# A Preliminary Study on Chinese Learners' Written Errors Based on an Error-Tagged Learner Corpus


**Ting-Yu Yang, Hui-Mei Yang, Wei-Jei Lee, Chen-Yu Liu, and Howard Hao-Jan Chen**

**Abstract** With the development of technology, the need for compiling computer-based learner corpora has gradually gained more attention from language teachers and researchers. A learner corpus can reflect learners' authentic use of a target language, which provides useful information for language teachers, researchers, and textbook editors. Limitations of retrieving errors in learner corpora, however, still exist. For example, it is difficult to retrieve omission errors if a corpus is not error-tagged beforehand. To offer researchers an error-tagged learner corpus of Chinese, this study manually error-tagged the two-million-word Chinese Learner Written Corpus of National Taiwan Normal University. A preliminary analysis of errors tagged in the learner corpus shows a total of 48,266 errors distributed to 119 tags. These 48,266 errors are mostly distributed to the incorrect selection of words or the missing of necessary word-level components, and the misuse of nouns, action verbs, adverbs, and structural particles is especially common. Among the 119 tags, the top 12 common error tags (i.e., occurring more than 1,000 times) accounted for more than 50% of the total errors, and incorrect selections of nouns and action verbs together constituted more than 27% of the total errors. These 12 common error types, especially the wrong choice of nouns and action verbs, should thus be regarded to be particularly difficult for second language (L2) learners of Chinese to acquire. Analysis of the top 12 common errors also reveals that learners' misuse of verbs, adverbs, and structural particles were somewhat varied (i.e., involving different types of target modification, such as missing, redundant, and incorrect selection), whereas their misuse of nouns mostly resulted from an incorrect selection. A comparison between the top 10 common error types in this study with those in Lee et al. (2016) reveals that, regardless of some discrepancies in ranking, 90% of the top 10 error tags overlapped in the two studies, suggesting that these error types are indeed difficult for L2 Chinese

T.-Y. Yang (✉) · H.-M. Yang · W.-J. Lee · H. H.-J. Chen
Department of English, National Taiwan Normal University, 162, Section 1, Heping East Road, Taipei City 106, Taiwan
e-mail: christiney37@gmail.com

C.-Y. Liu
English Language Center, Ming Chuan University, 5, De Ming Road, Taoyuan City 333, 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_4

learners to acquire and should be investigated further. Based on the findings yielded in this study, suggestions for further research on L2 Chinese learners' errors are provided.

**Keywords** Chinese teaching · Learner corpus · Error-tagging · Error analysis

# 1 Introduction

## *1.1 The Development of Learner Corpus*

The concept of error analysis was firstly introduced by Corder (1967), who pointed out the significance of analyzing language learners' erroneous output to understand the linguistic features and developmental process of their interlanguage. Since then, analyses of language learners' errors have been one of the main research areas in the field of second/foreign language (L2) learning (Pan & Liu, 2006). Early studies on language learners' errors were mostly based on language teachers' reports on learners' erroneous sentences observed in their teaching, which often included a limited number of language learners' errors. The problem with small-sized samples stems from the fact that no statistical analysis can be performed to formulate rules of learners' interlanguage (Corder, 1967; Nemser, 1974; Selinker, 1972). Thus, the limited number of errors identified in early studies makes it difficult for researchers to systematically establish the causes of learners' errors and to obtain more generalizable results to point out their linguistic features.

The importance of collecting and analyzing a large quantity of learner errors to gain more generalizable results urges the establishment of a learner corpus. Learner corpora are electronic collections of authentic linguistic output by L2 learners. They consist of data larger than the types (e.g., output from elicitation tasks) commonly used in second language acquisition (Granger, 2003), and therefore afford researchers the confidence to report significant recurrent patterns or errors produced by L2 learners (McEnery et al., 2019). In addition, the electronic format of learner corpora allows researchers to extract target language structures from a large number of data for further analysis with a wide range of software tools, saving researchers more time and effort in the manipulation of the data (Granger, 2003).

With the wide application of learner corpora in research and the compilation of teaching/learning materials, more and more research institutes and publishers are involved in the building of learner corpora. The first learner corpus, Longman Learners' Corpus, was compiled by Longman Publishing Group in the late 1980s, which contains 10 million words of English learners' essays and exam scripts worldwide. In 1990, Sylviane Granger started building International Corpus of Learner English, and she continues to expand its size to more than 5.5 million words, which consists of learners' written data from 25 first language (L1) backgrounds. Since the 1990s, the number of learner corpora has been rapidly increasing. According to a survey by Centre for English Corpus Linguistics of Louvain-La-Neuve (2020),

there are more than 180 learner corpora around the world, consisting learners' written/spoken data from more than 20 target languages. Currently, more than half of the corpora target the output of English learners, and around 25 of them contain more than 1 million words.

The growing trend of teaching/learning Chinese as a Second/Foreign Language (CSL/CFL) also encourages the development of Chinese learner corpora. To the best of our knowledge, the biggest learner corpora of learners' Chinese is Jinan Chinese Learner Corpus, a 6-million-character corpus containing exam scripts and assignments by learners from over 50 different L1 backgrounds (Wang et al., 2015). The second largest corpus is the 4.24-million-character HSK Dynamic Composition Corpus, which covers more than 11,000 compositions by exam takers of Hanyu Shuiping Kaoshi (HSK). The third largest corpus is the Continuity Corpus of Chinese Interlanguage of Character-error System, a 2-million-character corpus consisting of learners' sentence-makings and essays (Zhang, 2013). While the three corpora deal with simplified Chinese, attempts have also been made to build learner corpora of traditional Chinese. For example, Chinese Learner Written Corpus of National Taiwan Normal University collects more than 2 million characters of writings in traditional Chinese by learners from more than 60 different L1 backgrounds. Another corpus dealing with traditional Chinese is the 1.5-million-character TOCFL Learner Corpus, which collects 4,567 exam scripts from the Test of Chinese as a Foreign Language (TOCFL).

With these resources, researchers have employed these learner corpora to investigate Chinese learners' interlanguage and yielded some insightful results. For example, Zhang (2010) examined Chinese learners' use of 把 *bǎ*-sentences from HSK Dynamic Composition Corpus and discovered that the learners' avoidance of 把 *bǎ*-sentences was not as obvious as indicated in previous studies. Also based on HSK Dynamic Composition Corpus, Wang (2010) investigated Russian CSL learners' erroneous use of the particle 了 *le* and reported that missing 了 *le* was the most frequent error in these learners' writing. Hu's (2012) investigation of CSL learners' use of the adverb 都 *dou* revealed that low-level learners tended to misuse 都 *dou* significantly more often than both intermediate-level and advanced-level learners. In addition to the use of HKS Dynamic Composition Corpus, studies based on Chinese Learner Written Corpus were also conducted to examine learners' interlanguage. Wang et al. (2013) investigated Chinese learners' uses of two sets of synonymous verbs: 幫 *bang*, 幫忙 *bang-man*, 幫助 *bang-zhu*, and 變 *bian*, 變得 *bian-de*, and 變成 *bian-cheng*, and findings of their study showed that learners often wrongly replace 幫忙/幫助 *bang-man*/*bang-zhu* with 幫 *bang* and 變得/變成 *bian-de*/*bian-cheng* with 變 *bian*. Lin et al. (2014) examined the use of directional complement 起來 *qilai* based on Chinese Learner Written Corpus, and they discovered that the learners had great difficulty in using the stative meaning of 起來 *qilai*, which was mostly attributable to misformation.

Construction of these existing Chinese learner corpora provides a considerable amount of learner output for researchers to explore CSL/CFL learners' interlanguage with quantitative statistics; however, some error types, such as omission errors, might

not be easily retrieved by the direct use of these corpora. To better resolve this problem, further processing of learner data with error-tagging is suggested.

## *1.2 The Development of Error-Tagged Learner Corpus*

Learner corpus researchers (e.g., Díaz-Negrillo & Domínguez, 2006; Jia, 2007; Tono, 2003) have been advocating the importance of annotating learners' grammatical errors to provide useful information for the development of L2 research and/or teaching (Brook & Hirst, 2012; Granger, 2015; Swanson & Charniak, 2013; Wang & Seneff, 2007). Error-tagged learner corpora, however, are relatively scant. With the help of computer programs, most of the current learner corpora are annotated with part-of-speech (POS) tags, which allow users to carry out meaningful searches of target linguistic features (e.g., nouns, verbs, and adjectives) rather than a single word form (McEnery et al., 2019). Nevertheless, annotation of learners' errors requires more time and effort since tagging learners' grammatical errors heavily relies on human judgment and can only be done manually (Lüdeling & Hirschmann, 2015). Thus, only few current learner corpora are error-tagged.

To the best of our knowledge, two of the largest error-tagged learner corpora are Cambridge Learner Corpus and Longman Learners' Corpus. Cambridge Learner Corpus, currently the largest error-tagged learner corpus, contains annotations of 30 million words, the error-tagging system of which was devised by Cambridge University Press. This error-tagged corpus has become one of the major resources for publishers to compile English teaching/learning materials and dictionaries (Nicholls, 2003). Longman Learners' Corpus, built by Longman Publishing Group, is composed of 10 million words with error-tagging and also serves as a useful reference for the publisher to compile dictionaries. The dictionary, Longman Dictionary of Common Errors, is in fact compiled based on the learner corpus. Other error-tagged learner corpora of English include the 1-million-word Chinese Learner English Corpus, the 2.5-million-word HKUST Corpus of Learner English, and the 700,000-word Japanese EFL Learner Corpus.

In addition to learner English, efforts have also been made to the annotation of learner Chinese. The HSK Dynamic Composition Corpus is currently considered the most comprehensive error-tagged learner corpus of simplified Chinese. In the corpus, errors are manually annotated and distributed into four major categories, namely character-level errors (11 cases), word-level errors (5 cases), sentence-level errors (28 cases), and discourse errors (1 case). Based on the error-tagged data, investigations on CSL/CFL learners' interlanguage have been conducted to reveal learners' overall error distribution (e.g., Hsu, 2011) or errors in specific linguistic forms (e.g., Han, 2016; Jin, 2011; Li, 2013; Zang, 2014). Other error-tagged learner corpora of simplified Chinese include the Jinan Chinese Learner Corpus and the Continuity Corpus of Chinese Interlanguage of Character-error System.

Regarding the construction of error-tagged learner corpus in traditional Chinese, the TOCFL Learner Corpus is one of the corpora that contains around 1 million

characters of manually annotated errors produced by learners of traditional Chinese, in which 2,837 out of 4,567 learner essays that are graded at least 3 are error-tagged. Errors in the corpus are also distributed into four major categories, which are somewhat different from the error category of the HSK Dynamic Composition Corpus. In the TOCFL Learner Corpus, a total of 36 error types are categorized into word-level errors (16 cases), grammatical function-level errors (11 cases), sentence pattern-level errors (7 cases), and mixture errors (2 cases). Based on the corpus, Lee et al. (2016) analyzed 33,835 grammatical errors in the 2,837 essays and reported the top 10 error types that account for 47% of the total errors as follows: incorrect selection of action verb (n = 3,809, 11.26%), incorrect selection of noun (n = 2,167, 6.40%), missing adverb (n = 1,755, 5.17%), missing aspectual particle (n = 1,602, 4.73%), missing auxiliary (n = 1,357, 4.01%), incorrect selection of adverb (n = 1,168, 3.45%), missing structural particle (n = 1,165, 3.44%), missing action verb (n = 1,040, 3.07%), redundant aspectual particle (n = 1,003, 2.96%), and incorrect selection of stative verb (n = 780, 2.31%). While an incorrect selection of action verb is the most common error type, half of the 33,835 errors are attributed to missing word-level linguistic components.

Although efforts have been made to construct error-tagged learner corpora of Chinese, most of the current corpora, however, are based on simplified Chinese. While Lee et al. (2016) have contributed to the building of error-tagged learner corpora of traditional Chinese, the size of which is comparatively smaller than the HSK Dynamic Composition Corpus and the Jinan Chinese Learner Corpus. To provide CSL/CFL researchers with more resources for the study of learners' inter-language around the world, the current study aims to annotate data in the Chinese Learner Written Corpus of National Taiwan Normal University and to reveal the common error types made by CSL learners in Taiwan, results of which could offer researchers useful insights for further research on CSL learners' common errors. In the next two sections, we will firstly describe how we annotated errors in Chinese Learner Written Corpus, and present common error types identified in the corpus.

## 2 Method

### *2.1 The Learner Corpus*

In this study, the Chinese Learner Written Corpus (http://kitty.2y.idv.tw/~hjchen/cwrite-mtc/main.cgi) was chosen as the target corpus for error-tagging. The corpus contains 4,288 essays (totally 2.14 million characters) written by CSL learners from 64 different countries at the Mandarin Training Center of National Taiwan Normal University during 2010–2012. All of the essays were take-home assignments hand-written by CSL learners and later manually typed as electronic files by the corpus builders. Genres of the essays include general epistle (e.g., a letter to your parents/siblings/friends), narrative (e.g., an unforgettable trip), argumentation

(e.g., a comparison between what you have in your home country and what we have in Taiwan), and application (e.g., your autobiography), written by learners across five proficiency levels (i.e., A2, B1, B2, C1, and C2 refer to the Common European Framework of Reference for Languages) and graded from 2 to 9.

## *2.2 Tagging of the Learner Corpus*

### 2.2.1 Error Domain and Category

To annotate errors in the Chinese Learner Written Corpus, we adopted the hierarchical tag sets of grammatical errors established by Chang (2017), an error classification system that combines both target modification taxonomy (TMT) and linguistic category classification (LCC). The TMT system is "based on the ways in which the learner's erroneous version is different from the presumed target version" (James, 1998, p.106), while the LCC system is carried out "in terms of where the error is located in the overall system of the target language based on the linguistic item which is affected by the error" (James, 1998, p.105). In the error classification system by Chang, an error is tagged simultaneously with a capital letter denoting target modification based on TMT and subsequent lowercase letters denoting the linguistic category of the error based on LCC. There are four error types of target modification, namely missing (M), redundant (R), incorrect selection (S), and word ordering error (W). As for linguistic category, there are totally 36 error types distributed into word-level error, grammatical function-level error, sentence pattern-level error, and mixture error (see Table 1). The advantage of using such a mixed error classification system is that the annotator can effectively assign an error to a specific tag without referring to the tagset each time. Once the annotator specifies how an erroneous surface structure deviates from the target language based on the four main types (i.e., M, R, S, and W), the annotator will only need to identify the problematic linguistic item of that error.

### 2.2.2 Error Marking Tool

In this study, we employed a software developed by a programming team led by Prof. Yuen-Hsien Tseng at NTNU to annotate errors in the learner corpus, the interface of which is shown in Fig. 1. The left column shows the text files of the learner corpus, and the other two columns present the running text of each selected file. Annotators can mark errors in a chosen text in the central column, and errors will be highlighted in red with error tags. The right column then presents the text corrected by annotators, and corrections will be highlighted in blue.

**Table 1** Tags of errors in linguistic category (adopted from Chang, 2017)

| Linguistic category | |
|---|---|
| Word-level (16 cases) | Action verb (v), auxiliary (aux), stative verb (vs), noun (n), pronoun (pron), conjunction (conj), preposition (p), numeral (num), demonstrative (det), measure word (cl), sentential particle (sp), aspectual particle (asp), adverb (adv), structural particle (de), question word (que), plural suffix (plural) |
| Grammatical function-level (11 cases) | Subject (sub), object (obj), noun phrase (np), verb phrase (vp), preposition phrase (pp), modifier (mod), time expression (time), place expression (loc), transitivity (tran), separable structure (vo), [numeral/determiner + measure] phrase (dm) |
| Sentence pattern-level (7 cases) | Complex noun clause (rel), 把 *ba*-sentence (ba), 被 *bei*-sentence (bei), 讓 *rang*-sentence (rang), 是 *shi*-sentence (shi), 有 *you*-sentence (you), other patterns (pattern) |
| Mixture (2 cases) | Formation (form), ambiguity of syntactic or meaning (sentence) |

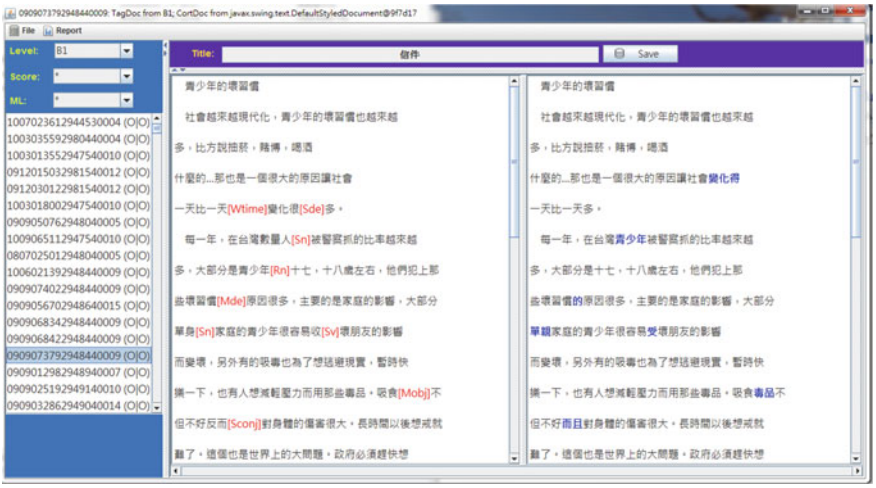


**Fig. 1** The interface of the error marking software

### 2.2.3 Principles of Error Marking

To ensure the consistency of the two human annotators' error identification and marking, the annotators would have to follow the annotation guidelines developed in this study. First, corrections of errors were made with two premises. The first premise was that annotators' corrections should not alter what learners intended to express. In addition, annotators should use words/phrases in accordance with learners' language proficiency. Secondly, annotators would firstly determine the target modification of an error (i.e., M, R, S, W) and then assign the erroneous element to the linguistic category.

# 3 Results and Discussion

## *3.1 Number and Distribution of All the Annotated Errors*

In the learner corpus, 48,266 errors were identified and annotated by the annotators, which were distributed into 119 error tags. The numbers and percentages of the 119 error tags are presented in Table 2. Many of the errors belonged to incorrect selection and missing linguistic components, which respectively took up 39.86 and 36.24% of the total errors. As for the linguistic category, 80.7% of the total errors belonged to the word-level, while errors at the other three levels took up less than 20%. In addition, incorrect selection of word-level linguistic components, missing word-level linguistic components, and redundant word-level linguistic components totally accounted for 77.42% of the total errors, whereas word ordering errors of word-level linguistic components took up only around 3%. Further examinations of errors at the word-level revealed that nouns, action verbs, and adverbs were the top three commonly misused linguistic components, all of which accounted for more than 13% of the total errors, and the fourth commonly misused linguistic components, structural particle (de), amounted to around 9% of the total errors. These four commonly misused components amounted to around 50% of the total errors.

In sum, the distribution of the 48,266 errors revealed that CSL learners have greater difficulties in choosing the right words or making correct sentences with necessary word-level components. These deficiencies were especially serious in their use of nouns, action verbs, adverbs, and structural particles. Since half of the total errors were in the four word classes, more investigations on words in these word classes should be further conducted to better understand how and why CSL learners misuse these components in their writing.

## *3.2 The Most Frequent Error Tags in the Learner Corpus*

To further understand the common error types in the learner corpus, error tags with more than 1,000 counts were identified for further discussion. Figure 2 illustrates the distribution of the top 12 error tags with more than 1,000 counts, which accounted for more than 50% of the total errors. The most common errors were attributed to the incorrect selection of nouns (Sn) and action verbs (Sv), the summation of which constituted 20% of the total errors; the other 10 error types, on the other hand, represented around 30%. Table 3 presents example sentences extracted from the learner corpus for the 12 error tags.

While the incorrect selection of nouns was the most frequent errors identified in the learner corpus, it was also the only one out of the 12 error types that related to the misuse of nouns. Among the top 12 error types, three resulted from the misuse of verbs (i.e., Sv, Svs, and Mv), three resulted from the misuse of adverbs (i.e., Madv, Sadv, and Radv), two resulted from the misuse of structural particles (i.e., Mde and Rde),

**Table 2** Distribution of the 48,266 errors among the 119 error tags

| | | M | | R | | S | | W | | Subtotal | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % | n | % | n | % |
| Word-level | v | 1,474 | 3.05 | 830 | 1.72 | 4,181 | 8.66 | 0 | 0 | **6,485** | **13.44** |
| | vs | 449 | 0.93 | 204 | 0.42 | 1,926 | 3.99 | 126 | 0.26 | **2,705** | **5.6** |
| | n | 463 | 0.96 | 577 | 1.2 | 5,556 | 11.51 | 330 | 0.68 | **6,926** | **14.35** |
| | aux | 1,741 | 3.61 | 337 | 0.7 | 508 | 1.05 | 88 | 0.18 | **2,674** | **5.54** |
| | pron | 69 | 0.14 | 140 | 0.29 | 251 | 0.52 | 0 | 0 | **460** | **0.95** |
| | conj | 940 | 1.95 | 527 | 1.09 | 1,021 | 2.12 | 59 | 0.12 | **2,547** | **5.28** |
| | p | 911 | 1.89 | 535 | 1.11 | 385 | 0.8 | 0 | 0 | **1,831** | **3.79** |
| | num | 84 | 0.17 | 36 | 0.07 | 49 | 0.1 | 0 | 0 | **169** | **0.35** |
| | det | 320 | 0.66 | 224 | 0.46 | 339 | 0.7 | 51 | 0.11 | **934** | **1.94** |
| | cl | 377 | 0.78 | 95 | 0.2 | 495 | 1.03 | 6 | 0.01 | **973** | **2.02** |
| | sp | 31 | 0.06 | 31 | 0.06 | 16 | 0.03 | 0 | 0 | **78** | **0.16** |
| | asp | 1,202 | 2.49 | 810 | 1.68 | 107 | 0.22 | 84 | 0.17 | **2,203** | **4.56** |
| | adv | 2,923 | 6.06 | 1,239 | 2.57 | 1,618 | 3.35 | 632 | 1.31 | **6,412** | **13.28** |
| | de | 2,478 | 5.13 | 1,129 | 2.34 | 603 | 1.25 | 189 | 0.39 | **4,399** | **9.11** |
| | que | 34 | 0.07 | 26 | 0.05 | 86 | 0.18 | 13 | 0.03 | **159** | **0.33** |
| | Subtotal | **13,496** | **27.95** | **6,740** | **13.96** | **17,141** | **35.51** | **1,578** | **3.26** | **38,955** | **80.7** |
| Grammatical function-level | sub | 347 | 0.72 | 127 | 0.26 | 0 | 0 | 201 | 0.42 | **675** | **1.4** |
| | obj | 293 | 0.61 | 12 | 0.02 | 0 | 0 | 219 | 0.45 | **524** | **1.09** |
| | np | 368 | 0.76 | 253 | 0.52 | 177 | 0.37 | 176 | 0.36 | **974** | **2.02** |
| | vp | 25 | 0.05 | 41 | 0.08 | 84 | 0.17 | 104 | 0.22 | **254** | **0.53** |

(continued)

**Table 2** (continued)

| | | M | | R | | S | | W | | Subtotal | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % | n | % | n | % |
| | pp | 19 | 0.04 | 40 | 0.08 | 9 | 0.02 | 243 | 0.5 | **311** | **0.64** |
| | mod | 33 | 0.07 | 67 | 0.14 | 0 | 0 | 24 | 0.05 | **124** | **0.26** |
| | time | 176 | 0.36 | 57 | 0.12 | 245 | 0.51 | 103 | 0.21 | **581** | **1.2** |
| | loc | 311 | 0.64 | 112 | 0.23 | 207 | 0.43 | 93 | 0.19 | **723** | **1.5** |
| | tran | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0.04 | **20** | **0.04** |
| | dm | 72 | 0.15 | 123 | 0.25 | 14 | 0.03 | 56 | 0.12 | **265** | **0.55** |
| | vo | 65 | 0.13 | 0 | 0 | 0 | 0 | 31 | 0.06 | **96** | **0.2** |
| | plural | 0 | 0 | 15 | 0.03 | 0 | 0 | 0 | 0 | **15** | **0.03** |
| | Subtotal | **1,709** | **3.53** | **847** | **1.73** | **736** | **1.53** | **1,270** | **2.62** | **4,562** | **9.46** |
| Sentence pattern-level | rel | 0 | 0 | 0 | 0 | 14 | 0.03 | 3 | 0.01 | **17** | **0.04** |
| | ba | 61 | 0.13 | 44 | 0.09 | 14 | 0.03 | 2 | 0 | **121** | **0.25** |
| | bei | 71 | 0.15 | 42 | 0.09 | 18 | 0.04 | 3 | 0.01 | **134** | **0.28** |
| | rang | 175 | 0.36 | 51 | 0.11 | 97 | 0.2 | 0 | 0 | **323** | **0.67** |
| | shi | 886 | 1.84 | 448 | 0.93 | 92 | 0.19 | 42 | 0.09 | **1,468** | **3.04** |
| | you | 504 | 1.04 | 385 | 0.8 | 88 | 0.18 | 29 | 0.06 | **1,006** | **2.08** |
| | pattern | 99 | 0.21 | 47 | 0.1 | 49 | 0.1 | 0 | 0 | **195** | **0.4** |
| | Subtotal | **1,796** | **3.73** | **1,017** | **2.12** | **372** | **0.77** | **79** | **0.17** | **3,264** | **6.76** |
| Mixture | form | 495 | 1.03 | 0 | 0 | 377 | 0.78 | 0 | 0 | **872** | **1.81** |
| | sentence | 0 | 0 | 0 | 0 | 613 | 1.27 | 0 | 0 | **613** | **1.27** |
| | Subtotal | **495** | **1.03** | **0** | **0** | **990** | **2.05** | **0** | **0** | **1,485** | **3.08** |

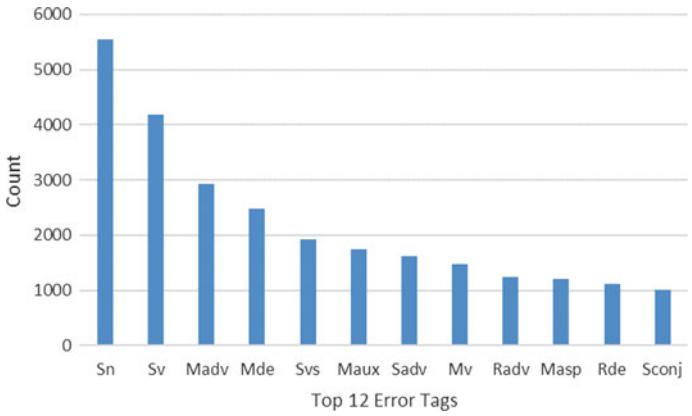


**Fig. 2** Distribution of the top 12 error tags

and the others related to the misuse of different word-classes. Based on the findings, it was obvious that the learners were prone to misuse verbs, adverbs, and structural particles in various ways. On the contrary, their misuse of nouns was mostly attributed to incorrect selections, suggesting that learners' incorrect use of nouns might result from their confusion of nouns with similar meanings or forms. Hence, research on CSL/CFL learners' misuse of nouns is suggested to specifically investigate learners' difficulties in acquiring and differentiating synonymous nouns. As for the misuse of verbs and adverbs, researchers are suggested to examine CSL/CFL learners' use of specific verbs/adverbs and uncover the causes of their misuse(s).

## *3.3 Comparison of Results in This Study and the Previous*

In addition to presenting the common error types in our learner corpus, we also compared findings yielded in our study with those in Lee et al. (2016). The reasons for drawing such a comparison are that both the two studies used the same error annotation system and investigated CSL learners' written production in traditional Chinese. Comparisons between the two studies might help us to identify the common errors produced by SL/FL learners of traditional Chinese. Table 4 presents the comparisons of the top 10 error tags in the two studies.

As shown in Table 4, nine out of the top 10 error tags in this study also appeared in Lee et al. (2016). The top 3 error tags in the two studies were an incorrect selection of nouns (Sn), incorrect selection of action verbs (Sv), and missing adverbs (Madv), though the top two were in reversed orders. From top 4 to top 10, however, rankings in the two studies were somewhat different. Discrepancies in the rankings of missing auxiliary (Maux), incorrect selection of adverbs (Sadv), and missing action verbs (Mv) in the two studies were small. Missing action verbs ranked eighth in both studies.

**Table 3** Example sentences and suggested corrections of the top 12 error tags

| Rank | Tag | Example sentence |
|---|---|---|
| 1 | Sn | *(a) 雖然青年人[Sn]吸毒已成為當前很多國家的社會問題。<br>*Suiran qingnianren* [Sn] *xidu yi chengwei dangqian henduo guojia de shehui wenti.*<br>(b) 雖然年輕人吸毒已成為當前很多國家的社會問題。<br>*Suiran nianqing ren xidu yi chengwei dangqian henduo guojia de shehui wenti.*<br>'Although youngsters' use of drugs has currently become a social problem in many countries.' |
| 2 | Sv | *(a) 老師教得很好,常使[Sv]我們複習,以便我們都會用新學到的生詞、句型等。<br>*Laoshi jiao de hen hao, chang shi* [Sv] *women fùxi, yibian women duhui yong xin xue dao de shengci, ju xing deng.*<br>(b) 老師教得很好,常幫我們複習,讓我們都會用新學到的生詞、句型等。<br>*Laoshi jiao de hen hao, chang bang women fuxi, rang women duhui yong xin xue dao de shengci, ju xing deng.*<br>'The teacher teaches very well, who often helps us review things we learned so that we can use the newly acquired words, sentence patterns, etc.' |
| 3 | Madv | *(a) 每次選擇的時候,有好悶的感覺<br>*Mei ci xuanze de shihou, [Madv] you hao men de ganjue.*<br>(b) 每次選擇的時候,都有好悶的感覺。<br>*Mei ci xuanze de shihou, dou you hao men de ganjue.*<br>'I feel so stuffy every time when I have to make choice.' |
| 4 | Mde | *(a) 我們唱歌要比誰唱[Mde]最好。<br>*Women changge yao bi shui chang* [Mde] *zui hao.*<br>(b) 我們唱歌要比誰唱得最好。<br>*Women changge yao bi shui chang de zui hao.*<br>'We sing to compete for the best singer.' |
| 5 | Svs | *(a) 在美國,家庭主婦越來越少,職業婦女越來越豐富[Svs]。<br>*Zai meiguo, jiating zhufu yue lai yue shao, zhiye funu yue lai yue fengfu* [Svs].<br>(b) 在美國,家庭主婦越來越少,職業婦女越來越多。<br>*Zai meiguo, jiating zhufu yue lai yue shao, zhiye funu yue lai yue duo.*<br>'There are less housewives yet more professional women in the United States.' |
| 6 | Maux | *(a) 他不但[Maux]說兩個語言而且會跳舞!<br>*Ta budan* [Maux] *shuo liang geyuyan erqie hui tiaowu.*<br>(b) 他不但會說兩個語言而且會跳舞!<br>*Ta budan hui shuo liang ge yuyan erqie hui tiaowu.*<br>'He can not only speak two languages but also dance.' |
| 7 | Sadv | *(a) 那時候,冬天好[Sadv]到了。每天的風景與變化對當時的我來說, 都很美麗。<br>*Na shihou, dongtian hao* [Sadv] *daole. Meitian de fengjing yu bianhua dui dangshi de wo lai shuo, dou hen meili.*<br>(b) 那時候,冬天剛好到了。每天的風景與變化對當時的我來說,都很美麗。<br>*Na shihou, dongtian ganghao daole. Meitian de fengjing yu bianhua dui dangshi de wo lai shuo, dou hen meili.*<br>'At that time, winter had just arrived. The everyday changing scenery was very beautiful to me at that time.' |

(continued)

**Table 3** (continued)

| Rank | Tag | Example sentence |
|---|---|---|
| 8 | Mv | *(a) 台灣的文化跟美國[Mv]起來完全不一樣。最大的差別是宗教的影響。<br>*Taiwan de wenhua gen meiguo* [Mv] *qilai wanquan bu yiyang. Zuida de chabie shi zongjiao de yingxiang.*<br>(b) 台灣的文化跟美國比起來完全不一樣。最大的差別是宗教的影響。<br>*Taiwan de wenhua gen meiguo bi qilai wanquan bu yiyang. Zuida de chabie shi zongjiao de yingxiang.*<br>'The culture of Taiwan is completely different from that of the United States. The biggest difference is the influence of religion.' |
| 9 | Radv | *(a) 不同的利益團體對於環保與經濟發展的價觀非常不同,而且非常[Radv]互不信任。<br>*Butong de liyi tuanti duiyu huanbao yu jingji fazhan de jia guan feichang butong, erqie feichang* [Radv] *hu bu xinren.*<br>(b). 不同的利益團體對於環保與經濟發展的價觀非常不同,而且互不信任。<br>*Butong de liyi tuanti duiyu huanbao yu jingji fazhan de jia guan feichang butong, erqie hu bu xinren.*<br>'Different interest groups have very different views on environmental protection and economic development, and they do not trust each other.' |
| 10 | Masp | *(a) 我是從日本來的。不過我想很多日本同學們介紹[Masp]日本。<br>*Wo shi cong riben lai de. Buguo wo xiang henduo riben tongxuemen jieshao* [Masp] *riben.*<br>(b) 我是從日本來的。不過我想很多日本同學們介紹過日本。<br>*Wo shi cong riben lai de. Buguo wo xiang henduo riben tongxuemen jieshaoguo riben.*<br>'I am from Japan. But I think many Japanese classmates have introduced Japan.' |
| 11 | Rde | *(a) 所以我每天不但要很早地[Rde]起來,還要乖乖地聽旅館裡的人的話。<br>*Suoyi wo meitian budan yao hen zao de* [Rde] *qilai, hai yao guaiguai de ting luguan li de ren dehua.*<br>(b) 所以我每天不但要很早起來,還要乖乖地聽旅館裡的人的話。<br>*Suoyi wo meitian budan yao hen zao qilai, hai yao guaiguai de ting luguan li de ren dehua.*<br>'Hence, I need to not only get up very early every day but also listen to staff in the hotel.' |
| 12 | Sconj | *(a) 只是[Sconj]這樣,他才能在挽救他的家庭告一段落後,進入人生並追求心裡上的啓示。<br>*Zhishi* [Sconj] *zheyang, ta caineng zai wanjiu ta de jiating gao yiduanluo hou, jinru rensheng bing zhuiqiu xinli shang de qishi.*<br>(b) 只有這樣,他才能在挽救告一段落後,進入人生並追求心裡上的啓示。<br>*Zhiyou* [Sconj] *zheyang, ta caineng zai wanjiu ta de jiating gao yiduanluo hou, jinru rensheng bing zhuiqiu xinli shang de qishi.*<br>'Only in this way can he enter life and pursue the quest of spiritual enlightenment after saving his family.' |

*Note* For each error tag, both erroneous sentence and suggested correction are provided. Erroneous sentences are labeled with * (a), and suggested corrections are labeled with (b)

Rankings of missing auxiliary and incorrect selection of adverbs were both one place higher in Lee, Chang, and Tseng. On the contrary, missing structural particles (Mde), incorrect selection of stative verbs (Ssv), and missing aspectual particles (Masp) ranked quite differently in this study and in Lee, Chang, and Tseng. Missing structural particles and incorrect selection of stative verbs ranked respectively the

**Table 4** Comparisons between the top 10 error tags in this study and those in Lee, Chang, and Tseng (2016)

| | This study | | | Lee et al. (2016) | | |
|---|---|---|---|---|---|---|
| Rank | Tag | n | % | Tag | n | % |
| 1 | Sn | 5556 | 11.51 | Sv | 3809 | 11.26 |
| 2 | Sv | 4181 | 8.66 | Sn | 2167 | 6.40 |
| 3 | Madv | 2923 | 6.06 | Madv | 1755 | 5.19 |
| 4 | Mde | 2478 | 5.13 | Masp | 1602 | 4.73 |
| 5 | Svs | 1926 | 3.99 | Maux | 1357 | 4.01 |
| 6 | Maux | 1741 | 3.61 | Sadv | 1168 | 3.45 |
| 7 | Sadv | 1618 | 3.35 | Mde | 1165 | 3.44 |
| 8 | Mv | 1474 | 3.05 | Mv | 1040 | 3.07 |
| 9 | Radv* | 1239 | 2.57 | Rasp* | 1003 | 2.96 |
| 10 | Masp | 1202 | 2.49 | Svs | 780 | 2.31 |

*Note* Tags with asterisks (*) are overlapped items in the two studies

fourth and the fifth in this study, yet they only ranked the seventh and tenth in Lee, Chang, and Tseng. In contrast, missing aspectual particles, the tenth common error tags in our study, ranked fourth in Lee, Chang, and Tseng. In addition, errors of redundant adverbs (Radv) ranked ninth in our study, whereas it was not included in the top 10 common error tags in Lee, Chang, and Tseng. The ninth common error tag in their study was redundant aspectual particles (Rasp), while this error type ranked the seventeenth out of the total errors in our study.

Findings of the comparison revealed that 90% of the top 10 error tags in our study overlapped with those in Lee et al. (2016), suggesting that these error types are indeed common in CSL learners' written production and should be further investigated in future research. Regardless of the 90% coverage of the top 10 error tags, rankings of the overlapped items in both studies were sometimes different, such as missing structural particles, incorrect selection of stative verbs, and missing aspectual particles. In addition, errors of redundant adverbs were also not listed in the top 10 error types in Lee, Chang, and Tseng. For these discrepancies, two possible explanations are provided. The first explanation lies in the different contexts where data in the two annotated corpora were gathered. Data in our study consisted of learners' writing assignments, while those in Lee et al. (2016) consisted of exam scripts. Exam scripts might better reflect learners' language proficiency in a way that no consultation of resources was allowed within the context of examination (Yang, 2003); nevertheless, the pressure learners experienced during the test might somewhat negatively influence their actual language use and thus cast doubt on the authenticity of the learner data. As a result, the contextual difference between the two sets of data might contribute to the different rankings of the top 10 error tags in the two studies.

Another explanation for the discrepancies lies in the proficiency levels of the learners in the two corpora. Our learner corpus contains writing assignments

produced by learners from the basic level to the advanced level. The learner corpus in Lee et al. (2016), however, comprises exam scripts that scored at least 3 or higher, which represent learners at the intermediate to advanced levels. The different range of language proficiency of the two datasets might be the cause of the ranking differences of some error types in the two studies. Some error types ranking higher in our study but lower in Lee, Chang, and Tseng might be errors that are more often made by learners at the lower level (e.g., missing structural particles and incorrect selection of stative verbs). On the other hand, errors ranked higher in their study yet lower in ours might be difficult for even higher-level learners to acquire. However, since the cross-level comparison was not the focus of the current study, the potential influence of proficiency levels on the two studies' different findings could not be confirmed. More research should be done to examine the distribution of each error tags at different proficiency levels.

## 4 Conclusion and Suggestions for Future Research

This study was set out to annotate errors in the Chinese Learner Written Corpus of National Taiwan Normal University and to present an overview of CSL learners' common error types. Manual annotation of the corpus yielded 48,266 errors distributed into 119 error tags, and more than 75% of the total errors belonged to incorrect selection or missing linguistic components. Among the four linguistic categories, around 80% of the total errors were caused by the misuse of word-level linguistic components, and about 50% of the total errors resulted from the misuse of nouns, action verbs, adverbs, and structural particles. Among these four commonly misused word classes, noun-based errors were mostly made by incorrect selection, whereas verb-based (including action verbs and stative verbs), adverb-based, and structural particle-based errors were committed in more diverse ways (i.e., incorrect selection, missing, and redundancy). Comparisons of the top 10 error tags in the current study and the previous one revealed that nine out of the top 10 error tags overlapped in the two studies, while rankings of the nine error tags in one study were somewhat different from the other. Regardless of the ranking difference, the 90% overlapping rate of the top 10 error tags in the two corpora suggests that these errors are indeed commonly misused items in CSL learners' writing and should be further investigated in future research.

Based on the findings yielded in this study, suggestions for future research are offered. First, CSL learners' use of nouns, verbs, adverbs, and structural particles should be extensively investigated, since these four word classes took up more than 50% of the total errors. Investigations on learners' use of these components might better reveal learners' difficulties in acquiring them and further provide useful information for effective material writing and teaching. In addition, since noun-based errors were mostly attributed to incorrect selection of other nouns, further examinations of CSL learners' perceptive and productive knowledge of synonymous nouns are also recommended to uncover how and why CSL learners made such type of

errors. In addition to targeting the four word classes, research on the common error types made by CSL learners at different proficiency levels is also suggested. Comparisons between findings in our study and those in the previous one have indicated that language proficiency might play a role in CSL learners' production of different error types; cross-level comparisons of common errors made by learners at different proficiency levels are hence recommended to discover whether longitudinal changes occur in CSL learners' making of errors.

## References

Brooke, J., & Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *Proceedings, 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul.

Centre for English Corpus Linguistics. (2020). Learner Corpora around the World. Louvain-la-Neuve: Universite catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Chang, L.-P. (2017). TOCFL 學習者語料庫的偏誤標記 [The error annotation of TOCFL Learner Corpus]. In H.-J. Chen (Ed.), *Corpus and teaching Chinese as a second language* (pp. 159–196). Taipei: Taiwan Higher Education Press Co.

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

Diaz-Negrillo, A., & Dominguez, J. F. (2006). Error tagging systems for learner corpora. *Revista Española De Lingüistica Aplicada, 19*, 83–102.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

Han, L. (2016). 基于HSK 动态作文语料库的连词"而且"偏误分析 [An error analysis of the conjunction *erqie* based on the HSK Dynamic Composition Corpus.] *Journal of Lanzhou Institute of Education*, *32*(3), 18–19.

Hu, R.-J. (2012). 试析留学生范围副词"都"的偏误 [An error analysis of *dou* used by foreign learners of Chinese]. *Foreign Language and Literature, 1*, 72–73.

Hsu, H.-M. (2011). 基于HSK动态语料库研究留学生作文中的偏误现象 [*A research on the errors of foreign students' composition: Based on the HSK Dynamic Corpus*]. Master's thesis: Liaoning Normal University.

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

Jia, X.-H. (2007). 日本人汉语学习者语料库的建立与语法偏误分类法 [Establishment and grammar errors categories of learner corpus in Chinese learner of Japanese]. *Research of Applied Linguistic, 1*, 12–16.

Jin, L.-J. (2011). 韩国留学生使用介词"在"的偏误分析 [The error analysis of Korean student's use of the preposition *zai*]. *Journal of the Graduates at Sun Yat-Sen University (social Sciences), 32*(4), 6–11.

Lee, L. H., Chang, L. P., & Tseng, Y. H. (2016). Developing learner corpus annotation for Chinese grammatical errors. In *Proceedings of the 20th International conference on Asian language processing* (pp. 254–257). Tainan, Taiwan.

Li, J. (2013). 基于HSK 动态作文语料库的泰国学生"有" 字句的习得考察 [A study on the second language acquisition of *you*-sentence of Thai students based on HSK dynamic composition Corpus]. *Overseas Chinese Education*, (4), 388–397.

Lin, Y.-T., Chen, H.-J., & Wang, C.-C. (2014). 從學習者語料庫探究趨向補語[起來]之偏誤情形及教學建議 [A learner corpus-based study on Chinese directional complement *Qilai*]. *Journal of Chinese Language Teaching, 11*(4), 73–109.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge University Press.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics, 39*, 74–92.

Nemser, W. (1974). Approximative systems of foreign language learners. *Error analysis: perspectives on second language acquisition*, 55–63.

Nicholls, D. (2003). The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference* (vol. 16, pp. 572–581).

Pan, P., & Liu, L. (2006). 学习者语料库与外语教学研究 [Learner corpus and foreign language teaching research]. *Journal of Beijing International Studies University, 4*, 53–55.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching, 10*(1), 209–232.

Swanson, B., & Charniak, E. (2013). Extracting the native language signal for second language acquisition. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 85–94).

Tono, Y. (2003). Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference* (pp. 800–809). Lancaster: University Centre for Computer Corpus Research on Language.

Wang, H.-C. (2010). 俄罗斯留学生使用"了"的偏误分析 [An error analysis on *le* used by Russian students]. *Chinese Language Learning, 3*, 99–104.

Wang, Y.-T., Chen, H.-J., & Pan, I.-T. (2013). 基於中介語語料庫之近義動詞混用情形調查與分析-以[幫],[幫助],[幫忙]及[變],[變得],[變成] 為例 [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.

Wang, C., & Seneff, S. (2007). Automatic assessment of student translations for foreign language tutoring. In *Human language technologies 2007: The conference of the north American chapter of the association for computational linguistics; proceedings of the main conference* (pp. 468–475).

Wang, M., Malmasi, S., & Huang, M. (2015). The Jinan Chinese learner corpus. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 118–123).

Yang, H.-Z. (2003). 中国学习者英语语料库 *[Chinese Learner English Corpus]*. Shanghai Foreign Language Education Press.

Zang, W.-T. (2014). 对外汉语教学中带有标记词的强调句研究 [*Research on the emphatic patter with token words in teaching Chinese as a foreign language*]. Master's thesis, Senyan Normal University.

Zhang, B.-L. (2010). 回避与泛化——基于"HSK 动态作文语料库"的"把" 字句习得考察 [Avoidance and overgeneralization —an investigation of acquisition of the *ba*-Sentence based on the HSK Dynamic Composition Corpus]. *Chinese Teaching in the World, 2*, 263–278.

Zhang, R.-P. (2013). 三个汉语中介语语料库若干问题的比较研究 [A comparison study on some problems in three Chinese interlanguage corpora]. *Institute of Applied Linguistic, 3*, 133–140.