The Study of Error Types of Chinese Learners' Written Texts: A Chinese Written Corpus-Based Study

Jia-Fei Hong, Hsin-Tzu Jen, and Yao-Ting Sung

Abstract The present study aims to tackle the issue of error in written texts by Chinese learners from a macro perspective. Although previous research has demonstrated the significance of positive feedback and effective correction in the realm of Second Language Acquisition (SLA) (Fathman and Whalley, 1990; Ashwell, 2000; Ferris and Robers, 2001; Chandler, 2003), little consensus has been reached regarding its practical implementation in pedagogy. In particular, writing holds a crucial role among the four basic language skills for its complex construction and meaning in written language. However, with the rise of corpus linguistics, new approaches and perspectives have been added to the study of Chinese as a Second Language (CSL) writing (Chang et al., 2015; Hong et al., 2018). In hopes of improving the teaching of writing in an integrated way, this study adopts methodologies from SLA and corpus linguistics to broaden the scale of interdisciplinary research. Through the lens of error analysis, this study examines data from learners with diverse backgrounds in Chinese Written Corpus and analyzes learners' error types with reference to the categorization proposed by Dulay et al. (1982). The results of this analysis identify possible contributing factors of various types of errors, such as native language and level, which can then be further analyzed and may account for learners' error patterns. The present study's findings yield significant insights in outlining the distribution

This work was supported by National Taiwan Normal University's Chinese Language and Technology Center. The center is funded by Taiwan's Ministry of Education (MOE), as part of the Featured Areas Research Center Program, under the Higher Education Sprout Project.

J.-F. Hong (✉)
Department of Chinese As a Second Language, National Taiwan Normal University, 162, Section 1, Heping East Road, Taipei City 106, Taiwan
e-mail: jiafeihong@ntnu.edu.tw

H.-T. Jen
Department of East Asian Languages and Literatures, University of Hawaii at Manoa, 2500 Campus Road, Honolulu, HI 96822, USA

Y.-T. Sung
Department of Educational Psychology and Counseling, National Taiwan Normal University, 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan

H. H.-J. Chen et al. (eds.), *Learner Corpora: Construction and Explorations in Chinese and Related Languages*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-19-5731-4_8

of errors in CSL writing and provide teachers and future researchers with practical advice on the study of teaching strategy, instructional setting, and teaching sequence.

Keywords Error analysis · Chinese Written Corpus · Corpus linguistics · CSL

1 Introduction

With the emerging number of Chinese learners worldwide, Chinese has become a dominant language in the twenty-first century and is gradually becoming one of the most popular languages besides English. According to data from the Department of Statistics at the Ministry of Education in R.O.C, the number of international students entering Taiwan to learn Chinese is growing exponentially, which rose from 8,182 to 18,645 between 2005 and 2015.[1]

In an attempt to help language learners develop well-rounded language competence, learners tend to be exposed to exercises that are focused on four fundamental language skills: listening, speaking, reading, and writing. During the process of learning a second language, learners tend to have difficulty with speaking and writing skills. Specifically, due to the nuanced meanings and the rather complex sentential structures of written language, writing is considered to be more difficult than speaking for second language learners. Students often need to put more effort into the process of writing, and teachers are also required to invest more time in providing feedback. The tendency of Chinese learners' error types described in this study, which is drawn from comprehensive and objective data, is provided to the current teachers and learners of a second language. To learners, the key to using a language fluently and communicating well is to understand grammar and develop language competence (Nassaji & Fotos, 2011). When learning a new language, obstacles in the acquisition of grammar often produce ungrammatical sentences. Theories of Second Language Acquisition (SLA) identify the benefits of positive feedback in helping learners to develop second language competence (Fathman & Whalley, 1990; Ashwell, 2000; Ferris and Robers, 2001; Chandler, 2003). Through both theoretical studies and practical settings, it has been discovered that learners tend to struggle more with speaking and writing than with listening and reading. Enlightened by further exploration, writing actually plays a more intractable role than speaking. Writing skills require learners to master sentential structures that are more complex, as well as be proficient in the nuance of meaning in the written text. Therefore, students must invest more time in learning. Furthermore, the teacher also needs to put more effort into correcting vocabulary and grammar. Due to the difficulties of learning a second

[1] According to statistic data in the report "number of university international students in degree programs and language programs". The report is excerpted from "important statistic data in education" that published on the website by the Department of Statistics in the Ministry of Education in R.O.C. http://depart.moe.edu.tw/ED4500/cp.aspx?n=002F646AFF7F5492&s=1EA96E4785E6838F#.

language, it is relatively hard for foreign learners to have a noticeable improvement in their writing performance (Buckingham & Pech, 1976).

In the field of Chinese as a Second Language (CSL), there are unsolved problems between theory and practice. Due to a lack of research that analyses the application and teaching strategies of CSL teaching, while accounting for learners' backgrounds and levels, theoretical perspectives often fail to address the actual challenges of learning a second language. Additionally, the existing research that studies errors of Chinese learners tends to solely concentrate on learners speaking a particular native language, learners at a particular level, or learners using a particular linguistic form. Although the outcome of these studies can indeed provide insight into the phenomenon of particular learners, a comprehensive view of learners' error types remains unseen. Considering the diverse backgrounds of CSL learners, distinct patterns of errors may emerge from individual native languages. Also, learners at different levels tend to have varying kinds of errors and learning difficulties. The current solution for students with different backgrounds is to assign them to different learning tracks, such as regular class, intensive class, theme-based class, and so forth, according to their native language or level. The drawback of this system is that the placement is solely based on the student's class level, and no attention is paid to the influence of the learner's native language. Even though the same course material, class arrangement, and teaching procedure can be provided, the influence of a student's native language may still influence the kinds of errors that are made and the different language levels.

In an attempt to address the aforementioned gaps in research, this study will take a top-down perspective to investigate students' learning and discuss the distribution of errors from learners of distinctive backgrounds in terms of native language and level. Furthermore, different error types will be analyzed to understand the pattern of grammatical errors in hopes of facilitating the instructional design and teaching strategies.

The Chinese writing corpus used in this study includes 43 written texts from learners of diverse backgrounds and levels and is built according to the framework of the ACTFL writing proficiency test (ACTFL, 2012). With help from the corpus, this study retrieves specific data based on the different "native languages" and "levels" of learners; it is then able to determine if the error types correlate with the grammatical attributes of a learner's native language via their authentic written text. The result of the current study suggests that understanding errors from learners of different levels not only offers implications for the instructional and material design of CSL (Hong and Sung, 2017), but could also improve a learner's overall performance and help them to express their thoughts in writing more effectively (Hong et al., 2018). Notwithstanding the achievement of the Auto-correct Chinese Written Text System, which has 65% accuracy in Auto-detecting Grammar System (Chang et al., 2015) and 88% accuracy in Auto-correcting Written Text Grading (Hong et al., 2014a, 2014b), the information that lies in the pattern of grammatical errors is a critical factor for further breakthrough accuracy.

Considering theories in SLA, corpus linguistics, the application of natural language processing (NLP), and perspectives from second language learners, this

study discusses how to incorporate the findings from common grammar mistakes and error types by Chinese learners in the field of CSL. Moreover, in light of interdisciplinary design, this study seeks to identify applications for the result of this study and further development. In order to achieve the goal of nationality-based differentiated instruction both accurately and effectively in a comprehensive, systematic, and objective manner, this study examines the error types of CSL learners in the written text through research methods in corpus linguistics using "Chinese Written Corpus." Meanwhile, this study also categorizes the error types from learners of different backgrounds and levels and constructs a framework of error patterns through cross-checking. When teaching a second language, teaching materials, methodologies, and teaching strategies should all be differentiated according to an individual student's native language and level. Hence, the corresponding differentiation is an inevitable question in this study. If the data of grammatical errors can be described and analyzed in a comprehensive and objective way based on learners' native languages, levels, and the linguistic forms they use, it would offer CSL teachers, learners, and textbook writers effective strategies for language learning and teaching. Thus, the present study aims to construct a framework of error patterns that is relevant to teaching Chinese writing and to accurately identify the mistakes in a written text by cross-checking grammatical errors in the corpus. These error patterns can thereby provide CSL teachers with advice on how to design teaching materials and give feedback to Chinese learners for self-learning, as well as provide strategies for teaching Chinese learners that speak different native languages and are at different levels. With the aid of this framework, the effectiveness and efficiency of learning and teaching Chinese writing would significantly improve.

2 Literature Review

This section will review the existing literature relating to second language acquisition, error distribution of CSL learners, CSL pedagogical grammar, and corpus-based methodologies.

2.1 Sla

Second language acquisition, psychology, cognitive psychology, and education are all closely related. Different approaches and theories have proposed different perspectives to account for the factors that influence language acquisition and the application of effective pedagogy. The following section includes discussions that are related to theories of language acquisition, types of errors, and the causes of errors.

2.1.1 Theories of Language Acquisition

Since 1990, cognitivism has gradually become the dominant theory in the field of language acquisition. In Universal Grammar (Chomsky, 1995), it is stated that the human brain is equipped with a device that enables humans to acquire grammar and language. This device adopts a universal principle that formulates certain language structures, which embodies diverse forms and causes the distinction between languages. Studies in cognitive linguistics also emphasize the psychological process of learning and processing information. The emergence of Universal Grammar and cognitive theories consequently put error analysis in a crucial position in the study of language acquisition and teaching.

In the article "The significance of learner errors" (Corder, 1967), Corder suggests that teachers should pay close attention to the errors that students are unaware of. Likewise, the concept of interlanguage, which was proposed by Selinker (1972), emphasized that the transition from a learner's native language to a target language is systematic and analyzable. The value of the study of interlanguage lies in the prediction of possible errors by students and the prevention of learners' fossilization. Thereafter, studies on linguistic errors have gradually received recognition and have led to an increase in methodologies, such as error analysis, contrastive analysis, and so forth. These methodologies are all dedicated to the investigation of systems and types of errors by students at different levels and aim to develop particular strategies to facilitate the teaching of a second language. Many recent studies have also discovered that there is considerable disparity in possible difficulties and error types between beginners, intermediate learners, and advanced learners.

In cognitive structure migration theory, Ausubel (1968) indicated that the existing learning experience contributes significantly to the ongoing process of learning. He stated that the existing learning experience and the ongoing learning process would interact with each other and ultimately form a new cognitive structure. A similar phenomenon can be seen in the acquisition of language. Several types of transfers between languages can be categorized as interlanguage transfer and intra-language transfer based on their source, and positive transfer and negative transfer based on their influence on the learning process. The errors that learners make when learning a new language may be a negative transfer derived from the grammatical rules of their native language. Thus, in the field of CSL, the study of a learner's native language and its influence on a second language holds a central place among various research topics. Many studies have collected, analyzed, and categorized the errors from learners speaking different native languages and have proposed corresponding teaching strategies.

From the studies above, it can be concluded that a learner's level and the different kinds of transfer from their native language are both crucial factors that lead to errors when learning a new language. Apart from the research of language acquisition and cognitive psychology, social and cultural factors are included in the study of language teaching and learning as well. Furthermore, with the rapid development of digital technology, the study of language teaching has not only had a substantial breakthrough in data processing and analysis, but has also been closely connected

with digital content. Since the teaching of language is inevitably oriented by these aspects, it should focus not exclusively on errors due to linguistic influence, but should also take into account the difficulties drawn from cultural factors, social factors, and teaching strategies.

2.1.2 Types of Errors

The terminology "error" in SLA refers to an unconscious mistake that correlates to a learner's native language when they are using the target language. In reference to the errors of learners at different stages when learning a target language, Corder (1976) categorized errors into three types: pre-systematic error, systematic error, and post-systematic error. He further explained that a learner's errors would decrease progressively as their grasp on the grammar system of the target language grew. Amidst the continuum, errors that are produced during the period of pre-system and post-system are the most systemic for learners who have not yet mastered the grammar system of the target language.

From a linguistic point of view, Dulay et al. (1982) discussed learners' error types and divided them into the categories of lexical error and syntactic error. After inspecting learners' output based on the disparity in sentential structures from their target language, the structural errors can be further categorized into four types: omission, addition, misformation, and misordering. Omission indicates that the learner left out a necessary part of the sentence or discourse. Addition refers to the error resulting from a redundant grammatical unit in a sentence or discourse. Misordering references a situation where a grammatical unit is misplaced in a sentence or discourse. Misformation refers to the embedding of an inappropriate grammatical unit in certain structures, namely, an error due to misuse of a grammatical unit. Many studies (James, 1998; Zhou et al., 2007) have analyzed error types through the framework of this categorization.

2.1.3 Cause of Errors

The cause of an error when using the target language demonstrates a learner's tendency to approach the new language with the grammar system of their native language, along with a gap in linguistic knowledge toward the target language. Selinker (1972) suggested that the emergence of interlanguage is drawn from five factors: linguistic transfer, overgeneralization, the impact of pedagogy, learning strategies, and communication strategies. In learning transfer, errors are likely influenced by negative transfers from the native language, a lack of knowledge of the target language, cultural factors, learning environment, teaching strategies, drilling methods, or strategies of interpersonal communication.

Limuria (2014) and Okuno (2018) examined errors in *bei* sentences by Chinese learners from Indonesia and Japan, respectively. Limuria (2014) discussed the difficulties that Indonesian learners encounter when learning *bei* sentences in Chinese

and discovered the cause of the errors through the lens of contrastive analysis and error analysis. In Limuria's research, it was found that addition caused the highest percentage of errors, followed by misordering and misformation. Omission was the least prevalent among the four types. Okuno (2018) also inspected the difference in *bei* sentences in Chinese and Japanese and the error types of learners. The results showed that the errors are mainly caused by the distinction in verb form in Chinese and Japanese. The second reason is the semantic discrepancy in the passive voice between Chinese and Japanese. The third reason is "empathy," which compels Japanese learners to focus on human subjects rather than putting a lifeless object as the subject of the sentence. Furthermore, the study also discovered some errors due to the omission of verb complements and the misuse of psychological verbs. Beyond the typical interference from a native language, some Chinese learners from Japan tend to interchange *rang* and *bei*, or omit *bei* in sentences.

From the studies above, universal errors can be found in learners speaking different native languages. Thus, through the contrast between Chinese and a learner's native language, researchers and teachers can target learners speaking a specific native language and then design specific pedagogy and learning strategies to prevent the possible occurrence of errors, and therefore, improve learning effectiveness.

2.2 Error Analysis of Chinese Learners

There is some research that is concentrated on the error analysis of Chinese learners based on their level, nationality, and knowledge of the four language-learning skills. The results of this research are used to develop corresponding teaching strategies.

2.2.1 Error Analysis of Chinese

In studies related to different levels of learners, Hung (2013) attempted to address the difficulties of potential complements for intermediate learners. The "Interlanguage Corpus of Potential Complement for Learners" used in the study is built with data collected from a self-designed questionnaire. The types and percentages of errors from learners are analyzed through the utilization of an interlanguage corpus relating to the acquisition of potential complements by Chinese learners. On par with the percentage of errors, the frequency, complexity, surface structures, and internal semantic structure of complements are jointly considered for the recommended arrangement of pedagogical grammar. Instructional design and teaching strategies are thereby developed to meet the needs of intermediate Chinese learners exclusively. Finally, the study proposes advice and gives recommended revisions pertinent to the design of and strategies for teaching potential Chinese complements through practical techniques in the classroom.

Huang (2014) spent two academic years collecting data from Chinese-language beginners from Japan. The pilot study analyzed the learners' systemic errors in

monosyllable words in the first year and continuously monitored learners' errors in both monosyllable and two-syllable words in the second year. The results of the research showed that, among monosyllable words, the third tone had the highest percentage of error, followed by the second tone, the first tone, and the fourth tone. As for errors in two-syllable words, the highest percentage is found in the tonal combination that begins with the third tone. Huang (2014) then designed a teaching plan based on the outcome of the research. Firstly, it incorporated the concept of pitch to help learners distinguish different tone values in Chinese, then it compared similar stresses and intonations in Japanese and Chinese, and finally, it included drilling exclusive to the third tone.

Huang (2018) inspected the common errors of intermediate Chinese learners from Korea and English learners from the United States in the construction of "one + classifier." The findings of this research indicate that learners from the United States have a stronger tendency toward using the structure of "one + classifier." Surprisingly, learners from Korea remained rather conservative with their use of the structure "one + classifier." This study highlights that errors are derived from a lack of teaching on how to identify the noun phrase in discourse when teaching classifiers, and the reference of a noun phrase is directly connected with the use of the structure "one + classifier."

To understand the impact of a learner's native language, Chen (2011) examined the reason for Thai-speaking Chinese learners' erroneous use of the structural particle "de" by collecting interlanguage data from questionnaires. The study classifies the Chinese structural particle "de" into "de1" and "de2," with eight subgroups based on pedagogical implications. According to the results of this study, the lack of similar structures, such as "pseudo-genitive" and "separable word," in their native language is the main cause of errors by Chinese learners from Thailand.

Similarly, Chuyen (2015) researched the difficulties that Chinese-language learners from Vietnam encounter when learning alternative question sentences from the aspect of grammatical structure. The study conducted a contrastive analysis of sentences in Chinese and Vietnamese with a postulation: sentence forms that are similar in two languages are rather easy to acquire, while sentences that differ in structure cause potential obstacles. With this postulation, Chuyen (2015) collected data from the questionnaire and discovered the distribution of errors made by Vietnamese learners of Chinese alternative question sentences: omission (65%), addition (17%), misformation (12%), and misordering (6%). The causes of these errors are due to the negative transfer from a native language, influence from teaching materials and pedagogy, intervention from the questionnaire, or a lack of linguistic knowledge of Chinese.

As for the teaching of writing, Wang (2011) studied the acquisition of directional complements of Chinese learners whose native language is German by analyzing students' written text. Questionnaires and error analysis were conducted based on the contrastive analysis of Chinese and German and the discussion of teaching materials. Except for misuse among different directional complements, the findings suggest that aspect markers in Chinese, for instance, *le* and *zhe*, jointly contribute to these interlanguage errors.

Liu (2016) conducted an error analysis on the use of sentential conjunctions in writing by Chinese learners from France. By contrasting the correct sentences and sentences with errors in the scope of a compound sentence, paragraph, and discourse, the study looked into the cause of errors in terms of the semantics, pragmatics, and function of each sentential conjunction. In addition to theoretical explanations, the study also provided an instructional model instantiating "ye" and "temporal conjunctions" on par with the textbook used in teaching "An Easy Approach to Chinese" and "Intermediate Chinese Vol. 1" for practical reference.

Tang (2018) retrieved and examined the use of punctuations in interlanguage sentences by learners speaking English and Japanese in TOCFL Learner Corpus, compiled diagnostic tests and related topics with reference to the standard punctuation systems of Chinese, English, and Japanese, and classified various types of misuse by native speakers. The study discovered that errors from native speakers tend to be from related punctuations, such as "" and "", while errors from learners tend to be unrelated punctuations, such as · and ◦. As specific usage often collocates with certain semantic attributes, both native speakers and learners could misapply punctuation due to the uniqueness in its form or meaning. Indeed, the form and meaning of punctuation from a speaker's native language tend to transfer to the target language. The study listed four situations in different punctuation systems that are particularly difficult for learners: punctuation that is similar in shape but has a restricted meaning, punctuation that exists in a particular language system, punctuation with the same meaning but a different shape, and punctuation with a similar shape but a different meaning. Thus, a language teacher should emphasize the correlation between punctuational attributes and linguistic content, as well as their collocation from an integrated perspective.

2.2.2 Teaching Strategies

Liang (2008) conducted research on the acquisition of Chinese classifiers by adult learners. A total of 68 participants (29 native speakers of Korean, 29 native speakers of English, and 10 native speakers of Taiwanese or Chinese) were asked to complete three types of tests (pairing up classifiers and nouns, pairing up classifiers and pictures, and sequencing classifiers based on concreteness). The results of this study showed that native speakers of Korean performed better than native speakers of English in the experiment. The reason for this is rooted in the similarities between Chinese and Korean. More specifically, classifiers also exist in Korean and the cognitive association with classifiers in Chinese and Korean overlaps. In the test of classifiers that are conceptually connected to shape, the most common images provided by native participants are also the most common images from participants with other native languages. In other words, with reference to the different systems of learners' native languages, different pedagogies should be incorporated when teaching Chinese classifiers to adult learners. Likewise, learners are also expected to have different responses to the pedagogies in terms of levels, learning progress, and types of classifiers.

Cai (2014) investigated the errors in character writing by Chinese learners from Japan through the contrastive analysis of characters in Chinese and Japanese. The study analyzed the errors of 10 Chinese learners from Japan in an advanced Chinese summer program at a university in Taiwan and then offered advice on the textbooks and teaching methods that target Chinese learners from Japan. The findings of this research identified six types of errors that are caused by the negative transfer from Japanese characters: (1) errors of same characters; (2) errors of different characters, but same meanings; (3) errors of same characters, but different meanings; (4) errors of non-Chinese characters; (5) errors of non-Japanese characters; and (6) errors of inverted co-morpheme phrases. As for the advice on teaching, "targetization" must be taken into account; concurrently, teachers should have a rather low tolerance level for errors, and they should remain vigilant in identifying them. Furthermore, with regard to the development of textbooks, materials for Chinese learners from Japan should be based on the contrast of characters in Chinese and Japanese, as well as the distinction between the two writing systems.

Chen (2016) discussed the discrepancy of errors between multilingual learners in international schools and ordinary Chinese learners from Thailand. By inspecting the source of errors from multilingual learners through the application of error analysis and the Principle of Temporal Sequence (PTS), Chen (2016) proposed the Lexical Chunk Approach as the solution to the errors in word order. With four months of practice, errors relating to word order decreased significantly, especially with the use of temporal and spatial adverbial modifiers.

2.3 Studies on Chinese Pedagogical Grammar

2.3.1 Pedagogical Grammar

The discussion of pedagogical grammar has long been central to the field of language teaching. Expanding on the foundation of grammar, pedagogical grammar is regarded as a prescriptive form of language for L2 learners to acquire the grammar of a target language in an integrated and logical way. Through progressive learning, learners are able to process information using the logic of the target language and, as a result, reach accuracy and proficiency. Through examining the performance of individual learners and their errors in written text, information can be provided on their ability to communicate in the prescriptive linguistic form.

While learners face many different challenges when learning a second language, writing is considered to be a relatively difficult skill to acquire. In order to produce written language, a learner must integrate grammar and vocabulary based on correct linguistic knowledge, as well as produce a coherent discourse by combining transitional clauses and sentences. Any error in the incorporation of these factors contributes to the production of ungrammatical sentences. Therefore, it is crucial to incorporate pedagogical grammar in the study of CSL. The present study has identified that pedagogical grammar sets out to address the practical needs of CSL in

order to facilitate a student's acquisition of Chinese grammar and leaves the theoretical aspect to linguistics (Zhou, 2002). As emphasized by Nassaji and Fotos (2011), grammar is rooted in every language system, and as such, language cannot function without grammar.

The theoretical value of pedagogical grammar was first recognized by Odlin (1994), who provided theoretical and systematic evidence for the significance of progressive teaching steps of grammar with reference to syntactical and grammatical theories. Pedagogical grammar is a student-centered approach that requires practicality and prescriptivity to address the factors that influence learners, such as intention, competence, and cognition. The goal of pedagogical grammar is to help learners acquire the target language systematically and efficiently so that they are able to communicate in an authentic context. Since the acquisition of linguistic knowledge and grammatical structure of the target language provides CSL learners with the ability to communicate clearly in all skills (writing, especially), the merit of pedagogical grammar in the study of CSL is great and deserves further recognition.

The theoretical systems that systematically extract collections of grammar have tremendous value to researchers and educators; as such, they ought to be viewed as a corpus that allows for the retrieval of needed information. Lv (2008) offered two suggestions for the choosing and arranging of grammar in CSL textbooks. Firstly, considering practicality and concision, a textbook should only include the basic and frequently-used constructions that are necessary for communication and should eliminate constructions that are unnecessary for the preliminary stage of learning through statistics of frequency. Secondly, regarding the shift of paradigm in pedagogy, a more detailed explanation should be attached to topics, vocabularies, and constructions that have been newly added to textbooks for advancing essential communication skills, such as non-subject sentences and single-word sentences. Furthermore, constructions that are more frequently used in written text, rather than in a colloquial context, ought to be removed from textbooks completely. Lv (2008) argues that the implementation of these suggestions would provide value and enhance the learning outcomes of CSL learners.

Pedagogical grammar is a very important element of CSL learning, and it is critical to helping learners to acquire knowledge. In Yang (2000), he indicated that CSL pedagogical grammar is programmable and that it is not arbitrary or orderless. Therefore, pedagogical grammar can be conducted in accordance with progressive steps, and it remains highly applicable for the instructional setting being sequenced from basic to advanced.

In order to progress the application of pedagogical linguistics, Lu (2000) offered three perspectives relating to the content of pedagogical grammar. The first perspective centers on the essence of Chinese linguistics. Specifically, it seeks to address the question, "What grammar is the most needed and necessary for students?" The second perspective elaborates on the difference between learners' native language and Chinese. Namely, it seeks to address the following questions, "What do the two languages have in common? And what is the difference? What kind of difference would influence the acquisition of Chinese?" The third perspective discusses the role of grammatical errors in language acquisition. It attempts to answer the question,

"What are the most common mistakes students make when learning Chinese?" Lu (2000) also insisted on the implementation of unplanned learning at the preliminary stage of grammar teaching and the necessity of summative "basic grammar consolidation" after learners have reached a high level. With respect to this teaching method, two suggestions are proposed by Lu (2000). Firstly, choosing and arranging teaching materials should not solely depend on the content. Instead, the text should incorporate characters, vocabularies, and grammar that need to be acquired by learners. Nonetheless, the arranging of grammar in a text should be highly regulated. Secondly, a summative "basic grammar consolidation" is necessary once students reach a certain level. All of these suggestions have been proposed with the goal of improving learners' acquisition of Chinese.

2.3.2 The Application of Chinese Pedagogical Grammar to Writing

Several studies have discussed the topic of pedagogical grammar in CSL. Hong et al. (2018) presented a student-centered learning sequence in the cluster of grammatical structures. Additionally, Hsieh (2009), Chen and Lin (2003), and Peng (2003) suggested that communication and writing competence can be cultivated by enhancing a learner's knowledge of grammar. Considering that the incorporation of pedagogical grammar in writing skills and written text is developed from a learner's awareness and metacognition, it is well-accepted that pedagogical grammar plays a crucial role in a learner's use of target language and holds a central place in the study of CSL writing.

The current technology of automatic grading systems of Chinese writing can detect 65% of grammatical errors (Chang et al., 2015) and reach 88% accuracy on the automatic revising system (Hong et al., 2014a, 2014b); however, the accuracy of the automatic grading system of writing remains relatively stagnant. The main reason is the detection of grammatical errors (Chang et al., 2015). Specifically, because the system lacks the grammar that CSL learners need, the precision of identifying errors is unable to make much progress. The appropriateness or difficulties of grammar is closely correlated with the learner's level. Thus, in order to contrast the common grammatical errors made by learners, the present study seeks to categorize and construct the structures of learners' grammatical errors based on different types of errors from the data and expects to further the application in the teaching of writing, as well as the evaluation of learners' writing competence.

2.4 Corpus-Based Studies

Although many language teachers tend to incorporate corpus into the study of language teaching, most of the existing research focuses on analyzing a single grammar rule; only a few among them are integrated studies. These studies can be divided into two kinds. Some studies summarize the frequency of grammar and

offer teaching advice through the utilization of corpus data collected from native speakers. The other studies categorize learners' error types and sequence the difficulty of grammar through the learner corpus, as well as provide advice on pedagogy.

2.4.1 The Application of Corpus in CSL

Chang (2005) implemented Sinica Treebank Version 1.1 (http://treebank.sinica.edu.tw/) to sort out linguistic forms that contain the function of comparison and discovered that "presentative comparison sentences" are the most common form. The study retrieved the frequency, collocation, and mutual information of "*bi*" in Sinica Corpus and lists several frequently used "*bi*" sentences, as well as provides teaching steps for "*bi*" with reference to theories of pedagogical grammar. Chang (2014) observed how learners at different levels and with different native languages (English, Japanese) acquire Chinese relative clauses through data in the learner corpus and offered advice regarding instructional design.

Lin et al. (2014) extracted data that contained the Chinese directional complement "*qilai*" from Chinese Learner Corpus by National Taiwan Normal University (NTNU) and analyzed learners' error distribution to discover possible difficulties and offer advice for the instructional setting.

In order to identify discrepancies in language use, as well as to extract usages that are either completely identical or completely different, Hong and Huang (2013) used WordNet, Chinese WordNet, and the Chinese Concept Dictionary. The study utilized Chinese Word Sketch Engine to examine data from the cross-strait area in Chinese Gigaword Corpus and analyzed the distribution in the corpus. The findings revealed an interesting phenomenon; distinction and mutual influence are restored in the usage of words in the cross-strait areas.

2.4.2 The Application of Corpus on Error Analysis

Wang et al. (2013) put forth that near-synonyms often cause difficulty in teaching, and thus, should be closely examined. Furthermore, they stated that with extensive data from learners' interlanguage, vocabulary errors would be tractable and analyzable. The study opted for the "Chinese Learner Corpus" by NTNU to differentiate the use and error distribution of two groups of near-synonyms, "*bang*," "*bangzhu*," "*bangmang*," and "*bian*," "*biande*," "*biancheng*." The study produced insights on instructional steps in the teaching of near-synonyms by examining the connection between textbooks and learners' errors.

Further research on the acquisition of transition words has been conducted by Tseng and Hsieh (2013). Specifically, they utilized Sinica Corpus and TOCFL Learner Corpus (http://tocfl.itc.ntnu.edu.tw/) to compare the acquisition of the transition word "*er*" by Chinese learners and native speakers of Chinese. The findings showed that, with higher language levels, the conjunctions that learners deploy in discourse appear to transfer from intra-sentence to inter-sentence. Additionally, the

cause of errors is derived from the learner's unawareness of grammatical and semantic restrictions governing different conjunctions. This study demonstrates the usefulness of studying specific aspects of grammar, as it provides tangible and actionable data that can impact student learning.

In light of the lack of research that concentrates on computer-based correction of Chinese word order, Cheng (2014) used "HSK Dynamic Composition Corpus" by Beijing Language and Culture University to collect sentences with errors by foreign learners. Then, a revised corpus was established based on the misordering marking in sentences; misordering was marked by two researchers who speak Chinese as their native language. The study extracted feature engineering from Google Chinese Web 5-g Corpus after retrieving the data set from HSK Dynamic Composition Corpus. The study then generated a series of available combinations that could contain correct sentences by using CRF to detect the possible sections of misordering in sentences. These combinations were then sequenced according to the possibility of correct word order. The research found 83.4% accuracy for identifying sectional misordering and 85.8% accuracy for correcting misordering. The findings of the study are applicable to future research, and the accuracy can be improved by expanding the database.

Further research utilizing the Chinese Learner Corpus was conducted by Tung et al. (2015); they analyzed data from A2 learners and B1 (referring to CEFR proficiency levels) learners, whose native language is English, and calculated the error distribution of "*le*" sentence. The findings of this research provided advice on teaching steps, as well as information that could be used for further examination.

Derived from the aforementioned studies related to corpus linguistics, the corpus provides us with valuable information on the attributes of vocabularies and grammar. The frequency of certain types of sentences and the wording difference in cross-strait areas can all be observed from the corpus data. In addition, through the data from "learners' interlanguage corpus," existing errors have become analyzable and serve as a reference to help understand the possible difficulties that CSL learners may encounter. The findings can also be utilized in future studies and offer implications for practical use.

3 Methodology

3.1 The Learner Corpus

"Chinese Written Corpus" (CWC) (http://140.122.63.128/Index.aspx) is a CSL/CFL written corpus that discovers error patterns from the same written text by learners at different levels. The collected data are then used to construct the self-evaluation system and the feedback system, as well as for the exploration of how a self-evaluation system can be applied to the study of CSL/CFL-based writing (Hong et al., 2014a, 2014b).

The corpus provides information on grade band and error marking from the post-evaluated text and also provides the error sentence and the revised sentence that are applicable in research and teaching, as shown in Figs. 1 and 2.

The grading system used in CWC is in accordance with the proficiency guidelines for writing (ACTFL, 1987, 2012) by the American Council on the Teaching of Foreign languages (ACTFL) and developed from the framework "Rating Scale of Testing Chinese Writing" by Sung et al. (2012). The assessment is composed of four elements: content, grammar, vocabulary, and punctuation. All of the texts are then classified into five levels: excellence, good, advanced, intermediate, and beginner. A total of 11 bands are employed within the level of beginner, intermediate, and advanced as low (band 1–3), medium (band 4–6), and high (band 7–11). The text is then given a score based on the performance of the four elements during the human assessment. When assessing each text, consistency and accuracy are assured by the

Fig. 1 The home page of CWC

Fig. 2 The search result of CWC

Table 1 The distribution of band score from the four texts

Text ACTFL band score	Text 1	Text 2	Text 3	Text 4	Total
Band 3	16	61	106	102	285
Band 4	195	184	267	233	879
Band 5	349	238	159	127	873
Band 6	139	140	94	77	450
Band 7	63	66	33	31	193
Band 8	9	20	6	10	45
Band 9	4	4	1	2	11
Total	775	713	666	582	2736

program monitoring grading criteria, sample texts, the trial assessment by the grader, alignment of the trial assessment, alignment of the assessment, and alignment after the assessment. The goal of this design is to produce meaningful and accurate results.

Most of the data in CWC are collected from Chinese learners of different native languages in the Mandarin Training Center (MTC) at NTNU and 11 other CSL/CFL institutes from September 2010 to December 2016. The existing data in the corpus have been documented with detailed information, such as the title of the text, the learners' name in Chinese and English, nationality, the learners' native language, institute, and so forth; the data has also been restored in the form of a text file or an image file. There are four texts that have been marked and graded and that are utilized in this analysis: "a place worth going," "the beach in summer," "a letter to my family," and "introducing my country." Samples that were completely off-topic or unanswered were deleted during the compilation of the database. The total number of texts is 2,736, the individual number for text 1, text 2, text 3, and text 4 is 775, 713, 666, and 582, respectively, as shown in Table 1.

The present study utilizes four texts, "a place worth going," "the beach in summer," "a letter to my family," and "introducing my country," in Chinese Written Corpus (CWC), with a distribution of grade from band 3 to band 9. The texts are composed of foreign language learners who speak 43 different native languages. Among the data collected from the learners, the number of text are arranged in descending order according to native language; the top five groups are listed as follows: Japanese, English, Vietnamese, Korean, and Indonesian. In light of the diverse background of learners and the disparity of data, the present study only analyzes and discusses the five groups of learners with the highest number of texts (see Table 2).

The error marking system in CWC is supported by WeCan (Chang et al., 2012a, 2012b; Chang et al., 2012a, 2012b) and is able to provide functions such as word segmentation, tagging parts of speech, error marking, and so forth. The system can then export files to be used with programs to support related studies and future development. As for the tagging of parts of speech, the study selects a total of 48 simplified markers that represent 46 simplified markers classified by the Chinese

Table 2 The number of texts from the five groups of learners classified by their native languages

Native language	Text 1	Text 2	Text 3	Text 4	total
Japanese	209	161	160	174	704
English	131	120	154	83	425
Vietnamese	117	113	81	61	378
Korean	99	112	55	60	360
Indonesian	38	54	53	60	175

Knowledge and Information Processing group (CKIP), as well as the items Nominalized Verb (Nv) and Unknown (b) that are manually added by this study. Regarding error marking, the study divides learners' errors into two parts: surface structure and linguistic form. Surface structure refers to "addition," "omission," "misformation," and "misordering," and linguistic form refers to "character," "word," and "punctuation" (see Fig. 3).

The following are the error sentences found in written texts, which are classified into four types of surface structures:

(1) Addition (a place worth going/ACTFL band 7)

*我 已經 離開 家 也 快 十年 了。

* I already left home already almost ten years AM

我 離開 家 也 快 十年 了。

I left home already almost ten years AM

(2) Omission (the beach in summer/ACTFL band 6)

*沙灘 上 有 好多 的 人 曬太陽。

*beach P have many de people bask (in) sun

沙灘 上 有 好多 的 人 在 曬太陽。

beach P have many de people AM bask (in) sun

(3) Misformation (the beach in summer/ACTFL band 5)

*而且 福隆 海邊 是 海水 跟 河水 見面 的 河口。

*And fulong beach SHI sea with river meet de estuary

而且 福隆 海邊 是 海水 跟 河水 相會 的 河口。

And fulong beach SHI sea with river join de estuary

*476 名 的 乘客 中 只 146 名 救助 了。

*476 C de passenger P only 146 C help AM

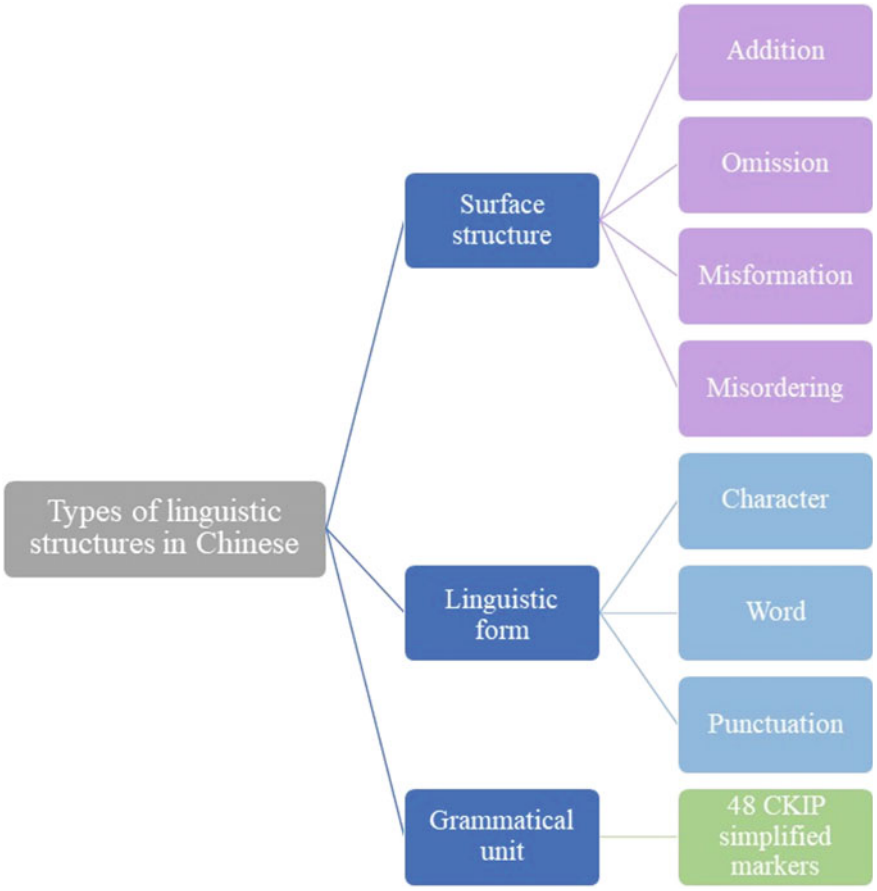

Fig. 3 The types of linguistic structures in Chinese

476 名 的 乘客 中 只 146 名 獲救 了。

476 C de passenger P only 146 C rescue AM

(4) Misordering (a place worth going/ACTFL band 7)

*讓 你 回來 以後 再 想 去 一次。

*Let you come back after again want go once

讓 你 回來 以後 想 再 去 一次。

Let you come back after want again go once

The research steps for this study are divided into two parts: fundamental studies and applied studies. These two categories are then divided into four additional subsections. Fundamental studies are divided into information on learners' errors

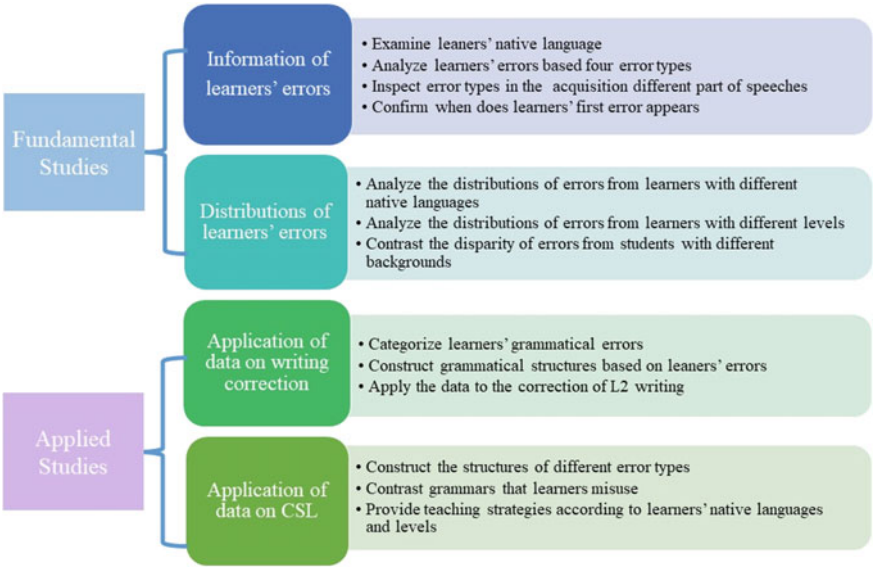

Fig. 4 Research steps of the present study

and distribution of learners' errors. Applied studies are divided into application of data in writing correction and application of data on CSL. The research framework is illustrated in Fig. 4.

3.2 The Reference Corpora

3.2.1 Sinica Corpus

"Academia Sinica Balanced Corpus of Modern Chinese version 4.0" (Chen et al., 1996, http://asbc.iis.sinica.edu.tw/), abbreviated as Sinica Corpus, contains more than ten million word tokens collected from 1981 to 2007. The database is mainly comprised of written language, and each word is segmented and tagged with part of speech. The data are retrieved from texts related to literature, social science, science, philosophy, arts, and so forth, and represent different linguistic modes (written text, manuscript), different writing styles (narrative, essay), different media (newspaper, textbook, audiovisual media), and different themes (science, literature). The corpus has collected 19,427 texts, and has 1,396,133 sentences, 11,245,330 word tokens, 239,598 word types, and 17,554,089 character tokens.

In order to examine the use of written language by native speakers with systematic tagging of parts of speech and to ensure the exclusive use of traditional Chinese in order to maintain the rigor of research, the present study retrieves data from native

speakers from Sinica Corpus. Since CWC and Sinica Corpus have the same tagging system for parts of speech, the present study can conduct a contrastive analysis through the comparison of the written text in CWC and data from native speakers in Sinica Corpus.

3.2.2 The Digital Platform of Chinese Grammar (DPCG)

"The Digital Platform of Chinese Grammar version 4.3.3." (DPCG) (http://203.64.95.103:8089/SyntaxSystem/) seeks to integrate "teaching" and "learning" in theory and practice. For teachers, it provides insight into possible obstacles that learners may encounter. For learners, the platform offers information on learning steps based on the frequency of different elements of grammar. For the development of textbooks, the platform merges teaching steps and error frequency to facilitate the compiling of teaching materials for CSL. Future research can conduct experiments pertaining to the teaching of written language and incorporate CWC as a resource and target in the study of CSL (see Fig. 5).

The DPCG brings together perspectives from native speakers, L2 learners, and textbook development by combining Chinese Gigaword Corpus (LDC, 2009) and CWC for the frequency of grammar that native speakers deploy on a daily basis and data from Chinese learners to accurately analyze the use of grammar and error frequency by learners at different levels. Through cross-checking the results and the illustration of the frequency quadrants, the platform presents a thorough analysis of the arrangement of grammar in the four textbooks that are commonly used in CSL learning: "A Course in Contemporary Chinese" (2015), "Road to Success: Threshold" (2008), "Practical Audio-Visual Chinese" (2007), and "New Practical Chinese Reader Textbook" (2002). The results that are presented in the platform offer evidence-based advice on the teaching of frequently-used grammar, as well as

Fig. 5 The home page of DPCG

Fig. 6 The frequency quadrants and sample sentences in DPCG

sentences from native speakers and error sentences from learners. Furthermore, the results are used to study the development of frequency quadrants of CSL learners (see Fig. 6).

A comparison of the data in Chinese Gigaword Corpus and CWC has led the present study to classify four quadrants that correspond to a learner's learning progress using frequency in Chinese Gigaword Corpus as the X-axis and error frequency in CWC as the Y-axis: "commonly used, high error frequency," "commonly used, low error frequency," "seldom used, high error frequency," and "seldom used, low error frequency." The four quadrants are designed to determine the appropriate steps that should be taken when teaching grammar. For example, if a grammatical construction appears in the quadrant of "commonly used, high error frequency" after comparing frequency in the two corpora, it should be taught prior to other constructions and vice versa. Likewise, teachers can understand the use of each construction by native speakers and learners and decide if certain constructions should be emphasized or underemphasized in teaching. The platform also provides error sentences by learners for instructional purposes. Overall, the four quadrants are designed to provide actionable information to teachers and learners.

4 Result and Discussion

4.1 Overall Distribution of Error Types in the Learner Corpus

The number of error sentences in the text is roughly 100,000. Among all four types of errors, misformation accounts for about 50% of the errors, which is significantly higher than other error types.

The reason for the disproportionate percentage of misformation is due to the vagueness of near-synonyms and the difficulties that arise in teaching (Hong and Sung, 2017). The semantic vagueness not only causes miscomprehension and confusion, but also leads to misuse in practice. Furthermore, misformation is prevalent among all texts by learners from different levels, which indicates that the problem of misformation is not alleviated by a learner's advancement in language competence (Cai, 2010). Hence, miscomprehension of near-synonyms ultimately gives misformation a rather salient portion of the four error types.

The possible applications of the data collected from CWC include analyzing learners' error types in written text based on the surface structure of language and examining the distribution of errors according to grammatical features, namely, parts of speech. The parts of speech of data in the present study are tagged in accordance with the 48 CKIP simplified markers in Sinica Corpus. The major categories are noun (N), verb (V), adjective (A), conjunction (C), adverb (D), interjection (I), postposition (P), particle (T), "*de*, *zhi*, *de*, *de*" (DE), "*shi*" (SHI), and foreign word (FW). Generally speaking, colloquial context and written language are primarily composed of units such as noun, verb, adjective, conjunction, adverb, and so forth. Particularly, in light of the uniqueness of its grammatical structure, *shi* not only holds a special place in the study of Chinese linguistics, but is also categorized as a transitive verb in the tagging by Sinica Corpus. Furthermore, based on observations from learners' writing proficiency, *shi* remains one of the most frequently-used linguistic errors at all levels (Hong and Sung, 2017). The words found in these six main categories tend to be the most commonly used on a daily basis. Thus, the present study aims to inspect the number of error sentences based on the parts of speech by conducting a cross-checking analysis. From the statistic results shown in Table 3, it can be seen that with addition and omission, most errors occur in the learning of adverbs, and the number of errors in the noun category is the second. As for misformation and misordering, the number of errors in the noun category dominates in both types. The second highest in terms of the number of errors in misformation and misordering are verb and adjective, respectively.

Table 3 The statistics of the error types in CWC

Types of structural errors	Number of error	Percentage of error (%)
Addition	22,496	21.61
Omission	28,874	27.74
Misformation	**48,355**	**46.46**
Misordering	4352	4.18
Total	104,077	100.00

Table 4 The statistics of error types based on parts of speech

Part of speech / Error types	Noun	Verb	Adjective	Conjunction	Adverb	*Shi*	Total
Addition	4698	2946	653	955	6547	1094	16,893
Omission	4061	2331	386	1047	7690	932	16,447
Misformation	9383	7019	2888	1791	5100	319	26,500
Misordering	563	362	443	80	159	25	1632
Total	18,705	12,658	4370	3873	19,496	2370	61,472

4.2 Distribution of Error Types Among Different Learner Variables

Many studies (Chen, 2011; Hung, 2013; Limuria, 2014; Okuno, 2018; Huang, 2018; Tang, 2018) have revealed that learners' errors tend to appear in different aspects. The present study aims to analyze the distribution of learners' errors in terms of learners' native language, level, and the use of parts of speech.

4.2.1 Native Language as the Variable

Despite classifying learners into different groups based on their native languages, according to the statistics result, the top five groups of learners (Japanese, English, Vietnamese, Korean, and Indonesian) have the same distribution and tendency for errors. As shown in Table 4, the most common type of error is misformation, followed by misordering. This suggests that, in spite of the diverse background of native languages, learners' errors in surface structure appear to be highly consistent. In addition to the impact of individual native language, the study also accounts for the reason and distribution of errors to form an integrated perspective.

4.2.2 Proficiency Level as the Variable

As with the distribution of errors by learners speaking different native languages, misformation dominates in the number of errors and remains as the main error type in all of the incorrect sentences with proficiency level as the variable. On the contrary, the number of misordering is remarkably lower than the other three error types. Addition and omission present less discrepancy in the total number of incorrect sentences. From the data in Tables 5 and 6, a universal trend can be seen in that the distribution of the four error types remains the same, regardless of a learner's native language or proficiency level.

Table 5 The statistics of errors based on learners' native languages

Native language / Error types based on surface structure	Japanese	English	Vietnamese	Korean	Indonesian
Addition	5486	2749	4044	3931	1629
Omission	6845	3438	6459	3984	1762
Misformation	11,575	6251	8165	7606	3391
Misordering	1221	494	742	686	272

Table 6 The number of sentences with different error types in different bands[2]

ACTFL band score / Error types based on surface structure	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9	Total
Addition	2204	7055	7070	3990	1725	385	58	22,487
Omission	3546	9446	9097	4511	1831	375	61	28,867
Misformation	4576	14,960	15,575	8817	3476	768	164	48,336
Misordering	563	1375	1363	723	250	64	14	4352
Total	10,889	32,836	33,105	18,041	7282	1592	297	104,042

4.2.3 Part of Speech as the Variable

Apart from a learner's native language and proficiency level, parts of speech as the variable have the potential to provide valuable information on the overall distribution of error types to provide a holistic view of a learner's performance. Based on the data retrieved from CWC, this study will discuss how the six parts of speech, noun, verb, adjective, conjunction, adverb, and *shi*, present in the four types of errors in surface structure in the following section.

In the distribution of the first error type, addition/adverb appears to be the part of speech that is easily misused in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than in other parts of speech. Regarding other parts of speech, texts with the highest mean of sentences with the addition of noun, adjective, and conjunction are found in band 6. Also, the addition of verb and *shi* in sentences are particularly noticeable in band 7. However, the most dominant mean of sentences with the addition of adverb exists in band 8, rather than at the intermediate level. The distribution of data reveals that learners at the intermediate level tend to insert redundant units into sentences.

[2] The statistics in Table 3 are retrieved from CWC directly and constitute incorrect sentences from band 1 to band 9, and thus different from the statistics shown in Table 6, which includes data from band 3 to band 9 only. Due to the exclusion of band 1 and band 2, the number of incorrect sentences differs slightly in addition, omission, and misformation. However, the number remains identical in misordering because students in band 1 and band 2 are not exposed to long sentential structure, but instead short phrases of survival language. Hence, the error type of misordering does not exist in band 1 and band 2.

In the distribution of the second error type, omission/adverb appears to be the part of speech that learners most commonly misuse in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than in other parts of speech, which aligns with the tendency in the first error type, addition. In regards to other parts of speech, texts with the highest mean of sentences with the omission of nouns are found in band 7. The omission of verbs is particularly excessive in band 5, and the omission of adjectives is prominent in band 8. As for conjunctions, band 6 and band 7 both have the highest number of sentences with incorrect omissions. The omission of adverbs, on the other hand, is discovered to be most salient in band 6. Lastly, the omission of *shi* is particularly noticeable in band 7 and band 8. The distribution of data indicates that the error of omission is more obvious among learners at the intermediate and advanced levels.

In the distribution of the third error type, misformation/adverb appears to be the part of speech that learners most commonly misuse in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than other parts of speech, which aligns with the tendency of the aforementioned error types. As for other parts of speech, texts with the highest mean of sentences with the omission of nouns are also found in band 7. The misformation of verbs is detected to be excessive in band 5, and the misformation of adjectives is relatively noticeable in both band 6 and band 7. The texts with the highest mean of sentences with the misformation of adverbs are found in band 6. Finally, the misformation of *shi* is particularly dominant in band 7 and band 8. The distribution of data indicates that the error of misformation, similar to the error of omission, should receive extra attention among learners at the intermediate and advanced levels.

When examining the error of misordering, this study discovers that it appears to be the most divergent in terms of distribution among the four error types. The misordering of nouns is found to be most salient among learners from band 4 to band 6. Nevertheless, for beginner and advanced learners, the misordering of adjectives dominate in number. With respect to detailed information, the highest mean of sentences with misordering of nouns is found in the text of band 5. For the misordering of verbs and conjunctions, the highest means of sentences in the texts both appear in band 9. The misordering of adverbs, however, is relatively noticeable in band 6 and band 7. Lastly, the misordering of *shi* is especially pronounced in band 6. In conclusion, the error of misordering appears to be particularly significant among advanced learners.

The overall pattern of error distribution based on each part of speech is depicted in Table 7, which shows the mean of sentences in a text with incorrect parts of speech in different band scores and error types.

5 Conclusion

In general, sentences in written text, compared to colloquial data, appear to be more complex in terms of linguistic form and are expected to adhere to the framework

Table 7 The mean of sentences in a text with incorrect parts of speech in different band scores and error types

		Band 3	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9
Addition	Noun	1.74	1.66	1.68	1.89	1.77	1.67	1.00
	Verb	1.02	1.11	1.07	1.06	1.12	0.84	0.64
	Adjective	0.25	0.24	0.23	0.26	0.24	0.07	0.09
	Conjunction	0.37	0.31	0.34	0.42	0.38	0.36	0.18
	Adverb	2.17	2.33	2.31	2.62	2.75	3.00	2.00
	Shi	0.35	0.43	0.37	0.42	0.46	0.40	0.09
Omission	Noun	1.55	1.45	1.52	1.44	1.59	1.31	0.82
	Verb	0.81	0.79	0.94	0.90	0.76	0.69	0.55
	Adjective	0.16	0.13	0.15	0.14	0.12	0.22	0.00
	Conjunction	0.34	0.34	0.41	0.44	0.44	0.24	0.18
	Adverb	2.76	2.75	2.88	2.90	2.81	2.38	1.73
	Shi	0.32	0.29	0.36	0.37	0.42	0.42	0.36
Misformation	Noun	3.47	3.14	3.42	3.82	3.82	3.80	4.82
	Verb	2.27	2.38	2.55	3.03	3.03	2.89	3.36
	Adjective	1.00	0.95	1.02	1.28	1.27	0.91	1.64
	Conjunction	0.61	0.62	0.71	0.59	0.75	0.78	0.55
	Adverb	1.76	1.83	1.84	2.04	1.94	1.73	1.00
	Shi	0.11	0.12	0.12	0.11	0.10	0.02	0.18
Misordering	Noun	0.20	0.19	0.22	0.23	0.16	0.18	0.18
	Verb	0.09	0.13	0.16	0.14	0.10	0.11	0.18
	Adjective	0.11	0.16	0.18	0.16	0.18	0.11	0.27
	Conjunction	0.02	0.03	0.04	0.03	0.01	0.02	0.09
	Adverb	0.04	0.06	0.05	0.07	0.07	0.02	0.00
	Shi	0.00	0.01	0.01	0.02	0.01	0.00	0.00

of prescriptive grammar. Ideally, in a practical context, teachers would only teach grammar that is confined to certain norms, and students would, therefore, be exclusively exposed to prescriptive usages. However, in the texts used in this study, various errors are spotted in vocabulary and grammar. Thus, the present study seeks to assist teachers in discovering students' potential grammatical errors by identifying the types and patterns of errors with the support of data from CWC. Apart from examining the existing errors, this study also attempts to improve the effectiveness of error identification. The previous research has yielded little progress in identifying errors by comparing students' written text with reference to correct grammar. Hence, this study contrasts students' written texts with the structures of grammatical errors categorized in the research and further discovers the distribution of learners' errors on parts of speech in hopes of advancing the effectiveness and efficiency of the error

identification system. The findings of the present study reveal two universal distributions in learners' error types. Firstly, among all four error types, misformation appears to be the most common error, while misordering is the rarest, regardless of a learner's background. Secondly, based on the observed association between error types and parts of speech, it appears that learners often have difficulty with adding and omitting adverbs in a sentence, and therefore, have a tendency to misform nouns and verbs.

Furthermore, since a learner's native language and level often play a crucial role in organizing teaching activities, one element of CWC is its error marking system and graded texts. Through the application of the error marking system and graded texts, future studies can conduct cross-checking based on the existing data and design teaching strategies for learners speaking different native languages or at different levels. Through the error analysis of learners' texts, as well as contrasting the distribution and frequency of various grammar errors in CWC, the present study constructs different error types and identifies shared error types among learners at different levels. The findings of the study offer insights into the implementation of teaching strategies as well as methodologies at different levels.

References

ACTFL Chinese Proficiency Guidelines. (1987). *Foreign Language Annals, 20(5),* 471–481.

ACTFL Proficiency Guidelines 2012–Writing. (2012). Retrieved from http://actflproficiencyguidelines2012.org/writing.

Ashwell, T. (2000). Patterns of teacher response to student writing in a multi-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing., 9*(3), 227–257.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Buckingham, T., & Pech, W. C. (1976). An experience approach to teaching composition. *TESOL Quarterly, 10*(1), 55–65.

Cai, B. G. (2010). An investigation and analysis of errors in using synonymous action verbs in interlanguage. *Chinese Teaching in the World, 4*, 526–535.

Cai, Q. Y. (2014). The analysis of chinese character writing and words usage errors made by Chinese-Japanese learners and suggestions. *Chung Yuan Journal of Teaching Chinese as a Second Language, 17*, 53–77.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296.

Chang, L. P. (2005, July). Qiantan jiaoxueyufa–cong nijiaoju tanqi. Paper presented at the 8th Forum of Chinese Teaching. China, Beijing. [張莉萍. (2005, 7月). 浅談教學語法–從比較句談起. 第八屆漢語教學討論會. 中國北京.].

Chang, L. P. (2014). Salient linguistic features of chinese learners with different L1s: A corpus-based study. *International Journal of Computational Linguistics and Chinese Language Processing, 19*(2), 53–72.

Chang, T. H., Sung, Y. T., & Hong, J. F. (2015). Automatically detecting syntactic errors in sentences writing by learners of chinese as a foreign language. *International Journal of Computational Linguistics & Chinese Language Processing, 20*(1), 49–64.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012a, November). *A Chinese word segmentation and POS tagging system for readability research.* Paper presented at the 42nd Annual Meeting of the Society for Computers in Psychology (SCiP 2012), Minneapolis, MN.

Chang, T. H., Sung, Y. T., Lee, Y. T., & Hsieh, G. S. (2012b, October). Zhongwen duanci zhi keduxing yingyongyanjiu. Paper presented at The 51st Annual Conference of Taiwan Psychology Association. Taichung: Asia University. [張道行, 宋曜廷, 李垚曒, & 謝冠生 (2012, 10月). 中文斷詞之可讀性研究應用. 發表於台灣心理學會第五十一屆年會. 台中, 亞洲大學.].

Chen, C. L. (2011). An error analysis and pedagogical study of DE in Mandarin Chinese for thai learners. (Master's thesis). Department of Chinses as a Second Language at National Taiwan Normal University, Taipei.

Chen, H. H., & Lin, C. H. (2003, October). Yunyong wangluhudong jinxing huayuwen xiezuo xuexi zhi tantao. Paper presented at International Conference on Internet Chinese Education (ICICE). Taipei: Overseas Community Affairs Council. [陳懷萱, & 林金錫 (2003, 10月). 運用網路互動進行華語文寫作學習之探討. 第三屆全球華人網路教育研討會. 臺北: 僑委會.].

Chen, K. J., Huang, C. R., Chang, L. P., & Hsu, H. L. (1996, December). Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation* (pp. 167–176).

Chen, P. (2016). A study of chinese word order errors and its pedagogy for multilingual learners-a case study in an international school in Bangkok, Thailand. *Journal of Language and Literature Studies, 29*, 191–232.

Cheng, S. W. (2014). Chinese word ordering errors detection and correction for non-native Chinese language learners. (Master's thesis). Department of Computer Science and Information Engineering at National Taiwan University, Taipei.

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1–61.

Chuyen, V. T. (2015). The alternative question in Chinese Vietnamese language: Using Contrastive Analysis as a Reference Design in Grammar Teaching. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Corder, S. P. (1967). The significance of learner's errors. *IRAL-International Review of Applied Linguistics in Language Teaching, 5*(4), 161–170.

Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). *Language two.* New York: Oxford University Press.

Fathman, A. K., & Whalley, E. (1990). Teacher Response to Student Writing: Focus on Form versus Content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–190). Cambridge: Cambridge University Press.

Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161–184.

Hong, J. F., Chiu, S. W., Sung, Y. T., & Chang, T. H. (2018). Applying Chinese hierarchical grammar bank to the evaluation of chinese writing instruction. *Journal of Technology and Chinese Language Teaching, 9*(2), 40–60.

Hong, J. F., & Huang, C. R. (2013). Cross-strait lexical differences: A comparative study based on chinese gigaword corpus. *Computational Linguistics and Chinese Language Processing, 18*(2), 19–34.

Hong, J. F., & Sung, Y. T. (2017). Huayu xiezuo yuliaoku jianzhi yu fenxi. In H.-J. H. Chen (Ed.), *Corpus and teaching Chinese as a second language* (pp. 197–229). Taipei: Taiwan Higher Education Press Co. [洪嘉馣, & 宋曜廷. (2017). 華語文寫作語料庫建置與分析. In 陳浩然 (Ed.) 語料庫與華語教學 (pp. 197–229). 臺北: 高等教育出版社.].

Hong, J. F., Chang, J. Y., Chang, T. H., & Sung, Y. T. (2014a, October). Huayu wei waiyu zhi xiezuo zidongpinggu yu jiaoxue. Paper Presented at the 1st CLTA-ISCLTL. U.S., Indiana. [洪嘉馣, 張人懿, 張道行, & 宋曜廷 (2014a, 10月). 華語為外語之寫作自動評估與教學. 全美中文教師學會第一屆中文教學國際研討會 (CLTA-ISCLTL). 美國印第安納.].

Hong, J. F., Sung, Y. T., & Chang, T. H. (2014b, November). Yi fenji gainian jianzhi huayuwen xiezuo yuliaoku. Paper presented at The 10th Taiwan E-Learning Forum, 2014 (TWELF). Taipei:

National Taiwan Normal University. [洪嘉馏, 宋曜廷, & 張道行 (2014b, 11月). 以分級概念建置華語文寫作語料庫. 第十屆臺灣數位學習發展研討會. 臺北: 臺灣師範大學.].

Hsieh, C. Y. (2009). Application of controlled writing to teaching in Chinese writing: Organization of writing. *Journal of Applied Chinese, 5*, 225–262.

Huang, H. C. (2014). An analysis of the strategies of teaching mandarin tones to Japanese learners of Chinese at the basic level. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Huang, T. G. (2018). Teaching "Yi+Classifier" to native speakers of English and Korean in intermediate Chinese classes: Error analysis and the designing of a pedagogical decision tree. *Taiwan Journal of Chinese as a Second Language, 17*, 153–183.

Hung, J. W. (2013). Interlanguage analysis and a study of the teaching strategies of Chinese potential complement for the intermediate Chinese learners. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

James, C. (1998). *Errors in language learning and use: Exploring error analysis.* London, UK: Addison Wesley Longman.

Liang, N. S.-Y. (2008). The acquisition of Chinese shape classifiers by L2 adult learners. In Marjorie K.M. Chan & H. Kang (Eds.), *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20), 1,* (pp. 309–326). Columbus, Ohio: The Ohio State University.

Limuria, R. (2014). An error analysis on chinese passive voice produced by indonesian-speaking learners and suggested teaching notes. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Lin, Y. T., Chen, H. J. H., & Wang, C. C. (2014). A Learner Corpus-based Study on Chinese Directional Complement "Qilai." *Journal of Chinese Language Teaching., 11*(4), 73–109.

Liu, S. C. (2016). An analysis of french learner's conjunction errors in writing with pedagogical implications—based on méthode de chinois I–II. (Master's thesis). NTU Graduate Program of Teaching Chinese as a Second Language, Taipei.

Liu, X. (Ed.). (2002). *New practical Chinese reader* (Vol. 1). Beijing: Beijing Language and Culture University Press.

Lu, J. M. (2000). Duiwai hanyu jiaozhong de yufa jiaoxue. *Language Teaching and Linguistic Studies*, *3*, 1–8. [陸儉明. (2000). 對外漢語教中的法學. 語言教學與研究. 3, 1–8.].

Lv, W. H. (2008). *Duiwai hanyu jiaoxue yufa tansuo.* Beijing: Beijing Language and Culture University Press. [呂文華. (2008). 對外漢語教學法探索. 北京: 北京 語言大學出版社.].

Nassaji, H., & Fotos, S. (2011). *Teaching grammar in second language classrooms: Integrating form-focused instruction in communicative context*. New York: Routeledge.

Odlin, T. (1994). *Perspectives on pedagogical grammar.* Cambridge University Press.

Okuno, A. (2018). Error analysis of "Bei" construction by Japanese learners. (Master's thesis). NTU Graduate Program of Teaching Chinese as a Second Language, Taipei.

Peng, N. S. (2003). A study of chinese texts, reading and writing from the systemic-functional linguistics perspective. *Journal of University of Taipei, 44*(2), 33–62.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 209–231.

Tang, N. P. (2018). An error analysis of the interlingual Chinese users from Japan, English-speaking Regions, and Chinese Speech Communities including a Contrastive Analysis of Chinese, English, and Japanese Punctuation (Master's thesis). Department of Chinese as a Second Language at National Taiwan Normal University, Taipei.

Tseng, Y. C., & Hsieh, M. L. (2013). Second language acquisition of Chinese conjunction "*er*" (and): A corpus-based study. *Taiwan Journal of Linguistics, 11*(1), 125–172.

Tung, T. Y., Chen, H. J. H., & Yang, H. M. (2015). The error analysis of "Le" based on "Chinese Learner Written Corpus." *Computational Linguistics and Chinese Language Processing, 17*, 76–95.

Wang, Y. C. (2011). Interlanguage analysis of directional complements for german learners of Chinese. (Master's thesis). Department of Chinese as a Second Language at National Taiwan Normal University, Taipei.

Wang, Y. T., Chen, H. J. H., & Pan, Y. T. (2013). Investigation and analysis of chinese synonymous verbs based on the Chinese learner corpus: Example of “bang”, “bang-zhu”, “bang-mang” and “bian”, “bian-de”, “bian-cheng.” *Journal of Chinese Language Teaching, 10*(3), 41–64.

Yang, J. Z. (2000). Duiwai hanyu jiaoxue chuji jieduan yufa xiangmu de paixu wenti. *Language Teaching and Linguistic Studies*, *3*, 9–14. [楊寄洲. (2000). 對外漢語教學初級階段語法項目的排序問題. 語言教學與研究, 3, 9–14.].

Zhang, H. (Ed.). (2008). *Road to success: Threshold*. Beijing: Beijing Language and Culture University Press.

Zhou, X. B. (2002). Chinese as a foreign language: Characteristics of instructional grammar. *Journal of Sun Yat-sen University (Social Science Edition), 42*(6), 137–142. [周小兵. (2002). 漢語第二語言教學語法的特點. 中山大學學報(社會科學版), 42(6), 137–142.].

Zhou, X. B., Zhu, Q. Z., & Zheng, X. Y. (2007). Waiguoren xue hanyu yufa piawu yanjiu. Beijing: Beijing Language and Culture University Press. [周小兵, 朱其智, & 鄭小宇. (2007). 外國人學漢語語法偏誤研究. 北京: 北京語言大學出版社.].