Using Medical Academic English Corpus for Postgraduates Students Academic Writing Training

Feng Zhang
Binzhou Medical University, Yantai, China

Yuanhua Zheng
Binzhou Medical University, Yantai, China

Li Li
Binzhou Medical University, Yantai, China

Abstract—This paper reports on the use of personal do-it-yourself medical academic corpora by postgraduate students for academic writing. Thirty-five graduate students in medicine attended our course in which they learn to build and use their self-compiled corpora of research articles to train themselves in academic writing. At the end of the course, they completed some questionnaires, which asked about their use of corpora in and after class, and six months after the course, they completed some similar email questionnaires, which asked about their corpus use six months later. Some personal interviews were also conducted for users and nonusers. This paper also discusses reasons behind this change in the use of DIY corpora, and advocate some measures should be taken to encourage the long-term use of corpora and some of the challenges to be overcome for more widely use of the approach in the future.

Index Terms—do-it-yourself corpora, English for medical purposes, academic writing training

I. INTRODUCTION

The use of corpora for language teaching and learning has become a trend in the past decades, and some particular focus has been on the field of academic writing in English at university level (Boulton, 2010; Yoon, 2011; Hyland, 2005; Hunston & Thompson, 2000; Mark, 2013). The corpora used can be roughly classified as three types: online or locally installed large general corpora, such as British National Corpus (BNC) and online BYU corpora; medium specialized ones, usually only some individual course or discipline involved; small do-it-yourself (DIY) corpora, constructed by researchers or students for their personal use. During their research, they wanted to know, to what extent, and under what circumstances students can get involved in the corpus data when learning to write their academic papers. To date, most of the focus has been on large general corpus, but there is a growing interest in the use of specialized and DIY corpora.

The application of general corpora to teaching and learning of grammatical and lexical items involves both quantitative and qualitative approaches. Some investigated whether the employment of corpora consultation had a demonstrable effect on learning of lexical and grammatical problems (Boulton, 2010; Creswell, 2007). Other studies suggested that most of the applicants mastered the techniques of using corpora to solve language problems (Gaskell & Cobb, 2004; Gilmore, 2009; Todd, 2001). And some qualitative researches showed that most of the students being interviewed had positive attitudes towards the use of corpora in learning and teaching (Granath, 2009; Varley, 2009), and the results also indicated that their English proficiency, training and support available to some extent were decisive factors to their enthusiasm of using corpora.

Most of the studies are mainly paying their attention on student attitudes or the evaluation of corpus work in or immediately after the classes, though there are some focusing on long-term use (Charles, 2014; Yoon, 2008), the aim of the present study mainly focus on the comparison of immediate use and long-term use to find out why the student enthusiasm of using corpora to learn academic writing when firstly involved in this course vanished as compared to the low usage of long-term investigation.

II. BACKGROUND

The research is based on the data of the corpora-based academic writing project integrated into their Academic English Course in 2015 and 2016 of first-year postgraduate students and their online reports on their corpus use six-month later. The project is designed for first-year postgraduate medical students to improve and revise their academic paper writing with the help of consultation of personal do-it-yourself corpora. Though the project is integrated...
into the Academic English Course, it is by itself a non-assessed and open-access project. Groups of 20 to 23 postgraduate students are trained in a multi-disciplinary class every year, and the work of training students to use corpora lasted over eight weeks, with one two-hour session every week. Students are required to take their own laptop computers with them and all the corpora work and software were done or installed on their own computers.

**The Project**

The project is designed to meet postgraduates’ urgent needs of writing academically acceptable papers by offering them the chance to compile do-it-yourself corpora in their own discipline and learn to practice the skill of retrieving and interpreting the information their personal corpora can provide. In and after class, participants are trained to work with their own corpora to find answers for their lexical and grammatical queries and explore discourse issues.

The project is divided into three sessions. In the first session, we offer some fundamental basics about corpora and popular corpora applications in language teaching and learning, which helps the participants understand what is corpus and what corpus can do for us, and then we show the participants how to install and use the freely available concordancer AntConc (Anthony, 2014), which is a popular corpus software to consult a corpus for dealing with lexical and grammatical queries. The first session is focused on the general understanding of corpora and getting familiarized with corpus software.

In the second session, participants are required to compile their own personal corpus with the guidance of instructors. Firstly, they are expected to select at least 100 research articles (all in PDF editable file) from important academic journals representing the academic performance of their fields, 50 of which are written by native speakers and 50 by non-native speakers. Secondly, the participants are shown how to convert the PDF file into plain text files, and clean the texts by removing parts that is not the running text, such as, the references part, graphical elements, etc. This procedure makes it easier to read concordance lines and renders the statistical data more reliable.

The third session is focused on the analysis of information that corpus data provide. Each corpus software function has some linguistic purposes to perform. Each step provides detailed questions which ask participants to explore a specific discourse function in their personal corpus. For example, in argumentation part, participants use the Word List tool to explore the frequency and sentence position of linking adverbials such as “however” in their own field.

By the end of the project, participants have mastered the use of concordancing and other functions of the software, and they know how to use Word List to examine words in their corpus, and use Collocates and Clusters to retrieve collocations. Some participants have become proficient at interpreting corpora data, and all have achieved the basic competence in using corpora to solve lexical and grammatical problems.

One of the stated aims of this project is to provide students with a custom-built resource which will be available throughout their future academic career, and with the know-how and skills to use it effectively. The rationale for building a personal DIY corpus is that its use will help students to move away from an over-reliance on outside agencies such as teachers and proofreaders and to become more independent and autonomous in their production of academic texts.

### III. **Methodology**

The background details of participants were collected at the beginning of the course; the size data of the participants’ DIY corpora were collected when participants finish compiling their DIY corpora. A questionnaire about their attitudes towards corpora use in academic writing training and their own performance in this course was conducted immediately after the course was finished. And a follow-up questionnaire was conducted by email and QQ approximately six months after completion of the course, which was to investigate participants’ corpus use during the preceding 6 months, if their attitudes toward corpus had changed comparing with that of six months ago, and the reason why they still use corpora or why they gave up using corpora. Three personal interviews were followed after the questionnaire. The surveys consisted of 20 closed and open questions, which were sent to participants, completed and returned electronically. Where necessary, there was email or QQ correspondence to clarify or amplify the responses from the participants.

The participants are first year postgraduate students from a provincial medical university. The project was conducted in Grade 2015 and Grade 2016 following the same procedure. The participants were all volunteered to join the project and data from 35 valid participants were finally collected, 16 of 2015 and 19 of 2016. Their academic disciplines include: Anatomy, Biochemistry, Biomechanics, Biostatistics, Cytology, Embryology, Genetics, Histology, Immunology, Microbiology, Molecular biology, Pharmacology, Physiology, Toxicology. As there were no substantial differences between the two courses, the data from the two year-groups were combined, giving a total of 35 sets of data.

The personal DIY corpora were compiled by the participants, which was roughly divided into two sub-corpora, one for native writers and another for non-native writers. Every participant selected at least 20 papers from several important journals representing the academic performance of each medical field, 10 for each sub-corpus. To avoid that several participants may select the same paper, the monitors of the two groups coordinated the papers participants selected. If the average length of each paper, after cleaning or partial cleaning, is about 6,000 words, so every participant will have a corpus of more than 120,000 words. And then we collected all the participants’ corpora to form a bigger one, so in total we have a medical academic English corpus of 4.2 million words. Though it may not be a balanced corpus, we can still use it to conduct lots of research.
IV. Results and Discussion

A. Participants

The participants’ personal information showed that 25 participants (71.43%) were female and 10 (28.57%) were male. Participants studied in 13 different disciplines: Anatomy (3, 9%), Biochemistry (2, 6%), Biomechanics (2, 6%), Biostatistics (1, 3%), Cytology (2, 6%), Embryology (2, 6%), Genetics (4, 11%), Histology (2, 6%), Immunology (5, 14%), Microbiology (3, 9%), Molecular biology (4, 11%), Pharmacology (2, 6%), Physiology (2, 6%), Toxicology (1, 3%). (See Fig 1)

B. Size of the DIY Corpora

The size of the DIY corpora participants compiled is varied according to their fields they worked with and the journals they selected. They selected five or six journals in their own research fields, and there are 13 disciplines and roughly 70 peer-viewed world-famous journals they used. The publishing time ranged from 2008 to 2016. (See Fig 2)

The participants chose the journals and selected research articles and converted to text format, optionally cleaned and added to the corpus individually. Because of large disciplinary differences in the length of research articles, the size of the DIY corpora is not normalized, but all the compiled corpora have exceeded the required size as we planned. The largest was constructed in the field of Immunology (570,383), because it held the largest share in participants (5, 14%), by contrast, the smallest corpus in number of words was in Toxicology (126,900), because we have only one participant in this field. The average size of the research article was 6,223 words, and the average size of participants’ DIY corpora was 309,413 words. (See Fig 3)
C. Use of the Personal Corpus

Questionnaires were conducted immediately after the course was completed and six-month after the course. The immediate survey showed that 94.2% (33 out of 35) of the participants had mastered and used their self-compiled corpora in and after class, and most of the users consulted their corpora for checking lexical collocation or grammatical usage while writing and revising their papers, and 85.7% (30 out of 35) held very active and positive attitude toward the use of corpora to train their writing and 95% thought it was helpful and corpus use had improved their academic writing. But, the six-month-later follow-up questionnaire asked whether students had used their personal corpus at any time since the course ended, the results showed that only 5.7% were regular users (2 out of 35, once or more every week), 14.3% irregular users (5 out of 35, once every month or seldom) and 80% non-users. (See Fig 4)

![Fig 4 the use of corpora](image)

This immediate survey result, which is highly encouraging, suggests that after a relatively short period of training, participants can compile and use their corpora independently in the absence of further input or help from a corpus specialist. But the six-month-later one was not encouraging, after further interviewing with the participants, we learnt that non-users did not necessarily reject corpus use; they may simply not be at a stage of their work where the corpus would be useful, a point also made by Yoon (2011) and Charles (2014). And some non-users said they intended to use their DIY corpora when they began to write their papers, when the time the questionnaires were conducted, those medical postgraduates were doing their experiments.

D. Purpose of Use

To explore how personal DIY corpora were incorporated into their writing practice, participants were interviewed how they use their corpora. The survey suggests that most of the participants were consulting corpora for lexical and grammatical problems. When participants wanted to know expressions of certain meanings, or not sure the usage of some words or grammatical structures, they turned to corpora for help. When they wrote some papers, they looked up something in the dictionary and found several similar expressions and then they came to corpora to compare the contexts of each expression and found the best ones.

Such responses are consistent with those reported in earlier research (Charles, 2015; Yoon, 2011), and indicate that participants usually consult corpora for sentence level difficulties and addressing lexical and grammatical concerns is participants’ most pressing need (Lee & Swales, 2006). However, as the participants had relatively little experience with corpus techniques and tools, lexical and grammatical problems were easy to tackle and was the first step to access more complicated discourse ones.

E. Problems of Personal DIY Corpora

There were some problems participants encountered when using DIY corpora. These problems or potential disadvantages severely compromised the enthusiasm of using corpora to help writing and suggested us ways to improve the popularity of corpora use.

1. Accessibility

Participants complained of corpora software lacked of accessibility. It is not very convenient to use, installation has to be performed when changing computers or something wrong with the local computer. The software is not very easy to use, and you have to take tutorials or training courses to learn how to use it. The processing speed of the software varies according to different configuration of the computer, that is, if you want high speed, the configuration parameters of your computer must be taken into consideration. One possible solution is to set up a “web-based interface that allowed the personal corpus to be stored in the ‘cloud’ and consulted online from any computer” (Charles, 2015). Another solution is to use the BYU online corpora registered version, but there are more tutorials to watch and if you want the full function one you have to pay some money. But if the web-based online corpora are not that convenient in that the processing speed is always decided by the internet speed and computing compacity of the servers.

2. Size

Size is always the concern of corpus linguist, as John Sinclair puts it “small is not beautiful”, “texts are so different as
you put a lot together” (Sinclair, 2004). Some participants raised the same concerns about the small size of their DIY corpora. As we all know, general corpora should be bigger than ESP corpora, while specialized corpora can be comparatively small, but some of our DIY personal corpora were undoubtedly too small to be representative and unable to address certain kinds of problems. The solution to the size problem is to encourage participants to form a habit of adding new articles to their corpora whenever they read something new and useful. To add one article to corpus is just like to store a doc to a folder, not too much manual work to do.

One point worth mentioning is that if we want our participants to form a habit of using corpora as consulting tool, we have to let them feel the usefulness of corpora, that is, help them to see the “beauty of corpora”. As ESP learners, most of their time have to be dedicated to the own disciplinary knowledge learning, so they have limited time available to learn English writing and even limited time to learn to use corpora and undertake substantial corpus construction, so “they need to be convinced that the utility of the resource justifies the time taken to build it” (Charles, 2015). Therefore, when setting up teaching plans, it is necessary to devise tasks that offer opportunities to tackle problems that are easy to access and can meet their most pressing needs even with very small corpora. Working in groups also serves as a good solution, which can help participants share the load in finding research articles and cleaning the texts, and supplement each other with their individual findings.

3. Reliability

Some participants worried about the reliability which is another problem caused by small size. To some specialized corpora, the research results may be similar compared with results from large corpora, but some participants were still reluctant to trust their data from DIY corpora, even when it was large enough. Several reasons may contribute to this: firstly, participants are not confident enough to trust their own finding. As in Chinese education system, Chinese student are accustomed to trusting external authoritative sources, such as expert opinions or dictionaries. They are not used to trusting their own findings. It takes a longer time and some proof to develop confidence in their own finding, judgments and interpretations. Therefore, some confirmatory tests can be designed into the teaching plans to help build their confidence, that is, participants can use their own DIY corpora to testify certain conclusions that have been conducted and proved true by many researchers and teachers themselves.

Secondly, some papers published in even some high-profile journals are not always perfectly written. Because we will use the articles in the corpora as writing examples, participants are expected to choose good quality research articles from well-regarded journals that may provide appropriate lexical and grammatical evidence to meet their writing needs. The problem is that some journals may value ideas or disciplinary importance more than linguistic perfectness, and language is always not the first priority. So that is the reason why we will compile two sub-corpora: one is written by native speaker authors and another is written by non-native speaker authors. We can also consult the same queries in the two corpora and compare the results to see the differences, which may offer another perspective for the participants to avoid the mistakes the non-native speaker authors made in their research papers. Another solution is to enlarge the corpora. Select more native speaker author articles and technically decrease the percentage of the less perfect ones.

V. CONCLUSIONS AND FUTURE CHALLENGES

Results show that most of the participants had mastered how to use their self-compiled corpora for checking lexical collocation or grammatical problems while writing and revising their papers, and most of them held very active and positive attitude toward the use of corpora to improve their academic writing. Though six-months after the course, the frequent irregular users drop dramatically due to their disciplinary emphases and timing, most of the participants expressed their willingness to use DIY corpora in the future, and their confidence of using corpora independently to meet their language needs. It may be concluded that Participants have incorporated corpora tool into academic writing training and considered it a valuable tool, so it is a worthwhile undertaking to teach students to compile and consult DIY corpora with a brief introductory course. Furthermore, we can see that using corpora to teach academic writing is also a practical tool for individualized teaching, that is, there is no need for teachers to select different disciplinary materials for academic training, the students themselves will take the responsibility to do that, and the process itself is effective methods to improve academic reading and writing. We can cultivate more functions of corpora in language teaching and learning.

Based the individualized needs of the participants in the present study, some challenges entail in maximizing the potential of personal corpora. The irregularity of participants’ academic writing requirements raises issues concerned with the timing and provision of corpus courses. So the handling and the arrangement of the course can be various. We may hold one formal course in their first year and then we can set up refresher sessions, drop-in clinics, on-line, on-demand courses or other means of just-in-time support, which would be responsive to student needs as they arise.

Another one is concerned with people’s difficulty in installing and using the software. It is the challenge to software engineers. We hope we can just login the account online or in the cloud, and then we have everything we need and we used before. No need to worry about technical circumstances, and easy accessibility, friendly interfaces and timely support online will facilitate anyone who is interested in personal DIY corpora.

The present study is just the first step in the field of corpora aided language teaching and learning. More topics and perspectives will be considered for further research, and more challenges will be encountered. With our further research
in this field, the real beauty of DIY personal corpora will reveal themselves, and more students and people interested will benefit from the use of corpora.

REFERENCES


Feng Zhang, was born in Shandong, China in 1978. He is currently an associate professor in the School of International Studies, Binzhou Medical University, Yantai, China. His research interests include corpus linguistics and language teaching technology.

Yuanhua Zheng, Professor, Deputy Dean of the School of International Studies, Binzhou Medical University, Yantai, China. Her research interests include second language acquisition, language testing, cross-culture communication and corpus linguistics.

Li Li, was born in Shandong, China in 1980. She is currently lecturer in the School of International Studies, Binzhou Medical University, Yantai, China. Her research interests include second language acquisition, cross-culture communication and corpus linguistics.