ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 7, No. 10, pp. 868-873, October 2017
DOI: http://dx.doi.org/10.17507/tpls.0710.07

Using Medical Academic English Corpus for Graduates 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*—In this paper, we discussed the use of personal do-it-yourself (DIY) corpora by medical graduate students for academic writing. Thirty-five graduate students in internal medicine school at a Chinese medical university attended this course in which they learn to build and use the self-compiled corpora of research articles to train themselves in academic writing. At the end of the course, they were asked to complete questionnaires and attend interviews about their habits of using DIY corpora in and after class, and then a follow-up questionnaire was administered. This paper also investigated possible reasons of habit changes in using DIY corpora, and gave some suggestions on how to encourage long-term use of corpora for wider application of DIY corpora in academic writing training.**

***Index Terms*—do-it-yourself corpora, English for special 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 corpora, usually only for some individual course or discipline; 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. General corpora have been popular for the recent decade, "but there is a growing interest in the use of specialized and DIY corpora (Charles, 2014, p.30)".

The application of general corpora to teaching and learning of grammatical and lexical items involves both quantitative and qualitative approaches. Some research explored what corpora can do, such as whether consulting corpora can significantly impact vocabulary and grammar learning (Boulton, 2010; Cresswell, 2007). Other studies investigated the effectiveness of learning how to use corpora, suggesting the majority of participants acquired the skills to address language problems (Gaskell & Cobb, 2004; Gilmore, 2009; Todd, 2001; Charles, 2011). Additionally, some qualitative research triangulated results from quantitative research, showing that most of the interviewed students had positive attitudes towards corpora use (Granath, 2009; Mizumoto & Chujo, 2015), and further indicated that factors, such as English proficiency, proper training, and technical support level, were decisive in enhancing 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 graduate students and their online reports on their corpus use six-month later. The project is designed for first-year graduate 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 graduate 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 graduates' 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) and Wordsmith (Scott, 2012), 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 files by removing matter which is not part of the running text (e.g., references, graphical elements), which makes it easier to read concordance lines and renders the statistical data more reliable (Charles, 2014, p.31)".

The third session is focused on the analysis of information that corpus can provide. When studying how language works in a more formal setting, it involves a few different steps. In each step, people are given questions that help them understand better how certain language elements help in getting a message across in the corpora. The process starts with a general look at the texts to find common patterns. Then, it moves on to focus on specific ways language is used, for instance, how people express doubt or how they link ideas together. In the end, corpora text analysis gives people the skills to better understand how language is put together and how it helps us communicate our ideas with people in academic writings (Charles, 2014).

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.

The initial objective of this project is to facilitate students with a custom-built resource designed to be an enduring asset throughout their academic pursuits (Charles, 2014). Do-It-Yourself (DIY) corpus is intended to serve as a foundational instrument that fosters the development of expertise and contextual deployment of language. The impetus for the creation of an individualized DIY corpus is rooted in the pedagogical strategy of diminishing the dependency on external entities, such as teachers and proofreaders (Charles, 2014), in favor of promoting a more self-directed and self-sufficient approach to the generation and refinement of academic discourse (Charles, 2014).

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 about six months after completion of the course(Charles, 2014), 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 the questionnaire. The surveys include 20 questions, which were administered to the participants electronically (Charles, 2014). If further investigation needed, there were QQ connections to clarify or amplify responses from the participants.

The participants are first year graduate 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. The comparison results yielded no significant difference between the two courses, the combination of the two-year data facilitated a more robust statistical evaluation with a total of 35 sets of data (Charles, 2014).

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)

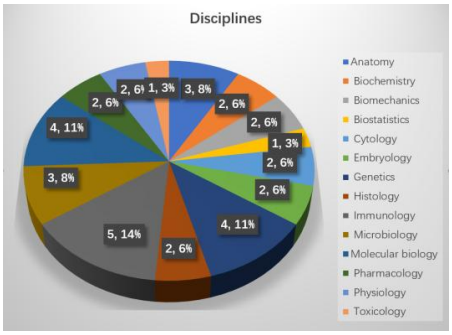

Fig 1 Participants Disciplinary Distribution

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)

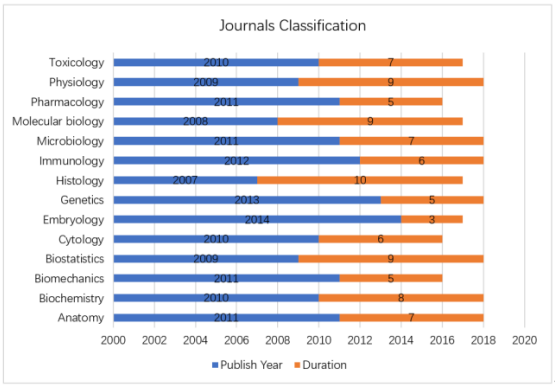

Fig 2 Journal Classification and Details

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)

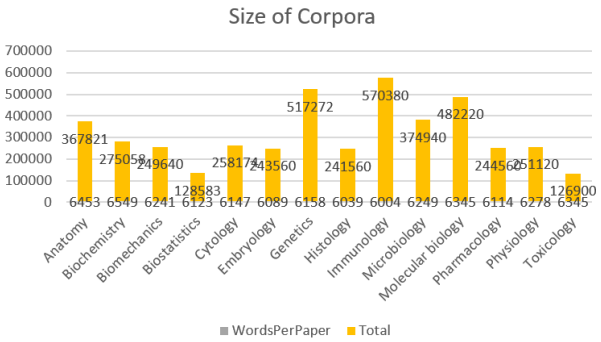

Fig 3 Size of corpora

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. However, in the subsequent inquiry conducted six months after the instructional program, a follow-up questionnaire was administered to the participants with the objective of ascertaining the post-curricular use of their individually DIY corpora (Charles, 2014). We found 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)

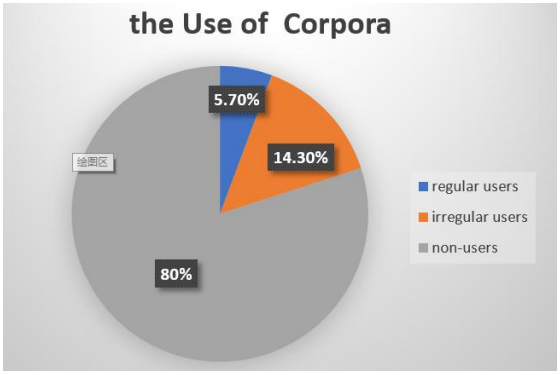

Fig 4 the use of corpora

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 found that the non-engagement of DIY corpus, was not necessarily a rejection of corpus as a research tool, but rather a contextual misalignment between their current working context and the potential utility of corpus resources (Yoon, 2008; 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 graduates 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.

These results resonate some previous studies (Charles, 2012; Yoon, 2011), which found that participants usually consult corpora for sentence level difficulties, such as addressing lexical issues and grammatical concerns (Lee & Swales, 2006; Charles, 2014), which were easy to tackle and appropriate to be the very first step to access before working on more complicated discourse questions.

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 must be performed when changing computers or something wrong with the local computer. The software is not very easy to use, and you must 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 should be taken into consideration. Some previous studies (Charles, 2014, p.35) also suggested "web-based interface" to facilitate users with high processing speed and cloud storage. Another solution is to use the BYU online corpora registered version, which provides tutorials and storage allowance, but you must pay if you want the full function.

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, p.64). 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 must let them feel the usefulness of corpora, that is, help them to see the "beauty of corpora". As ESP learners, most of their time was dedicated to learning disciplinary knowledge, and they have limited time available for learning how to write, and even limited time to learn and build corpora, so "they need to be convinced that the utility of the resource justifies the time taken to build it" (Charles, 2014, p.36). 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 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. Therefore, 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 provide follow-up support, such as "refresher sessions, drop-in clinics, on-line, on-demand courses or other means of just-in-time support" (Charles, 2014, p.36), addressing their requirements as they emerge.

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 filed, the real beauty of DIY personal corpora will reveal themselves, and more students and people interested will benefit from the use of corpora.

REFERENCES

[1] Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University.
[2] Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3), 534–572.
[3] Charles, M. (2011). Using hands-on concordancing to teach rhetorical functions: Evaluation and implications for EAP writing classes. In *New Trends in Corpora and Language Learning*, A. Frankenberg-Garcia, L. Flowerdew & G. Aston (eds), 26–43. London: Continuum.
[4] Charles, M. (2012). "'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building". *English for Specific Purposes*, 31, 93–102.
[5] Charles, M. (2014). Getting the corpus habit EAP students' long-term use of personal corpora. *English for Specific Purposes*, 35, 30–40.
[6] Cresswell, A. (2007). Getting to 'know' connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 267–287). Amsterdam: Rodopi.
[7] Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32, 301–319.
[8] Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal*, 63(4), 363–372.
[9] Granath, S. (2009). Who benefits from learning how to use corpora? *Corpora and language teaching*, 33, 47.
[10] Hunston, S. & Thompson, G. (2000). Evaluation in Text. Oxford, UK: Oxford University Press.
[11] Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7, 2, 173–192.
[12] Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, 25(1), 56-75.
[13] Mark, E. (2013). Student satisfaction and the customer focus in higher education. *Journal of higher education policy and management*, 35(1), 2-10.
[14] Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. English *Corpus Studies*, 22, 1–18.
[15] Scott, M. (2008). Developing Wordsmith. *International Journal of English Studies*, 8(1), 95–106.
[16] Scott, M. (2012). Wordsmith Tools 6.0. Liverpool: Lexical Analysis Software.
[17] Sinclair, J. (1997). Corpus evidence in language description. In *Teaching and Language Corpora*, A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (eds), 27–39. London: Longman.
[18] Sinclair, J. (2004). Trust the text: Language, corpus, and discourse. Routledge.
[19] Todd, R. W. (2001). Induction from self-selected concordances and self-correction. *System*, 29, 91–102.
[20] Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning and Technology*, 12(2), 31–48.
[21] Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10, 130–139.

Feng Zhang, was born in Shandong, China. 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. 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.