International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 3, No. 2, 2022

A Corpus-Based Analysis of Chinese Character Errors

Ming Zhang [1], Yang Zhang [2]

[1] Shandong Vocational College of Science and Technology, Weifang Shandong 261053, China
[2] College of Liberal Arts, Jiangxi Normal University, Nanchang Jiangxi 330022, China

Abstract: Based on the corpus, this research conducts a systematic statistical and analysis on the types of Chinese characters errors of Chinese learners from different countries. It is found that errors of traditional Chinese characters are the main types of learners from Philippines, Indo-European language countries and Sino-Tibetan language countries. Learners from Japan, South Korea and North Korea have the highest frequency of wrongly written characters. To know the frequency of various types of errors in learners with different native language backgrounds could provide reference for the teaching of Chinese characters as a second language.

Keywords: Error analysis, Interlanguage, Chinese characters, Corpus.

1. Introduction

Interlanguage is the term coined by Selinker to refer to the systematic knowledge of a second language which is independent of both the learner' first language and the target language [1]. Interlanguage is a dynamic language system, which is constantly moving from the departure level to the native-like level. The errors of a second language learner, which are the gaps between the interlanguage and the target language, are not random, but are in fact systematic.

With the development of computer technology, the corpus has become an important research tool for second language teaching and research. The construction of Chinese interlanguage corpus provides resources for interlanguage research and error analysis. Error analysis of Chinese characters is a very distinctive research field in teaching Chinese as a second language, and some research results have been achieved in recent years. The research mainly focuses on the types of Chinese writing errors, cognitive strategies and Chinese character teaching. Scholars such as Xiqiang X., Shiwei N., Ting G., Fengqin L., Deyin C. and Xuming Y. have studied the characteristics of Chinese characters errors of learners from different countries [2-6]. Zhiping Z., Zhengyu S., Yu L., Yang H. and Li S. and other scholars have conducted research on Chinese character course books and Chinese character teaching [7-10]. Based on a relatively large-scale corpus, this research conducts a systematic statistical and analysis of the types of Chinese characters errors of Chinese learners from different countries, which can provide reference for Chinese character teaching as a second language.

2. Source of Corpus

The corpus data used in this study is mainly based on the "HSK Dynamic Composition Corpus" of Beijing Language and Culture University. We obtained 11,554 pieces of composition from the database, with a total of 3.25 million characters of effective corpus. The candidates come from 95 countries and regions besides China. Considering the influence of mother tongue on learners, we classify candidates according to language family, and take the first official language of the country where the candidates are located as the basis for language family division, ignoring other common languages temporarily. If there are two or more official languages, which belong to the same language family, they will be merged; if they do not belong to the same language family, the language families of each language shall be listed separately. The classification is based on the classification system of world languages by Tongqiang X. and Jicheng H. from Peking University.

Since there is a great controversy over the division of language families of Japanese and Korean, moreover Japan, South Korea and North Korea are all countries in the Chinese character culture circle, which have certain particularity, we list them separately. In addition, the official languages of Philippines and Singapore, which have a large number of candidates' compositions, are more than one, and belong to different language families. It is difficult to classify the candidates into a certain language family, so the data of these two countries are also listed separately.

In the statistics, countries with less than 5 compositions, Chinese candidates and candidates with wrong nationality information are excluded. The total number of compositions in final is 11,253.

3. Data Analysis

The types of Chinese character errors that this research focuses on are mainly wrongly written characters, typos, traditional characters and variant characters. Unrecognized characters are not strictly a kind of errors, because it is impossible to determine which character the learner wrote, and such errors are not universal and systematic. Pinyin errors which mainly depend on the learner's learning habits and vocabulary are also not universal and systematic. Both types of these errors are outside the scope of this research.

Wrongly written characters refer to the characters written by the candidates do not exist in Chinese, which is due to the wrong strokes or radicals. Typos are errors of using the character A as the character B. Typos are characters that exist in Chinese. Native Chinese speakers also often make such mistakes in character recognition when using Chinese. However, learners of Chinese as a second language are more likely to make this mistake and show certain regularities. Traditional Chinese characters correspond to simplified Chinese characters. The clear distinction between simplified Chinese characters and traditional Chinese characters began with the "General List of Simplified Chinese Characters"

formulated in March 1964. Generally speaking, for the same Chinese character, the strokes of traditional characters are more than simplified characters. China regards simplified characters as standard Chinese characters. If the learners use traditional Chinese characters, we also regard it as a type of errors. Variant characters, in a broad sense, refer to the characters with same pronunciation and meaning but different shapes. In the narrow sense, variant characters refer to the characters that are homophonic and synonymous with the normative normal characters but in different shapes. The variant characters discussed in this research are in a narrow sense.

We made statistics on the corpus, and the average frequency of different character errors of learners from different language family (or country and region) is shown below from Table 1 to Table 4. The error frequency refers to the number of errors in one composition.

3.1. Wrongly Written Characters

The types of wrongly written characters include stroke errors, radical errors and structure errors. Stroke errors mainly include wrong strokes in quantity, or in shape, or in location. Radical errors mainly include misusing radicals and wrong radicals in quantity, or in shape, or in location. Structure errors mainly include loose structure and disordered structure.

Table 1. Average Frequency of Wrongly Written Characters

	Language Family (or Country and Region)	Wrongly Written Characters
1	Altaic languages	1.54
2	Austroasiatic languages	1.69
3	Austronesian family	1.24
4	Indo-European languages	2.81
5	Semito-Hamitic languages	7.79
6	Sino-Tibetan languages	2.90
7	Japanese	3.81
8	Korean	4.90
9	Philippines	3.63
10	Singapore	1.76
	Overall average	3.61

It can be seen from Table 1 that the learners whose mother tongues are Semito-Hamitic languages have the highest average frequency of errors, with an average of about 7.79 mistakes per composition, followed by learners from South Korea and North Korea, and learners from Japan also have a high frequency of errors, whose average frequencies are all higher than the overall average. The learners whose mother tongues are Austronesian family have the lowest frequency, with an average of only 1.24 mistakes per composition. That the average frequency of wrongly written characters among learners from Japan, South Korea and North Korea is high indicates that the negative transfer of Chinese characters in their mother tongue to their Chinese character learning may exceed the positive transfer.

3.2. Typos

If learners make a mistake about the relationship among the pronunciation, form and meaning of a Chinese character, they may write another character. By analyzing the typos errors in the corpus, we found that the learners' errors can be mainly divided into four categories: errors of typos with similar structure, errors of typos with similar pronunciation, errors of typos similar both in structure and pronunciation, errors of synonyms.

Errors of typos with similar structure refer to the errors that confuse two characters that are not similar in pronunciation but similar in form. Errors of typos with similar pronunciation refer to the errors that confuse two characters that are not similar in form but similar in pronunciation. Errors of typos that are similar both in structure and pronunciation refer to the errors that confuse two characters that are similar both in form and pronunciation but not synonyms. Errors of synonyms refer to the errors that confuse two characters that are not similar in form or pronunciation but similar in meanings.

Table 2. Average Frequency of Typos

	Language Family (or Country and Region)	Typos
1	Altaic languages	2.46
2	Austroasiatic languages	3.16
3	Austronesian family	2.57
4	Indo-European languages	3.48
5	Semito-Hamitic languages	3.53
6	Sino-Tibetan languages	2.97
7	Japanese	1.56
8	Korean	3.11
9	Philippines	2.92
10	Singapore	2.58
	Overall average	2.59

It can be seen from Table 2 that the learners whose mother tongue are Semito-Hamitic languages and Indo-European languages have a high average frequency of typos, with an average of about 4 typos per composition. Xin J. (2003) pointed out that the readers with the background of alphabetic writing first obtain phonetic information through visual input information, and activate semantic representation through phonetic representation, and phonetic coding is the main way [11]. Influenced by the cognitive processing mode of their mother tongue, they may pay more attention to the memory of phonetic symbols in pictophonetic characters, while ignoring the semantic symbols, which leads to the errors of typos. Learners from Japan have a low frequency of typos, with an average of less than 2 typos per composition. However, learners from South Korea and North Korea have a high frequency of typos, with more than 3 typos per composition. That is to say, although South Korea and North Korea are in Chinese character culture circle and the learners from these two countries have been exposed to ideographic characters, it does not significantly reduce the occurrence frequency of typos. The specific reasons for this phenomenon need further research.

3.3. Traditional Chinese Characters

It has been observed that the vast majority of candidates with errors in traditional Chinese characters use traditional characters throughout the composition, while a small number

of candidates mainly use simplified characters, occasionally interspersed with traditional Chinese characters. Errors in traditional Chinese characters can be divided into two categories: complete errors and component errors. A complete error means that the whole character is written in traditional Chinese, while a component error means that a certain part or radical of the character is not simplified, that is, it is half simple and half traditional, with elements of traditional Chinese characters.

Traditional Chinese characters have a history of thousands of years and are the carrier of traditional Chinese culture. Until the 1970s, most of the materials published in China were still mostly in traditional Chinese characters. The history of simplified Chinese characters is far less than that of traditional Chinese characters.

Table 3. Average Frequency of Traditional Characters

	Language Family (or Country and Region)	Traditional Characters
1	Altaic languages	0.20
2	Austroasiatic languages	3.94
3	Austronesian family	4.72
4	Indo-European languages	17.96
5	Semito-Hamitic languages	0.00
6	Sino-Tibetan languages	12.69
7	Japanese	0.78
8	Korean	1.25
9	Philippines	27.16
10	Singapore	3.90
	Overall average	3.77

It can be seen from Table 3 that the learners from Philippines have the highest average frequency of traditional Chinese characters (27.16), followed by learners whose mother tongue are Indo-European and Sino-Tibetan languages, with an average frequency of more than 10. This shows that traditional characters still have great influence today.

3.4. Variant Characters

The common errors of variant characters are mainly due to the different composition or relative position of the components. There is more than one variant character corresponding to an individual Chinese character in the corpus.

It can be seen from Table 4 that the learners from Singapore have a high average frequency of variant characters. The learners from Chinese character circle are influenced by Chinese earlier and more profoundly, and are prone to nonstandard writing of Chinese characters. There are some Chinese characters in both Japanese and Korean. Most of these characters are ancient Chinese characters, which are different from the current standard Chinese characters. However, from the statistical data, the average frequency of variant characters learners from Japanese, North Korean and South Korean do not appear to have a high average frequency. This phenomenon may be due to that the learners had paid special attention to the writing of these Chinese characters in the process of learning. The specific reason needs to be researched on these learners.

Table 4. Average Frequency of Variant Characters

	Language Family (or Country and Region)	Variant Characters
1	Altaic languages	0.00
2	Austroasiatic languages	0.05
3	Austronesian family	0.06
4	Indo-European languages	0.05
5	Semito-Hamitic languages	0.00
6	Sino-Tibetan languages	0.05
7	Japanese	0.00
8	Korean	0.01
9	Philippines	0.06
10	Singapore	0.13
	Overall average	0.03

4. Conclusion

With different native language backgrounds, Chinese learners show different characteristics of various characters errors. Through analysis, it can be seen that the most frequent errors of Chinese learners from various countries and regions are traditional Chinese characters, which is more likely made by the learners from Philippines, the Indo-European language countries and Sino-Tibetan language countries. The errors in traditional Chinese characters become the most predominant type of errors, much higher than other type of characters errors. The learners from Japan, South Korea and North Korea have the highest frequency of wrongly written characters, followed by typos, and the average frequency of errors in traditional Chinese characters is low. To know the characteristics of various errors of learners with different mother tongue backgrounds, teachers can focus on different aspects in Chinese character teaching.

Acknowledgment

This work was supported by Social Science Planning Project of Jiangxi Province (No. 17BJ21).

References

[1] Selinker, L., Interlanguage, International Review of Applied Linguistics in Language Teaching, vol.3, 1972, pp. 209-231.

[2] Xiqiang X., An analysis of character errors of foreign learners of Chinese, Chinese Teaching in the World, vol.2, 2002, pp. 79-85+4.

[3] Shiwei N., Classification and Evaluation of Foreigners' errors in the Writing of Chinese Characters: A Case Study of Wrongly Written Characters by Portuguese Beginning-level Students, Journal of Yunnan Normal University (Teaching and Research of Chinese as a Foreign Language), vol.4, 2021, pp. 75-85.

[4] Ting G., Characteristic and Development of Errors Types of Chinese Strokes by Students from Non-Chinese Cultural Circles, Journal of Yunnan Normal University (Teaching and Research of Chinese as a Foreign Language), vol.2, 2019, pp. 19-24.

[5] Fengqin L., An Error Analysis of Chinese Characters among Intermediate and Advanced Level South Korean Students, TCSOL Studies, vol.3, 2013, pp. 28-33+40.

[6] Deyin C. and Xuming Y., Error Analysis of Chinese Characters Writing in Thailand Learners Based on Corpus and the Corresponding Teaching Strategies, Overseas Chinese Education, vol.4, 2018, pp. 41-49.

[7] Zhiping Z., Structural Theory of Chinese Character and Chinese Character Teaching, Language Teaching and Linguistic Studies, vol.4, 2002, pp. 35-41.

[8] Zhenyu S., Thoughts on Some Problems in Teaching Chinese Characters to Foreign Countries, International Chinese Language Education, vol.3, 2018, pp. 3-19.

[9] Yu L., Selection and Presentation of the Chinese Characters in Elementary Chinese Textbooks Used in American Universities, Overseas Chinese Education, vol.2, 2013, pp. 124-130.

[10] Yang H. and Li S., The Application of Philology Theory in Teaching Chinese Characters as a Second Language, Journal of Liaoning University of Technology(Social Science Edition), vol.5, 2010, pp. 126-129.

[11] Xin J., The Relationship between Knowing Pronunciation and Knowing Meaning of Chinese Characters among CSL Learners, Language Teaching and Linguistic Studies, vol.6, 2003, pp. 51-57.