A Corpus-based Study of Differences in Use of Structural Auxiliary Word 'de' in Native Chinese and Translated Chinese

Haoliang Chen
Zhejiang University
haelchan@zju.edu.cn

Abstract

This article explores the potential differences in use of structural auxiliary word *de* on the basis of two balanced comparable corpora of translated and native Chinese, namely the ZJU Corpus of Translational Chinese (ZCTC) and the Lancaster Corpus of Mandarin Chinese (LCMC). The results show that 1) in comparison with native Chinese, *de* is more frequently used in translated Chinese; 2) translated Chinese makes more frequent use of adjectives and pronouns before *de*; 3) the occurrences of some common words in native Chinese are even higher translated Chinese.

Key Words: comparable corpora, Chinese, structural auxiliary word

1 Introduction

The structural auxiliary word *de* (的) is the most frequently used word in Chinese. According to Modern Chinese dictionary, *de* has the following 6 usages: (1) used after the modifier; (2) used to construct *de*-structure without head-word; (3) used after the predicate verb to emphasize the agent, or the time, place, method; (4) used at the end of declarative sentence to show affirmation; (5) used between two same type of words or phrases to express 'and so forth'; (6) (oral) used between two quantifiers. As a consequence of these common usages, the role of *de* as a structural auxiliary word is of vital importance in Chinese.

In the translation of English to Chinese, the inappropriate use of *de* can lead to differences in various aspects. Some redundant uses of *de* weaken the readability of the text; some translational sentences may not agree with the Chinese expression habit. Since the effect of the source language on the translations is strong enough to make the translated language perceptibly different from the target native language. Consequently translational language is at best an unrepresentative special variant of the target language (McEnery & Xiao 2007). The distinctive features of translational language can be identified by comparing translations with comparable native texts, thus throwing new light on the translation process and helping to uncover translation norms, or what Frawley (1984) calls the "third code" of translation. In this study, we will reveal some features about *de*.

In addition, as Biber (1995: 278) observes, language may vary across genres even more markedly than across languages. Is the difference of the use of *de* varies across genres higher than those across languages? We will figure it out in our study.

This article first introduces the corpora we use in our research (Section 2). We will then cover the main part of our code when processing with these corpora (Section 3). Section 4 presents the result of our comparison study, and Section 5 concludes the article.

2 Corpus

The monolingual comparable corpus approach compares comparable corpora of translated language with the native target language in an attempt to uncover salient features of translations (Xiao 2010). To discover the difference between native Chinese and translated Chinese, we use two comparable corpora: the *Lancaster Corpus of Mandarin Chinese* (LCMC) for native Chinese, the *ZJU Corpus of Translational Chinese* (ZCTC) for translated Chinese.

2.1 LCMC

The *Lancaster Corpus of Mandarin Chinese* (LCMC) was designed as a Chinese match for the FLOB corpus of British English and the Frown corpus of American English. It is a one-million-word balanced corpus designed to represent native Mandarin Chinese (McEnery & Xiao 2004).

The LCMC includes a list of text categories: press reportage (A); press editorials (B); press reviews (C); religious writing (D); skills, trades and hobbies (E); popular lore (F); biographies and essays (G); miscellaneous (reports, official documents) (H); science (academic prose) (J); general fiction (K); mystery and detective fiction (L); science fiction (M); adventure and martial arts fiction (N); romantic fiction (P) and humor (R).

The LCMC corpus has also followed the sampling period of FLOB/Frown by sampling written Mandarin Chinese within three years around 1991. The LCMC model was slightly modified by extending the sampling period by a decade, i.e. to 2001, when the ZCTC corpus was built.

2.2 ZCTC

The ZJU Corpus of Translational Chinese (ZCTC) was created with the explicit aim of studying the features of translated Chinese in relation to non-translated native Chinese (Xiao 2010). It has modeled the LCMC. Both LCMC and ZCTC corpora have sampled five hundred 2,000-word text chunks from fifteen written-text categories published in China, with each corpus amounting to one million words. The two corpora are roughly comparable in terms of both overall size and proportions for different genres.

The corpus is annotated using ICTCLAS2008, the latest release of the *Chinese Lexical Analysis System* developed by the Institute of Computing Technology, the Chinese Academy of Sciences. Part-of-speech annotation is given in Extensible Markup Language (XML) format, with the POS attribute of the *w* element indicating its part-of-speech category.

3 Information Extracting

There are various corpus tools available to process LCMC and ZCTC. Some are free, like AntConc and #LancsBox; some are paid, such as WordSmith and PowerGREP. Most of them focus on the target word and provide common functions dealing with that word. Apart from those functions, in this study we need to pay more attention to the concordance of the target word *de*. To be more specific, the previous word and next word of *de*, their tags and frequencies correspondingly. However, those corpus tools fail to meet our demand. Therefore, to extract information we want, we just use Python to process our corpora directly, without relying on any corpus tools.

3.1 Extracting tags and tokens

The body part of both LCMC and ZCTC corpus files contains part-of-speech annotation in XML markup, with the POS attribute of the element - *w* used in ZCTC and *w* or *c* used in LCMC - indicating its part-of-speech category. With such format, it is easy for us to extract all the tokens using regular expression. To make it easier for later processing, we store the information of tags and tokens in two `list` (a built-in data structure in Python) separately, using the following two lines of code.

```
tag = re.findall(r'<[wc] POS="(\w{0,4})">.{,10}</[wc]>', read_data)
token = re.findall(r'<[wc] POS="\w{0,4}">(.{,10})</[wc]>', read_data)
```

Here we treat not only words but also punctuations as tokens. Since the word *de* could occur at the end of the sentence, by treating punctuations as tokens, we can determine that *de* occurs at the end of the sentence if its next tag is marked as *ew*, i.e., sentence-final punctuation.

3.2 Indexing the word *de*

With all word tokens extracted, it is necessary to index our target word *de*. Here we use some tricks to get all the indices we want.

```
indices = [i for i, x in enumerate(tokens) if x == '的']
```

Once we get the indices of *de*, it is easy to get the previous and next tokens along with tags of *de*, with two parallel list we created before. We can get four separate lists with the following four lines of code.

```
pre_tokens = [tokens[i-1] for i in indices]
next_tokens = [tokens[i+1] for i in indices]

pre_tags = [tags[i-1] for i in indices]
next_tags = [tags[i+1] for i in indices]
```

3.3 Frequency count

The frequency is one of the most important statistics that we need for analysis. Fortunately, Python's module `collections` contains `Counter`, which helps us to count the frequency easily. For example, the following code counts the number of tokens, and then print the 10 most common word types in the corpus.

```
import collections
counter1 = collections.Counter(tokens)
print(counter1.most_common(10))
```

4 Result

This section presents the differences in use of *de* in translated Chinese as represented in the ZCTC corpus in comparison with native Chinese represented in the LCMC corpus. We will first compare the differences of frequencies of *de* in an holistic view (Section 4.1). Then we compare the differences of previous and next tags of *de* separately (Section 4.2 and 4.3). Finally we focus on the features of most common words in previous words of *de* (Section 4.4).

4.1 Overview of the frequency of *de*

The structural auxiliary word *de* is the most frequently used word in Chinese. Is there any difference between native Chinese and translated Chinese? Does the category of genres have an impact on the frequency of *de*? In this section we will cover these questions.

We first examine the frequencies of *de* in native Chinese and translated Chinese. By counting the number of tokens and the number of target word *de* in each genre, we get the frequencies of *de* in different genres, as shown in Table 1 and Table 2.

Table 1. Frequencies of *de* in different genres in LCMC

Genre	Tokens	Occurrences of *de*	Frequency of *de* per 1M
A	88009	4108	46677
B	53587	2967	55368
C	34147	2288	67004
D	34083	2107	61820
E	76244	3596	47164
F	88093	4667	52978
G	154331	7138	46251
H	60466	2865	47382
J	160134	11149	69623
K	58115	2464	42399
L	48111	2134	44356
M	12043	694	57627
N	58107	1719	29583
P	58050	2739	47183
R	18182	505	27775
Total	1001702	51140	51053

Table 2. Frequencies of *de* in different genres in ZCTC

Genre	Tokens	Occurrences of *de*	Frequency of *de* per 1M
A	87984	5538	62943
B	54143	3757	69390
C	34061	2289	67203
D	35116	2632	74952
E	76569	4865	63537
F	89607	4516	50398
G	155433	10738	69084

H	60261	3745	62146
J	164445	11946	72644
K	60503	3116	51502
L	48904	2439	49873
M	12256	543	44305
N	58898	3119	52866
P	59000	2528	42847
R	19059	1003	52626
Total	1016339	62774	61765

Figure 1 shows the frequency of *de* in the fifteen genres covered in the LCMC and ZCTC corpora as well as their mean frequency. As can be seen, the mean frequency of *de* (61765) in ZCTC is considerably higher than that in LCMC (51053). It is also clear from the figure that in most genres the frequency of *de* is significantly higher, while in the genres of popular lore F, science fiction M and romantic fictions P, the use of *de* is lower in translated Chinese.

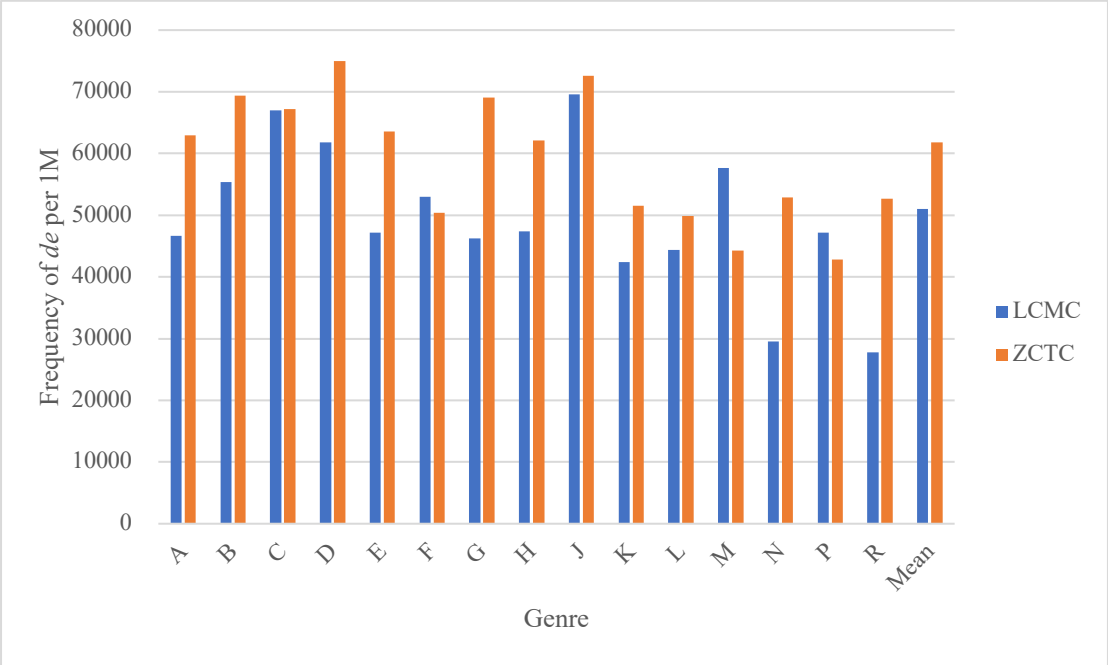

Figure 1. Frequency of *de* in LCMC and ZCTC

Table 3 gives us the result of the log-likelihood test for difference in each genre. While the LL score should be non-negative, here we use the negative sign to indicate the situation where the use of *de* is higher in native Chinese. As is shown in Table 3, while the frequency of *de* is significantly higher in translated Chinese in most genres, there are also genres in which the use of *de* is more common in native Chinese (namely popular lore F, science fiction M and romantic fiction P). The frequency of *de* is almost the same in the category of press reviews C.

Table 3. Log-likelihood tests for frequency of *de* in ZCTC and LCMC

Genre	LL score
A	213.18
B	85.06
C	0.01
D	43.66
E	185.69
F	(-)5.72
G	704.59
H	120.51
J	10.41
K	52.36
L	15.68
M	(-)21.22
N	392.27
P	(-)12.23
R	144.96

4.2 Differences in previous tags of *de*

As discussed in Section 1, *de* has 6 different usages. What is the most frequent usage in Chinese? What type of words are most frequently used before *de*? In this section, we will first examine the previous tags of *de.*

Table 4. Proportions of previous tags of *de* in academic prose J

Tag	Count	Proportion
n	4375	0.3924
v	1910	0.1713
a	1610	0.1444
vn	644	0.0578
f	530	0.0475
r	456	0.0409
b	292	0.0262
w	203	0.0182
l	171	0.0153
ng	117	0.0105
t	106	0.0095
i	85	0.0076
m	79	0.0071
k	75	0.0067
ns	71	0.0064
nr	62	0.0056
q	61	0.0055
nz	51	0.0046

u	45	0. 0040
an	33	0. 0030
s	33	0. 0030
z	32	0. 0029
d	30	0. 0027
vg	22	0. 0020
j	16	0. 0014
nx	16	0. 0014
ag	7	0. 0006
p	5	0. 0004
c	3	0. 0003
rg	3	0. 0003
tg	3	0. 0003
g	2	0. 0002
y	1	0. 0001

Table 4 shows a typical distribution of previous tag of *de*. While a variety of categories of tags could occur before *de*, the tags of *n* (noun), *v* (verb) and *a* (adjective), sometimes together with *r* (pronoun), make up a large proportion, in this academic prose J the ratio is 70.8%.

Here for the sake of simplicity, we ignore the variants of nouns (noun morpheme, etc.), verbs (verb morpheme, etc.) and adjectives (adjective morpheme, etc.) and focus on their simplest forms only.

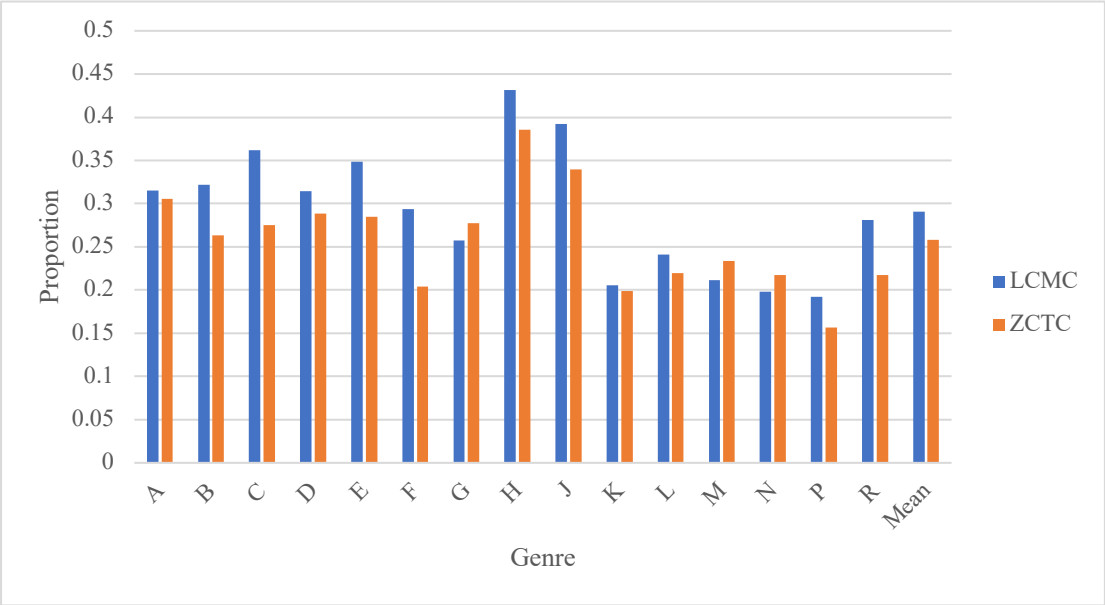

Figure 2. Proportion of *n* in previous tag of *de*

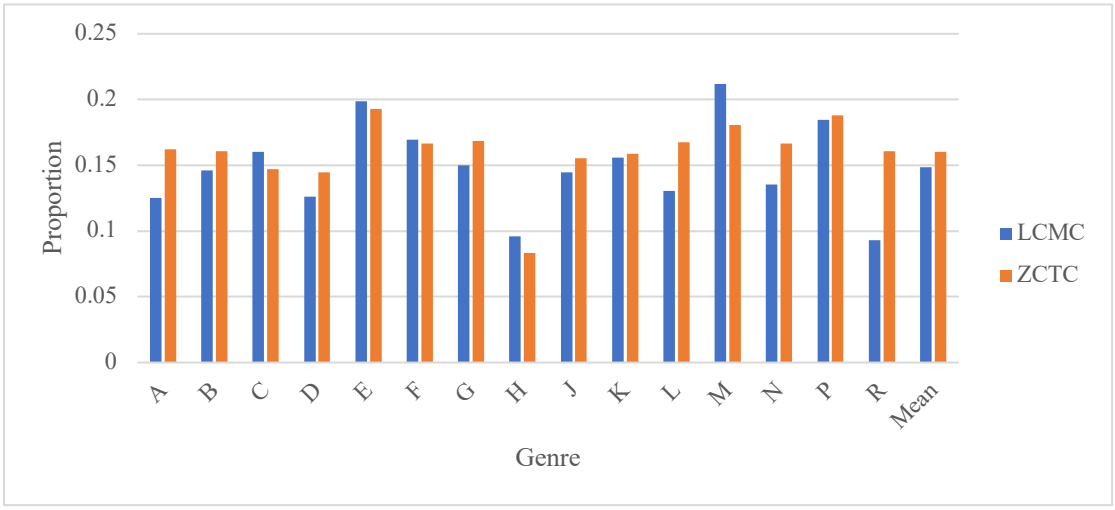

Figure 3. Proportion of *a* in previous tags of *de*

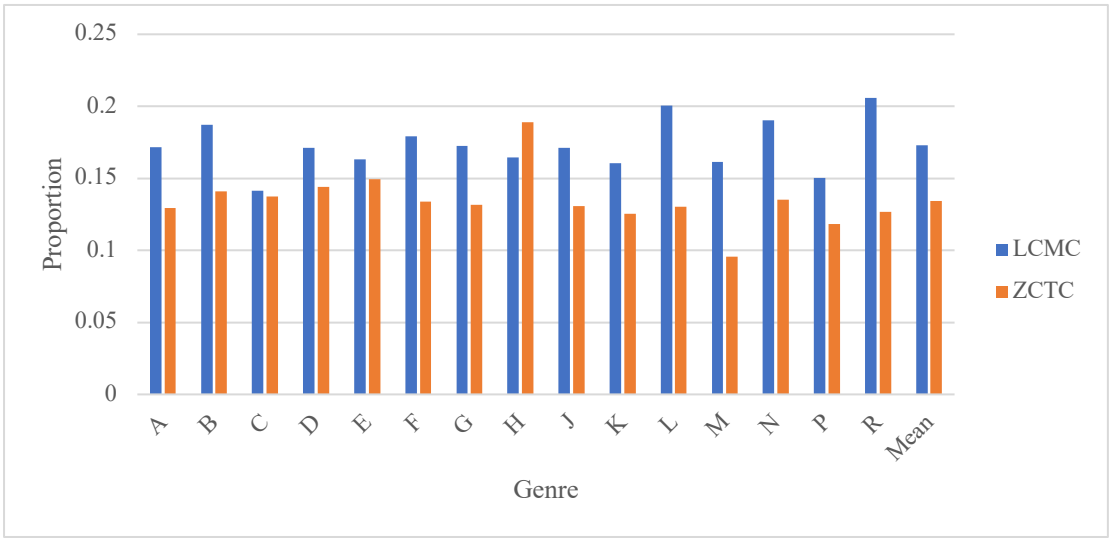

Figure 4. Proportion of *v* in previous tags of *de*

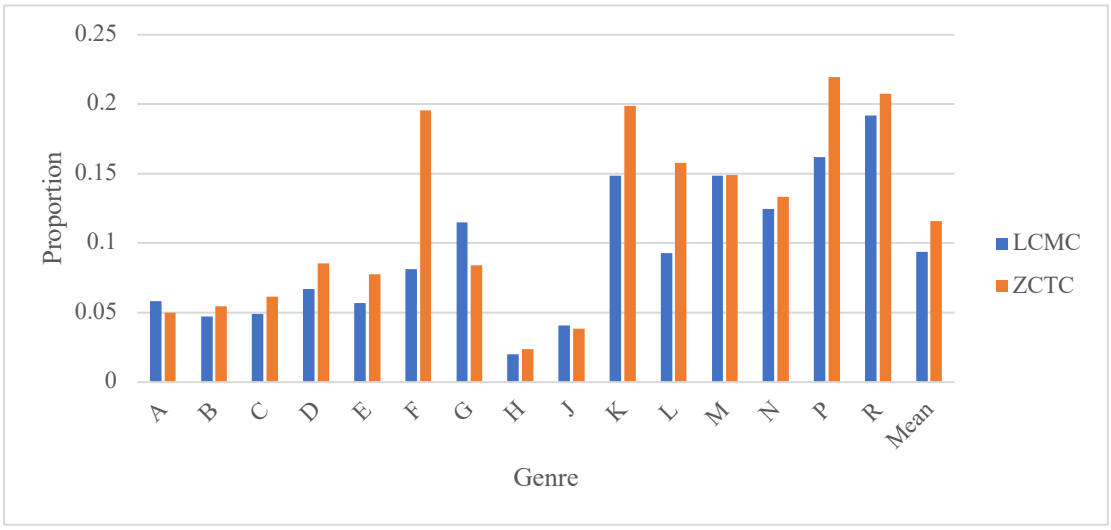

Figure 5. Proportion of *r* in previous tags of *de*

While the use of nouns and verbs before *de* tends to be higher in native Chinese (shown in Figure 2 and Figure 4), the use of adjectives and pronouns is more frequent in translated Chinese (shown in Figure 3 and Figure 5).

4.3 Differences in next tags of *de*

Unlike the diversity in previous tags of *de*, the word after *de* is less diversified. The probability that noun comes after *de* is higher than 50%. As shown in Figure 6, the proportion in translated Chinese is slightly higher than that in native Chinese.

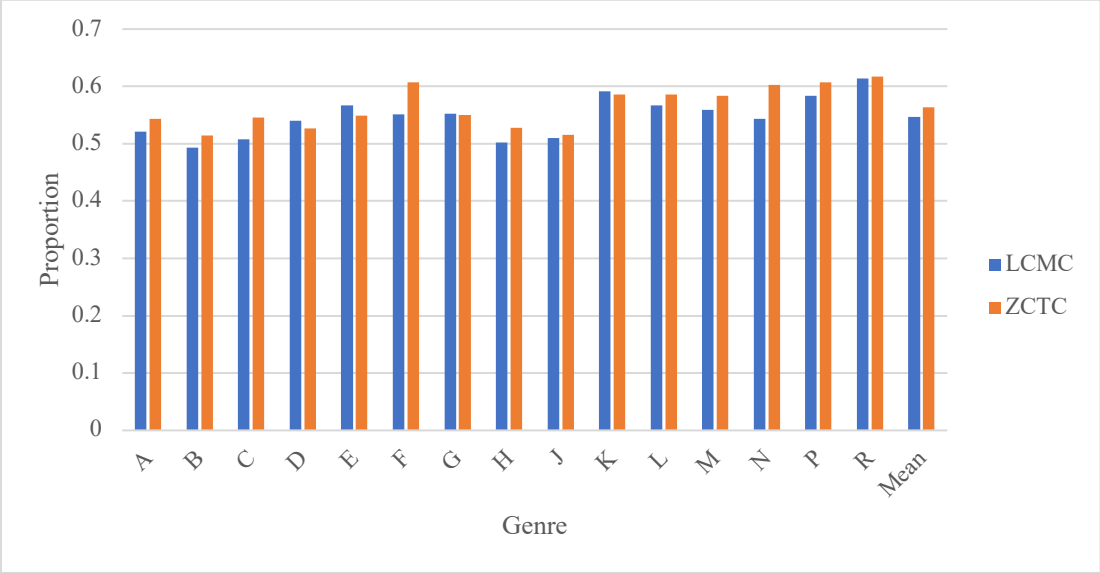

Figure 6. Proportion of *n* in next tags of *de*

4.4 Differences in previous words of *de*

In this section, we will focus on the previous words of *de*. Though the tag such as *p, r* do not make up a large proportion in the previous tags of *de*, their limited words will stand out when we turn to words rather than tags. Table 5 shows us the 20 most frequently used word before *de*.

Table 5. 20 most frequently used words before *de*

	LCMC		**ZCTC**	
Rank	**Token**	**Number of Time**	**Token**	**Number of Time**
1	他	714	他	1549
2	自己	571	我	889
3	中	501	她	743
4	我	495	自己	650
5	上	480	中	611
6	人	429	上	556
7	新	322	你	504
8	她	308	他们	488
9	大	291	大	444
10	你	238	新	442

11	我们	215	公司	437
12	他们	210	人	392
13	这样	208	重要	356
14	发展	202	多	338
15	它	200	这样	338
16	重要	195	它	334
17	工作	190	我们	314
18	不同	179	好	299
19	来	178	美国	231
20	好	173	不同	223

While the ranking may be a bit different, the most frequently used words are almost the same. 17 most frequently used words in LCMC also appear in ZCTC, including: 他 "he", 自己 "self", 中 "in", 我 "I", 上 "on", 人 "people", 新 "new", 她 "she", 大 "great", 你 "you", 我们 "we", 他们 "they", 这样 "such", 它 "it", 重要 "important", 不同 "different", 好 "good". The exclusive words - 发展 "develop", 工作 "work", 来 "to" in LCMC, 公司 "company", 多 "more", 美国 "America" – are mainly result from the corpus sampling.

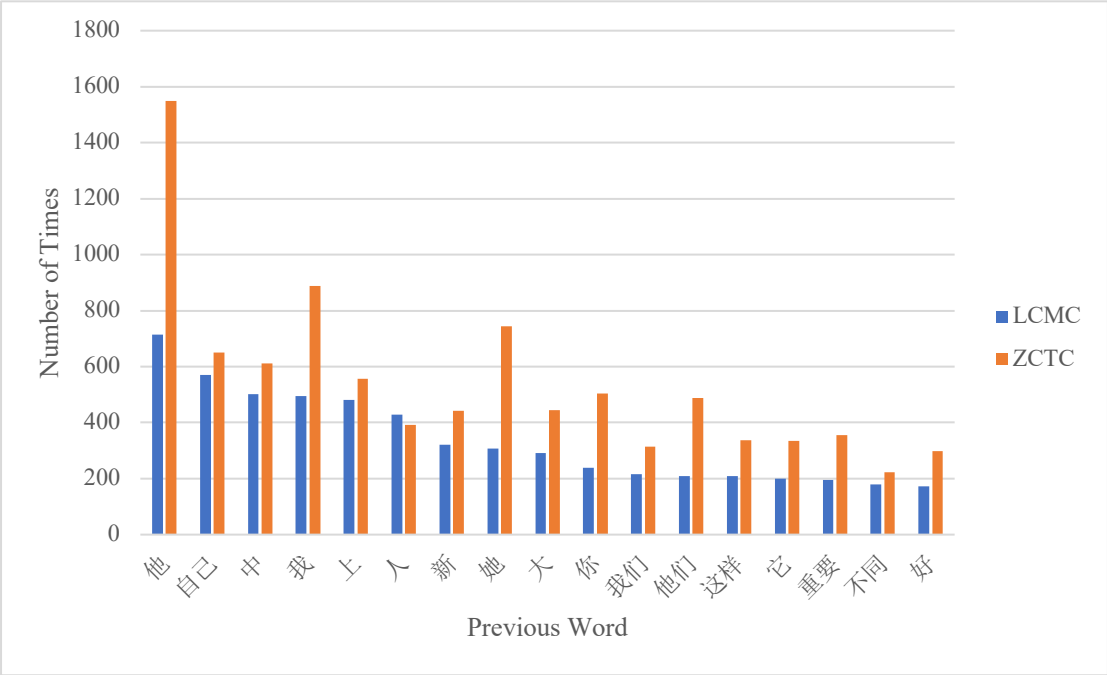

Figure 6. 20 most frequently used words before *de*

Figure 6 shows that the use of these common words is much higher in translated Chinese. Further examination reveals that the use of singular pronouns (他, 我, 她, 你, 它) in translated Chinese is about twice of that in native Chinese, and the use of positive adjectives (新, 大, 重要, 好) is about 1.5 times, as indicated in Table 6 and Table 7. Why these common words appear more often in translation? To some extent, the use of pronouns are essentially higher in translated Chinese (Wang & Hu, 2010). Besides, translators may tend to use simple structures and ignore the looming of *de* (Xu, 2011), which is a feature of native Chinese, in in the process of translation.

Table 6. Number of singular pronouns before *de*

Word	LCMC	ZCTC	Ratio
他	714	1549	217%
我	495	889	180%
她	308	743	241%
你	238	504	212%
它	200	334	167%

Table 7. Number of positive adjectives before *de*

Word	LCMC	ZCTC	Ratio
新	322	442	137%
大	291	444	153%
重要	195	356	183%
好	173	299	173%

5 Conclusion

This article applies Python programming language when processing with LCMC and ZCTC. Our study has shown that 1) in comparison with native Chinese, *de* is more frequently used in translated Chinese; 2) translated Chinese makes more frequent use of adjectives and pronouns before *de*; 3) the occurrences of some common words in native Chinese are even higher translated Chinese. The result can be meaningful for translators and helps to improve the readability of translational texts.

For lack of diachronic comparable corpus of native Chinese and translated Chinese, our study is confined to the Chinese used around 1990. The use of *de* can vary both in native Chinese and translated Chinese along with time. The improvements should be made once diachronic comparable corpus is available.

References

[1] McEnery, T. & Xiao, R. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In M. Lino, M. Xavier, F. Ferreire, R. Costa & R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004.* Lisbon: Centro Cultural de Belem, 1175-1178.

[2] Richard Xiao. How different is translated Chinese from native Chinese? *International Journal of Corpus Linguistics* 15:1 (2010), 5-35.

[3] Kefei Wang & Xianyao Hu. 2010. The explicitation and deviation of personal pronouns in Chinese literary translation. *Foreign Languages in China,* 2010, 7(04): 16-21

[4] Xiaohong Ma. 2014. Study on *de* in modifier-head construction when translating from English to Chinese. *Dissertation, Nanchang University*.

[5] Zhonghua Xiao, Guangrong Dai. 2010. In pursuit of the “third code”: A study of translation universals based on the ZCTC corpus of translational Chinese. *Foreign Language Teaching and Research,* 2010, 42(01): 52-58+81.

[6] Yangchun Xu. 2011. Chunking, prominence and usage of “de”. *Language Teaching and Linguistic Studies,* 2011(06): 76-82.

[7] Biber, D. 1995. Dimensions of Register Variation: A cross-linguistic comparison. *Cambridge: Cambridge University Press.*