SGML too complex to program for in a Web environment are removed (http://www.ucc.ie/xml/).

Put simply, SGML is the mother tongue that has been used to describe thousands of document types in many different fields of human activity; XML is an abbreviated version of SGML, which makes it easier for a user to define his/her own document types, and easier for programmers to write programs to handle them; and HTML is the most frequently used SGML or XML application used on the Web. This allows information that is usable by all to be dispersed on a network that connects many different types of computers (http://www.ucc.ie/xml/).

4.3 The corpora

With regard to the database, either flat or relational, utilized by the two commercial corpora discussed in this survey, no information is available. In addition, no information concerning the markup language used for CLC was ascertained. On the other hand, the data for LLC, as with all Longman dictionary titles, is stored in XML.

For ICLE, the PostgreSQL, a sophisticated object-relational database management system that supports almost all SQL (Structured Query Language) constructs, is utilized. In ICLE, there exists one database per participant, and one database is a relational database, ie PostgreSQL, which, for historical reasons, consists of only one main table. Each entry in the table is comprised of a learner profile that contains all learner information with the exception of the participant's essay. The learner essays are stored as text files (ASCII format) beside the actual database and are available as a link from HTML search result pages. Finally, the database runs on a Solaris Sun (Ultra) Sparc Station platform.

FileMaker Pro, relational database software for Macintosh computers, was used to organize JPU. In addition, the files in the corpus are in ASCII (American Standard Code for Information Exchange), or text-only, format with no markup language added.

The organization for PELCRA is not yet fully developed. With regard to the markup language used, indications are that the corpus has been enhanced with some SGML.

USE is a collection of individual text files. At the present time, the files are in ASCII (.txt) format with minimal markup (eg <docid=NNN> </doc>). Titles are marked (<title></title>). The file identifier is an ‘id’ that is connected to each contributor (=student) with an extension to indicate what kind of essay it is (eg .a2 = first term (=a), second essay (=2)).

HKUST is generally queried in text format. Some SGML tags have been employed, but most of the tagging follows the conventions used in the CLAWS tagset.

TSLC is organized through the use of concatenated files, which are in ASCII format with no markup language added.

With regard to the organization of JEFLL, sometimes the data is exported into MS-Windows Excel, a spreadsheet, or MS-Windows Access, a relational database management system that can be used to create simple and small-scale database applications. Usually, all the information for JEFLL is kept in XML format.

The files that comprise MELD are connected by Perl scripts. At this time, no markup language is being used.

5 Tools to use the corpora

In most cases, the best tool to use with a learner corpus is text retrieval software that can retrieve data and carry out various statistical analyses, and that also includes a concordancer. Various types of mathematical analyses can be performed on the data, such as calculations for the actual number of words per sentence, the average number of words per sentence, the type-to-token ratio, ie the number of different words in the essay, the number of paragraphs and the number of words per paragraph, and the number of errors per sentence. A tool that is used extensively with learner corpora is a concordancer. It is a simple, but extremely useful, tool with many applications.

Various linguistic analyses of learner corpora can be carried out by means of concordances. They are useful for the study of morphology, lexical semantics, collocations, and to some extent, syntax and discourse analysis. By using a concordancer, it is also possible to undertake comparative studies by gender, age, dialect, etc. In addition, patterns of error in the writing of second language learners can be identified, and comparisons can be made between students’ writing and the writing of native-speakers. In order for a concordancer to work properly, however, a corpus must be in machine-readable form, which usually translates into ‘plain ASCII format’.

Besides text retrieval software, other computer programs can be used to work with learner corpora. These can be pedagogical tools that have been created by the researchers involved in the corpus specifically to address the needs of learners.

The authors, editors, and lexicographers who use CLC work with sophisticated Windows-based software that enables them to perform a wide range of searches and concordances. Developed at Cambridge University Press, Cambridge Corpus Tools is a state-of-the-art software package that is used in the development of learners' dictionaries and other publications. One such dictionary is the *Cambridge Learner's Dictionary* (http://uk.cambridge.org/elt/cld/book/intro.htm). CLC has allowed the identification of the most troublesome areas for learners, which are illustrated through the 'Usage Notes' included in the dictionary (http://www/teflfarm.com/teachers/reviews/monthly/1/jan_feb/CUPlearndic.htm).

By querying the corpus, exploratory work is also being done to find out collocational information on words. In addition, the University of Cambridge Local Examinations Syndicate hopes to develop a comprehensive lexicon for use in its exams with the lexicon being validated by referring to CLC and other publicly available corpora (http://www.cambridge-efl.org/rs_notes/0001/ rs_notes1_6.cfm).

Longman uses proprietary XML corpus retrieval software to analyze the data in LLC. In fact, this software allows concordancing, wordlists, and statistical analysis using mutual information, T-score, and binomial log likelihood statistics. However, any technologically advanced software could be used to run searches and concordances on the data.

Once the process of tagging and parsing has been carried out on ICLE, the corpus can be analyzed by using standard text retrieval software (Granger 1998). The researchers who work with ICLE primarily utilize WordSmith Tools for their analyses.

For his dissertation that included an analysis of JPU, Horvath utilized Conc 1.7, the Mac-based Summer Institute of Linguistics concordancer, for text retrieval and analysis. It is a Macintosh application that processes text files and is only limited by the size of a computer's hard disk and memory allocation (Horvath 1999: Section 2.3.2). Like most other concordancing software, Conc 1.7 can process data saved as ASCII, or text-only files, which is the case of JPU. However, although it has a simple user interface, a limitation of Conc 1.7 is that it can only be used with small texts (http://www.georgetown.edu/cball/preprints/microconcord.html).

The WordSmith Tools package is also used to work with the PELCRA corpus (http://www-gewi.kfunigraz.ac.at/talc2000/Htm/menu.htm).

With the help of text retrieval software, Lenko has employed the PELCRA corpus to trace the second language vocabulary acquisition of Polish learners of English and to compare the reliability and validity of three measures of lexical richness. Two groups of essays, one containing 100 essays and the other containing 69 essays, were measured for lexical richness along three dimensions. These were the ability to understand a word's meaning, the ability to produce it in an elicitation task, and the ability to use it in free uncontrolled production. For both essays, three measures of lexical richness were calculated: the type/token ratio, the mean type/token ratio, and the lexical frequency profile. Analysis of the three measurements indicates that the lexical frequency profile is the most reliable instrument to distinguish between learners at different proficiency levels (http://www-gewi.kfunigraz.ac.at/talc2000/Htm/menu.htm). The knowledge gained from this study has an application in the field of language teaching and learning.

Mainly the WordSmith Tools package is utilized to work with USE, but since the files are in plain text, any text retrieval program that reads '.txt' files can be used. In addition, parts of the corpus have been indexed so it can be used with Qwick, but that is more as a side activity, and not really in the corpus itself. Qwick, created by Oliver Mason and maintained at the University of Birmingham in the UK, is software used to do corpus analyses for concordances and collocations.

Milton has used relational databases, eg Microsoft Access, only incidentally in his research on HKUST. Microsoft Access was primarily used for organization and because it imports Microsoft Excel files, ie spreadsheet files, easily. At various points, Milton also used Microsoft Excel quite extensively to tabulate word and n-gram counts.

In addition, Milton used askSam quite extensively because it allows more flexible handling of text. askSam is user-friendly, free-form database software that allows the organization and search of structured or unstructured data. Moreover, Milton also used Unix scripts quite extensively. However, much of his work was based on text searches, eg concordances (Milton 2000). One of the measures determined through analysis that is of particular value to HKUST corpus researchers is the type/token ratio, which provides a comparison of the range of vocabulary employed by Chinese students in their examination scripts and in their untimed assignments. They have made extensive use of other statistical procedures as well, for example, log-likelihood (Milton 2000).

Moreover, based on HKUST, Milton has developed the *AutoLANG* and *WordPilot* computer programs for use with the corpus. Both were created to help students become better and more effective writers.

The *AutoLANG* program is a self-access interactive English tutorial. When students access an *AutoLANG* exercise passage, a different combination of errors is viewed, with each line having only one error. If students can learn to locate and correct these errors in their own writing, it is expected that their writing will become more accurate and effective. Students can click on ‘Hints and Answers: the English Grammar Guide’ which opens a hypertext file called ‘the English Grammar Guide’ that provides an explanation to the specific error being worked on (http://home.ust.hk/~autolang/whatis_AL.htm).

WordPilot, on the other hand, is a writer’s assistant that can help students compose and proofread their writing, while at the same developing and acquiring an effective vocabulary of words and phrases. While composing, students can use *WordPilot* to find examples of any word or phrase from professionally written text libraries that are sizable, see common phrases, ie collocations, and check their writing. In the acquisition of a better vocabulary, students can look up definitions and word relationships for any word; study business and professional expressions as well as commonly confusing words; create their own personalized dictionary; and test themselves on their ability to use words and expressions correctly (http://home.ust.hk/~autolang/whatis_WP.htm). The off-campus shareware version of *WordPilot* can be downloaded for a 30 day free trial from (http://home.ust.hk/~autolang/download_WP.htm). Milton wrote this concordance program for use by students (Milton 1999).

The text retrieval software used for querying TSLC includes the concordancing features of WordSmith Tools, as well as the concordance programs, *MicroConcord* and *Monoconc Pro*. *MicroConcord* is a concordance package that was developed with the language teacher and student in mind (http://info.ox.ac.uk/ctitext/resguide/resources/m125.html). Distributed by Oxford University Press, *MicroConcord* is a well-designed basic concordancer that is useful for a variety of applications, and it is robust and simple at the same time, thus making it suitable for use on authentic texts by novices and in the classroom (http://www.georgetown.edu/cball/preprints/microconcord.html). *Monoconc Pro*, a Windows-based concordancer developed by Michael Barlow, was designed for researchers, language teachers, and language students, that is, anyone who works with texts. An important feature of the program is that it allows users to extract patterns easily from either untagged texts or texts that have been annotated with mark-up or tags (http:///www.athel.com/mono.html). In addition, TELEC is currently developing a web concordancer called ‘PatternFinder’, which will be available for use by teachers on the *TeleNex* website. A tutorial for novice users is also being developed.

Researchers working with JEFLL utilize the WordSmith Tools package to retrieve data and perform various analyses on the data (http://users.ox.ac.uk/~talc98/tono.htm), as well as Perl (Practical Extraction and Report Language) (http://www-gewi.kfunigraz.ac.at/talc2000/Htm/menu.htm). Perl is an interpreted language that scans arbitrary text files, extracts information from those files, and prints reports based on that information. It uses very sophisticated pattern matching techniques that can scan large amounts of data very rapidly (http://www-2.cs.cmu.edu/Web/People/rgs/pl-intro.html).

Tono is, in fact, in the process of using JEFLL to investigate the collocation patterns of Japanese EFL learners of English. In particular, his objective is to reveal some overuse and underuse patterns and to identify the possible causes of those uses by identifying them as either developmental patterns in the L2 itself or L1 native tongue influence. After the basic lemmatization process is completed, collocation patterns will be accessed through the use of n-gram extraction programs written by Perl. Tono expects that the results of his study will undoubtedly illuminate the nature of interlanguage development and propose a new approach to the study of learner language (http://www-gewi.kfunigraz.ac.at/talc2000/Htm/menu.htm).

The researchers involved with MELD use Linux tools and Perl programs for text retrieval and analysis of the data. Readily available from the data are such statistics as the number of words per sentence, the average number of words per sentence, the number of errors per sentence, and the type-to-token ratio.

6 Conclusion

The purpose of this study has been to provide a survey of existing learner corpora along with wide-ranging details. In order to carry out a comparison of the features of these corpora, it was necessary to consult the current literature available on the topic of learner corpora. In addition, an extensive internet search was required to access the most up-to-date information available on each learner corpus. Moreover, with the exception of one commercial corpus, the primary person involved with each corpus was contacted via e-mail in order for me to obtain certain information and/or verify the information that was gathered by the researcher.

The unique nature of this study is that it presents compiled information about learner corpora that is not only extensive, but that is also not available elsewhere. It provides detailed information about various learner corpora that is useful to researchers and educators, as well as learners. For example, with this knowledge, a researcher can explore a particular linguistic aspect of learners’ written language by determining which corpus lends itself to the type of research to be conducted. Likewise, educators can identify a corpus that can provide examples that are directly related to a particular lesson to be taught, thereby enhancing the lesson. Lastly, learners can ascertain which corpus affords them information that is useful in learning English more effectively, particularly if they use it in conjunction with a concordancer.

It is clear from this survey that there is no one standard for the compilation, tagging, and organization of a learner corpus, or for the tools used to access it. In fact, each of the corpora has been designed and created for different purposes. Since each corpus seeks to describe learner language in a way that suits the needs of the corresponding researcher(s), educators, and learners, decisions have been made on an individual basis regarding the purpose of the corpus, the size of the corpus, and the accessibility of the corpus to outside researchers, for example. In addition, the more advanced features of each corpus have been selected to assist in addressing the particular linguistic aspect of learner language that is to be investigated. It is important to note that, while a corpus has been designed for the researchers involved in the individual corpus, other researchers can use the corpus differently by performing their own specific analysis on the data. Furthermore, educators can utilize a learner corpus to design innovative approaches to teaching learners of English a variety of aspects of the language.

Researchers can use learner corpora to carry out language research or pedagogical research. For example, by isolating the texts of learners in terms of first language background, proficiency level, gender, or age, a specialized study can be carried out addressing a particular linguistic feature. On the other hand, researchers can utilize a learner corpus to create various ELT tools to assist educators and learners, such as *Electronic Language Learning and Production Environment* tool, *AutoLANG*, *WordPilot*, or the tools that make up the *TeleNex* network, ie *TeleGram* and *TeleTeach*, in addition to ELT dictionaries and coursebooks. Educators can use a concordancer with a learner corpus to not only identify common errors in learners' writing, but also to serve as a basis for remedial exercises for students (Tribble & Jones 1997). Finally, through the use of concordancers, a learner who is interested in searching an accessible corpus of learner writing involving learners who share with him/her the same native language background can, for instance, view their writing and begin to recognize certain collocational patterns; or he/she can compare samples of learner writing with native-speaker text in order to find language items that are common to both as well as those that are expressed differently (Tribble & Jones 1997). These are novel ways for learners to gain knowledge of English and potentially acquire the language more effectively.

Notes

1. Interlanguage is a dynamic language system that changes constantly as the learner progresses through a theoretically infinite number of states of grammatical development along a continuum between the native language on one end and the target language on the other end.
2. ESL refers to the role of English as a subject that is taught in schools within countries where it is widely used as the language of instruction at school, as the language of business and government, and as the language for everyday communication. EFL, on the other hand, refers to the role of English in countries where it is not used as a medium of instruction nor as a language of communication (eg, in government, business, or industry).
3. In text analysis, a lemma is a root morpheme. A lemmatizer removes the affixes and leaves the root.

References

Allan, Quentin. Forthcoming. The TELEC Secondary Learner Corpus: A resource for teacher development. In S. Granger, J. Hung, J. Petch-Tyson, and S. Benjamins (eds). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*.

Axelsson, Margareta Westergren. 2000. USE – The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal*, 24:155–157. Available online at http://www.hit.uib.no/icame/ij24/use.pdf.

Berglund, Ylva. 1999. ‘You’re *gonna*, you’re not *going to’*: A corpus-based study of colligation and collocation patterns of the *(BE) going* to construction in Present-day spoken British English. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 161–192.

Berglund, Ylva. 2000. The influence of external factors on learner performance. *TALC Papers*. Available online at http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm.

Cambridge Learner's Dictionary. 2001. Cambridge: Cambridge University Press.

Flaska, Jan. 1999. Sentence Length in ICLE Students’ Essays. Available online at http://kvt.ujep.cz/~flaskaj/icle/sentleng.htm.

Flowerdew, Lynne. 1998. A corpus-based analysis of referential and pragmatic error in student writing. *TALC Papers*. Available online at http://users.ox.ac.uk/~talc98/flowerdew.htm.

Gillard, Patrick and Adam Gadsby. 1998. Using a learners' corpus in compiling ELT dictionaries. In S. Granger (ed), 159–171.

Granger, Sylviane. 1998. Computer-aided error analysis. *System,* 26:163–174.

Granger, Sylviane (ed). 1998. *Learner English on Computer.* London: Addison Wesley Longman Limited.

Granger, Sylviane. 1998. The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (ed), 3–18.

Granger, Sylviane. 2000. Learner corpora and English language teaching. *TALC Papers.* Available online at http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm.

Horvath, Jozsef. 1999. Advanced Writing in English as a Foreign Language, A Corpus-based Study of Processes and Products. Online Ph.D. dissertation. Available online at http://www.geocities.com/writing_site/thesis/.

Kaszubski, Przemyslaw. 1999. Origin of PICLE, How it all got started… Available online at http://main.amu.edu.pl/~przemka/l-95louv.html.

Kaszubski, Przemyslaw. 1999. Lexical profiling of English (learner) corpora: Can we measure advancement levels? In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 249–286.

Leech, Geoffrey. 1998. Preface. In S. Granger (ed), xiv–xx.

Lenko-Szymanska, Agnieszka. 1999. Passive and active vocabulary knowledge in advanced learners of English. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 287–301.

Lenko-Szymanska, Agnieszka. 2000. How to trace the growth in learners' active vocabulary? A corpus based study. *TALC Papers*. Available online at http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm.

Lewandowska-Tomaszczyk, Barbara, Agnieszka Lenko-Szymanska, and Anthony McEnery. 1999. Lexical problem areas in the PELCRA Learner Corpus of English. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 303–312.

Lewandowska-Tomaszczyk, Barbara and Paul Melia (eds). 2000. *PALC '99: Practical Applications in Language Corpora*. New York: Peter Lang.

Longman Dictionary of Common Errors, Second Edition. 2001. London: Longman.

Longman Essential Activator. 1997. London: Longman.

Longman Language Activator. 1993. London: Longman.

McNeill, Arthur. 1994. A corpus of learner errors: Making the most of a database. In L. Flowerdew and K.K. Tong (eds). *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 114–126.

Meunier, Fanny. 1998. Computer tools for the analysis of learner corpora. In S. Granger (ed), 19–37.

Milton, John and Nandini Chowdhury. 1994. Tagging the interlanguage of Chinese learners of English. In L. Flowerdew and K.K. Tong (eds). *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 124–143.

Milton, John and Robert Freeman. 1996. Lexical variation in the writing of Chinese learners of English. In C.E. Percy, C.F. Meyer, and I. Lancashire (eds). *Synchronic Corpus Linguistics. Papers from the Sixteenth International Conference on English Lanugage Research on Computerized Corpora*. Amsterdam: Rodopi, 121–131.

Milton, John. 1996. Exploiting L1 and L2 Corpora for CALL design: The role of a hypertext grammar. In S.P. Botley, J. Glass, A. McEnery, and A. Wilson (eds). *Proceeding of Teaching and Language Corpora (TALC '96)*. UCREL Technical Papers 9, Lancaster University, 233–243.

Milton, John, Ian Smallwood, and James Purchase. 1996. From word-processing to text-processing. In R. Pemberton, E. Li, and H. Pierson (eds). *Taking Control: Autonomy in Language Learning*. Hong Kong: Hong Kong University Press, 233–248.

Milton, John. 1997. Providing computerized self-access opportunities for the development of writing skills. In P. Benson and P. Voller (eds). *Autonomy and Independence in Language Learning*. Harlow: Longman, 237–248.

Milton, John and Ken Hyland. 1997. Qualification and certainty in L1 and l2 students' writing. *Journal of Second Language Writing,* 6 (2):183–205.

Milton, John. 1998. Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (ed), 186–198.

Milton, John and Ken Hyland. 1999. Assertions in students' academic essays: A comparison of L1 and L2 writers. In R. Berry, B. Asker, K. Hyland, and M. Lam (eds). *Language Analysis, Description and Pedagogy.* Hong Kong: HKUST, 147–161.

Milton, John. 1999. Lexical thickets and electronic gateways: Making text accessible by novice writers. In C. Candlin and K. Hyland (eds). *Writing: Texts, Processes & Practices*. Harlow: Longman, 221–243.

Milton, John. 2000. *Research Report: Elements of a Written Interlanguage: A Computational and Corpus-based Study of Institutional Influences on the Acquisition of English by Hong Kong Chinese Students*. Hong Kong: HKUST.

Minovska, Vladimira. 1999. Czech subcorpus of the International Corpus of Learner English (ICLE). In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 313–321.

Tono, Yukio. 1998. Learner corpora and SLA research: Morpheme order studies revisited. *TALC Papers*. Available online at http://users.ox.ac.uk/~talc98/tono.htm.

Tono, Yukio. 1999. Using learner corpora for L2 lexicography: Information of collocational errors for EFL learners. Available online at http://www.lancs.ac.uk/postgrad/tono/userstudy/LEXICOS6.html. Originally published in LEXICOS 6 (AFRILEX SERIES 6), 116–132, Stellenbosch: Universitet van Stellenbosch, 1996.

Tono, Yukio. 1999. A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 323–339.

Tono, Yukio. 2000. Japanese EFL learners' colligation/collocation patterns: Multiple comparison. *TALC Papers*. Available online at http://www.gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm.

Tribble, Chris and Glyn Jones. 1997. *Concordances in the Classroom. A Resource Guide for Teachers*. Houston, Texas: Athelstan.

Uzar, Rafal and Jacek Walinski. 1999. A comparability toolkit: Some practical issues for terminology extraction. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds), 445–457.

Appendix

Websites for learner corpora discussed in this survey

Cambridge Learners Corpus (CLC)

http://uk.cambridge.org/elt/reference/clc.htm

http://esl.cup.org/cdae/dictionaries/clc.html

http://www.cambridge-efl.org/rs_notes/0001/rs_notes1_6.cfm

Hong Kong University of Science and Technology (HKUST)
No website available.

International Corpus of Learner English (ICLE)
http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm

http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html

http://www.abo.fi/fak/hf/enge/icle.htm

Japanese English as a Foreign Language Learner (JEFLL)
http://leo.meikai.ac.jp/~tono/index.html

Janus Pannonius University (JPU)
http://www.geocities.com/jpu_corpus

http://www.geocities.com/writing_site/thesis/

Longman Learners' Corpus (LLC)
http://www.longman.com/dictionaries/corpus/lccont.html

http://www.longman-elt.com/dictionaries/corpus/lclearn.html

Montclair Electronic Language Database project (MELD)
http://www.chss.montclair.edu/chss/linguistics/MELD/index.html

Polish Learner English Corpus (PELCRA)
http://www.lodz.pl/pelcra/index.htm

http://www.lodz.pl/pelcra/corpora.htm

TELEC Secondary Learner Corpus
http://www.TeleNex.hku.hk

Uppsala Student English (USE)
http://www.engelska.uu.se/use.html

Contact information

Cambridge Learners Corpus (CLC)
No contact information available.

Hong Kong University of Science and Technology (HKUST)
John Milton, lcjohn@ust.hk

International Corpus of Learner English (ICLE)
Sylviane Granger, granger@lige.ucl.ac.be

Japanese English as a Foreign Language Learner (JEFLL)
Yukio Tono, y.tono@meikai.ac.jp

Janus Pannonius University (JPU)
Jozsef Horvath, joe@btk.pte.hu

Longman Learners Corpus (LLC)
Steve Crowdy, steve.crowdy@pearsoned-ema.com

Montclair Electronic Language Database (MELD)
Eileen Fitzpatrick, fitzpatricke@mail.montclair.edu
Milton S. Seegmiller, seegmillerm@mail.montclair.edu

Polish Learner English Corpus (PELCRA)
Agnieszka Lenko-Szymanska, LENKO@krysia.uni.lodz.pl
PELCRA@krysia.uni.lodz.pl

TELEC Secondary Learner Corpus
Quentin G. Allan, qgallan@hkucc.hku.hk

Uppsala Student English (USE)
Margareta Westergren-Axelsson, use@engelska.uu.se