ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 5, No. 3, pp. 613-619, March 2015
DOI: http://dx.doi.org/10.17507/tpls.0503.22

Error Type Analysis in CAT

Dongmei Zhou
School of Foreign Languages, Wuhan Textile University, Wuhan, China

Abstract—**This paper conducts an empirical investigation on the errors in English-Chinese translation memories in computer-aided translation (CAT) and shows that the intention of error occurrence and the statistical difference in three common types: fixed expressions, omissions and symbols. It reveals that the translation errors in sentence pairs of translation memories reach as high as over 14%, and among all the errors in target texts punctuation errors account for nearly 46%. These findings are of great significance in improving the target text quality as well as in lowering the cost in CAT.**

Index Terms—**computer-aided translation, error types, translation memory**

I. INTRODUCTION

Steven (2009) says computer-aided translation(CAT) is an emerging industry in the translation market, which is a trend to make full use of artificial intelligence and develop through cooperation cross fields. It is distinct from both traditional translation and machine translation (James & Anastasia, 2000). Wherein human translators play key parts and the computer serves as a powerful assistant to improve translation efficiency as well as quality. Undoubtedly, translators in CAT projects are likely to make some typical errors due to the computer input procedure. As a result, more and more attention has been put on the quality check or post-editing of the target texts in CAT. Quality check or quality management has already become an essential and key part in CAT which directly decides the overall quality, efficiency and cost of the whole translation project in CAT.

Considering these important facts, this research is focused on the quality check of target texts in English-Chinese CAT translation memories. By doing a practical manual contrastive analysis of the 15000-word translation samples, we obtained a particular classification of some frequently occurring translation errors and compared them with those in traditional English-Chinese translation in order to help make the specific error features under CAT circumstances and achieve automatic identification of some typical errors in Chinese text without going back to and comparing with the original text or looking it up in the build-in dictionaries of the CAT software. This study can help find a clearer clue in target (Chinese) text quality check in CAT translation memories.

II. RESEARCH DESIGN

Unlike machine translation, Bowker (2002) says computer-aided translation (CAT) is more practical and accurate and a storage and retrieval operation which is carried out on line with a computer through out the whole translation procedure. Considering the fact that most of the current widely used CAT softwares and systems are designed to be one-way operations which means the texts imported to and exported from the CAT software are both monolingual. In order to better complete the quality check process and meanwhile guarantee the accuracy as well as effectiveness of the results, we have to select Chinese which is our mother language as the target text language. As a consequence, the subject of this study is about 15000 words of English-Chinese translation sentence pairs. All the samples are selected from 10 Trados translation memories uploaded by translation students both senior and MTI students in CAT class in a certain university. Given consideration to the proportion of different sorts of texts in all the translation memories, this research selects 6 novel translation and 4 article translation.

Based on the comprehensive contrastive analysis, 10 translation samples are picked randomly out of more than 40 translation memories and sorted according to their text type. Then 1500 word's Chinese texts are carefully checked without comparing with the original English texts in respects of wrongly written or mispronounced characters, sentence structure, punctuations. Then we go back to the original English texts for more detailed semantic and lexical quality check. The feature of each error type will be analyzed after collecting and counting various errors and making specific error classifications with SPSS 20.0. Then regular expression is used in text editor in order to approach automatic identification of some common error types out of the whole category list. Finally, the specific error types and the statistics are compared with that in some papers on traditional English-Chinese translation for finding out its uniqueness under CAT circumstances The whole process of this study is shown in Fig.1.

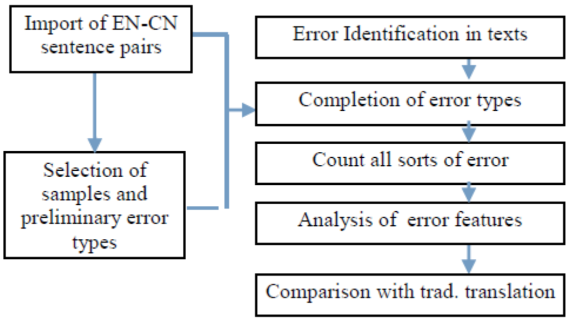
Import of EN-CN sentence pairs
Selection of samples and preliminary error types
Error Identification in texts
Completion of error types
Count all sorts of error
Analysis of error features
Comparison with trad. translation

Figure 1 Process flow

III. Finding and Discussion

After manual work on checking through all the 1000 English-Chinese sentence pairs selected from 10 translation memory samples, we have set the general error types (Error Level 1)including Critical Words, Non-Critical Words, Fixed Expressions and Symbols. Then through collecting and carefully analyzing the features of various translation errors in the target texts, we have eventually classified them into 11 specific error types (Error Level 2) under the general patterns. And the specific error types are shown in Table 1.

When working on the complete error classification, we have taken some English-Chinese translation error analysis and even some post-editing rules as references, including China's National Standard for Translation (GB/T 1968-2005), English-Chinese translation error types in Wang Jianjia's (2013) research based on corpus of 150 translation test paper, and machine translation errors by Luo and Li (2012) from Tongji University .

Table 1
CAT error type list

Error Types (L. 1)	Critical Words	Non-Critical Words	Fixed Expressions	Symbols
Weight	3	1	0.5	0.25
Sub-types (Level 2)	Critical semantic error; Omission of sentences, words and Numbers	Polysemy; POS; Prepositional phrase; Input error	Terminology; Brand ames& Names; Abbreviations	Punctuations; Units

When having done the sentence-by-sentence quality checking work, we collect all the errors appeared in the target texts and count their frequencies while sorting them into the error list shown above. (Table 1)

After classifying all the errors in the target texts, we first count the error rate of the whole samples. it soudns reasonable to take one English-Chinese sentence pair as a single unit instead of putting every single word in the text into the statistic results. When a sentence pair includes at least one of the error types listed in Table 1, we count it in the "WRONG PAIRS". Besides, when a sentence pair includes more than just one error type in it, we do not count repeatedly while calculating the translation error rate of the whole 1000 English-Chinese sentence pairs. The error rate, we calculate it as "WRONG PAIRS" / "TOTAL PAIRS" *100%. Among all the 1000 sentence pairs, we count 147 pairs as "WRONG PAIRS" and 853 as "CORRECT PAIRS".

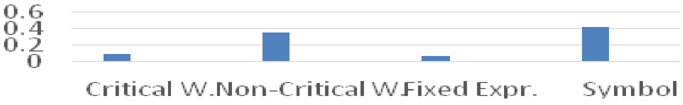

Figure 2 General error percentage

When we count the frequency of the four general error types occurring in the target texts, every single error is counted to calculate the total error number. Even as in some situations one sentence contains more than one error types or one error type occurring more than once. By doing this way, we have worked out the total error number at 273 and the frequencies of each general error type are shown in Fig. 2. From Fig. 2, it is obvious that elements of critical words, non-critical words, fixed expressions and symbols can result in errors. We will analyze them one by one.

A. *Critical Words*

It is quite clear that mistranslation of some critical words in the sentence may have a great influence on the understanding of the whole sentence. And that is why in most of the quality assessment standards "Critical Semantic Error "is weighted most. Considering the weight and influence of critical words, translators would spare no effort to avoid critical word translation error in order to ensure their target texts are up to the quality standard. In Fig. 3, among all the 273 translation errors, we find only 28 critical word errors. That is just about 10.26% of the total error number.

While under the general type “Critical Words”, we also work out statistics of frequencies and proportion of each specific error type. 15 critical semantic error, 5 omission of sentences and words, 8 Numeral& quantifier mistranslation.

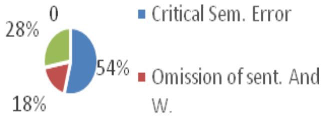

Figure 3 critical words

Here, we will use some examples of translation errors to go through a more detailed analysis of every specific error type.

First of all, in all the 28 Critical Words errors, we find 15 errors belong to critical semantic error. And the vast majority of these 15 errors appear in 5 out of the 6 novel samples. In fact just 3 errors appear in 2 out of the 4 informative articles. And these 3 errors are all mistranslation of key words. To be more detailed, the 15 errors are basically two sorts: mistranslation of double negative sentences and mistranslation of rhetorical questions.

a) Double negative sentence

EN: “I didn’t betray nobody,” Tracy cried, “and you are setting me up.”

CN: “wo shi you guo bei pan,”te lei xi han dao, “ke ni xian zai jiu shi zai gei wo xia tao.”

In this example, the translator might be confused by the double negative sentence “didn’t betray nobody” and translated it as positive. But it is actually a common practice among western people especially black people when they are trying to emphasize the point. They would frequently use “no” to replace “any”, like nobody, nothing for anybody and anything. So the first sentence in this example actually means “I didn’t betray anybody”. And here, just a slight mistranslation of words changed the meaning of the whole sentence.

b) Rhetorical question

EN: “Don’t you miss her, Ted?” “Yes. And I bet she will answer this way too.”

CN: “ni nan dao bu xiang ta ma, te de?” “dang ran xiang.er qie wo da du ta ye xiang wo.”

This example is a very typical mistranslation of rhetorical question. Obviously in the answer “yes” means “I don’t miss her”. And this is also the exact evidence of how translators lack language skills and influence the whole context of translation. And these two examples should represent “critical semantic error” well. As for omission of sentences and words, we find the reasons for the 5 omission errors basically result from two factors. One is careless operation of translators. When the sentence pair is rather short, like one word, the translator might ignore it. The other reason is misspelling of words in the original English texts, so that the translator left it untranslated. Fortunately, the majority of omission errors appeared in the samples can be identified by the QA checker of Trados system. Numeral& quantifier translation errors also account for a certain proportion though very small in “Critical Words” general error type, because the accuracy of translation of numbers is quite significant and sometimes determines the quality of whole translation. While among the 8 numeral translation errors in the samples, it is clear that the translators attach much importance of the accuracy in numbers when translating, which leads to the mistranslation. Here are two examples (c) and (d):

c) EN: “John and I, we have known each other for decades.”

CN: “yue han he wo yi jin ren shi you shi ji nian le.”

Obviously in this example “decades” was mistranslated. It means not “over ten years” but “several ten years”.

d) EN: Nowadays, computers work seven times faster than they did just 3 years ago.

CN: ru jin，ji suan ji yun xing su du bi qi jin jin 3 nian qing，jiu kuai le 7 bei.

Actually, this is a typical numeral translation error which has been discussed by many scholars. Here when westerners use “seven times faster than”, they really mean exactly “seven times as fast as”. And if Chinese people read only the translation, we would definitely regard it as “eight times as fast as”. And this special kind of numeral translation error is unacceptable in informative texts or government reports

The above four specific error types constitute the error type of “Critical Words”. It is quite reasonable that the frequency of this general type shows fairly low, just 28 out of all 273 errors. Though it doesn’t appear in a large number, it is an important error type. Once it occurs, it hurts the translation quality. What’s more, in order to identify and even solve this type of error, we have to take advantage of the existing CAT system with build-in dictionaries to carry out word-by-word comparison and even go back to the original English texts for careful manual check of sentence structure and meaning.

B. *Non-critical Words*

When we come to “Non-Critical Words”, we find it is of such a big proportion of all translation errors. In Fig.4 it is clear that among all 273 errors, we count 97 into this general error type. And to be more detailed, 97 errors in this type are made up of 13 polysemy errors, 7 prepositional phrase errors and 77 input clerical errors. The result is in Fig. 4. As a fact, the vast majority of English words are polysemant. Sometimes it is difficult for Chinese translators to pick out their specific meaning in a particular context. And even some of the frequently used words would adopt uncommon meanings in some situations. If translators are not aware of those uncommon use of words or only make choices among common meanings, mistranslation of words are almost unavoidable. Here are two examples (a) and (b):

a) EN: "I need a wrench and some nuts."

CN: "wo xu yao yi ba ban shou he yi xie jian guo."

Here "nuts" does not mean the food. It refers to a sort of tool used to bite the screw tight. In Chinese it should be translated as "luo mu".

b) EN: "I was cross with him at that time."

CN: "na shi wo zheng hao he ta ca jian er guo."

In this sentence, "cross" does not mean "pass by" but "mad or angry".

As for the 7 prepositional phrase errors, nearly all of them are mistranslated into their look-like phases or "twin phrases". In this type, we encounter "in case of" (in the case of), "out of question"(out of the question), "at a word" (in a word), "at no time"(in no time). Here is an example:

c) EN: "See? The engine lies right in the front of the wing."

CN: "kan jian mei？ ying qing jiu zai ji yi de qian fang."

In the sentence, the translator confused "in the front of" with "in front of". The correct translation should be "ji yi qian duan".

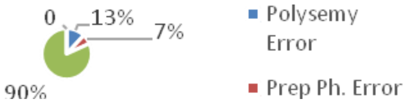

Figure 4 Non-critical words

In terms of the input clerical errors, there are 77 different careless Chinese character input errors in total. We present some most frequent ones in Table 2.

TABLE 2
CHINESE CHARACTER INPUT ERROR AND FREQUENCE

Wrong (Correct)	Frequency
… mention (say)	11
he (she)	15
slow ···(slowly)	28

It is quite clear that the last three errors account for the vast majority percentage of this specific error type. Although the " he (she)" error and the rest of low-frequency error can only be identified with going back to the original sentence pairs and carrying out careful manual check, the "mention (say)" and "slow (slowly)" two sorts of errors could be automatically identified with the help of regular expression in text editor to some extent. Due to the clear Chinese grammatical and structural feature of these two errors, we have the chance to achieve automatic identification and only deal with the target (Chinese) texts.

Firstly, when we use "mention", we must attach more contents after it. But when we use "say", it is always followed just by a comma, a period or a colon. So the pattern of this error is: "a number of Chinese characters" + "shuo dao" + "a comma, a period or a colon". For the first part we could use "[\u4e00-\u9fa5]*" to represent "any number of Chinese characters. And for the last part, it could be written in "[，。：；]". And considering the whole text in Chinese, we could even simplify it as "shuo dao[，。：；]" . After checking out this way in the text editor, we pick out all of the 11 "mention (say)" errors within the target texts.

Secondly the Chinese character "de (di)" error is a very common Chinese input slip-up. In fact, we find among the 28 errors there are 25 which should be "di" instead of "de". And this is typical misunderstanding in the use of Chinese adverbs and adjectives. Though there are some of the errors showing no common features, which means if we want to identify or even correct them, we have no choice but check carefully through the target texts, there are still over half of the 25 errors having something in common. That is before Chinese character "di", there are often two same characters or so-called reduplication. We can certainly take use of this common feature and try to identify this error. Although the research result might be larger than we need and not that accuracy, it can conclude most of this sort of error. And the trial method we put forward now is "([\u4e00-\u9fa5])\1[de di])".

C. *Fixed Expressions*

Compared to the first two general error types, the third general type "Fixed Expressions" occurs in quite low frequency. And the situations are much simpler. Here we have 13 name errors (missing of separatrix in foreign names), 5 Terminology errors and 4 abbreviation errors. And these errors of course we could solve it by setting detailed rules for QA Checker and improving our terminology database. Among them, the name errors takes 59%, the terminology error takes 23% and the abbreviation errors takes 18%.

D. *Symbols*

When dealing with the last general error type, we find that under this general type, all the errors we collected can be just sorted into two specific types: punctuations and units. Although the type number is quite small, the total error number in this general type is the biggest among all the 4 types. 119 punctuation errors and 7 unit errors. In total 126 here out of all 273 errors. The 7 unit errors are not rare, most of them just results from careless typing. Like “less than 12 miles” into “bu dao shi er li lu”. The types are listed in Table 3.

TABLE 3
PUNCTUATION ERROR TYPE DISTRIBUTION

No.	Type	Ratio
1	Missing of period	18%
2	Missing of book title quotes	5%
3	Missing of quotation marks	13%
4	Incompletion of bracket	6%
5	English punctuations in Chinese texts	7%
6	Quotation marks reverse	14%
7	Incompletion of quotation marks	37%

These 7 more detailed error types in punctuation shows very clearly that the translators using CAT softwares are even more likely to commit typing errors in the beginning or the end of the translated sentence than in doing the translation itself. And if we put together some of the similar errors in the list above, we will find that “missing or incompletion of punctuations” reaches as high as nearly 80% of all the 126 punctuation errors. Although it seems not that important in assessment weight. However when little things come in huge number, they make a great difference. Ignorance on this less-weighted error type could shake the stem of the whole quality of translation project. It is obvious that the translator should not only pay much more attention to the translation of words but also things around them. Unfortunately, most of these punctuation errors could be left to the QA Checker of Trados system. Such as “missing of period” and “English punctuations in Chinese texts”, what we should do is just making more strict rules for the checker and choose more regulations in full stop. To turn to the CAT system for help can no doubt save us from much heavy manual checking work, but there are still some punctuation errors we have to handle them ourselves such as the book title quotes. More or less, there will be some manual work in translation and post-editing no matter how advanced the information technology and artificial intelligence become. Computer Aided Translation has its limitations.

IV. ERROR COMPARISON WITH TRADITIONAL EC TRANSLATION

The study on common error types in traditional English-Chinese translation has attracted plenty of attention for quite a long time. And there have been some researchers doing various experiments and researches on different aspects of English-Chinese translation errors. In terms of error type list, there also have been some distinct results achieved by some people. According to Wang (2011) in his English-Chinese translation research based on corpus of 150 translation test paper, he collected all the translation errors appeared and sorted them into four different level: Lexical level, Phrasal level, Syntactic level and Textual level. And for each general level he set a few error types. And he sorted all the translation errors into different types. While in the corpus-based study of Liang (2004) on English-Chinese translation from non-English major postgraduates, they checked nearly 30000 words’ English-Chinese translation and concluded all the errors into 6 types. And in the research paper of Wang in 2011 on Chinese students’ common translation errors in English-Chinese translation, he classified the translation errors according to their causes into ten types. Taking their results as references, we can conclude a preliminary error type list for traditional English-Chinese translation. It is shown as follows:

TABLE 4
TRADITIONAL ERROR LIST

Error Type	Frequency	Proportion
Key Words	84	14.92%
Smoothness of Sentence	94	16.70%
Mismatch of Words	59	10.48%
Part of Speech	108	19.18%
Omissions	83	14.74%
Redundance	105	18.65%
symbols	30	5.33%
Total Error Number	563	
Total Word Number	29067	

If we compare the list above with our specific error type list in CAT, we can easily find out that there are some common types in both of them but still a lot of apparent differences. First of all, we can easily tell that in the traditional error list it focuses frequently on the structure and expression of the whole sentence in quality check. Smoothness of Sentence and Redundance error types are not inclusive in the CAT list, while input Chinese error which takes a huge part in the CAT list here is even not in the traditional list. Compared these two lists, we find that the error types in CAT list are more word-focusing and detailed. We think the reason for this obvious difference in error types has basically two aspects. One is selection of samples from different translators. In CAT study, we pick all the 10 translation samples from English major and MTI students. To some degree, they are quite skillful in both languages. They might have very little difficulty in understanding and sentence expression. As a consequence, the errors appeared in CAT translation memories are more subtle. While in the traditional error list, the statistics of samples are mostly from non-English major postgraduate students. So they made more errors in terms of expression and understanding. And the other factor we consider is technical the translator uses. In traditional translation, the translator's language skills more or less has an impact on the producing high-quality translation. While in CAT, it is not the case at all. Like we checked in the study, as the translators were of comparatively high level of language and translation skills, they did make very few errors in understanding. But they encountered a huge number of errors in choice of Chinese characters and punctuations due to the unfamiliarity of input method of keyboard and mis-operation of Chinese pinyin input softwares. Then when we come to those common error types in both the two lists, we find that the frequencies of them vary greatly. For instance, the "Omissions" in traditional list reaches as high as about 15% of total error amount. However in CAT list, it accounts for less than 2%. It shows the great advantage of computer-aided translation system. In CAT softwares, the whole passage or paragraph is divided strictly and clearly into separate sentence pairs. When the translators are working on a text, they do it pair by pair instead of taking it as a whole target, which often make the translators ignore something unconsciously. What's more, the CAT system will usually make a notice if a sentence is left unfinished. As a result, it is relatively easier for translators to avoid omissions under CAT circumstances. In another similar case, we can compare the "Mismatch of Words" (10.48%) in the traditional list with the "Fixed Expressions" (less than 5%) in CAT list. It also shows how much improvement CAT system has brought to translation quality. In CAT systems, there are build-in dictionaries, terminology database, and most of all translation memories. All these instruments help greatly in translation of terms, abbreviations, fixed phrases, and also the consistency of translation in target text. So the error rate in this part is reasonably lower than that in traditional translation. However, CAT systems are still far from perfection. It even causes new troubles for its users. A very clear clue is that the "Symbols" which is mainly punctuation errors accounts for over 46% of all the errors. While "Symbols" in traditional translation takes even less than 6%. This is a sharp difference. And among all the punctuation errors in CAT list, three types stand out: missing of period, incompletion of quotation marks, misuse of English punctuations in Chinese text. We think that the cause for these phenomena is to some extent the sentence separation (Gao, 1996). When the translator finishes one sentence pair translation, he must slid to the next line to see the following content, which means he would often forget to complete the sentence with a correct and suitable punctuation for there is nothing just behind his translation in this line. As for the misuse of English punctuations, we can simply tell it is mis-operation of key board input. In order to solve these "little" errors, we think a smart algorithm should be applied in the exploration of Chinese pinyin input. For example, the input software should be able to automatically identified and even replace English punctuations in Chinese text. And to prevent incompletion of quotation marks, the input software could insert a complete quotation mark and set the mouse between the quotation marks even when we type just the left side of the quotation mark. Another same kind of situation is the input or choice error of Chinese characters. It accounts for nearly 30% of the errors in CAT error list. In fact, it is the second most frequent error type among all types in the CAT list. It is an apparent as well as a unique phenomenon in computer-aided translation. We believe the reason for this error lies in the "association and smart match of characters" function of current pinyin input software. As is known to all, current pinyin input software will always predict more phrases than we need. And it ranks the characters or phrases according to the frequency of utilization in web database. So the character comes out first is not always what we want. Especially when we are type some Chinese character with exactly the same pronunciation in pinyin, we have to pick out the right character carefully. With the fast development of translation industry, CAT is applied more and more widely. It is urgent that translators input the target text with a fairly high speed and get high efficiency of translation .But obviously high speed of input will no doubt challenge the quality of Chinese translation. In order to further improve the translation quality and efficiency, we have to cooperate with researchers in input software industry. A better and truly smarter input software can definitely help avoid careless input error meanwhile produce high-quality translation with great efficiency.

V. Conclusion

Computer-aided translation has achieved great progress in recent years especially in terminology database and translation memory. In this research, we contrastively analyse the specific error classification list of typical translation errors in CAT and put forward some preliminary solutions and ideas for improvement of CAT performance. We believe it would be of some contribution to the later research and development in certain aspects of CAT. By analyzing the features of various error types and getting a clearer clue of the high-frequency error trend, the translators or users of CAT system could be much more careful to avoid careless input errors. Only through improving the accuracy as well as efficiency of both computers and human translators, can we expect to reduce the cost of translation projects and raise the whole industry onto another tide. Of course, completely automatic identification of translation errors in CAT will remain a great challenge even in the short future.

Acknowledgment

This search is partially supported by National Social Science Fund of China (11CYY030).

REFERENCES

[1] Bowker, L. (2002).Computer-aided translation technology: A practical introduction. Ottawa: University of Ottawa Press, 157.

[2] Gao Fenjiang. (1996). Punctuation marks in English Chinese translation. *Chinese Translation*, 25-27(1).

[3] James, M & Anastasia, G. (2000). Error types in the computer-aided translation of tourism texts. *Database and Expert Systems Applications*, 138-142(2000).

[4] Luo Jimei&LiMei, (2012). CAT text error analysis, *Chinese Translation*,.84-89(5).

[5] Liang, S. Y. (2004). A comparative analysis of machine translation and computer-assisted translation .*Computer-assisted Foreign Language Education*, 42-43(100).

[6] Steven, B; Ewan, K & Edward, L. (2009). Natural language processing with python. Canada: O'Reilly Media, 251.

[7] Wang, H., A. (2011). Computer-aided translation technology and translation teaching .*Overseas English*, 158-161(12).

[8] Wang Jianjia, (2013). A corpus-based empirical study of quantitative evaluation of college English E-C translation quality. *Foreign Language Learning Theory and Practice*, 55-57(4).

Dongmei Zhou was born in Yongzhou city, Hunan province, in 1974, and graduated from School of Foreign Languages, Hunan Normal University, Changsha city, China in 1997, and got the Bachelor's Degree in language teaching. In 2004, she got the Master's Degree in applied linguistics in School of Foreign Languages, Central China Normal University, Wuhan city, China.

Form1997 to 1999, as a Teacher, she worked in Shanghu Vocational and Technic School. From 1999 to now, she has worked in School of Foreign Languages, Wuhan Textile University, Hubei, China, as a teacher. During these years, she has published some books and journals. In 2012, as a vice-editor, Fast Reading of College English (Book1-4) were published by Shanghai Jiao Tong University Press. Now her interests focus on language teaching and cross-cultural communication.

Ms Zhou hasn't participated in any professional committees.