CSCanada
Studies in Sociology of Science
Vol. 6, No. 6, 2015, pp. 78-82
DOI:10.3968/8078

ISSN 1923-0176 [Print]
ISSN 1923-0184 [Online]
www.cscanada.net
www.cscanada.org

The Application of Corpus in English Writing and Its Influences

ZHAO Lili[a],*

[a]Department of Foreign Language Teaching, Inner Mongolia University for the Nationalities, Tongliao, China.
*Corresponding author.

Supported by the Scientific Research Project of Inner Mongolia University for the Nationalities "a Survey of Mongolian Students' English Writing Based on Corpus Resources" (NMD1366).

Received 6 October 2015; accepted 16 December 2015
Published online 26 December 2015

Abstract
Corpus plays a very important role in English teaching and learning. This paper undertakes an empirical study on the errors made by Mongolian learners in their English writing based on the resources in the corpus. The paper first gives a brief introduction of corpus, and then gives illustration about the relevant theory of the research. Afterwards, with the purpose of finding out the rules of these errors and exploring the causes of them, a study is carried out based on the automated scoring system provided by corpus. In the next section, a discussion about the strengths and weaknesses of corpus is made hoping to improve the service of corpus. Finally, the author attempts to make predictions about the trends of the corpus-based approach to English writing.

Key words: Corpus; English writing; Data-driven learning; Prospect

Zhao, L. L. (2015). The Application of Corpus in English Writing and Its Influences. *Studies in Sociology of Science, 6*(6), 78-82. Available from: URL: http://www.cscanada.net/index.php/sss/article/view/8078
DOI: http://dx.doi.org/10.3968/8078

INTRODUCTION

Writing is probably the linguistic skill that is least used by most people in their native language. In the universities, most of the students have difficulty in writing a composition. This is particularly true for the Mongolian students. Influenced by Mongolian language and Chinese, Mongolian students find it really hard to write well. The negative language transfer often confuses them a lot. Mongolian students are inclined to make a variety of errors in their writing. They lack in sufficient writing skills to finish a good composition. Lado, in his book *Linguistics Across Cultures* notes:

Individuals tend to transfer the forms and meanings, and the distribution of forms and meanings of their native language and culture to the foreign language and culture both productively when attempting to speak the language and to act in the culture, and respectively when attempting to grasp and understand the language and the culture as practiced by natives.

The main way of investigating L2 acquisition is by collecting and describing samples of learner language . The description may focus on the kinds of errors learners make and how these errors change over time, or it may identify developmental patterns by describing the stages in the acquisition of particular grammatical features such as past tense, or it may examine the variability found in learner language.

Different from the traditional way of examining students' writing abilities, the author uses the data collected in corpus to evaluate students' compositions. Corpus is a collection of writings, conversations, speeches, etc., that people use to study and describe a language. One of the most popular corpora of English writing in China is Juku automated scoring systems. It has strong functions in making an analysis of students' writing. It can grade students' writing automatically and point out the common mistakes in students' writing after they hand in their compositions online. Corpus offers rich online language resources and is a good means for researchers to study the linguistic performance of other people. In the past few years, corpus has been developed quickly .A variety of corpora arises both at home and abroad. Corpus

and corpus linguistics have great influence on English writing and research. In this paper, special attention is paid on the investigation of students' writing status based on the resources offered by the corpus. Some theories concerning corpus and writing are mentioned in the following parts, such as: constructivism learning theory, data-driven learning, and process writing. With a case study of the application of corpus in Mongolian students' writing assignments, the author analyzes the common problems in students' written work and then makes a conclusion. After that, a further discussion of the strengths and weaknesses of corpus in the application is made and the prospect of corpus in future use is predicted.

1. LITERATURE REVIEW

1.1 An Introduction of Corpus

A "corpus" is a large collection or database of machine-readable texts involving natural discourse in diverse contexts (Bernardini, 2000). With the expectation that facilitation of the writing process would lead to more and better writing, a large body of research has focused on the number of words produced, students' motivation to write, and the assessed quality of the writing when examining the use of computers to teach writing (e.g. Berninger et al., 2009; Goldberg et al., 2003; Owston & Wideman, 2001). Corpus is a collection of materials that has been made for a particular purpose, such as a set of textbooks which are being analyzed and compared or a sample of sentences or utterances which are being analyzed for their linguistic features. (Richard, Platt, & Platt, 2000, p.110).

There are two major advantages to use of text corpora for linguistic analysis (Biber et al., 1994):

a) They provide a large empirical database of natural discourse, so that analyses are based on naturally occurring structures and patterns of use rather than intuitions and perceptions,· which often do not accurately represent actual use.

b) They enable analyses of a scope not feasible otherwise, allowing researchers to address issues that were previously intractable. This is particularly true of computer-based text corpora, which can be analyzed using (semi-) automatic techniques. Such analyses can examine much more language data than otherwise possible, including more texts, longer texts, a wider range of linguistic characteristics, and the systematic co-occurrence patterns among linguistic features. In addition to quantitative analyses previously not possible, corpus-based approaches thus allow investigation of issues such as register variation and discourse factors influencing the choice among structurally related variants (e.g. adverb placement, active vs. passive, etc.).

Some readily available corpora include the British National Corpus (BNC,http://www.natcorp.ox.ac.uk), which contains 100 million words from written and spoken language in a variety of contexts, the Michigan Corpus of Academic Spoken English (MICASE, http://micase.elicorpora.info), which features 1.8 million words of speech in various academic contexts, and The Corpus of Contemporary American English (COCA), with 410 million words(http://www.americancorpus.org).

Though corpus linguistics is a new emerging discipline, it has received more attention in second language teaching and learning. The teacher usually regards corpus data as resources that provide descriptive insights relevant to how people use language and as tools that enable students and instructors to analyze both how people use different language forms at various levels of formality and how language fulfills multiple speech functions across contexts. Corpus data suggest that individuals often do not use language as specified in grammar books and that word meanings vary across contexts and users (Biber & Reppen, 2002).

1.2 Constructivism Learning Theory

Constructivism learning theory is a philosophy which enhances students' logical and conceptual growth. The underlying concept within the constructivism learning theory is the role which experiences-or connections with the adjoining atmosphere-play in student education. The constructivism learning theory argues that people produce knowledge and form meaning based upon their experiences. Two of the key concepts within the constructivism learning theory which creates the construction of an individual's new knowledge are accommodation and assimilation. Assimilating causes an individual to incorporate new experiences into the old experiences. This causes the individual to develop new outlooks, rethink what were once misunderstandings, and evaluate what is important, ultimately altering their perceptions. Accommodation, on the other hand, is reframing the world and new experiences into the mental capacity already present. Individuals conceive a particular fashion in which the world operates. When things do not operate within that context, they must accommodate and reframing the expectations with the outcomes. The role of teachers is very important within the constructivism learning theory. Instead of giving a lecture the teachers in this theory function as facilitators whose role is to aid the student when it comes to their own understanding. This takes away focus from the teacher and lecture and puts it upon the student and their learning. The resources and lesson plans that must be initiated for this learning theory take a very different approach toward traditional learning as well. Instead of having the students relying on someone else's information and accepting it as truth, the constructivism learning theory supports that students should be exposed to data, primary sources, and the ability to interact

with other students so that they can learn from the incorporation of their experiences. (http://www.teach-nology.com/currenttrends/constructivism/)

1.3 Data-Driven Learning

A key pedagogical approach focusing corpora in language teaching and learning is‘ data-driven learning’(DDL), which emerged in the mid-1980s. Johns (1991a) writes that the “ . . . language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data -- hence the term ‘data-driven learning’ (DDL) to describe the approach”. Data-driven learning studies vast databases of English text (corpora) with software programs called concordancers, which isolate common patterns in authentic language samples. The approach also draws from product teaching in that it provides authentic language material for study.Data-driven learning brings to the class abundant examples of authentic language samples that can be studied and exploited in many ways.

2. METHODOLOGY

2.1 Participants

The participants are Mongolian students from Inner Mongolia University for the Nationalities. They are ethnic minority students of preparatory education. They are clinical medicine majors. 72 students are assigned the writing task. Some of them are language beginners and have just learned English for only one or two years before. Quite a few of the students have learned English in middle schools and they can communicate with others in simple English. Many students have difficulty in writing English well.

2.2 Research Design

The research is designed for the purpose of finding out the influences of corpus in the process of writing. A writing task with the title *On the Low-Carbon Life* is assigned by the teacher to the students with the purpose of investigating the main problems that students make in their English writing .With the samples provided in the corpus , the teacher intends to make an analysis of the errors students have made and find the reason. Then the teacher tries to get a conclusion.

2.3 Instruments

The instrument is the Juku automated scoring system which has a strong function of collecting data and analyzing data. It is one of the famous learners’ corpora in China to help second language learners to write English well. It is a good assistant for the teacher and it in some way plays an important role in students’ writing process. It can mark the writing samples automatically after the students hand in their compositions. Contrastive analysis and error analysis are employed in the system to help the teacher get a clear clue of students’ writing status. Examples of errors and forms of the statistics of students’ average grades are shown in the corpus. The tables concerning students’ common errors and similarity statistics are presented in the system as well.

2.4 Procedure

The survey is made in the following steps.

(a) The teacher arranges writing assignments in Juku automated scoring system. Writing requirements and the deadline of composition submission are made clear in the automated scoring system. Because of the rich linguistic resources online, no copy or stickup is allowed in the writing process.

(b) Students are firstly asked to sign up in Juku automated scoring system .Then students are required to finish the writing task before the deadline. After students upload their samples of composition, their will get the scores graded by Juku automated scoring system. The teacher makes his own comments and evaluation according to the data and scores provided by Juku automated scoring system.

2.5 Data Analysis

The average score they get from Juku automated scoring system is 71.9. The highest mark is 83.6, the lowest mark is 48.1. There are altogether 379 mistakes found in students’ compositions. The most common mistakes are Misplaced Modifier, Sentence Fragments, Dangling Modifiers, Misuse of Parts of Speech, Troubles in Diction, Redundancy, and Incoherence. The following Figure illustrates clearly the problems that exist in students’ writing. Look at the Figure below.

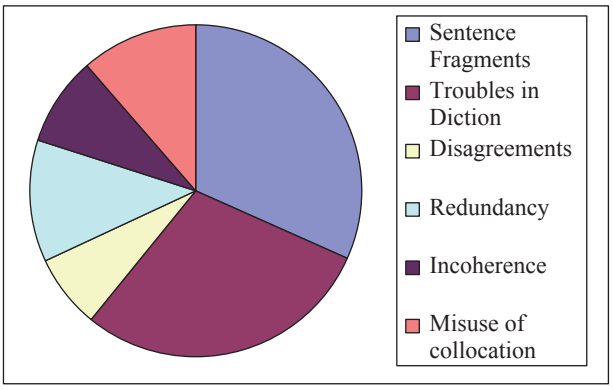

Figure 1
The Errors in Students' Writing

3. RESULTS AND DISCUSSIONS

The above data is provided by Juku automated scoring system. Juku automated scoring system is an online service system to revise students’ writing samples. It is

one of the largest and famous corpora of writing in China. It can effectively reduce the workload of the teacher and improve students' writing abilities. Compared with the traditional way of manually marking , automated scoring system has the following features:

a) It can automatically recognize the simple errors in the written work and give suggestions for revision.

b) It can give immediate reaction to students' submission and mark the composition.

c) There is no need to edit the errors repeatedly because the system has stored the data and the prior experiences of the teacher to avoid doing the same work for several times.

d) It points out the lexical, grammatical and syntactic errors directly and gives corresponding advices of revision. In addition, the system provides specific knowledge for students' training and good examples are recommended. With a large amount of language input , the system provides students with comprehensive writing guidance to help students to learn independently.

e) Plagiarism detection can help the teacher check students' composition for plagiarism. The system will take advantage of the data and resources it has stored to check if there are repeated contents, which serves as a very useful reference for the teacher.

f) The system offers progress report of students' writing abilities. It is of beneficial use for the teacher to grasp the overall vulnerabilities and individual vulnerabilities of the students and get to know how well students have mastered the knowledge. It facilitates the teacher with data support.

4. THE STRENGTHS AND WEAKNESSES OF CORPUS

As has mentioned above, the corpus has its unique features different from the traditional way of grading writing. Computerized corpus linguistics continues to have a profound effect on language studies, both descriptive and applied. It is natural that there should also be interested in carrying out the computational analysis of learner corpora, of English, to which we now turn. We shall consider two projects on the computer analysis of learner language, with particular attention being paid to their potential relevance to error description and classification (James, 2001). From the analysis in the prior part, the strengths and weakness of corpus can be demonstrated in the following section.

4.1 The Strengths of Corpus

The advantages of making use of corpus as a means to do some research work in writing are as follows. The advantage of this method is that it is objective and reliable. Corpus can provide vast amount of easy access and real data for language teaching. It provides learners autonomy and increase their own interests in reading corpora. It helps to develop students' learning autonomy and students can get immediate feedback on their written work. What's more, students are able to formulate language rules themselves and avoid making some errors according to the marks and comments shown in the automated scoring system.

4.2 The Weaknesses of Corpus

Though corpus has offered writers rich language resources, it has its disadvantages. The Mechanical Accuracy method looks only at the form of what is written. Using this method the test writers decide how many marks will be given for correct punctuation, spelling and grammar. The scorers then go through the scripts, deducing one mark from the total for each error made. It may be technically challenging and time consuming for the teachers. In order to make full use of the resources in corpus, the teacher needs to be equipped with advanced information technologies. And the analysis and statistics provided in corpus are only references for the teacher. It can not replace the role of the teacher. Though the corpus provides authentic materials and data, it can not check out whether the composition has digressed from the subject. The disadvantages of the corpus are obvious. It ignores the content of what has been written. Whether what has been written is understandable or not, and whether it is relevant to the topic is not considered. In addition, language tests which mark using this method encourage students and teachers to view accuracy of form as more important than the ability to express meaning in written language. Furthermore, in any circumstances, the teacher's grades are more reliable than the scores marked by a machine. Though the automated scoring system may mark a written work as quickly as possible, it can't tell if a composition has strayed from the point. The automated scoring system can't recognize some typical Chinglish expressions as well. Moreover, the corpus is in need of an index for the teacher to use easily and conveniently.

5. THE PROSPECT OF THE DEVELOPMENT OF CORPUS

Since corpus-based method relies on the extensive use of computer-aided software, the exploration and analysis of language features in a great number of texts is efficient and reliable. If corpus-based analysis is employed, the text corpora can provide an empirical database of natural discourse and assist researchers to examine more texts than would otherwise be possible with (semi-) automatic techniques (Biber, Conrad, & Reppen, 1994). In the future, spoken language corpus and bilingual corpus will

play a more and more important role. Spoken language corpus can provide people with more information, and can reveal the inherent characteristics of the real communicative language and rules. It is imperative to establish spoken language corpus because it collects basic information for further study of discourse analysis. Linguistic corpus counts for translation research and the training of translators (Cui & Wang, 2013). The main focus of research on corpus will shift from the setting-up of corpus to the application of corpus. Meanwhile, foreign language teachers should participate actively in the development and application of learners' corpus and communicate actively with each other about the advanced scientific research achievements. The corpus will make great contribution to English writing if we make full use of the resources in it.

CONCLUSION

Corpus provides rich linguistic data for the researcher to evaluate the general standard of learners, and make an analysis of the language transfer of mother tongue. The development of corpus is a reference to all kinds of English tests and essay scoring. The automated scoring system developed by the technology of corpus, has improved the efficiency of writing to a large extent. In the present data-driven learning circumstances, research on the influences of corpus on English writing is gaining more and more attention. How to use corpus appropriately is of upmost importance for the teacher and students. In addition to necessary technical skills, the teacher and the students should take advantage of its powerful functions and abandon its disadvantages. The study of corpus remains a project that needs to be further explored, which is worthwhile and useful for most of the researchers investigating in the field of writing.

REFERENCES

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective*. Frankfurtam Main: PeterLang.

Berninger, V., Abbott, R. D., Augsberger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123-141.

Biber, D., Conrad, S., & Reppen, R.. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics, 15*(2).

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching ? *Studies in Second Language Acquisition, 24,*199–208.

Cui, W. C., & Wang, J. S. (2013). The development and trend of discipline-specific corpus in China. *Journal of Xi'an International Studies University, 21*(1), 55-58.

Constructivism Learning Theory. (n.d.). Retrieved from http://www.teach-nology.com/currenttrends/constructivism/

James, C. (2001) *Errors in language learning and use: Exploring error analysis* (p.124). Beijing: Foreign Language Teaching and Research Press.

Johns, T. (1991a). Should you be persuaded: Two examples of data-driven learning. In T. F. Johns & P. King (Eds.), *Classroom concordancing* (pp.1-13). Birmingham: ELR.

Lado, R. (1957) *Linguistics across cultures applied linguistics for language teachers.* Ann Arbor, Michigan: University of Michigan

Long, M. (1988). Instructed interlanguage development. In L. Beebe (Ed.), *Issues in second language acquisition: Multiple perspectives*. Rowley, Mass.: Newbury House.

Richards, J., Platt, C. J., & Platt, H. (2000). *Longman dictionary of language teaching applied linguistics.* Beijing: Foreign Language Teaching and Research Press.

