Modern English Education, Vol. 9, No. 3, Winter 2008

Computer Learner Corpora and Their Pedagogical Applications

Dong-Ju Lee
Sanggye High School

Lee, Dong-Ju. (2008). Computer learner corpora and their pedagogical applications. *Modern English Education, 9*(3), 83-101.

In the past couple of decades, computer corpora and analysis tools or software programs have become increasingly accessible, and a great number of corpus-based studies have become progressively common in the field of applied linguistics. In line with this, there has been a growing interest and awareness that it is useful to investigate learner language by compiling a large amount of learner performance data on computer. Thus, a variety of studies on computer learner corpora (henceforth, CLC), which can be defined as electronic collections of written and/or spoken texts produced by non-native speakers of English, have started to play an outstanding role in the development of corpus linguistics for language teaching and learning. This paper describes the current state of CLC compilation and CLC-based studies regarding: (i) available collections of learner corpora and their major characteristics; (ii) learner corpus analyses and their essential techniques or methodologies; and (iii) pedagogical applications of CLC-based research in language teaching and learning. This study also mainly focuses on providing an idea of the wide range of applications of CLC and CLC-based studies to areas as interlanguage research and language teaching, especially within the data-driven learning (hereafter, DDL) approach.

[contrastive interlanguage analysis/computer-aided error analysis/data-driven learning/중간언어 대조분석/컴퓨터보조 오류분석/언어자료중심 학습]

Ⅰ. INTRODUCTION

With the rapid development of computer technology, a number of different types of corpora have been constructed in the form of machine-readable empirical language data (Gabrielatos, 2005; Hunston, 2002). Corpora come in many shapes and sizes to serve different purposes. Corpora, in general, can be classified into

two kinds with reference to size and design: general corpora and specialized corpora (Hunston, 2002). A general corpus (e.g., the *British National Corpus*, the *Bank of English*) is mainly designed for general descriptive linguistic purposes. In contrast, a specialized corpus is compiled from particular types of texts for specific research or teaching purposes (e.g., the *Michigan Corpus of Academic Spoken English*, the *International Corpus of Learner English*).

Corpus linguistics usually refers to the "empirical study of language relying on computer-assisted techniques to analyze large, principled databases of naturally occurring language" (Conrad, 2000, p. 548). Corpus linguistics is mainly concerned with the study of the meanings of words across registers, the distribution and function of grammatical forms and categories, the investigation of lexico-grammatical interconnections, discourse features, and register variation. When it comes to learner corpora, on the other hand, the areas of language acquisition, interlanguage development and learner corpus-based material/syllabus design have been focused on.

This study discusses the present state of the field of CLC construction, CLC-based research into language studies and the practical implications for language pedagogy. It starts by introducing available collections of CLC and their distinctive features. Drawbacks and limitations of their usefulness are also discussed, including some suggestions for the further development of CLC compilation and CLC-based research. A section follows on the existing studies of CLC and CLC-based research, which have employed such new techniques as contrastive interlanguage analysis (henceforth, CIA) and computer-aided error analysis (henceforth, CEA). The present study closes with a detailed discussion of how CLC and CLC-based studies can be applied to language pedagogy, especially for the development of materials and language learning tools and CLC-based data-driven learning (DDL) approach. Although the study does not provide a full picture of CLC-based language studies and their pedagogical implications for the Korean EFL context, it may be valuable in that it can offer an opportunity to overview the development of a new way of investigating learner languages and creating more learner-friendly English language teaching materials.

II. COLLECTIONS OF COMPUTER LEARNER CORPORA

According to Granger (2004), CLC are developed on many scales and for

various purposes. She distinguishes two major categories of CLC: commercial and academic CLC. The commercial CLC, such as the *Longman Learners' Corpus* and the *Cambridge Learner Corpus*, have been developed by famous publishing companies, and are very large in size, consisting of over ten million words. The academic corpora, on the contrary, may be the work of large international projects or more modest in scope. Such work has been mainly conducted by individual researchers. For example, one of the most useful corpora which is available for academic research is the *International Corpus of Learner English* (henceforth, ICLE) based on the project coordinated by Granger in Belgium (Granger, 1998, 2003, 2004). The ICLE consists of sub-corpora of English produced by learners who have 21 different mother tongue varieties, and comprises over three million words of EFL writing. Another large corpus is the *HKUST Learner Corpus* (the *Hong Kong University of Science and Technology Learner Corpus*, Milton & Chowdhury, 1994) consisting of 25 million words of Chinese-speaking undergraduates' writing. Examples of smaller learner corpora include the *Montclair Electronic Language Database* (Fitzpatrick & Seegmiller, 2004) which contains only 100,000 words, and the *Korean Learner Corpus* (Dong-Ju Lee, 2007) which comprises only about 20,000 words (but a fully error-tagged corpus) of Korean secondary school Grade 10 students' writing. Another is the small corpus for academic purposes discussed by Reppen (2001), which is a collection of writing by elementary Navajo students and native speakers.

As Granger (2004) points out, however, most existing CLC are limited in their usefulness to learner language researchers, representing for the most part the output of EFL rather than ESL learners. And although commercial corpora tend to include output by learners from a wide range of mother tongue backgrounds, academic corpora comprise output from learners of only one mother tongue background, excepting the ICLE. With regard to the learners' proficiency, most corpora focus on the intermediate-advanced level, excepting the *Korean Learner Corpus* collected from the high beginner/elementary level of Korean secondary school students. In addition, there is a lack of spoken learner language corpora compared to those focusing on the written language, which may be due to the difficulty of collecting and transcribing spoken data. This is a problem researchers also confront when constructing native speaker corpora. Another shortcoming is that since there are few longitudinal corpora, where data from the same informants are collected over a long period of time, it has not been easy to inquire into interlanguage development. Finally, there is a necessity to develop

annotated learner corpora (e.g., POS-tagged and error-tagged) for more sophisticated investigations of learner language rather than being obliged to depend on untagged learner corpora.

Due to these drawbacks and limitations, therefore, the current generation of CLC does not fully represent the wide diversity of learner language being produced. In order for CLC to be more widely exploitable in investigating learner language from various perspectives, developing reference works and teaching materials, it is necessary to collect and build additional learner corpora of various types suggested above.

Ⅲ. STUDIES OF COMPUTER LEARNER CORPORA

1. Contrastive Interlanguage Analysis: CIA

Studies of CLC usually involve either CIA or CEA (Granger, 2004). The method most frequently used so far is CIA in which learner corpora are used to compare different languages. CLC-based comparisons of different languages are considered to be a new type of contrastive analysis (CA) (Selinker, 1989). Péry-Woodley (1990) explains that the new approach consists of "comparing/contrasting what non-native and native speakers of a language do in a comparable situation" (p. 43).

According to Granger (1996a, 1998), there are two major types of comparison in CIA: (i) comparison of native language (NL) and learner language (interlanguage, IL); and (ii) comparison of different interlanguages. In the former case, Granger (1998) argues that NL vs. IL comparisons can identify the characteristics of non-nativeness of learner language, which "not only involve plain errors, but differences in the frequency of use of certain words, phrases or structures, some being overused, others underused" (p. 13). That is, this type of research makes it possible to identify both qualitative differences (misuse) and quantitative differences (overuse and underuse) between learner language and native language. Since these distinctive features of interlanguage could not be systematically explored by employing traditional contrastive analyses, this type of CIA is valuable and has significant implications for language teaching and learning.

The other type of CIA, IL vs. IL, involves comparing interlanguages of the

same language or of different languages. Since this type of study can focus on various learner variables, such as age, proficiency level, first language background, etc., it enables researchers to distinguish between more universal interlanguage developmental patterns and more transfer-related (or L1-influenced) ones.

Most of the studies using this CIA approach have been carried out using the ICLE database, mainly consisting of advanced learners' argumentative writing. Although there are few studies using low-level learners' corpora which can be more useful for the Korean EFL context, almost all of the major areas of language have been investigated to some degree. That is, a wide range of patterns of misuse, underuse and overuse in learner lexis, grammar, lexico-grammar and discourse have been revealed. For example, regarding lexis, Ringbom (1998) compared vocabulary use in essays by several different European L1 groups included in the ICLE and uncovered that although there are few L1-influenced errors in general, learners tend to use the same lexical words (e.g., the verb *think)* more frequently than native speakers. In terms of discourse features, Altenberg and Tapper (1998) found that Swedish learners of English are likely to use some informal connectors (such as *but* or *still*) more frequently, but use other formal connectors (such as *therefore, however, thus, though* and *yet*) less frequently than native speakers do.

There are some noteworthy studies based on annotated learner corpora (mainly POS-tagged) which compare and contrast various grammatical features across native and learner languages (Aarts & Granger, 1998; de Hann, 1999; Granger & Rayson, 1998; Tono, 2000). Using annotated learner corpus data, Granger and Rayson (1998) uncovered that French learners' essay writings display most of the features typical of native-speakers' speech, but few of those features typical of the natives' academic writing. These findings confirm the speech-like nature of learner writing and reinforce the fact that underuse of more formal and overuse of more informal words is a typical phenomenon relating to almost all word classes.

Another interesting example is a study of morpheme acquisition order by Tono (2000). He analysed morpheme accuracy in a small corpus of writing (300,000 words) by Japanese ESL learners and compared the findings to those identified in the early morpheme studies by Dulay and Burt (1974). His findings indicated that articles and plural *-s* seems to be acquired later and possessive *-s* seems to be acquired earlier than revealed in the Dulay and Burt study. His study is valuable in that it shows how CLC-based empirical studies can provide the means for

re-examining previous research findings in second language acquisition. More studies on CLC based on the ICLE and other learner corpora can be found in the bibliography compiled by the Centre for English Corpus Linguistics in Belgium (available at http://cecl.fltr.ucl.ac.be/references.html). As previously pointed out, however, it is to be expected that more studies employing relatively low-level learners' corpora will be exploited.

2. Computer-aided Error Analysis: CEA

Another type of learner corpus research relates to CEA. Many studies of CIA have found out differences in frequency patterns between learner and native corpora. However, since most of the CIA studies are mainly based on unannotated learner corpora, such studies are not able to provide a full picture of errors a certain learner/student group commonly produces in their language use (Granger, 2004). Hence, in spite of the time-consuming and labour-intensive process involved, error annotation or error-tagging is obviously necessary if we wish to obtain a comprehensive CEA.

CEA research puts much focus on errors in interlanguage and makes use of computer tools to tag, retrieve and analyse them. Dagneaux, Denness and Granger (1998) opened up a new approach to the analysis of learner errors using computerized learner corpora. They employed error tags and an error editor to analyse the rate of progress made by French intermediate and advanced level students of English regarding a range of grammatical and lexical variables. They suggest that their approach can be used to generate comprehensive lists of peculiar error types, count and sort them in various ways, and view them in context alongside instances of non-errors. They also claim that materials designers would be able to create more useful pedagogical tools by using such techniques.

In this respect, Dong-Ju Lee's (2007) study on a computer-aided error analysis is worth noting. He carried out a CEA of a corpus of writing in English produced by EFL Korean secondary school students. His CEA adopted the error taxonomy developed by Dagneaux et al. (2005), which classifies errors mainly according to their linguistic descriptions. The CEA identified the most common errors the students made using the error-tagging software tool, *UCLEE* (*Université Catholique de Louvain Error Editor*, Hutchinson, 1996) and the text retrieval software program, *WordSmith Tools* (Scott, 2004). The results of the CEA showed that the Korean secondary school Grade 10 students tend to make grammar-related errors most frequently, followed by errors of punctuation, form

and lexis, etc. The ten most common errors the students made in writing were also identified, such as errors involving articles, spelling, punctuation, verb tenses, and so on. The major findings of the CEA highlight the advantages of the error-tagging CEA method compared to the text retrieval error analysis approach which did not utilize fully error-tagged corpora (e.g., Dong-Ju Lee, 2004; Seung-Min Lee, 2004). The most significant benefit of the method adopted in the CEA is that a fully error-tagged learner corpus makes it possible to identify the most common errors of any given learner population (in this case, the Korean secondary school Grade 10 students). On the basis of the CEA findings, Dong-Ju Lee (2007) designed corpus-based materials as DDL teaching lessons to remedy the students' most common and frequent error types, and tried out the lessons on Korean secondary school English classes.

Ⅳ. PEDAGOGICAL APPLICATIONS OF CLC AND CLC-BASED RESEARCH

It has been argued that native speaker English corpora are being increasingly exploited in language pedagogy thanks to the fact that a variety of native speaker corpora (e.g., the *British National Corpus*, the *Brown Corpus*, the *Bank of English*, etc.) are already available, and a lot of research has already been undertaken into native varieties of English. On the contrary, CLC are now opening up a new field of language study and CLC-based studies have not yet had a significant impact on language teaching and learning. However, although the advantages of the use of native corpus data in teaching have been introduced, this type of corpus cannot provide us with certain types of information, such as the degree of difficulty of words and structures for learners. In this respect, evidence from CLC research, such as learner misuse, underuse and overuse, is capable of helping materials designers and classroom teachers determine and prioritize language teaching material at a particular proficiency level.

1. Applications for the Development of Materials and Language Learning Tools

CLC information is used to a significant extent in ELT learner dictionaries. The first dictionary incorporating CLC data was the *Longman Essential Activator*

(1997). The dictionary was informed by the *Longman Learner Corpus* which provided details of typical learner errors. This error information was employed at the so-called 'help boxes' designed to alert learners to take particular care when producing certain types of language (Gillard & Gadsby, 1998). The *Longman Learner Corpus* was also used as a basis for the *Longman Dictionary of Common Errors* (Turton & Heaton, 1996). Learner corpus data have been used to produce general learners' dictionaries, such as the *Longman Dictionary of Contemporary English* (2003) and the *Cambridge Advanced Learner s Dictionary* (2003), and each includes language notes based on their respective learner corpora. The notes in each dictionary are provided to help users avoid making common errors. It is to be expected that this kind of CLC-informed dictionary will become more specialized, being aimed at particular groups of learners who have different mother tongue backgrounds.

With regard to learner reference grammars, while its original design was not significantly influenced by CLC studies, Allan's (2002) web-based *TeleNex* network, which has been developed for secondary-level English teachers in Hong Kong, is worth noting. *TeleNex* provides teaching material (through *TeleTeach*, a teaching resources database) and gives teachers the opportunity to ask questions and get answers from other teachers and experts (through various *Conference Corners*). Of particular importance for our purposes, however, another main facility of the *TeleNex*, *TeleGram*, an electronic grammar database, offers grammatical descriptions of English based on the *TeleNex* corpus which consists of both a native corpus of modern English and a corpus of Hong Kong students' writing (see http://www.telenex.hku.hk). The most pedagogically valuable thing about the network is that it tackles students' problems by consulting a learner corpus and provides teaching materials which can help teachers deal with these problems in the classroom.

Although CLC analyses have not yet made an immediate impact on the development of printed teaching material, such as textbooks, some suggestions have been made as to how textbooks could benefit from them. For example, in his CLC-based study on writing textbooks, Kaszubski (1998) claims that most ESL/EFL textbooks tend to provide learners with limited lexical and stylistic help, so they contain such shortcomings as: few lists of common errors; coverage of collocation and phraseology is very scarce; and specific genres such as an academic essay are rarely treated. So even advanced-level EFL learners have difficulty in these areas. Kaszubski therefore proposes that it is necessary to diagnose problematic areas of target learners using CLC-based analysis and to

design L1-sensitive materials whose content and form are more customized to the learners' needs and interests.

On a more positive note, various types of CALL programs based on CLC analyses are now being developed. Milton's (1998) *AutoLANG* is one of the pioneering works in this field. He investigated the problem areas for Hong Kong Chinese EFL learners in writing and developed the program to help them. The program consists of error recognition exercises, an online grammar hypertext and a database of underused lexical and grammatical features. Another program, *WordPilot*, was also designed by Milton (2001) in order to integrate concordancing into L2 essay writing for relatively low-level learners (see http://home.ust.hk/~autolang/download_WP.htm) These kinds of CALL software programs are of importance for actual language teaching and learning, mainly because they can provide remedial exercises targeting a particular learner group's attested difficulties and feature writing aids which help the learners consult native corpora of specific text types.

Cowan, Choi, and Kim's (2003) *ESL Tutor* courseware tool is another notable CALL program which was designed to assess how effectively Korean students were able to correct their common and persistent grammatical errors using the software and evaluate the effectiveness of the various types of error feedback provided by the program. They reported that the error-correction courseware is useful in helping learners recognize errors that persist in the L2 learner interlanguages and propose that it is a promising tool for investigating the effectiveness of feedback and learner noticing.

The web-based writing environment developed by Wible, Kuo, Chien, Liu and Tsao (2001) is another enchanting writing tool which provides learners with immediate feedback on their writing and allows them to have access to lists of errors they commonly produce. The tool also provides teachers with a large database of learner data which can help them create targeted exercises. Likewise, CLC results are now being actively applied to the development of CALL-based or web-based pedagogical writing tools.

In brief, although it is at an early stage, the incorporation of results from CLC studies into pedagogic material is of importance for several reasons. Firstly, once more CLC studies which can provide L1-specific information become available, they will help teachers and materials designers decide what items should be particularly focused on in teaching based on the learner group's attested difficulties. Secondly, CLC results can suggest how certain language features may be taught more effectively. Granger (1999), for instance, analysed advanced

learners' tense use and proposed that tenses need to be taught at discourse-level rather than sentence-level. Thirdly, findings on a certain learner group's developmental sequences can be a basis for deciding the order of language features that should be taught. Ellis (1994) is convinced of this idea, arguing that language acquisition takes place better when following the developmental sequence. Finally, CLC data can provide language examples of typical errors - misuse - and particular cases of overuse and underuse to assist the design of reference materials and to help the teaching of these language features directly.

2. Applications for the Classroom

To date, whilst there have been many studies on the use of native speaker corpus data in the language classroom, only a few actual implementations of learner corpus-based DDL have been exploited (Dong-Ju Lee, 2007; Flowerdew, 2001; Milton & Hyland, 1999; Ragan, 2001). DDL or classroom concordancing can be defined as a practical use of concordancing in which corpus data are used in the language classroom in the form of concordance examples or samples which can help both teachers and learners investigate language in use (see Johns, 1991a, 1991b).

Learner corpus-based DDL relies to some extent on the usefulness of 'negative evidence' in language learning. Although some researchers (e.g., Carroll & Swain, 1993; Ellis, 1994; Granger, 1996b) argue negative evidence is not always useful, Granger (1996b) suggests that 'focused negative evidence' can be especially useful for advanced learners and for working on the language forms that are prone to become fossilized. She goes on to claim that this type of DDL can have beneficial effects in a number of areas. For example, activities involving recognizing errors and/or looking for differences in learner and native speaker language can improve learner autonomy and develop students' learning skills, especially the discovery learning skill. Furthermore, as Fan, Greaves and Warren (1999) reported, such a learning process has a positive impact on learners' affect: that is, students tend to find it motivating to identify and examine mistakes. The following suggestion by Nesselhauf (2004) also focuses on the affective dimension:

> Data-driven learning with learner data is probably particularly useful for points which have already been covered in the classroom, possibly even repeatedly, but which the learners nevertheless still get wrong. In this way, instead of

> being told once again that what they are doing is wrong, learners have the opportunity to get something right, namely to identify and explain the mistake in question. (p. 140)

However, there are some differences in pedagogical procedures when using learner corpus data for DDL rather than native corpus data. For one thing, when using negative evidence from learner corpora, it is necessary to make students aware that they are investigating incorrect forms of language so as to prevent confusion. Thus, students always need to be provided with corrective evidence taken from comparable native or reference corpora corresponding to the negative evidence presented. As Milton and Hyland (1999) claimed, it may be dangerous for learners to remember the negative language uses rather than the correct ones. To avoid this, Granger (1996b) suggests that follow-up exercises or activities for consolidating the correct uses should be provided.

The basic procedure for designing DDL activities on the basis of CLC data is as follows (Hunston, 2002; Nasselhauf, 2004). At first, likely sources of errors are identified with the aid of comparisons between the learner corpus and native corpus. Then, comparable concordance lines taken from the two corpora are presented to encourage the students to identify the differences between native and non-native usage. At this phase, specific questions or directions can be provided to allow the students to work out the differences successfully. The learners are also encouraged to notice the different patterns and change their own erroneous usage so that it is more like the native-speaker usage. Finally, further exercises or activities are presented to help the students consolidate correct usage and make the insights longer-lasting.

Following the procedure mentioned above, Dong-Ju Lee (2007) created teacher-designed DDL teaching lessons based on the results of a CEA, and tried them out on Korean secondary school English classes. In the DDL lessons, error-recognition and correction exercises were designed using both erroneous concordance examples taken from the *Korean Learner Corpus* and attested/correct concordance lines taken from a reference corpus. Writing exercises were also included to help the students confirm appropriate usage. This study, focusing on the design and evaluation of the CLC-based materials and DDL approach, resulted in pedagogically useful findings. That is, both the students and teachers who participated in the DDL classes felt relatively positive about the DDL approach. The positive evaluation of the materials is sufficiently encouraging for materials designers and teachers to develop and implement DDL approach in

low-level (not just for intermediate or advanced) class settings, like the EFL Korean secondary classroom.

As Nesselhauf (2004) argues, some language areas may be more effective and easier to treat using learner corpus-based DDL than others. It is suggested that co-occurrences of words, especially if the co-occurring words are adjacent, such as prepositions or complementation of verbs, nouns and adjectives may be best suited to the approach. She exemplifies the approach using concordance lines taken from the German sub-corpus of ICLE and the corresponding native speaker corpus of LOCNESS (the *Louvain Corpus of Native English Essays*). This material can be useful in showing German-speaking learners that they tend to use the verb *suggest* and following complementation incorrectly. That is, they frequently use the pattern, '*suggest* + *to* + infinitive' (e.g., *Russell suggests to go on holiday*), which does not correspond to native-speaker usage.

Similarly, another example can be found in Granger and Tribble (1998). They introduce an example task focusing on the patterns following the verb *accept*. The concordance lines from the native-speaker writing show that *accept* is followed by a noun (e.g., *Why not accept the difference as an intentional* ⋯) or by a *that*-clause (e.g., *Hugo cannot accept that the party line has changed*), while the non-native concordance lines include some instances of '*accept* + *to* + infinitive' (e.g., *Feminists have to accept to be treated as men.*). In this case, the learners are required to notice the different patterns and are expected to understand what the correct usage is through observing both concordance lines from native and non-native speaker writing. For this task, students are provided with questions and instructions such as: *What grammatical structures appear to follow the verb 'accept ?* and *Do any grammatical forms only appear in the non-native speaker examples? If this is the case, check if the students are using an acceptable form.*

Additionally, Granger and Tribble (1998) illustrate another example DDL task involving overuse and underuse of particular words. Concordance examples from a native speaker corpus based on words such as *critical, crucial, major, serious* or *vital*, are presented. Then the concordance examples of *important* taken from the learner corpus are offered. The target word, *important*, is edited out, and students are asked to replace it with one of the alternatives presented. This example is based on the idea that a comparison of learner and native speaker concordance examples of a general word (in this case, *important*) and more specific words with a similar meaning (such as *critical, crucial, major, serious* or *vital*) can indicate that the general word is overused and the more specific words

are underused by the learners.

However, DDL with learner corpora may not be so useful when teachers wish to focus on more general grammatical areas. For instance, when focusing on tense or aspect, it may not be easy for learners to search for and identify the language points, mainly because the erroneous use is not always clear from the immediate context or the immediate context does not necessarily show whether a certain instance is correct or not. Furthermore, when treating the use of single words, it may be also difficult to determine if they appear too rarely or too frequently in the given corpus when the corpus is not large or representative enough. In line with this, Granger (1996b) and Nesselhauf (2004, 2005) argue that classroom teachers should be careful when introducing learner corpus concordancing especially if the learners are required to carry out the concordancing by themselves. It is suggested therefore that since the learners tend to be confused by some of the occurrences or by their very number in learner language data, it may often be better to edit concordance lines and provide them on paper rather than allowing the learners to work directly with the corpus.

In sum, DDL tasks based on either learner or native corpora can help learners become aware of common and persistent errors which are likely to be fossilized in their interlanguage. Tasks, exercises and activities can be developed which are motivating for students as they are targeted at common grammatical or lexical problems students are likely to have, as well as at students' own attested difficulties. Moreover, such DDL materials can be a positive way of offering corrective feedback because not only instances of erroneous use but also native use are provided in any exercise or task.

Ⅴ. CONCLUSION

This study has introduced available collections of CLC, CLC-based studies and their pedagogical applications up to the current state-of-the-art. Although there are many kinds of available CLC developed by publishing companies and individual researchers, it is necessary to construct additional learner corpora of different types, such as spoken learner language corpora, longitudinal corpora, and annotated or error-tagged corpora. When it comes to CLC-based studies, CIA and CEA techniques have been employed for investigating a variety of learner language patterns of misuse, underuse and overuse. However, more studies utilizing low-level learners' corpora, which can be more useful for EFL contexts,

should be exploited.

CLC and CLC-based research findings have been used to a significant extent in ELT on-/off-line dictionaries and CALL-/web-based teaching-learning facilities. In addition, CLC and/or native speaker corpus data have been increasingly used in the language classroom in the form of concordance lines within the DDL framework. DDL materials are used in two ways: the weak version involves teacher-designed concordance materials in the form of worksheets; while the strong version relates to learners undertaking independent concordancing themselves. It is hoped, however, that more empirical studies on the use of various DDL materials with different learners in different contexts will be conducted to validate the pedagogical effectiveness of DDL.

Although some disadvantages and limitations are issued, it is pedagogically useful to make use of either CLC or native speaker corpora, or both of them as well as corpus linguistics techniques in language teaching and learning. However, as Leech (1997) warned, it should be kept in mind that a corpus does not by itself lead directly to a great advancement in learning, but is only a facilitator which can enable the learner to observe language phenomena, and to explore, investigate, generalize, and verify hypotheses. It should also be remembered that corpus-based DDL materials are not always interesting and helpful for every student. That is, the materials are sometimes difficult and tedious depending on individual learners' language proficiency and preferred learning styles.

Even though the development of native speaker corpora, computer learner corpora and corpus-based techniques has led to many insightful and useful pedagogical proposals and tasks, it is obvious that learners will need to be trained in using corpora and familiarized with these new materials and tasks. A step-by-step introduction to corpora and corpus-based DDL materials rather than a sudden immersion will probably be advisable for most learners, with the teacher taking into account their local and contextual norms and expectations.

REFERENCES

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132-141). London: Longman.

Allan, Q. (2002). The TELEC secondary learner corpus: A resource for teacher development. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer*

learner corpora, second language acquisition and foreign language teaching (pp. 195-211). Amsterdam: Benjamins.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80-93). London: Longman.

Cambridge University Press Staff. (2003). *Cambridge advanced learner s dictionary*. Cambridge: Author.

Carroll, S., & Swain, M. (1993). Explicit and negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, *15*, 357-386.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*(3), 548-560.

Cowan, R., Choi, H. E., & Kim, D. H. (2003). Four questions for error diagnosis and correction in CALL. *CALICO Journal, 20*(3), 451-463.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System, 26*, 163-174.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J., & Thewissen, J. (2005). *Error tagging manual version 1.2.* Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

de Hann, P. (1999). English writing by Dutch-speaking students. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 203-212). Amsterdam: Rodopi.

Dulay, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning, 24*, 37-53.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Fan, M., Greaves, C., & Warren, M. (1999). Identifying characteristic patterns in students' writing using a corpus of learner data. In R. Berry, B. Asker, & K. Hyland (Eds.), *Language analysis, description and pedagogy* (pp. 176-188). Hong Kong: Language Centre HKUST.

Fitzpatrick, E., & Seegmiller, M. S. (2004). The Montclair electronic language database project. *Language and Computers, 52*, 223-238.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 363-379). Amsterdam: Benjamins.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding

bells? *TESOL-EJ, 8*(4), 1-35.

Gillard, P., & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), *Learner English on computer* (pp. 159-171). London: Longman.

Granger, S. (1996a). From CA to CIA and back: An integrated approach to computerised bilingual and learner corpora. In K. Aijmer, B. Altenberg & M. Johansson (Eds.), *Language in contrast* (pp. 37-51). Lund, Sweden: Lund University Press.

Granger, S. (1996b, August). *Exploiting learner corpus data in the classroom: Form-focused instruction and data-driven learning*. Paper presented at TALC 1996, Lancaster, UK.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London: Longman.

Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 191-202). Amsterdam: Rodopi.

Granger, S. (2003). The international corpus of learner English: A new resources for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*, 279-304.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 123-145). Amsterdam: Rodopi.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). London: Longman.

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199-209). London: Longman.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hutchinson, J. (1996). *UCL error editor*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Johns, T. (1991a). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *ELR journal 4: Classroom*

concordancing (pp. 1-16). Birmingham: CELS, University of Birmingham.

Johns, T. (1991b). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *ELR journal 4: Classroom concordancing* (pp. 27-46). Birmingham: CELS, University of Birmingham.

Kaszubski, P. (1998). Enhancing a writing textbook: A national perspective. In S. Granger (Ed.), *Learner English on computer* (pp. 172-185). London: Longman.

Lee, Dong-Ju. (2004). A computer-aided research into error analysis of Korean secondary school students' writing. *Foreign Languages Education, 11*(4), 133-160.

Lee, Dong-Ju. (2007). *Corpora and the classroom: A computer-aided error analysis of Korean students writing and the design and evaluation of data-driven learning materials*. Unpublished doctoral dissertation, University of Essex, Colchester, UK.

Lee, Seung-Min. (2004). Learner corpora-based error analysis and its application in intermediate English classes. *Foreign Languages Education, 11*(3), 33-57.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). New York: Addison Wesley Longman.

Longman Publishing Staff. (2003). *Longman dictionary of contemporary English* (4th ed.). Harlow, Essex: Author.

Longman Publishing Staff. (1997). *Longman Essential Activator*. London: Author.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer* (pp. 186-198). London: Longman.

Milton, J. (2001). *Describing and overcoming environmental limitations on the interlanguage of Hong Kong Chinese learners of English: A computational and corpus-based methodology*. Unpublished doctoral dissertation, Lancaster University, UK.

Milton, J., & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In L. Flowerdew & K. Tong (Eds.), *Entering text* (pp. 127-143). Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Milton, J., & Hyland, K. (1999). Assertions in students' academic essays: A comparison of English NS and NNS student writers. In R. Berry, B.

Asker, & K. Hyland (Eds.), *Language analysis, description and pedagogy* (pp. 147-161). Hong Kong: Language Centre HKUST.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125-152). Amsterdam: Benjamins.

Nesselhauf, N. (2005). *Collocations in learner corpus*. Amsterdam: Benjamins.

Péry-Woodley, M. (1990). Contrasting discourses: Contrastive analysis and a discourse approach to writing. *Language Teaching, 23*, 143-151.

Ragan, P. (2001). Classroom use of a systematic functional small learner corpus. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 207-236). Amsterdam: Benjamins.

Reppen, R. (2001). Writing development among elementary students: Corpus-based perspectives. In R. Simpson & J. Swales (Eds.), *Corpus linguistics in North America* (pp. 211-225). Ann Arbor, MI: Michigan University Press.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). London: Longman.

Scott, M. (2004). *WordSmith Tools 4.0*. Oxford: Oxford University Press.

Selinker, L. (1989). CA/EA/IL: The earliest experimental record. *IRAL, 27*, 267-291.

Tono, Y. (2000). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In L. Bernard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 123-132). Frankfurt: Peter Lang.

Turton, N., & Heaton, J. (1996). *Longman dictionary of common errors* (2nd ed.). Harlow, Essex: Longman.

Wible, D., Kuo, C-H., Chien, F-Y., Liu, A., & Tsao, N-L. (2001). A web-based EFL writing environment: Integrating information for learners, teachers, and researchers. *Computers and Educations, 37*, 297-315.

Dong-Ju Lee
Sanggye High School
Geum-ho Apartment 102-808,
Junggyebon-dong, Nowon-gu, Seoul, 139-930, Korea
Tel: (02) 933-2707 / C.P.: 010-2053-2707
Email: maydjlee@hanmail.net

Received in October 3, 2008
Reviewed in October 14, 2008
Revised version received in December 10, 2008