Auto-analysis of writing database based on corpus model

Qin Wang
Shaanxi Normal University, Xi’an, China

Corresponding author and e-mail: Qin Wang, qqwang@snnu.edu.cn

Abstract. The paper presents corpus-based writing database analysis and its application to college English writing classes. A small writing database was built by collecting 88 students’ 264 writing samples in one semester. Corpus model was used in the research for analysing connector using, cluster score and word frequency. The contrastive connector analysis reveals the difference between the collected samples and the existing Corpus SWECCL. The cluster and word frequency analysis introduces a new approach to analyse writing samples. The paper also details some classroom activities designed from research findings. Students’ in-class performance and after-class feedback both indicate that they are interested in the analysis-based classroom activities. At the beginning and the end of the semester, the study collected students’ English writing scores in the tests and adopted SPSS to assess their achievements. The paper ends with a short reflection on the enlightenments and limitations of the research.

1. Introduction
As an important composition of language proficiency, writing is an essential indicator of output competence. By contrast, English writing is a tough issue in college English teaching and learning for second language learners. The reality of overlarge class and disproportionate time allocation unavoidably leads to teacher-centered skill lectures, limited assignments with fixed structures and grammar-centered evaluation.

Much efforts are devoted into improving teaching and learning of English writing. Many papers have been published on methodology, influencing factors, testing and evaluation. Among which methodology research accounts for the largest proportion (Li Zhixue, Li Shaoshan, 2003) [1]. When some researches were shifting to learners, their studies were flawed by inappropriate application of quantitative and empirical methods (Zhu Yanyan, 2011) [2]. With the development of computer, Internet and processing technology, recent years it has been witnessing the interest into electronic texts collection and analysis in various forms. Computer-assisted database is becoming a hot issue in writing research. Many researchers discuss issues on automated evaluation, validity, writing progress, weaknesses and difficulties with the help of database (Tang Jinlan, Wu Yi’an, 2011)[3].

Corpus, a kind of database, also gets attention by taking authentic language elements into language research. This performance-based approach has large contribution to promoting teaching and learning, like Lexical Approach and Data-driven Learning. Application of Corpus to writing research makes great achievements since BNC 1994. In China, CLEC (Chinese Learner English Corpus) was released in 2003(Gui Shichun, Yang Huizhong, 2003) [4]; SWECCL (Spoken and Written English Corpus of Chinese Learners) was completed in 2008(Wen Qiufang, Wang Jinquan, 2008) [5]. Many researches based on domestic and overseas corpus were published (Xu Jiajin, 2003; Liang Maocheng, 2003; He
Anping, 2004; Yang Huizhong, 2005; Wang Lifei, 2006; Liang Maocheng, 2009; Gui Shichun et al., 2010). Those publications made broad discussion on contrastive analysis, data-driven learning and how to build corpus [6-12].

Although many achievements in corpus research could be listed, it is still, to some extent, unknown or challenging to many college English educators due to the obvious misunderstanding of “size”. Many teachers and researchers hold the belief that small number of text samples is equal to inappropriate representativeness, which barriers some researches or even leads certain researches to wrong conclusions.

Partly due to the belief mentioned above, corpus related approaches are not widely applied in college English writing classes and studies. In addition, many educators tend to use corpus are also challenged by inaccessibility of some large-scale corpus, technological and professional requirements for building corpus and so on.

This paper comes from the researcher’s personal experience and reflection on teaching writing to non-English majors. The researcher has a basic understanding of necessity and possibility of corpus application in college English teaching and learning. Based on the writer’s teaching experience and the interest in corpus, this paper focuses on accessible small database and its application in college English writing class. Since the self-build database used in the research is not professionally tagged, the paper refers it as database rather than corpus.

2. Methodology and research process

The research introduces how to apply computer assisted analysis to self-build writing database based on corpus model and discusses how useful it is for college English writing teaching and learning.

From September 2020 to January 2021, for three times, the researcher altogether collected 264 writing samples from three random 2020 classes, each having around 30 students with similar English learning background and English proficiency. All the writing assignments, with the word requirement of 120-150, are based on 2017-2019 CET-4, which is an authoritative test in China for evaluating college students' English learning. To avoid plagiarism, all the writings were finished in class without the possibility of online copying or reference consulting. Since the research does not focus on spelling, all the spelling mistakes were corrected by computer-based auto-checking and rechecked by the researcher herself. Then, all the after-checking writings were transformed into texts and stored as a small-scale writing database. After the preparation of the database, SWECCCL was used in contrastive analysis and KH Coding was used for cluster and word frequency analysis. With the data and findings from computer-assisted analysis, related class activities were designed and implemented for one semester. At the beginning and end of the semester, in September 2020 and January 2021, students’ English writing scores were collected from two tests and used in SPSS analysis.

3. Computer-assisted contrastive analysis of connectors

Contrastive analysis is a common approach in corpus research (Xu Jiajin, Liang Maocheng, 2011) [13]. For the consideration of accessibility, the paper makes contrastive analysis of SWECCCL (Spoken and Written English Corpus of Chinese learners), which collects writing samples of English majors, and self-build writing database, from non-English majors, to discuss the effectiveness and possibility of using contrastive research in college English writing teaching.

Connectors are known as logical connectors, linking signals, conjunctions and some more. This research does not make distinction among these terms and adopts the term “connector” in the whole paper to avoid confusion and disagreement. Many researches on connector starts with Halliday’s Cohesion in English (Michael A K H, 1976) which classifies logical connectors [14]. It is widely accepted that connectors play important role in English writing and second language learners are not good at using connectors in writing. Luo Yi (2003) points out Chinese postgraduates majoring in English use more connectors than natives. He also notes that they use more connectors for contrasting and cause-effect, less for confirming [15]. Pan Pan (2004) reports overusing of some connectors based on his corpus of CET-6 passers [16]. Lei Lei (2012) finds out doctoral students’ overuse of 33 connectors in
their dissertations [17]. Based on these achievements, this paper adopts the connector view to explore database analysis and its application in college English teaching.

With the help of computer, it is not difficult to extract the top 10 connectors from SWECCL written corpus. They are: and, but, or, for example, because, at least, in fact, in addition, however, for instance. The words and their frequency are shown in Table 1.

Table 1. Connectors and Frequency in SWECCL.

| Connector     | Frequency | Connector     | Frequency |
|---------------|-----------|---------------|-----------|
| and           | 2480      | at least      | 210       |
| but           | 770       | in fact       | 180       |
| or            | 600       | in addition   | 150       |
| for example   | 500       | however       | 140       |
| because       | 250       | for instance  | 140       |

According to Table 1, “and”, “but” and “or” ranks top 3 among all the connectors in terms of frequency. These most common connectors fall into three categories: connecting (and, or, in addition), adversative (but, however) and cause-effect (because, therefore, for instance). The SWECCL data reveals that the selected written texts from English majors have certain variety in connector using in writing, which illustrates their good understanding of semantic relations and connecting in English writing. At the same time, they have the tendency of overusing some high-frequency connectors, like “and”.

To understand the students’ connector using and compare with the connector using in SWECCL, the following contrastive analysis is done.

The researcher took the top 10 high-frequency connectors in Table 1. to search in the writing database of her students. In the searching process, the researcher noticed that some connectors, which are not in Table 1, were active in her students’ database. Therefore, the finding with 16 connectors is shown in Table 2.

Table 2. Connectors and frequency in self-build database.

| Connector              | Frequency | Connector     | Frequency |
|------------------------|-----------|---------------|-----------|
| and                    | 900       | at least      | 0         |
| but                    | 333       | in fact       | 0         |
| or                     | 18        | in addition   | 18        |
| for example            | 9         | however       | 72        |
| because                | 81        | for instance  | 0         |
| therefore              | 27        | first(second, third) | 63 |
| so                     | 171       | moreover      | 3         |
| actually               | 0         | furthermore   | 0         |

Even the in-use database does not have the same size as SWECCL, the contrastive analysis could explain some aspects of students’ connector using. According to Table 1. and Table 2., we may notice the following differences:

The most common connectors in Table 2. include the category of time-and-order(first), which does not appear in Table 1. The possible reason is the requirement of CET-4 writing and emphasis of orderly
linking in college English teaching. Since the students are not clear about logical relationship, they tend to use the simple linking -- “first, second, third” to list reasons, suggestions, etc. Table 2. also reveals frequency and frequency ranking of connectors, which represents variety difference in connector using. Considering the small size of self-build database, some connectors, like “and (900)”, has the tendency of overusing. Contrast to this overuse, Table 2. also clearly illustrates the missing of some connectors. For example, “in fact” and “for instance” both have zero frequency. In Table 2, the connector “so” is very noticeable for its high frequency (171), which is not appropriate for academic writing. This phenomenon may, to some extent, mirror the students’ misunderstanding of some connectors’ contextual connotations. In addition, the most widely used connectors in Table 2. Tell us that simple connectors, like “and, but”, account for the largest proportion of connector using in college English writing.

The research and analysis explain that this kind of contrastive analysis based on a small database could also bring some valuable reflection and enlightenment on English writing teaching and learning. By doing computer assisted analysis like this, teachers could be clear of students’ writing modal, logical understanding and some other important issues in English writing, like connector using. Then, the reflective data and authentic materials from writing database could be used in well-organized class activities to guide students to be aware of their own connector overusing, logic mode, missing connectors, etc.

When doing this, both self-build small database and well-established large corpus like SWELL (Spoken and Written English Corpus of Chinese Learners), CLEC (Chinese Learner English Corpus), BNC (British National Corpus), COCA (Corpus of Contemporary American English) could be used for contrasting. When the contrastive analysis is done, the finding and the authentic materials could be used for assisting teaching and learning. Students’ writing could be improved by noticing, understanding, imitating and practicing.

4. Computer assisted K-H coding
Citespace and KH Coder are both text mining tools for multi-language research for visualizing and analyzing trends and patterns in scientific literature. In consideration of accessibility, KH Coder is adopted by the researcher.

KH Coder, by Higuchi Koichi, is designed for quantitative content analysis of text data. It helps to extract words from data and statistically analyze them to obtain a whole picture and explore the features of the data while avoiding prejudices of the researcher. It could be easily applied in analyzing English and Chinese written texts without the requirement of professional skills.

The database used in this KH coding comes from students’ 60 in-class writing samples. All the samples, with the same title “Why I study”, are transformed into text data as required by KH coding. Figure 1. is the general description of the data.

![Figure 1. Database description.](image-url)
Since the samples have the same topic, the researcher tries cluster analysis to find the most commonly used clusters around the topic. Table 3. is about scores of different clusters. As is shown in Table 3, the score difference of clusters extracted from writing texts clearly illustrates the students’ general understanding of the topic “why I study”.

| Cluster               | Frequency | Connector    | Frequency |
|-----------------------|-----------|--------------|-----------|
| social development    | 75.573    | job opportunity | 8.588    |
| modern society        | 26.456    | knowledge    | 52.447    |
| good job              | 64.464    | good life    | 29.212    |
| ideal job             | 8.588     | ideal life   | 5.189     |
| future job/career     | 31.849    | bright future | 7.326    |

By descending order, society and society related clusters rank first, which are followed by job and knowledge related clusters. Then comes life clusters. Bright future ranks fourth. When we compare the clusters in the table with the title, without consideration of vocabulary competence, we may conclude that the writers’ answer to “why I study” mainly fall in two categories – social and personal, which is different from the popular “selfish” understanding about the youth. Their writing shows society and knowledge are very common intentions for study. They also have consideration of personal development factors, like job, life and future.

By using the above analysis, the teachers could easily find the students’ interests and personal understanding around a certain topic in writing. Then, some related tasks could be designed for class discussions about topic mining, vocabulary expansion and so on. These kinds of tasks could guide students to improve their writing and understanding.

KH coding could also be used for word frequency. Word frequency could be counted and listed in a second. In Table 4. of Word Frequency, we could see the most commonly used verbs, nouns and adjectives in students’ writing samples.

| Word Type | Frequency |
|-----------|-----------|
| Verb      | study201, learn143, have107, do68, make51, get44, become32, improve31, want31, know26, keep24, need24, say24, give21 |
| Noun      | study122, life72, knowledge71, society48, way37, reason36, people33, time33, school29, world29, job27, thing27, future25, development23, person18, student18, horizon17 |
| Adjective | more93, only 45, better40, good35, so32, many29, hard25, also23, important23, different18, old18, just17, always12 |

In Table 4. some simple words, like “study” and “learn”, are repeatedly used in the students’ writing samples. Since the assigned title is “Why I study”, “study” appears more than 300 times in 60 samples, both as verb and noun. The word “learn”, which has similar meaning as “study”, also appears over 140 times. General verb like “have”, “do”, “make”, “get”, “become” could be seen in many sentences. If we take Adjective as an example, “more”, “only”, “better” and “good” are most common ones. They are used very often in students’ writing samples. Table 4. also reflects that the students have the habit of using general words instead of specific words. If we take the Table for a comparison with CET-4 and
CET-6 vocabulary, we could see that many academic words in College English Vocabulary are not used in the students’ writing.

KH Coding could be used for more explorations. If we take “study” for a detailed checking by KH Coding, it will list all the sentences and collocations of “study”. From the listing, we could find more about the word “study”:

The structures “S. + should/have to/must/can + study (V.)” and “study (N.) makes me/us + V.” are the most common ones in the sample database. Extracted from the database are also clusters like “deep study”, “study knowledge” and “study hard”, which are clearly influenced by students’ native language expressions.

5. Class activities based on database analysis
Database analysis is not adopted in the research just for the sake of analysis. The researcher is using the analysis for assisting writing teaching and learning. Based on the analysis like Connector Analysis and KH Coding, classroom activities could be designed with the help of achieved data and authentic language materials from the database.

5.1. Activity one: adjective frequency
Before writing the essay, the topic “culture” was given to the students. And the students were asked to write down all the adjectives in their mind for the topic. After all the answer sheets were collected in class, the writing assignment with the topic of “culture” was done and collected. When all the collected samples were stored in the database, the adjective frequency was extracted. Table 5. shows the Adjective Frequency in students’ writing.

| Word Type | Frequency |
|-----------|-----------|
| Adj.      | more93, only 45, better40, good35, so32, many29, hard25, also23, important23, different18, old18, just17, always12 |

A Frequency Table like Table 5. and the Answer Sheets the students finished in class were given to each group for in-class discussion. The discussion was guided and organized by the instructions like the following:

Are you familiar with these adjectives?
Do you use them in your writing?
Please circle them in your written script.
Please circle all the adjectives in your writing.
Please make a comparison of the adjectives in your mind and the adjectives in your writing.
Are there some adjectives you do not use in your writing?
Please check the following sentences, see if you could add some specific and vivid adjectives for editing them?
Study helps us stand out.
From my study experience, I hold the belief that……
We can live a better life through study.

5.2. Activity two: repetition checking
This Repetition Checking Activity was also designed on the base of analysis finding. When the repetition issue was noticed in data analysis, some paragraphs of repetition were extracted from the students’ writing database. These authentic paragraphs were employed in class activities to present Repetition to the students without lecturing. The students were asked to check the extracted paragraph like Table 6. In Table 6, “study” is used four times in a four-sentence paragraph. The students noticed the “repetition” immediately after they checked the paragraph.
Table 6. Repetition.

| Study makes me grow up to be a better self. Through studying poems, I gradually find happiness in daily life. After studying maths, I feel my thinking becomes more flexible. Study brings me both material and spiritual treasure. |

When the students read the paragraph and recognized the repetition of the word “study”, they were asked to have a discussion with his/her partners about the following questions:

- Please circle the word “study” and count how many times it is used in the sample?
- Can we change the word “study” to some other words/forms to avoid unnecessary repetition?
- Please check your own writing, see if you have several “study”?
- Please check your own writing, see if you have some other word repetition like “study”?
- What’s your understanding of “avoiding word repetition in writing”?

5.3. Reflection on analysis-based class activities

These kinds of activities were all based on computer assisted database analysis. When computer assisted analysis was applied in writing, it could help us notice the students’ real issues in writing. Authentic writing samples and data could also be used in class activity designing. When the researcher adopted data analysis in designing student-centred classroom writing activities, the initial thought was to trigger more and better student involvement. Writing could not be improved by theoretical understanding or notetaking. It relies on students’ own awareness and practice. When this thought was implemented in writing classes, students’ in-class performance and after-class feedback reflected that the students were interested in participating these kinds of data-based class activities. In those activities, they were becoming aware of their own writing flaws by checking, discussing and analysing their own writing texts. During this process, their writing was gradually improved. Generally speaking, the data-based class activities in the researcher’s writing classes did bring better involvement and satisfactory writing improvement.

The researcher’s three classes were involved in the one-semester writing classes during 2020 Fall Semester. During the semester, the researcher consciously used database and related class activities for assisting writing teaching. At the beginning and the end of the semester, the 88 students took two writing tests and the scores were recorded. Figure 2. shows the data from two writing tests. T1 refers to Test 1, writing score at the beginning of the 2020 Fall semester. T2 refers to Test 2, writing score at the end of the semester. The total score for the two writing tests is 15. Although the manual grading may have unavoidable disadvantage in fairness and standard, Figure 2. still reflects the students’ writing improvement. Figure 2. clearly illustrates a 2-point increase in Mean and a 0.4-decrease in Standard Deviation, which represents the students’ improvement in writing.

![Figure 2. Paired samples statistics.](image)

6. Conclusions

The researcher is discussing how to effectively use computer-assisted analysis of small writing database in college English writing teaching and learning. By doing this, writing class could have more student-
centred class activities and better student involvement, which may lead to students’ better performance and more achievements. All the data and samples in the research are from students’ writing samples collected by the researcher in 2020 Fall Semester. All the activities mentioned in the paper were successfully implemented in writing teaching in 2020 Fall Semester.

The Contrastive Analysis of corpus and the students’ database is not as hard as some educators imagine. A small writing database could effectively help us to have some basic and general understanding of students’ writing, like connector using. It could also be used for some other explorations. The contrastive could also lead to the discussion and exploration of more. KH Coder could also be used in writing teaching and learning for word frequency, cluster and some other writing issues. Based on the data analysis of students’ authentic written samples, better-involved class activities could be designed to help students achieve more in writing classes.

Modern technology like database analysis plays an important role in assisting writing teaching and learning in this computer era. Nowadays, the traditional understanding of “difficulty” is eclipsed by many user-friendly applications, which should be widely introduced to college English teachers. For teachers in this computer and Internet age, they may take advantage of these amazing technologies to ease their traditional responsibility in spelling and grammatical checking, collecting and distributing. In addition, for textbook, the achievements in corpus linguistics, like data and authentic materials, should be valued in textbook publishing.

In terms of limitation, they are the following: The research is originally triggered by the researcher’s observation in teaching practice, which still lacks theoretical and systematic understanding. The research is only based on the data of 88 students’ one semester classes, which still needs empirical data and contrastive discussion in the future. For writing assessment, all the samples are graded manually by different teachers, which may bring personal factors unfairly influencing the final scores.

Acknowledgement
Thanks to all the students and their cooperation in the research.

References
[1] Li Zhixue and Li Shaoshan 2003 China. FL World 6 55
[2] Zhu Yanyan 2011 China. FL World 6 56
[3] Tang Jinlan and Wu Yi’an 2011 China. J. FL Teaching and Research 2 115
[4] Gui Shichun and Yang Huizhong 2003 Chinese Learner English Corpus (Shanghai: Foreign Language Education Press)
[5] Wen Qufang and Liang Maocheng 2008 Spoken and Written English Corpus of Chinese Learners (Beijing: Foreign Language Teaching and Research Press)
[6] Xu Jiajin 2003 China. FL Ed. 6 6
[7] Liang Maocheng 2003 China. CAFL Ed. in China 12 42
[8] He Anping 2004 Corpus Linguistics and English Teaching (Beijing: Foreign Language Teaching and Research Press)
[9] Yang Huizhong and Wei Naixing 2005 Construction and Data Analysis of Chinese Learner Spoken English Corpus (Shanghai: Shanghai Foreign Education Press)
[10] Wang Lifei and Zhang Yan 2006 China. CAFL Ed. in China 8 36
[11] Liang Maocheng 2009 China. CAFL Ed. in China 5 8
[12] Gui Shichun and Feng Zhiwei 2010 China. MFL 4 419
[13] Xu Jiajin and Liang Maocheng 2011China. Contemp. FL Studies 10 6
[14] Michael A K H 1976 Cohesion in English (London: Longman)
[15] Luo Yi 2003 China. J. PLA UFL 1 59
[16] Pan Pan 2004 China. MFL 2 157
[17] Lei Lei 2012 China. J. En. Acad. Purposes 11 267