English Language Teaching; Vol. 9, No. 8; 2016
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Application of Learner Corpora to Second Language Learning and Teaching: An Overview

Qi Xu[1]

[1] Guangdong University of Foreign Studies, China

Correspondence: Qi Xu, Guangdong University of Foreign Studies, China. E-mail: xqmiracle@gmail.com

Received: May 15, 2016 Accepted: June 10, 2016 Online Published: July 11, 2016

doi: 10.5539/elt.v9n8p46 URL: http://dx.doi.org/10.5539/elt.v9n8p46

Abstract

The paper gives an overview of learner corpora and their application to second language learning and teaching. It is proposed that there are four core components in learner corpus research, namely, corpus linguistics expertise, a good background in linguistic theory, knowledge of SLA theory, and a good understanding of foreign language teaching issues (Granger, 2009). Based on the above components, the present paper first introduces learner corpora, then reviews literature concerning the application of corpus linguistics to SLA by means of contrastive interlanguage analysis, and at last discusses the relationship between learner corpora and foreign language teaching.

Keywords: learner corpora, contrastive interlanguage analysis, second language learning and teaching

1. Introduction

With the development of learner corpora and multilingual corpora since the 1990s, there has been a revival of corpus linguistics, especially its application to second language research by incorporating the use of learner corpora, see for instance, Keck (2004), Pravec (2002), Myles (2005). Through investigation of actual language use in learner corpora, it is easier for researchers "to understand how best to help students develop competence in the kinds of language they will encounter on a regular basis" (Biber & Reppen, 1998: 157).

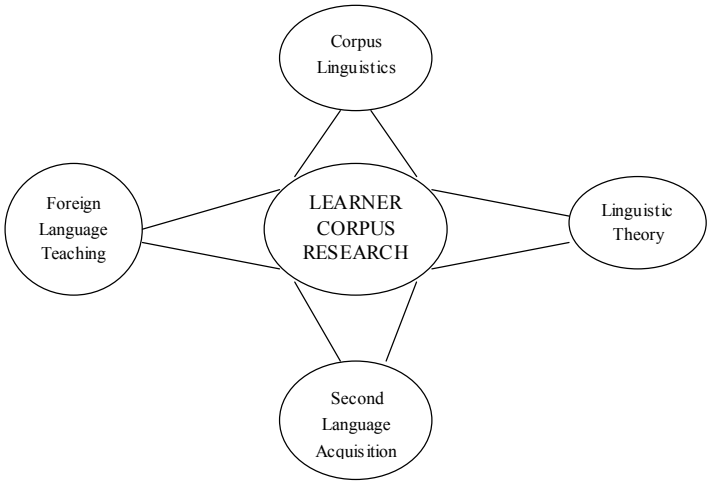

Figure 1. Core components of learner corpus research

(Adopted from Granger, 2009: 15)

Granger (2009) proposed that there are four core components in learner corpus research, namely, corpus linguistics expertise, a good background in linguistic theory, knowledge of SLA theory, and a good understanding of foreign language teaching issues, as shown in Figure 1. The present paper is aimed at

introducing learner corpora and their applications to second language learning and teaching. Based on the above components, in the present paper, we will look at 1) the introduction of learner corpora, 2) the application of corpus linguistics to SLA by means of contrastive interlanguage analysis, and 3) the relationship between learner corpora and foreign language teaching.

2. Literature Review

2.1 Introduction of Learner Corpora

Granger (2002: 7) provides a detailed definition of learner corpora:

Computer learner corpora are electronic collections of *authentic* FL/SL *textual data* assembled according to *explicit design criteria* for a particular SLA/FLT purpose. They are encoded in a standardized and homogeneous way and documented as to their origin and provenance.

There are a few keywords in this definition that are worth mentioning – "authentic", "textual data" and "explicit design criteria".

First of all, in terms of *authenticity*, it is almost impossible for the learner data to be completely natural, for the reason that foreign language teaching activities are inevitably involving some degree of "artificiality" (Granger, 2002: 8). As long as essay writing is conducted under authentic classroom circumstances, learner corpora of essay writing can be regarded as authentic written data. Besides that, learner corpora should contain *textual data* consisting of continuous "stretches of discourse", rather than, for example, lists of disconnected erroneous sentences. In addition, special attention must be paid to the *criteria* on which the learner corpus is built. Apart from the same compiling rules as the native corpora, factors like the characteristics of the learner and the task settings should also be taken into consideration.

A series of learner corpora have been released all around the world and made use of for research, for instance, *International Corpus of Learner English* (ICLE) which contains argumentative essays written by higher intermediate to advanced learners of English from various mother tongue backgrounds; *Louvain International Database of Spoken English Interlanguage* (LINDSEI), a spoken counterpart to ICLE containing oral data produced by advanced learners of English from several mother tongue backgrounds; *Chinese Learner English Corpus* (CLEC), a collection of English essays written by Chinese students ranging from senior middle school to university levels; *Cambridge Learner Corpus* consisting of exam scripts written by students taking Cambridge ESOL exams around the world, and many others.

2.2 Contrastive Interlanguage Analysis

It is advocated by Granger (1998) that it is of great significance to analyze learner corpora in second language acquisition studies. She maintains that learners' performance can be analyzed in corpus to infer the invisible mental process of SLA, and that previous hypotheses generated from the psycholinguistic approach can be tested through analysis of learner corpora.

When applying corpus linguistics to SLA, one type of methods is usually adopted, that is, Contrastive Interlanguage Analysis (CIA). CIA, both in quantitative and qualitative terms, refers to two different types of comparison: one between native language and learner language (L1 vs. L2), while the other between different varieties of interlanguage (L2 vs. L2) (Granger, 2009: 18).

In spite of the fact that controversies still exist with regard to L1 vs. L2 comparison, it is unreasonable to ignore its significance for describing the features of non-nativeness in learner writing and speech. A series of learner corpus studies have been conducted by taking a CIA approach, such as Altenberg and Granger (2001), Housen (2002), Nesselhauf (2005), Ädel (2006), Xiao (2007), etc., which have revealed a number of interlanguage features in various linguistic environments. Take the investigation of the high-frequency verb MAKE in Altenberg and Granger (2001) for instance, the findings suggested that learners even at the advanced level are still "at a risk of having a very crude knowledge of their grammatical and lexical patterning" without ruling out the "skeleton" entries held for high-frequency verbs (*ibid.*: 190).

One of the major limitations of traditional error analysis is that it "fails to provide a complete picture of learner language", and researchers "need to know what learners do correctly as well as what they do wrongly" (Ellis 1994: 67). It was also suggested by Schachter and Celce-Murcia (1977) that investigators should treat causes of error very cautiously, for in many cases, "what we see happening, however, is just the reverse". Current interlanguage research is different from traditional error analysis in that the current CIA approach treats learner performance data in its own right rather than in respect of merely decontexualized errors (Granger, 1998). Therefore, it is more likely to yield rewarding results by comparing second language learners' interlanguage with

the target native language. In my study, I will not concentrate on errors that learners have made, but focus on systematic analysis of the ditransitive constructions that they use.

3. Learner Corpora and Foreign Language Teaching (FLT)

Foreign language teaching has benefited from learner corpus linguistics research, and there is already general agreement that corpus data, especially the learner corpus data "opens up interesting descriptive and pedagogical perspectives" with "a profound and positive impact on the field of FLT" (Granger, 2002: 21). The two areas which gained most from corpus-based research are materials design and classroom teaching methodology. Literature in this section is not limited to learner corpora, but also covers the application of general native corpora to FLT.

3.1 Teaching Materials Design

Fast-paced progress has been made in developing such materials as EFL dictionaries, grammar references, and textbooks with the help of large-scale corpora, although the influence in dictionaries is more significant than in two other areas.

Compilation of Dictionaries

Recent years have shown an increasing trend that learners' dictionaries of English are compiled with reference to updated databases of language. By taking into consideration the frequency information from large-scale native corpora, a number of English dictionaries for advanced L2 learners have been compiled. Dictionaries of this type include *Oxford Advanced Learners' Dictionary*, *Collins Cobuild Dictionary*, and *Longman Dictionary of Contemporary English*, etc. (Leech, 2001: 329). These dictionaries can provide detailed information about the ranking of meanings, collocations, grammatical patterns, style and frequency (Granger, 2002: 21).

Gillard and Gadsby (1998) compiled *The Longman Essential Activator*, a dictionary consulting *The Longman Learners' Corpus* (LLC), with the aim of helping L2 learners of English to accurately and naturally produce a wider range of words and phrases, rather than heavily rely on a limited number of common words. The authors generated frequency lists from LLC, which were used to help compilers make decisions of what should be included in the dictionary. In this dictionary, they gave very detailed information of each word accompanied by near-synonyms under about 1000 'concepts'. For instance, regarding the 'concept' of WALK, words like *stroll*, *stride*, *amble*, and *jog* are listed together with definitions and examples, for the purpose of making it easier for learners to distinguish these words. In addition, based on the frequently occurring errors common to all learners, they made use of 'help boxes' to remind learners not to make similar errors in their use of English. Gillard and Gadsby (1998: 163) believe that "by having constant access to a very large body of students' writing, lexicographers are sensitized to and reminded of the needs of their audience far more thoroughly than they could achieve through their previous teaching experience". Their practice provides much insight for dictionary compilers to take the features of learner English into account.

Enhancement of textbooks

Before the advent of learner corpora, teaching materials were mainly based on the English language teachers' experience and intuition in deciding what should be taught to students. It was therefore quite difficult for compilers to check whether the teaching materials could meet learners' needs (Guo, 2006: 233). In the past two decades, a series of corpus studies have been conducted to test the effectiveness of the materials used in foreign language teaching, including Grabowski and Mindt (1995) on irregular verbs, Barlow (1996) on reflexives, Mindt (1997) on future time expressions, Conrad (2004) on linking adverbials, Römer (2004) on modal verbs, etc. Abundant evidence has been found from these studies that "the language presented in textbooks is frequently still based on intuitions about how we use language, rather than actual evidence of use" (O'Keeffe, McCarthy, & Carter, 2007: 21). There is, therefore, a great need for "revised pedagogical language descriptions that take corpus findings into account and present a more adequate picture of language as it is actually used" (Römer, 2010: 22).

Römer (2005) made a comparison between the use of progressives in native spoken English corpora (*British National Corpus* and *Bank of English*) and in representations of spoken English used in German EFL textbooks. It was found that 30%-40% of progressives are used to indicate repeated actions or events in native corpora, while repeatedness is seldom expressed by progressives in textbooks, where more than 90% of the progressives refer to single continuous events, as in *What are you doing?* or *What have you been doing?*

These descriptions may "address the described imbalance of functions and contexts in which progressives are used in real conversations and textbooks, use authentic instead of invented examples, and focus on frequent instead of rarely attested patterns" (Römer, 2010: 24).

An early attempt of applying corpus linguistics to course books was *Collins COBUILD English Course* (CCEC), designed by Willis and Willis (1989). It was a 'lexical syllabus' focusing on "the commonest words and phrases in English and their meanings" (Willis, 1990: 124). Another pioneering and promising work was *Touchstone* series published by Cambridge University Press (McCarthy, McCarten, & Sandiford, 2005). This series of corpora-based EFL textbooks have incorporated research findings from the *Cambridge International Corpora*, and "present(ed) the vocabulary, grammar, and functions students need for effective conversations".

Based on the investigation into ICLE, Kaszubski (1998) made recommendations for the traditional writing textbooks used in Poland by providing specific information as below (*ibid.*: 183):

a. longer lists of synonymous items, accompanied with frequency band information, register/style description, and (gradable) overuse/underuse/misuse warnings (if applicable). In cases of misuse, Polish and NS contrasting samples could be given;

b. […] lists of common collocations, with additional information on contrasts between Polish and NS use;

c. listings of commonly misused words and phrases as well as examples of serious over- and underuse.

These suggestions are not only applicable to textbook writers in Poland, but also useful for textbook writers from other countries.

3.2 Classroom Teaching Methodology

As for the teaching methodology, data-driven learning (DDL) has been highly recommended by many researchers (e.g. Cobb, 1997; Johns, 2002; Johns & King, 1991). It mainly refers to "the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language and the development of activities and exercises" (Johns & King, 1991: iii). It is an inductive approach relying on an "ability to see patterning in the target language and to form generalisations" about language form and use (Johns, 1991: 2). DDL is characterized by great emphasis on the fields of lexis and lexico-grammar of the activities, and the idea that learners should be exposed to as much authentic native speaker data as possible. Johns (2002: 108) sees DDL as a process which "confronts the leaner as directly as possible with the data" "to make the learner a linguistic researcher".

Other advantages of DDL also include that learners may have access to the errors they have made and to what is correct and valid; DDL activities reinforce negotiation, interactivity and interaction (Meunier, 2002: 134). Cobb (1997) did a longitudinal study of vocabulary acquisition by means of concordance line tasks drawn from a specially designed corpus, and showed positive effect of DDL activities on L2 learning.

However, different voices have called this method into question. For example, it is time-consuming to design DDL activities; it requires a considerable amount of preparation on the part of teachers; various types of strategies may cause confusion or problems for students; and many researchers doubt the role of DDL in low-proficiency learners.

In spite of different perspectives toward DDL, it is commonly believed that DDL can be used wisely to facilitate language teaching in the classroom by raising language awareness (Hawkins 1984) and self-discovery. Among various methodologies, concordance-based exercises have been proved to be an effective complement to traditional teaching strategies (Granger, 2002, 2009; Meunier, 2002).

In terms of concordance-based exercises, not only native data concordancing, but also comparison between learner and native speaker data can be useful methods. As Nesselhauf (2004: 140) suggests, one of the advantages of using such comparison is "that asking learners to look for mistakes, or rather for differences in learner and native speaker language, can increase learner autonomy and train the learners' general ability to notice such differences. In addition, such a procedure might also lead to a more positive attitude towards mistakes, because mistakes are then no longer merely a feature that has to be corrected, but also a feature that can be discovered". Nesselhauf also calls for more empirical studies to investigate such issues as"for which areas, for which learners and with what procedures data-driven learning with learner corpora is most efficient" (*ibid.*: 144).

4. Discussion and Conclusion

The paper has given an overview of learner corpora and their application to second language learning and teaching. It can be seen that learner corpora can play an important role in second language learning research, and be of great use to teaching materials design and classroom teaching.

Exercises from the data-driven learning approach gives learners access to authentic language samples accompanied by rich contexts. Learners can do the exploration of the use of words and phrases under the

guidance of teachers, and become increasingly aware of native language use through better noticing. Considering a wide variation in terms of aptitude, motivation, cognitive style, and other factors among different learners, teachers should treat DDL exercises with due caution.

As Granger and Tribble (1998: 209) suggest, "concordances need to be carefully edited to help learners find the relevant features. If vast quantities of information is thrown at learners, there is a considerable risk that DDL activities can become time-consuming and frustrating for learners." Furthermore, concordance-based exercises "are by no means a replacement for, but could be viewed as complementary to, the traditional, continuous cloze passages, learning of vocabulary through semantic fields, analysis of common roots etc." (Packard, 1994: 221-222). Despite the word of caution, it still remains as an important task for corpus linguistics researchers to design more various types of DDL materials "that address particular language items (especially items which cause constant problems for learners) and that could be used directly in the EFL classroom" (Römer, 2009: 91).

Acknowledgements

The study is supported by Innovative School Project in Higher Education of Guangdong, China (GWTP-BS-2015-20).

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.24

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173-195. http://dx.doi.org/10.1093/applin/22.2.173

Barlow, M. (1996). Corpora for theory and practice. *International Journal of Corpus Linguistics, 1*(1), 1-37. http://dx.doi.org/10.1075/ijcl.1.1.03bar

Biber, D., & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clause. In S. Granger (Ed.), *Learner English on computer* (pp. 145-158). London; New York: Longman.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System, 25*(3), 301-315. http://dx.doi.org/10.1016/s0346-251x(97)00024-9

Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 67-85). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.12.08con

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

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; New York: Longman.

Grabowski, E., and Mindt, D. (1995). A corpus-based learning list of irregular verbs in English. *ICAME Journal, 19*, 5-22

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; New York: Longman.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3-36). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/lllt.6

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In A. Karin (Ed.), *Corpora and Language Teaching* (pp. 13-32). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.33.04gra

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; New York: Longman.

Guo, X. (2006). *Verbs in the Written English of Chinese Learners: A Corpus-based Comparison between Non-native Speakers and Native Speakers.* Unpublished PhD Thesis, The University of Birmingham, Birmingham.

Hawkins, E. (1984). *Awareness of Language: An Introduction*. Cambridge: Cambridge University Press.

Housen, A. (2002). A corpus-based study of the L2-acquisition of the English verb system. In S. Granger, J.

Hung, & S. Petch-Tyson. (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 77-118). Amsterdam: John Benjamins.

Johns, T. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria, 10*, 14-34. http://dx.doi.org/10.1017/cbo9781139524605.014

Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettemann, & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Linguistics* (pp. 107-117). Amsterdam: Rodopi.

Johns, T., & King, P. (Eds.). (1991). *Classroom Concordancing* (Vol. 4). Birmingham: 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; New York: Longman.

Keck, C. (2004). Corpus linguistics and language teaching research: Bridging the gap. *Language Teaching Research, 8*(1), 83-109. http://dx.doi.org/10.1191/1362168804lr135ra

Leech, G. (2001). The role of frequency in ELT: New corpus evidence brings a re-appraisal. *Foreign language Teaching and Research, 33*(5), 328-339

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone Student's Book 1*. Cambridge: Cambridge University Press.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 119-142). Amsterdam/Philadelphia: John Benjamins. http://dx.doi.org/10.1075/lllt.6.10meu

Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 40-50). London: Longman.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research, 21*(4), 373-391. http://dx.doi.org/10.1191/0267658305sr252oa

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. http://dx.doi.org/10.1075/scl.17.08nes

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.14

O'Keeffe, A., McCarthy, M., & Carter, R. (Eds.). (2007). *From Corpus to Classroom: Language Use and Lnaguage Teaching*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511497650

Packard, V. (1994). Producing a concordance-based self-access vocabulary package: Some problems and solutions. In L. Flowerdew, & A. Tong (Eds.), *Entering Text* (pp. 215-226). Hong Kong: The University of Hong Kong Science and Technology.

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal, 26*, 81-114

Römer, U. (2004). A corpus-driven approach to modal auxiliaries and their diacritics. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 185-199). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.12.14rom

Römer, U. (2005). *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.18

Römer, U. (2009). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 83-98). Amsterdam; Philadelphia: John Benjamins. http://dx.doi.org/10.1075/scl.33.09rom

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. C. Campoy-Cubillo, B. Belles-Fortuno and M. L. Gea-Valor (Eds.), *Corpus-based Approaches to English Language Teaching* (pp. 18-38). London: Continuum.

Schachter, J., & Celce-Murcia, M. (1977). Some reservations concerning error analysis. *TESOL Quarterly, 11*(4), 441-451. http://dx.doi.org/10.2307/3585740

Willis, D. (1990). *The Lexical Syllabus: A New Approach to Language Teaching*. London: HarperCollins.

Willis, D., & Willis, J. (1989). *Collins COBUILD English Course*. London: HarperCollins.

Xiao, Z. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian Journal of English Language Teaching, 3*(2), 1-23