Research on Constructing "Parallel Contrast Corpus of Grammatical Errors"*

Jiayi Yao
Beijing Normal University, Beijing, China

Hui Chen
Beijing Normal University, Beijing, China

Yuan Liu
Beijing Normal University, Beijing, China

Abstract—Error analysis and interlanguage are two cores in second language acquisition research. Researchers have conducted studies and established corpora from various perspectives based on Big Data. However, most of the existing interlanguage corpora provide no feedback for students, which resulted in the barrier of improving self-study efficiency. Additionally, interlanguage systems are influenced by nationalities, while there is a vacancy on the construction of divisional interlanguage corpora. Based on previous studies and error analysis of BNU-Cardiff Chinese College students, this study proposes an idea and model of "Parallel Contrast Corpus of Grammatical Errors" for native English speakers in Chinese learning.

Index Terms—Chinese as a foreign language, error analysis, interlanguage, parallel contrast corpus

I. INTRODUCTION

Chinese language and culture have appealed to more foreigners, and the number of international students in China has reached over half a million in 2019, half of which are diploma education students. With greater influence, along with *the Belt and Road Initiative*, China has maintained closer and deeper interactions with the world. Chinese, thus, is used more and more frequently. This leads to a pressing demand of epoch elites who proficient in Chinese language and culture. The established Confucius Institutes throughout the world have made great contributions to cultivate their language ability, however, without a diploma. Thus, many countries start to cooperate with Chinese Universities to train professionals in China-related affairs through diploma education, such as Beijing Normal University-Cardiff University Chinese College (hereinafter referred to as "Chinese College"), whose aim is to cultivate future Chinese-related talents with knowledge of Chinese language and culture. Therefore, it's urgent to improve their Chinese acquisition efficiency and set up an online platform for them to do self-corrections.

Language learners and researchers are concerned about non-standard expressions derived from the interior linguistical system, also known as ERROR (Corder, 1981). However, the learners pay attention to the common use of verbal communication, while the researchers focus on their erroneous expressions. Different focuses result in divergence between teaching and learning, which influences the Chinese acquisition efficiency. The existing Chinese interlanguage corpora also have many limitations, lying in the restricted access to teachers and researchers, while students are not provided with a platform for independent study and self-correction. Moreover, the insufficient corpus annotations make it impossible to conclude the characteristics of interlanguage systems of people under the same language background.

To propagate Chinese, the present study aims to construct "Parallel Contrast Corpus of Grammatical Errors" based on a contrast method, which helps to overcome the defects mentioned above. Every error originated from learners under the same native language is profoundly analyzed with parallel comparison over error form, correct form, translation form and relative explanations. From the contrasts, language learners can seek for the regulations by themselves.

According to previous studies and current corpora, this study focuses on Chinese learners who are native English speakers to deeply investigate their grammar errors and to filter marginal grammar errors. Meanwhile, specific grammatical errors from "Chinese College" students are chosen, corrected, synthesized and analyzed. The outcomes, therefore, not without justification to depict the learning features of this special group, supporting a practical platform for language learners to self-correct their outputs.

II. REVIEW OF CORPORA RELATED RESEARCHES

Previous research includes collecting all the grammatical error corpora from essays that refer to native English speakers. Focusing on grammatical errors in Chinese as a foreign language, aiming to make Chinese learners realize that "the spoken words and written sentences basically conform to Chinese grammar" (Lu, 1998, p.2), the materials are sorted out, forming "Chinese Grammatical Error Corpus" (hereinafter referred to as Corpus I).

A. Definition of "Grammatical Error"

Lu (1984) introduced the concept of "interlanguage" and "error" into China and applied it in teaching Chinese as a foreign language (1992). He also discussed the grammar (including morphology and syntax) in detail (1994). "Error" refers to "(in the speech or writing of a second or foreign language learner) the use of a linguistic item in a way which a fluent or native speaker of the language regards as showing faulty or incomplete learning." (Richards, 2000, p.157). However, Researchers have a divergence of views about its boundary. Some speak of it refers to "illegal" usages (according to the grammatical rules), while others believe it also includes unidiomatic expressions. This study concurs with the first opinion. The reasons are as follows: (1) "Illegal" and unidiomatic usages essentially belong to the grammatical and pragmatic perspectives accordingly. Researches on grammatical errors and the construction of relevant corpora should stick to grammar firmly. (2) Pragmatics can't be explained clearly in a few words due to the complicated Chinese culture. It is also hard for beginners to comprehend and acquire. (3) Language conventions are influenced by regional and personal factors, resulting in the subjectivity in pragmatic judgment. Therefore, this study sifts the unidiomatic expressions out to maximize the objectivity of the outcomes.

TABLE I
UNIDIOMATIC EXPRESSION

E.g.	Original Sentence	Correct Formation
1	píng guǒ duō shǎoqián měi jīn 苹果多少钱每斤?	píng guǒ duō shǎoqián yī jīn píng guǒ měi jīn duō shǎoqián 苹果多少钱一斤? /苹果每斤多少钱?
E.T[1]	How much are the apples per pound?	
CUS[2]	Native language transfer	
2	wǒ men bān yǒu shí qī shí bā gè xué shēng 我们班有十七、十八个学生。	wǒ men bān yǒu shí qī bā gè xué shēng 我们班有十七八个学生。
E.T	There are seventeen or eighteen students in our class.	
CUS	Insufficient knowledge of the target language	

The examples above are unidiomatic expressions, which are rarely heard from native Chinese. E.g.1 is clearly influenced by English. The word "per" stands for "měi 每" in Chinese. Through literal translation, "měi jīn 每斤" is placed at the end of the sentence as the phrase "per pound" is, causing the disorder in the Chinese sentence. However, it doesn't affect understanding and communication. Similar occasions also occur in native Chinese. For instance, people in Northern China always use Inversion Sentence, like "chī le ma, nín 吃了吗,您? (have had a meal, you?)", in contrast to the conventional way "nín chī le ma 您吃了吗? (you have had a meal?)". It doesn't belong to grammatical error for its using frequency in the dialect, and this kind of "illegal" usages will transform into "legal" ones if they are widely acknowledged.

E.g.2 refers to Chinese expressions of approximate numbers. The phrase "shí qī, shíbā gèxuéshēng 十七、十八个学生(seventeen or eighteen students)" makes a semantic error without conveying the idea of uncertainty about class size, hence the approximate number "shí qī bā 十七八(lit. ten seven eight)" should be used instead. According to the definition of grammar (Richards, 2000), the original form is correct.

To summarize, subjective experience is frequently used to define unidiomatic usages when analyzing corpus materials, leading to controversial explanations. Thus, it is necessary to prescribe the objects of this research, which are the "illegal" outputs with grammatical errors.

B. Questions on "Reached Achievements"

After collecting and analyzing independently of "illegal" corpora based on grammar rules, the outcome shows that some previous conclusions are controversial. These questioned corpora are categorized as "Questioned Errors", including inappropriate classification, lack of context, misplaced error and artistic treatment.

1. Inappropriate classification. It is mainly caused by partial analysis of multi-typed errors, referring to more than one error in a sentence, while researchers often focus on one grammatical point so that they might ignore others. Typical examples are listed below:

---

[1] "E.T" is the abbreviation of "English Translation".
[2] "CUS" is the abbreviation of "cause".

TABLE II
INAPPROPRIATE CLASSIFICATION

E.g.	Original Sentence	Correction by Researcher
3	rú guǒ bǎ wǒ néng cǎi yòng wǒ yuàn zuò zài nǐ men gōng sī de tuī xiāoyuán 如果把我能采用，我愿做在你们公司的推销员。	rú guǒ bǎ wǒ néng cǎi yòng wǒ yuàn zuò nǐ men gōng sī de tuī xiāoyuán 如果把我能采用，我愿做你们公司的推销员。
C.F[3]	rú guǒ néng bǎ wǒ lù yòng wǒ yuàn zuò nǐ men gōng sī de tuī xiāoyuán 如果能把我录用，我愿做你们公司的推销员。	
E.T	If I were hired, I would be willing to work as a salesman in your company.	
4	xiě dé jù zi zǒng shì tài cháng yǒu shí hou xiě cuò zì 写得句子总是太长 有时候写错字。	tā xiě dé jù zi zǒng shì tài cháng yǒu shí hou hái xiě cuò zì 他写得句子总是太长 有时候还写错字。
C.F	tā xiě de jù zi zǒng shì tài cháng yǒu shí hou hái xiě cuò zì 他写的句子总是太长，有时候还写错字。	
E.T	Sentences are too long, sometimes typos.	

The researcher takes the error in e.g. 3 as the redundant preposition "zài 在(on)" before the object, while other errors are ignored: (1) The disordering of a modal verb in "bǎ zìjù 把字句 (passive disposal sentence)". (2) Confusing "lù yòng 录用 (hire)" and "cǎi yòng 采用 (adopt)".

The most prominent error of e.g.4 is the lack of adverb "hái 还 (and also)". The researcher studies adverbs, but neglects the indistinct usage of "de (variants 的, 地, 得 according to different patterns)". Their sole focus results in inappropriate classifications of multi-typed errors. Although some of them make corrections, they don't give complete explanations.

2. Lack of context. Corpus without context is always ambiguous. Given a proper context, the error is corrected. During a certain process of survey, researchers always pick out the sentence with errors from a paragraph or discourse, detaching it from the context. Although it reduces workload and seems clear enough, the ambiguity can't be ignored to some extent.

Without context, e.g.5 can be explained in different ways: (1) The two characters have no blood relation, but "he" treats "me" as his own child. Under this circumstance, the original sentence is complete without modification. (2) If the narrator claims that he or she has grown up, then the conjunction "hái 还" should be added to express the complaining tone towards the other character, who still regards the narrator as a little kid. The researcher only analyzes the latter without giving a supportive context, thus the readers may feel confused. E.g.6 is also ambiguous due to different tenses: (1) September is in the future and "he" has not but will come to China. Thus, the usage of "zài 再" is appropriate, but "le 了" is redundant for collocation principles. (2) September has passed and "he" has already come to China. "yòu 又" should be used accordingly, collocating with the structural auxiliary word "le 了" to show the past tense. To improve corpora, the context is a significant clue to assess the error accurately and make appropriate modifications.

TABLE III
LACK OF CONTEXT

E.g.	Original Sentence	Correction by Researcher
5	kě shì tā bǎ wǒ dāng chéng hái zi yí yàng duì dài wǒ 可是，他把我当成 孩子一样 对待我。	kě shì tā hái bǎ wǒ dāng chéng hái zi yí yàng duì dài wǒ 可是，他还把我当成 孩子一样 对待我。
E.T	However, he (still) takes me (as a kid) and treats me like a kid.	
A.R[4]	The condition that he treats me as a kid is ongoing, thus "hái 还" should be added[5].	
6	jīn nián jiǔ yuè tā zài lái zhōngguó le 今年九月，他再来 中 国了。	jīn nián jiǔ yuè tā yòu lái zhōngguó le 今年九月，他又来 中 国了。
E.T	This September, he came to China again.	
A.R	The main difference between adverbs "zài 再" and "yòu 又" (both contain the meaning of "again") is "zài 再" is used when a continuous or repetitive action has not happened, while "yòu 又" for already happened[6].	

3. Misplaced error. This indicates grammatically correct corpora which are able to use in daily communication. The redundant error is placed and needless analysis is done because some researchers try to support their research without independent analysis of each corpus; some might be subjectively influenced by their regional and traditional culture.

---

[3] "C.F" is the abbreviation of "Correct Formation".

[4] "A.R" is the abbreviation of "analysis by researchers".

[5] The original text is Chinese. 把我当孩子看这种状态的持续进行,所以应加上"还". The English is translated by the author.

[6] The original text is Chinese. "再"和"又"的主要区别是表示动作重复或者继续时，"再"用于未实现的，"又"用于已实现的. The English is translated by the author.

TABLE IV
MISPLACED ERROR

E.g.	Original Sentence	"Correction" by Researchers
7	měi guó rén dōu jué de zuì hǎo cóng zì jǐ de cuò wù xué duì de bàn fǎ 美国人都觉得最好从自己的错误学对的办法。	měi guó rén dōu jué de zuì hǎo cóng zì jǐ de cuò wù xué xí duì de bàn fǎ 美国人都觉得最好从自己的错误学习对的办法。
E.T	Americans deem that it's the best way to learn from their own mistakes.	
8	tā bìng bú shì yī wèi hěn shòu de nǚ hái ér shì tā de shēn cái bǐ jiào fēngmǎn 她并不是一位很瘦的女孩,而是她的身材比较丰满。	tā bìng bú shì yī míng hěn shòu de nǚ hái ér shì tā de shēn cái bǐ jiào fēngmǎn 她并不是一名很瘦的女孩,而是她的身材比较丰满。
E.T	She is not very thin, but rather, quite plump.	

E.g.7 refers to a choice of monosyllabic or disyllabic words. The modification of "xué学" to "xuéxí学习" is needless because "xué学" can replace "xuéxí学习" as they both stand for study. For e.g.8, there are no differences in meanings and semantic preference between classifier "wèi 位" and "míng 名" that both indicating a person, hence it's unnecessary to make a distinction here.

4. Artistic treatment. Literary expressions focus on elaborating different words, artistic conceptions and elucidation of ideas. Therefore, grammar is not the main concern. A treatment that meant to create an artistic conception is acceptable and cannot be regarded as an error.

TABLE V
ARTISTIC TREATMENT

E.g.	Original Sentence	Correction by Researchers
9	bù zhī bù jué wǒ zài běi jīng yǐ jīng guò le liǎng gè qiū tiān liǎngnián de shí jiān 不知不觉，我在北京已经过了两个秋天，两年的时间， fā shēng le xǔ duōràng wǒ yú kuài ràng wǒ shāng xīn de shì qíng 发生了许多让我愉快，让我伤心的事情。	bùzhībùjué wǒ zài běijīng yǐjīng guò le liǎng gè qiūtiān zhèliǎngnián de shíjiān fāshēng le 不知不觉，我在北京已经过了两个秋天，这两年的时间，发生了 xǔduō ràng wǒ yúkuài ràng wǒ shāngxīn de shìqíng 许多让我愉快，让我伤心的事情。
E.T	Unconsciously, I've witnessed two autumns in Beijing, when many cheerful and sad things have happened.	

There is no need to add the pronoun "zhè这" before the phrase relates to time "liǎng nián de shī jiān 两年的时间". In addition, grammar errors in literary sentences are personalized and can only be applied to high-leveled language learners, hence these "errors" have little reference value.

These four types of mistakes above will highly impair the accuracy of analysis and the efficiency of corpus data. To use the corpora efficiently and make error analysis reasonable, we should: (1) List the correct formation in English. It helps to figure out the relationship between error and the negative influence of the native language during the comparison. Moreover, it is easier to approach what Chinese learners are trying to convey from the perspective of English expression, paving a way to analyze and modify errors accurately. (2) Annotate the necessary context. When doing error analysis, researchers should give brief explanations to avoid ambiguity and provide abundant information for readers to make independent analyses. (3) Make complete corrections. Take every possibility into consideration, especially when analyzing corpora with multi-typed errors. (4) Error corpora aim mainly at benefiting beginners and intermediate-leveled language learners. Sentences with literal modifications should be chosen carefully.

III. ERROR OUTPUTS OF "CHINESE COLLEGE" STUDENTS

This study mainly targets "Chinese College" students, gathers their grammar errors from daily quizzes throughout the whole semester, and forms the "Grammatical Error Corpus of Chinese College Students" (hereinafter referred to as Corpus II). The characteristics of this group can be concluded according to the results. Comparing to Corpus I, the similarities and differences between "Chinese College" students and other native English speakers can be illustrated.

A. Characteristics of Chinese College Students

This study selects junior-year "Chinese College" students who have a similar environment and cultural background. All of them, generally beginners, have been studying Chinese for less than three years. According to Corpus II, they have made similar errors, such as the confusion of the word "zhuān mén 专门 (specialize)" when making sentences with it.

According to Table 6, students generally believe that "zhuān mén 专门" is a noun, while it can only be applied as an adverb. When outputting Chinese, they fail to distinguish "zhuān yè专业 (major)", "zhuān zhí专职 (specific duty)" and "zhuān mén 专门". One possible reason is the insufficient input of Chinese, or it may lie on the negative influence of English. The word "major" can be used as a noun, a verb as well as adjective, with flexible positions in sentences. When translating "major" into Chinese, students misuse "zhuān mén 专门" to substitute "zhuān yè专业" and "zhuān zhí专职".

TABLE VI
CHINESE COLLEGE STUDENTS USE THE WORD "ZHUĀN MÉN 专门" IN SENTENCES

E.g.	Error Formation	Correct Formation
10	méi pó de zhuānmén shì zuò méi 媒婆的专门是做媒。	méi pó shì zhuānmén zuò méi de rén méi pó de zhuān zhí shì zuò méi 媒婆是专门做媒的人。/媒婆的专职是做媒。
E.T	Matchmakers are people who specialize in matchmaking.	
11	hàn yǔ shì wǒ men de zhuānmén 汉语是我们的专门。	wǒ menzhuānmén xué hàn yǔ 我们专门学汉语。
E.T	We specialize in Chinese.	

In accordance with Corpus II, the main causes of errors are the unfamiliarity with part of speech and the meaning of a word. As a result, for beginners, the difficulty lies in vocabulary accumulation. Advanced grammar rules are not their primary missions. To improve acquisition efficiency, graded teaching with proper learning materials is necessary, to which the construction of corpora with learner's Chinese levels and error categories is helpful.

B. Distinctions between Corpus I and Corpus II

1. Similarities. Beginners and intermediates, including "Chinese College" students, always have difficulty with everyday vocabulary, especially content words like verbs and nouns. They are also negatively influenced by their mother tongue as other native English speakers are.

TABLE VII
COMPARISON OF ERROR CORPORA ABOUT "JIÉ HŪN 结婚(MARRY)"

E.g.	Error Formation	Correct Formation
12 (Corpus I)	mǎ lì jié hūn le tāngmǔ 玛丽结婚了汤姆。	mǎ lì hé tāngmǔ jié hūn le mǎ lì jià gěi le tāngmǔ 玛丽和汤姆结婚了。/玛丽嫁给了汤姆。
E.T	Mary married Tom.	
13 (Corpus II)	tā dǎ suànmíngnián jié hūn tā de nǚ péngyǒu 他打算明年结婚他的女朋友。	tā dǎ suànmíngnián hé tā de nǚ péngyǒu jié hūn qǔ tā de nǚ péngyǒu 他打算明年和他的女朋友结婚/娶他的女朋友。
E.T	He plans to marry his girlfriend next year.	

"Marry" is a transitive verb, while "jiéhūn 结婚 (marry)" usually collocates with the conjunction "hé和 (with)". Besides, "jiéhūn 结婚" can be substituted with the intransitive verb "jià嫁 (a woman getting married)" and the transitive verb "qǔ娶 (a man getting married)", sharing the same English expression of "marry someone". Learners, especially beginners, translate "marry" directly into "jiéhūn 结婚" and "jià嫁" due to the negative transfer of native language, missing the conjunction "hé和" and the preposition "gěi 给" accordingly. It also causes disorder in Chinese sentences.

2. Differences. "Chinese College" students present their own group distinctions. (1) They have difficulty in dealing with some Chinese words or phrases as other native English speakers do. However, their erroneous outputs are supplements to the mentioned error expressions, see the comparison of the degree adverb "yuèlái yuè越来越 (more and more)" below. In Corpus I, the collocation with adjectives is grammatically correct, while in Corpus II, the erroneous usages are self-created by "Chinese College" students.

TABLE VIII
COMPARISON OF ERROR CORPUS INCLUDING "YUÈ LÁI YUÈ 越来越"

E.g.	Error Formation	Correct Formation
14 (Corpus I)	zài xiàn dài shè huì shàng rén menyuè lái yuèzhòng de gè rén zhǔ yì xīn lǐ 在现代社会上，人们越来越重的个人主义心理。	zài xiàn dài shè huì rén men de gè rén zhǔ yì xīn lǐ yuè lái yuèzhòng 在现代社会，人们的个人主义心理越来越重。
E.T	In modern society, people are becoming more and more individualistic.	
15 (Corpus II)	lái zhōngguó yǐ hòu wǒ yuè lái yuè rèn shi gèngduō de zhōngguó rén 来中国以后，我越来越认识更多的中国人。	lái zhōngguó yǐ hòu wǒ rèn shi le yuè lái yuèduō de zhōngguó rén 来中国以后，我认识了越来越多的中国人。
E.T	I've met more and more Chinese after I come to China.	
16 (Corpus II)	tā de míngshēngchéngwéi yuè lái yuè dī de le 她的名声成为越来越低的了。	tā de míngshēngyuè lái yuèchà le 她的名声越来越差了。
E.T	Her reputation becomes worse and worse.	

Despite the redundant noun of locality "shàng 上 (on)" after the preposition "zài 在 (a preposition indicates direction)" in e.g.14, there is also a word order mistake in the second clause. The collocation "yuèlái yuèzhòng 越来越 重 (heavier)" is without error, but this phrase cannot function as an adjective to modify the nominal phrase "gèrén zhǔ yìxīn lǐ 个人主义心理 (psychology of individualism)", after which it can be applied as a degree complement. In e.g.15 from Corpus II, the degree adverb "yuèlái yuè越来越" should modify adjective "duō 多 (many)" and the collocation "yuèlái yuèduō 越来越多" modifies the nominal phrase "zhōng guórén 中国人 (Chinese)", but it can't collocate with verbs, such as "rèn shi 认识 (know)".

As for e.g.16, "chéng wéi 成为 (become)" and "yuèlái yuèdī 越来越低 (lower)" are both predicates, causing

erroneous sentence structural. Only one predicate should remain. "chéng wéi 成为" is redundant for the subject "tā de míng shēng 她的名声 (her reputation)" is incompatible with it. In addition, the adjective "dī 低 (low)" can't modify the noun "míng shēng 名声 (reputation)", which should be replaced with "chà 差 (worse)" instead.

(2) Marginal error formations that haven't been mentioned by previous researchers. "Chinese College" students collectively produce some similar grammar errors that are not regarded as typical error outputs. In preposition studies, many researchers only pay attention to the misuse of "yóu yú 由于 (because of)", excluding studies of the phrase "yóu yú……yuán gù 由于......缘故 (because of)". However, "Chinese College" students fail to apply this phrase correctly.

There are differences between Chinese and English forms of causation words and phrases. In Chinese, "yóu yú 由于" shows a causal relationship, which can be substituted by paired phrase "yóu yú……yuán gù 由于......缘故" with exactly the same meaning. However, there is a lack of paired causation phrases in English, thus language learners often over-generalize the rules of paired conjunctive phrases in the target language. In e.g.17, the language learner may imitate the word order of a more frequently-used paired conjunctive phrase and apply it to the usage of "yóu yú……yuán gù 由于......缘故". While e.g.18 omits the auxiliary word "de 的" before "yuán gù 缘故", mainly because of their insufficient vocabulary of the target language.

TABLE IX
ERROR CORPORA OF "YÓU YÚ……YUÁN GÙ 由于……缘故" FROM CORPUS II

E.g.	Error Formation	Correct Formation
17	yóu yú tā de zhàng fu sǐ le yuán gù 由于她的丈夫死了缘故。	yóu yú tā de zhàng fu sǐ le de yuán gù 由于她的丈夫死了的缘故。
E.T	Because of the death of her husband.	
18	yóu yú tiān qì bù hǎo yuán gù qǔ xiāo bǐ sài 由于天气不好，缘故取消比赛。	yóu yú tiān qì bù hǎo de yuán gù qǔ xiāo bǐ sài 由于天气不好的缘故取消比赛。
E.T	Due to the bad weather, the match was canceled.	

According to the results above, it's necessary to set up a corpus for a certain group due to their own characteristics, especially for long-term projects. Moreover, when using a corpus to assist teaching, teachers should re-sift the material based on their nationalities, Chinese levels, study aims and so on. During the educating process, new types of errors are supplementary materials with a reference value to both researchers and other language learners. Thus, a tunnel for feedbacks between teachers and researchers should be constructed.

IV. THE CONCEPTION OF CONSTRUCTING PARALLEL CONTRAST CORPUS OF GRAMMATICAL ERRORS

A. The Background

Big data is regarded as "an effective mode of learning a language" (Jiao, 2013, p.11), and the current corpora have made huge progress. For example, HSK dynamic Composition Corpus has a function of searching nationalities, which classifies Chinese interlanguage systems according to different countries. Scholars also attempt to carry out divisional researches on African countries where HSK seldom refers, such as the conception of building an interlanguage corpus of a sole country (Huang, 2019). It bases on Chinese learners in Rwanda who are at different Chinese levels, compensating the current deficiencies of interlanguage corpora.

Nevertheless, they still have a long way to go (Zhang & Cui,2015). Li (2016, p.116) pointed out "corpus that contains abundant data but is not annotated and classified is of limited help to learners". The current corpora are demonstrated in a single language without correct formations, let alone involving the comparison between the target language and the mother tongue of the language learners. These single-language interlanguage corpora mainly open for researchers and teachers. Language teachers are mostly accessible to collect different error forms from students, but they always fail to do researches and analyze complicated linguistic phenomena due to the lack of time and ability. Instead, they focus more on teaching. In contrast, researchers have a lag on updating error forms and usually conduct studies on age-long data. Even if they intend to collect present corpora for their surveys, the access is limited. It virtually shows that there is a gap between teaching and research, and the research results have no direct reference to teaching activities, which restricts its practical value.

B. The Contents and Influence

Enlightened by Lü (1992) and Zhang (2019), this study aims at constructing a parallel corpus including Chinese, English and Chinese interlanguage especially for "Chinese College" students, which is also a reference to other native English speakers. Chinese formation, a correction to Chinese interlanguage, allows language learners to do self-learning after class and consolidate what they have learned through comparisons as well as homogeneous examples. English translation removes the language barriers by helping them comprehend the exact meaning efficiently. In conclusion, the parallel comparisons provide more ways to explain errors.

To collect authentic and representative corpora, the construction of the "Parallel Contrast Corpus of Grammatical Errors" (hereinafter referred to as Corpus III) is divided into two parts. (1) Review relevant research findings and collect all the grammatical error corpora produced by native English speakers, forming a basic framework of their

Chinese interlanguage system. (2) Collect error outputs by Chinese College students. Comparing them with the former results to identify their difficulties and sift the most common errors during Chinese learning.

TABLE X
EXAMPLES OF SEGREGATORY WORDS[7] IN VERB-OBJECT STRUCTURE FROM CORPUS III

E.g.	Chinese Formation	Chinese Interlanguage Formation
19 (Corpus I)	dǎ suànmíngnián jié hūn wǒ de nǚ péngyǒu 打算 明 年结婚我的女朋友。	wǒ dǎ suànmíngniángēn wǒ de nǚ péngyǒu jié hūn 我打算 明 年跟我的女朋友结婚。
E.T	I plan to marry my girlfriend next year.	
20 (Corpus II)	wǒ zài zhōng lóu zhàoxiàng tā 我在 钟 楼照 相 他。	wǒ zài zhōng lóu gěi tā zhàoxiàng 我在 钟 楼给他照 相 。
E.T	I took pictures of him in the bell tower.	
A.R	Foreign students whose native language is English regard intransitive segregatory words in verb-object structure as transitive verbs. They directly add objects but omit the necessary prepositions before the objects.	

Corpora in Corpus III are demonstrated in parallel language forms with professional annotations and error analysis. Examples of segregatory words in verb-object structure are listed above. The innovations lie in three aspects. (1) Special research objects. Corpora are mainly collected from "Chinese College" students. Meanwhile, negative influence from uncertainties including the Chinese level, cultural background, age is under control. (2) Fresh new achievements. This corpus provides a brand-new pattern of multilingual comparisons and analyses. It compensates for the deficiency of current single-language corpora without grammatical annotations. It can also test how auxiliary teaching tools work on Chinese learners. (3) Appropriate context. When collecting new error corpora that may cause misunderstanding, its context will be annotated, providing more referential information to readers.

C. Anticipated Effects

The anticipated effects of Corpus III mainly contain: (1) A clear boundary of "grammatical error". (2) Accurate error classifications. (3) Suitable materials for learners at a certain level. It can be directly applied to "Chinese College" students, providing a platform for self-correction and active learning. (4) The parallel contrast method helps Chinese learners dissect their own grammatical problems by comparing different languages.

V. CONCLUSIONS

This study aims at constructing a "Parallel Contrast Corpus of Grammatical Errors" based on previous research findings and error corpora from "BNU-Cardiff Chinese College" students. It bridges corpus and language learners with multilingual annotations and compensates for the demerits of previous studies. The attempt to construct a corpus for a certain group, especially for "Chinese College" students, is a student-centered movement, allowing students to do self-correction of their second language outputs. Researchers can also deduce the learning characteristics of this group from the results.

REFERENCES

[1]   Corder, S.P. (1981). Error Analysis and Interlanguage. Oxford: Oxford University Press.
[2]   Huang, Xiaoqin (2019). The conception of creating a single-country interlanguage corpus--Take the Rwandan students with multilevel Chinese as an example. *Chinese Character Culture* DOI:10.14014/j.cnki.cn11-2597/g2.2019.07.009.
[3]   Jiao, Binkai (2013). Research on the data-driven auxiliary mode of teaching Chinese as a foreign language. *Language Planning* DOI:10.16412/j.cnki.1001-8476.2013.29.056.
[4]   Li, Juan (2016). Corpus-driven teaching and application of Chinese as a foreign language. *Modern Chinese* 2016(08):116-118.
[5]   Lu, Jianji (1992). Error analysis and teaching Chinese as a foreign language. *Applied Linguistics* DOI:10.16499/j.cnki.1003-5397.1992.01.012.
[6]   Lu, Jianji. (1994). Analysis of grammatical errors of foreigners in learning Chinese. *Language Teaching And Research* 1994(1):49–64.
[7]   Lu, Jianji. (1984). Interlanguage theory and analysis of the phonetic errors of foreigners learning Chinese. *Language Teaching and Research* 1984(03): 44–56.
[8]   Lu, Jianming. (1998). Some problems we often need to think about in teaching Chinese as a foreign language. Language Application. *Applied Linguistics* DOI:10.16499/j.cnki.1003-5397.1998.04.001.
[9]   Lǚ, Bisong. (1993). On the research of Chinese interlanguage. *Applied Linguistics* DOI:10.16499/j.cnki.1003-5397.1993.02.005.
[10]   Richards, J.C. (2000). Longman Dictionary of Language Teaching & Applied Linguistics. Beijing: Foreign Language Teaching and Research Press.
[11]   Zhang, Baolin. (2019). On the application of Chinese inter-language corpora in TCFL. *Language Teaching and Research* 2019(02):18-27.
[12]   Zhang, Baolin. & Cui, Xiliang. (2015). On the standards of building a Chinese inter-language corpus. *Applied Linguistics* DOI:10.16499/j.cnki.1003-5397.2015.02.017.

---

[7] A segregatory word should be regarded as a word-phrase unit. It has both semantic cohesion and grammatical looseness. Its form has the characteristics of a phrase, and in the sense of a word, so it is an intermediate or transitional linguistic unit between a word and a phrase.

Jiayi Yao was born in Changsha, China in 1999. She currently studies in the School of Chinese Language and Literature of Beijing Normal University, China. Her research focuses on second language acquisition.

Hui Chen was born in Sichuan, China in 1999. She currently studies in the School of Chinese Language and Literature of Beijing Normal University, China. Her research focuses on Modern Chinese.

Yuan Liu was born in Weifang, China in 1999. She currently studies in the School of Chinese Language and Literature of Beijing Normal University, China.