International Journal of Language and Linguistics
2020; 8(2): 75-81
http://www.sciencepublishinggroup.com/j/ijll
doi: 10.11648/j.ijll.20200802.13
ISSN: 2330-0205 (Print); ISSN: 2330-0221 (Online)

Application of Self-Constructed Corpus in the English Writing Course

Wang Heyu

English Department, School of Foreign Languages, Guangdong University of Technology, Guangzhou, China

Email address:
wangheyu@gdut.edu.cn

To cite this article:
Wang Heyu. Application of Self-Constructed Corpus in the English Writing Course. *International Journal of Language and Linguistics*. Vol. 8, No. 2, 2020, pp. 75-81. doi: 10.11648/j.ijll.20200802.13

Received: March 1, 2020; **Accepted**: March 23, 2020; **Published**: April 14, 2020

Abstract: The corpus constituents are actually the miniature of language. As the computer technology continuously progresses, corpus-based research proves increasingly applicable to language teaching and learning. As a free corpus analysis software program designed specifically for use in the language classrooms, *AntConc* provides the users with a powerful toolkit with various features to analyze the loaded written texts. Basic functions like the concordance tool and cluster sections of *Antconc* demonstrate the possibility to assess the performance of students' language proficiency. Considering the fact that IELTS is widely recognized as a standardized test and a big number of Chinese students are preparing for it in order to study abrord, the present study managed to built a self-constructed corpus of students essays in the English writing course. By means of the sophisticated coding and tagging systems, we obtained a set of standard text materials attached with the label of part of speech. Based upon the data analysis of the students' IELTS writing samples against the modal essay corpus, we conducted the corresponding evaluation of students' lexical resource and offer further suggestions and tailor-made solutions in order to help students improve their writing ability. Under the help of the model essay corpus, students might obtain useful instructions, increase grammar diversity and accuracy, thereby become more competent writers. Our study added more evidence to the conclusion that corpus has a great potential to fundamentally change the ways we approach language education. Apparently, teachers working with highly proficient students from diverse backgrounds need to be informed about the latest corpus tools for a more fine-tuned or specifically tailored application in their writing course.

Keywords: Self-Constructed Corpus, English Writing, Lexical Resource

1. Introduction

Generally speaking，a corpus is a large and structured set of texts electronically stored and processed. Specifically, a corpus is a collection of texts constructed by certain sampling criteria to a specified end [1]. Importantly, the corpus constituents—the collection of texts stored in an electronic database—are fundamentally the miniature of language. Studying these typical examples helps to grasp the language patterns and traits. Biber Douglas [2], for instance, utilized a corpus-driven approach to find that the multi-word patterns typical of speech are fundamentally different from those typical of academic writing: patterns in conversation tend to be fixed sequences including both function words and content words whereas most patterns in academic writing are formulaic frames consisting of invariable function words with an variable slot filled by content words.

Since 1990s, along with the thriving development of corpus linguistics and computer science, corpus, which was once an uncharted territory to the vast majority of foreign language teachers and learners, has been gaining snowballing momentum not only in fields of linguistics, but also in the domain of mainstream foreign language teaching and learning. Dovetailing with corpus-related knowledge and software, the modern handy computer technology equips an average language researcher with the ability to better examine the teaching effect in a more quantifiable fashion. As such, scholars have been making effort to strengthen communication between corpus researchers and language practitioners in order to provide the support needed in course design and activities [3-5]. Recent literature has shown that a corpus-and-data-driven revolution is underway in applied linguistics as well as in language pedagogy [6-7]). These studies not merely address the practical application and the

theoretical basis in corpus use and corpus-based instruction, but also offer tools, corpora and online resources as well as corpus-based lessons and testing. The well-integrated use of corpus is constitutive for meeting students' specialized needs, enhancing their class engagement with the teaching materials, and enabling them to carry out hands-on practices, therefore it is more likely to initiate the learner-centred instruction or autonomous learning [8]. Significantly, the corpus-based study can assist practitioners to have an in-depth understanding about their students' language proficiency. In other words, a language teacher can gain a panoramic view of the students' symptoms in the language performance on a calculable scale. According to McEnery & Xiao [9], corpus has a great potential to "fundamentally change the ways we approach language education, including both what is taught and how it is taught".

Via the corpus programs Antconc and Treetagger, the present study aims to investigate the possiblity to assess the students' English level reflected by the output of their written language. On the basis of the self-constructed corpus data, some interpretation and corresponding measures could be offered to the students.

2. About the Multi-Functional l Corpus Software *AntConc*

Developed by Laurence Anthony, a professor at the Waseda University Japan, *Antconc* was originally a Windows-based program designed for simple concordance, but now has developed into the panoply of world-renowned multi-functional corpus software available. Since it was launched in 2002, this software has witnessed more than 20 releases with a few major upgrades. Some new versions of the program are also going to be released in the future to improve and perfect its functions. An assortment of experiments and research can be implemented, including the compilation of high-frequency keyword lists from all the texts loaded with *AntConc*.

As a corpus analysis software program designed specifically for use in the language classroom, *AntConc* is a freeware app, making it ideal for individual students and teachers, schools as well as colleges with a limited budget, and is applicable to both Windows and Linux/Unix based systems. Despite the fact that it has a freeware license, it includes an easy-to-use, intuitive graphical user interface and provides the users with a powerful toolkit with various features to analyze the loaded written texts.

One of the most important tools used in *AntConc* is the concordancer. The concordancer have been proved to be an effective help in learning a second or foreign language, facilitating the mastery of vocabulary, collocations, grammar and various writing styles. The Following Figure is the concordance interface of *AntConc* while the Concordancer Tool is under operation.

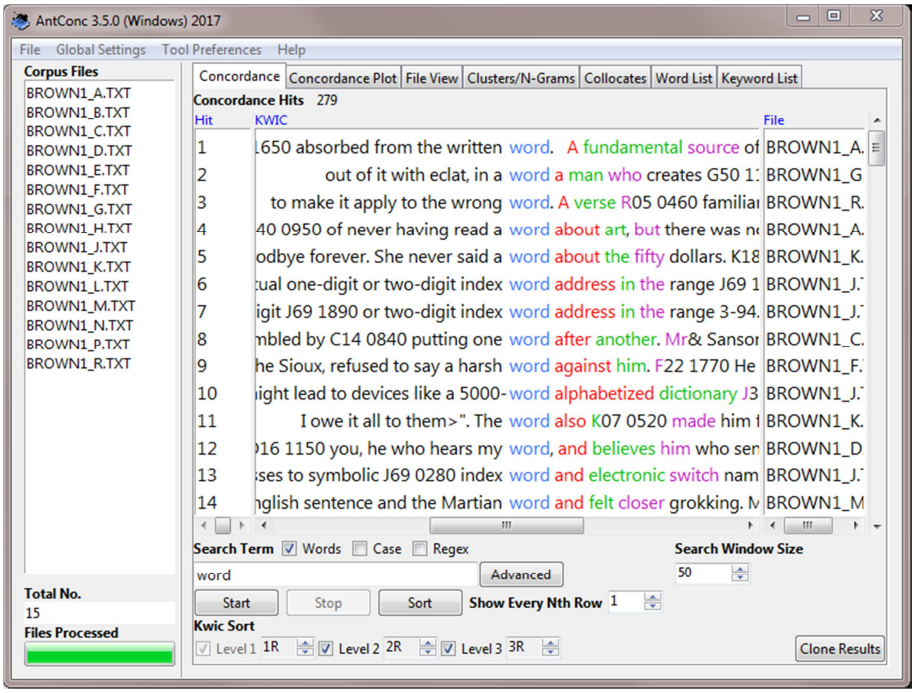

Figure 1. *The concordance interface of Antconc.*

Similar to all other tools, most of the common operations in *AntConc* are accessible on the main screen. In other words, the users of the concordancer do not need to open the pull-down menus and at the same time click some ectra windows. This feature, according to Lonfils & VanParys [10], is meant to be user-friendly, which is a vital element to take into account when a good software is designed.

To encapsulate, with an friendly interface, and a freeware license *AntConc* is a convenient corpus analysis toolkit that is simple and easy to use. It has been proven to be highly

effective and applicable in a classroom context [11]. Although it does not have all the tools and features that perefectly satisfy every specific need, it at least offeres the most important tools that are needed for the current corpora analysis. Such functions as the simple concordance and generation of keyword list, are merely the tip of the iceberg. We can effectively and clearly examine the figures it provides to introduce the quantitative research into the writing studies in language courses.

3. Construction of Students' Corpus

3.1. To Specify the Students' Target

As is the status quo with the diversified students, each of them is confronted with individual target exams like CET, TEM, and IELTS etc. Regardless of the fact that there is a substantial overlap between all the criteria concerning these popular tests, it is imperative to set a specific benchmark for the sake of the more accurate results of the study. Considering the fact that IELTS is widely recognized as a standardized test and a big number of Chinese students are preparing for it in order to study abrord, the Writing Task 2 of IELTS is selected as the target exam for the subjects in our study, which means the learners' corpus is IELTS-oriented.

3.2. To Collect the Students' Writing Samples

Typically, the sample materials can originate from students' daily exercise. In the two major categories, one is the pre-existing materials like their own term papers, homework; the other is the student-generated materials elicited by the instructor. An exemplary procedure is that with the assigned topics, students are expected to compose some corresponding essays and submit them in digital form before deadline. A more concrete process includes: (i) Contact the respondents and establish a telecommunication group via Wechat; (ii) Assign the IELTS topic and set a deadline; (iii) Receive and organize individual responses; (iv) Establish a proper corpus with the materials collected; and (v) Analyze the writers' characteristics via the corpus programs Antconc and Treetagger.

In light of the respondents' geographical disparity, it is safe to adopt Wechat as a bonding mechanism. As a convenient communication tool, wechat empowers every member to stay in touch and actively engaged. In stark contrast to the informality of Wechat and frequent overflow of information that ensues, e-mailing characterized by its formality figures as another vital platform for students. Email is an effective channel of disseminating formal information. As a matter of course, it is stipulated that all the student essays should be submitted in the Word document format via emails, facilitating the process of collection.

The Word document is featured by its excellent convenience for correction and marking. Nonetheless, as the critical link in the study, *Antconc* is mainly compatible with txt format documents only, necessitating that all the collected Word documents must be converted into txt format. Generally, the disparities in formatting can be effortlessly eliminated in txt format.

3.3. To Code and Tag the Students' Essays

During the process of conversion, it is noteworthy that the issue of misused formatting, like the inappropriate spacing and the mingling use of Chinese and English punctuation, should be handled with great dexterity. Far from being inconsequential, the existence of such inappropriate formatting will undermine the accuracy of the final data. With regards to the relatively large sum of essays, which is usually the case with a large-scale student corpus, we can resort to a program TextEditor developed by Fenglin instead of time-consuming manual modification.

Contrary to the popular belief that corpus is just a simple collection of texts, a full-fledged corpus is inextricably intertwined with sophisticated coding and tagging, which is the prerequisite for the categorization of the essay samples [12]. The tagging system adopted in this research is in agreement with that of the software *Treetagger* as Figure 1 demonstates. By utilizing the sophisticated tagging systems, we can obtain a set of standard text materials attached with the label of part of speech. To sum up, a student essay corpus is comprised of precisely processed electronic text materials with multiple tagging.

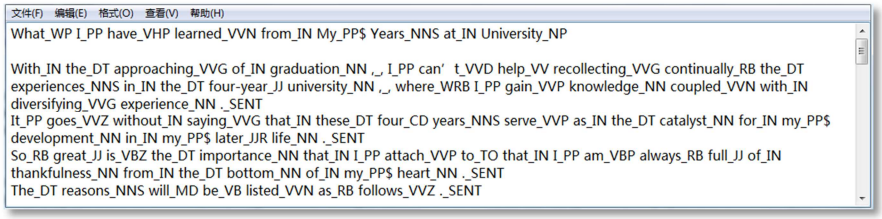

Figure 2. *A sample of students' tagged essays.*

4. Construction of Model Essay Corpus

In order to effectively evaluate students writing competence, suitable parallel corpus is needed. Parallel corpus has proved to be beneficial to translation teaching [13]. Likewise, parallel corpus could be used to facilitate students to improve their English writing. For the same reason, to construct a modal essay corpus needs no less technics than that of the student

corpus. Of course, therer are some differences in the source of text materials.

4.1. The Model Essay Corpus in a Broad Sense

Theoretically, almost all the comprehensive corpora with high-quality literature serve as the catalyst for a better performance of language. Usually the general corpus is a large-scale database with millions or billions of words, and the exemplary cases include *the British National Corpus* (BNC), *Bank of English* (BoE), and the *Corpus of Contemporary American English* (COCA). All of them contain concrete and exhaustive entries based upon the research on the overall English. Undoutedly, the general corpus can function as the reference corpus in order to manifest the characteristics of certain specialized corpus. Another exceptional corpus in the broad sense is the Google search engine, with millions of brand-new articles and texts engulfing the entire virtual world, fraught with the timeliest materials of phrases and expressions alongside with authoritative journalism outlets online. The Google search engine is by far the largest corpus. However, on account of the IELTS-oriented student corpus, an appropriate and restrictive framework of model corpus is helpful to boost the rate of progress.

4.2. The Model Essay Corpus in a Narrower Sense

Reading and writing are complementary processes that support each other. To be more effective, priority is given to the official publications *Cambridge IELTS Series—Examination Papers from University of Cambridge ESOL Examinations* [14-15] from which typical and idiomatic reading materials are derived. These high-caliber reading materials in the IELTS play a pivotal role in enhancing examine-takers' literacy so they gain a foothold in the model essay corpus.

It is a widely recognized idea that the stories and articles on the newspapers and magazines reflect a high degree of language proficiency plus a relatively greater formality and exactness than those of spoken register. For this reason, we add to our modal corpus articles from other important sources like *the Economist*, *TIME magazine*. Special attention should go to *the New York Times*' Editorial Column *Room for Debate*, in which reviews and editorials are concise and topic-centered with an average of 280 words, quite consistent with the requirement of the Writing Task 2 of IELTS [16]. Given their similarity in genre and size, it is highly advisable to integrate these editorials into the model essay corpus.

5. Assessment and Interpretation of Students' Writing

5.1. To Assess Students' Diction

In accordance with the requirement of IELTS Writing Task 2, "lexical resource" is one of the significant indexes of one's literacy (the rest of them is respectively task response, coherence and cohesion, grammar range and accuracy). Having established the students' writing corpora with sophisticated tagging, we can gain a clear insight into students' lexical preference and their flaws.

The procedure to generate a verb-only keyword list via *AntConc* is sketched as follows:

1. Load the students' tagged essays;
2. Select the cluster section of Antconc;
3. Adjust the maximum cluster size to 1;
4. Input the code "*_VV*" (meaning "any verbs in all forms") in the search bar;
5. Start generating the list.

After these five steps, we acquire a semi-finished product, but some data processing remains undone as Figure 2 shows:

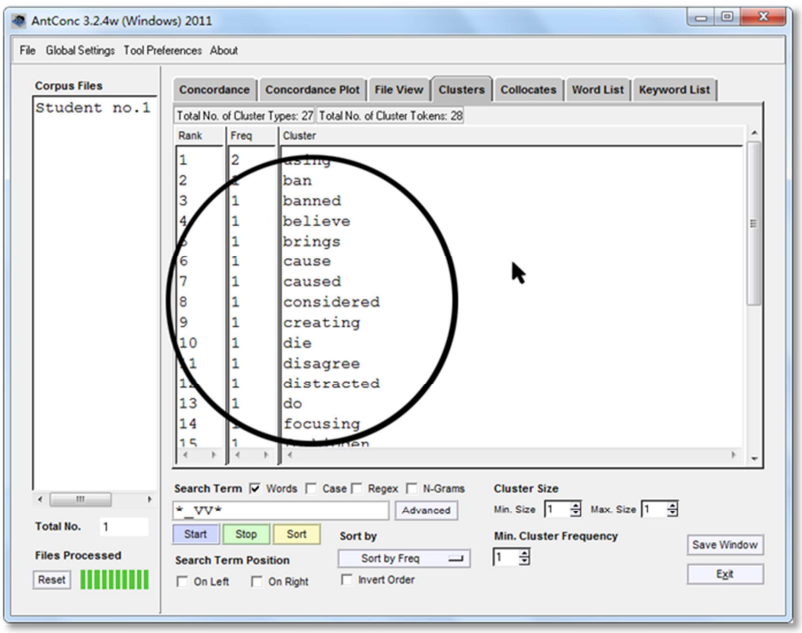

Figure 3. *The raw verbal list of No. 1 student.*

The inflected form of a word and its base form will be counted twice, like "cause" and "caused". As a consequence, we need to combine these terms in the final statistics. Eventually, three refined keyword lists are presented as in Table 1 (excerpts of verbal keyword lists in student essays on the discussion of mobile phone prohibition).

Table 1. *Excerpts of verbal wordlists of three students.*

Student 1			**Student 2**			**Student 3**		
rank	**freq**	**words**	**rank**	**freq**	**words**	**rank**	**freq**	**words**
1	5	take	1	3	use	1	2	take
2	5	ban	2	2	ban	2	1	annoy
3	3	want	3	2	cause	3	1	characterize
4	2	make	4	2	lead	4	1	demonstrate
5	2	leave	5	2	treat	5	1	detest
6	2	get	6	1	believe	6	1	encapsulate
7	2	find	7	1	bring	7	1	erupt
8	2	do	8	1	consider	8	1	exert
9	2	bring	9	1	create	9	1	impede
10	2	benefit	10	1	die	10	1	imprison

As is depicted, the maximum frequency of verbs witnesses a steady decline from 5 to 2. In details, student No. 1 tends to apply the monosyllabic verbs "take", "ban", "want" and "make" quite repeatedly, with the maximum frequency of 5. Comparatively, student No. 2 inclines to use verbs like "use", "ban", "cause" with the lower maximum frequency of 3. By contrast, student No. 3 employs polysyllabic words such as "annoy", "characterize", and "demonstrate" with the maximum frequency of 2.

Linguistically, students' mental lexicon varies a great deal from individual to individual. The maximum frequency of vocabulary functions as a indicator of one's capacity of commanding a wide variety of vocabulary. That is, the recurrent occurrence of certain expressions is attributable to their lack of flexible and sufficient lexical resource [17-18]. Such is the case of student No. 1 and student No. 2.

An experienced user of corpus analysis tools should know that word lists usually tell us little about how important a word is in a corpus. Accordingly, *AntConc* offers a Keyword List Tool in order to find which words appear unusually frequently in a corpus in comparison with those same words in a reference corpus also specified by the user. This, undoubtedly, helps teachers gain a better idea about students' lexical resource.

5.2. To Guide Students to Use Idiomatic Collocations

Research has shown that collocations and other multi-word units such as phrasal verbs, and idioms are particularly difficult for learners to acquire [19]. The importance of idiomatic collocations is even greater if the learner is working with very technical or scientific texts as the lexical unit is very often longer than a single word [20]. With *AntConc*, idiomatic collocations can be investigated by means of the Word Clusters Tool. This tool could show all the idiomatic expressions including the search term and put them in alphabetical or frequency order. The search term can be specified as a mopheme, a single word, a phrase or a fixed expression. In addition, the number of surrounding words to the left and right of the search term can also be specified. An alternative way to search for idiomatic collocations is to find lexical clusters which are equivalent to n-grams, where the range of n normally varies from 2 to 5 words. Fortunately, *AntConc* also includes lexical bundle searches as an option in the Word Clusters Tool.

If students' collocations are used inappropriately, an experienced instructor is obliged to encourage the students to identify the according idiomatic phrases in the model corpus. For example, in terms of the phrase "elevate the communication", students are guided to consult the collocation of both "elevate" and "communication" in the model essay corpus. Once again, COCA is a veritable reservoir of synonyms with alternative expressions.

Rank	Result	Freq
1	ELEVATED THE STATUS	11
2	ELEVATE THE STATUS	9
3	ELEVATE THE IMPORTANCE	8
4	ELEVATE THE LEVEL	8
5	ELEVATED THE IMPORTANCE	8
6	ELEVATING THE HEAD	8
7	ELEVATING THE STOCK	8

Rank	Result	Freq
1	IMPROVING THE COMMUNICATION	6
2	IMPROVE THE COMMUNICATION	6
3	BRIDGING THE COMMUNICATION	5
4	BRIDGE THE COMMUNICATION	5
5	INCREASE THE COMMUNICATION	3
6	OPEN THE COMMUNICATION	3
7	DISRUPT THE COMMUNICATION	2

Figure 4. *Excerpts of search results of collocations in the model corpus COCA.*

Figure 4 clearly shows that "elevate" collocates with "importance" and "status", ruling out its possibility of collocation with "communication". The verbs idiomatically collocating with "communication" should be "improve" or "facilitate". By the same token, the appropriate collocation with "family" should be "*family gathering*" instead of "*family assembling*". Ultimately, students are directed to conclude: "*improve/facilitate the communication*" is the proper usage of idiomtic English collocation.

The same method works equualy well with the usage of

other verbs and phrases. Importanly, students are instructed and encouraged to identify their own clichés and employ corresponding synonyms, thereby engineering a tailor-made developmental scheme of lexical diversity.

5.3. To Instruct Students to Use Sentences with Diversity and Accuracy

On top of a good command of lexical resources, students' grammar range and accuracy is defined as another paramount aspect in achieving high score. To liberate our students from those hackneyed and mind-numbing sentence templates, as well as to increase the grammar variety of sentences, we can take full advantage of the self-constructed model essay corpus. Both natural and sophisticated sentence structures and patterns which are geared to the grammar focus unfamiliar to foreign language learners can be derived from it. Due to the colossal architecture of English grammar, it is preferable to give prominence to some functional sentences of expressing opinions, exemplification, emphasis, and nominalization.

The formality of sentence is proportional to the percentage of nominalization. Nominalization is the use of a verb, an adjective, or an adverb as a noun, or as the head of a nominal phrase with or without morphological transformation. This term is also used to refer to the process of deriving a noun from another part of speech by adding affixes such as changing the verb *legalize* to the according noun *legalization*. Generally, nominalized sentences demonstrate a higher level of formality and conciseness than original sentences, notwithstanding their similarity in meanings. Formality, in particular, is usually deemed as an important dimension in academic writing. Examples with nominalization like "*a closer examination of the admittedly small sample of data Dr. Rockwell collected suggested two explanations for what is happening*" and "*the involvement of the participants has been essential to the development of relevant programs*" are easily found from *The Economist*.

Adhering to the student-oriented principle, we can start using part of students' favored verbs as the seeds of nominalization. For instance, student No. 2 used "consider" in her composition:"*If we carefully consider the noise and disturbance of mobile phone, we can see the distraction it exerts in resting places*." However, in the model essay corpus the verb "consider" is used in the nominalied form as in "*Careful consideration of our system of numeration leads to the conviction that…*" Then, this student is guided to improve her sentence as "*Our careful consideration of the noise and disturbance of mobile phone illustrates the distraction it exerts in resting places*."

To sum up, with the help of the model essay corpus, students are expected to obtain some corresponding instructions, tinker with their renditions with grammar diversity and accuracy, thereby becoming more competent writers of their own style.

6. Conclusion

In conclusion, a good mastery of corpus search and analysis methods empower language teachers to employ useful tools in their pedagogical practices. Apart from the apparent convenience of the corpus-based study in language teaching, the disadvantages and limitation are equally evident. First, the corpus-based studies are interdisciplinary, which undoubtedly increases the difficulties of its popularization and application. The construction of even a small-scale corpus necessitates at least months of efforts as well as a good command of computer technology. Furthermore, the access to corpus data also requires higher level of computer literacy. Secondly, we lack a convenient and systematic methodology to explore the internal attributes of the texts at the sentence or paragraph level though it can effectively analyze words and phrases. Thirdly, in the corpus-based research with myriads of data and statistics, the researchers might be overwhelmed by interpreting large piles of data. Last but not the least, it is worth noting that corpus linguistics is constantly updating with new tools, such as Graphcoll [21-22]), which allow researchers to look into more diverse aspects of linguistic and discourse features. Therefore, practitioners working with highly proficient students or those who have more diverse teaching demands need to be informed about emerging corpus tools for a more fine-tuned or specifically tailored application in their corpus-based instruction.

The corpus-based writing studies are far from being well-developed and will still be greeted with obstacles. Nevertheless, both teachers and students furnished with this powerful tool can work better. Despite the limitation and complexity, the studies are gaining increasing momentum and shedding new light in the quantitative analysis of data and application in the writing courses, injecting new vigor into the age-old language teaching and learnging.

References

[1] Alderson, J. (2009). Language Test Construction and Evaluation. Cambridge: CUP.

[2] Biber, D., (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. International Journal of Corpus Linguistics, 14 (3): 275-311.

[3] Conrad, S., (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? TESOL Quarterly, 34, pp. 548-560.

[4] Römer, U., (2011). Corpus research applications in second language teaching, Annual Review of Applied Linguistics, 31: 205-225.

[5] Friginal, E., (2018). Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities. New York: Routledge.

[6] Jiang, T., et al. (2019). Interlanguage: a perspective of quantative linguistic typology. Language Sciences 74: 85-97.

[7] Chen, H., & Xu H., (2019). Quantitative Linguistics approach to interlanguage development: a study based on the Guangwai-Lancaster Chinese learner corpus. Lingua, 230: 1-15.

[8] O'Keeffe, A., & Geraldine, M., (2017). The English Grammar Profile of learner competence: Methodology and key findings, International Journal of Corpus Linguistics, 22 (4): 457-489.

[9] McEnery, T., & Xiao, R., (2010). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 364-380), New York: Routledge.

[10] Lonfils, C. & Vanparys, J. (2001). How to design user-friendly CALL interfaces. Computer Assisted Language Learning, 14 (5), 405-417.

[11] Noguchi, J. (2004). A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers. English Corpus Studies, 11, 101-110.

[12] Liang M, et al. (2005). An Introductory Course of Corpus Application. Beijing: Foreign language teaching and research press.

[13] Dong. N., (2011). The Application of English-Chinese Parallel Corpus in Translation Teaching, Shandong Normal University.

[14] Cambridge ESOL., (2005). Cambridge IELTS 4 Student's Book with Answers: Examination papers from University of Cambridge ESOL Examinations. Cambridge: CUP.

[15] Cambridge ESOL., (2007). Cambridge IELTS 6 Student's Book with Answers: Examination papers from University of Cambridge ESOL Examinations. Cambridge: CUP.

[16] Cambridge ESOL., (2014). The Official Cambridge Guide to IELTS Student's Book with Answers with DVD-ROM. Cambridge: CUP.

[17] He A., (2013). The Phrase Concept of Corpus and Its Teaching Processing. Guangzhou：Guangdong Higher Education Press.

[18] Cai. J., (2005). Lecxical Choice and The Effect. Beijing: Foreign language teaching and research press.

[19] Nesselhauf, N. & Tschichold, C. (2002) Collocations in CALL: An investigation of vocabulary-building software for EFL. Computer Assisted Language Learning, 15 (3), 251-279.

[20] Bowker, L. & Pearson, J. (2002). Working with Specialized Language: A Practical Guide to Using Corpora. London/New York: Routledge.

[21] V. Brezina, et al., (2015), Collocations in context: A new perspective on collocation networks, International Journal of Corpus Linguistics, 20 (2): 139-173.

[22] M. Barlow, (2016), Linking corpus data and discourse structure, International Journal of Corpus Linguistics, 21 (1): 105-115.