# Universals of Translation: A Corpus-based Investigation of Chinese Translated Fiction

Yu YUAN

School of Languages and Cultures

Nanjing University of Information Science and Technology, Nanjing.

Fei GAO

College of Foreign Languages

Southwest Jiaotong University, Chengdu.

Abstract:

In the present study, all three of the above previously-studied recurrent features of translation are hypothesized and investigated, together with a fourth (leveling-out) will therefore be thoroughly explored in comparable corpora of Chinese translated fiction. We are motivated and committed to conducting the present study to make a contribution to the field of corpus linguistics, by gathering corpora of non-English texts, and by using self-built corpora to investigate all the four recurrent features of translation proposed by Mona Baker.

Keywords: Corpora; Normalization; Explicitation; Simplification; Leveling-out

## 1. Introduction

Translation studies has been provided with a number of relatively new theoretical questions, most notably the set of "universal features of translation" put forward by Baker (1993; see also Toury 1995 ). The discipline of Translation Studies (TS) has in the past decade seen a surge of interest in translation universals, a topic suited to the potentially large scale computerized corpora. According to the theory, translated texts are distinguishable from non-translated texts by certain recurrent features, which have been tested in recent contributions to Corpus-based Translation

Studies: several studies have already used corpus-based approaches to address various aspects of that particular theoretical problem (see Laviosa-Braithwaite 1996, 1997; Laviosa 1998; Øverås 1998; Baker 2000). As more can be seen in the special issue of META in 1998, it includes a collection of corpus-based translation studies attempting to outline the existing territory occupied by a new field of research in translation studies and show that the corpus-based approach is evolving, through theoretical elaboration and empirical realization, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation ( See Laviosa 1998). Meanwhile, Chinese scholastic explorations of the corpus-based approach started only from the early 21st century, which in nature are summaries and brief introductions to foreign corpus-based translation studies (See Liao 2000; Ding 2001; Zhang 2002; Ke 2002). A real empirical practice of corpus-based TS was not presented to Chinese academia until the year of 2004 on account of a lack of applicable corpora (See Qian 2004).

So far, major corpus-based studies have recently investigated three specific hypothetical recurrent features of translation (normalization, explicitation, and simplification). However, each of these research projects has touched upon only one recurrent feature of translation at a time, and using English and other European languages as the sole target language of the translated texts, few and incomplete investigation of all four features in Chinese translated texts have been done. In the present study, all three of the above previously-studied recurrent features of translation are hypothesized and investigated, along with a fourth (leveling-out), which has not been the subject of previous studies. Leveling-out will therefore be thoroughly explored in the present study. We are motivated and committed to conducting the present study to make a contribution to the field of corpus linguistics, by gathering corpora of Chinese fiction, and by using self-built corpora of translated fiction to investigate all the four recurrent features of translation proposed by Mona Baker.

## 2. Hypotheses

According to Baker (See also Olohan 2004: 91-100), there are four universal features of translation, namely simplification (the idea that translators subconsciously simplify the language or message or both), explicitation (an overall tendency to spell things out rather than leave them implicit in translation to make implicit information more explicit), normalization or conservatism (the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them) and leveling out (a hypothesis that translated language and translated texts 'steer a middle course between any two extremes, converging towards the centre', meaning that we may encounter less variance in textual features in a corpus of translations than in a corpus of non-translations) .

To be in line with Mona Baker and many other scholars who share the same ground, we agree that translation is a distinct linguistic behavior and thus translated texts inevitably have distinct features, which can be observed and measured as they are consistently recurring in the surface structures of translated texts. Following Shlesinger (1989), Baker (1996), Laviosa (1996), Kenny (1999; 2001), Olohan and Baker (2000), and Olohan(2001; 2004), we hold that normalization, explicitation, and simplification are supposed to be more prominent in (but may not necessarily exclusive to) translated texts and leveling-out, nevertheless, is deemed to be a feature exclusive to translated texts.

Built upon all the foregoing researches, our specific hypotheses are that if normalization is a recurrent feature of translation, fewer instances of unattested or "abnormal" usage, less foreignness, and lower frequencies of function words will occur in translated texts since translators tend to stick more closely to the prevailing norms governing written texts in target language, thus fewer instances referred to as "coinages"; if explification a recurrent feature higher use of more spelled-out syntax and optional structures like "*huan ju hua shuo*", "*ji*", and other means of

annotations so on, and of course longer sentence length would be distinctive; if simplification a recurrent feature lower type-token ratio, lower proportions of content words to running words and shorter sentence length can be detected among translated text; that if leveling-out recurrent a demonstration of similarity or closeness between translated texts in each translated corpus will be detected in comparison with non-translated corpora, say, these translated corpora display a homogeneous affinity.

Now all these hypotheses await tests in highly robust and representative corpora, and then new problems arise too as how to build a corpus of translated fiction and to what extent it can be used in analysis and comparison of translational features.

## 3. Methodology and Corpora Compilation

As defined by McEnery and Wilson (1996:21-24), a corpus is more viewed as a sample of authentic texts gathered in electronic format and used as a qualitatively representative reference for linguistic research. What are the basic elements in consideration when we build a translation corpus? In order to achieve the representativeness, balance and size (Kennedy 1998:60-70) of a corpus its builder must take into account the purpose of the corpus, its representativeness and balance, its size, selection of data, and many other elements including but may not be limited to text capture and markup, etc..

### 3.1 The Overall Principles Governing the Construction of CCTF.

Corpus design criteria depend on the envisaged use of a corpus in a given study (Williams 2005:67). We are supposed to differentiate translated and non-translated fiction by looking for features that may be considered distinctive. Corpora of Chinese Translated Fiction (hereinafter as CCTF) are designed to introduce the corpus-based research methodology into translation studies to enable a descriptive and empirical study of universals of translation. By comparison of lexical

and syntactic features of Chinese translated fiction and non-translated fiction we can testify Baker's hypotheses in Chinese translations. In this sense, CCTF can also be labeled as a corpus of special purpose.

CCTF is intended to represent the Chinese translated fiction as a whole, therefore translated fiction from other languages like English, French, Russian and Germany are collected to make a body of computerized texts exceeding one million words. What's more important, all the written texts have to be confirmed as authoritative translations of the original. Thus, All the translated novels are carefully chosen on the criterion that the work must be representative to both the original author and translator, published by leading press or publishing house, and stored in separate files. It is worth mentioning that the selection of a novel is purely random sampling, that is to say, we don't adopt a whole novel but pick up randomly certain chapters into our corpora.

### 3.2 Comparable Corpus

The Lancaster Corpus of Mandarin Chinese (LCMC) is designed as a Chinese match for the FLOB and FROWN corpora for modern British and American English. The corpus is suitable for use in both monolingual research into modern Mandarin Chinese and cross-linguistic contrast of Chinese and British/American English. The corpus sampled 15 written text categories including news, literary texts, academic prose and official documents etc published in P.R.China in the early 1990s (McEnery, A. & Z. Xiao. 2004). Thus, LCMC(especially categories from K to P, hereinafter LCMC(K-P)) can serve as a best comparable corpus in our study.

### 3.3 Compilation of CCTF and Corpus Tools

CCTF takes the same modules and structures of LCMC as they are designed to be comparable. We also established 5 individual corpus of general Fiction, mystery and detective fiction, science fiction, gangsters fiction(counterpart to LCMC's Martial Art fiction), and romantic fiction, which

are sampled between 1990s and 2000s from around 60 translated fiction available online. The LCMC corpus is marked up in XML format at five levels: text category, Sample file, paragraph, sentence and token, in addition to an informative corpus header. The data is tokenized and POS tagged, with an accuracy rate of ca. 98% (Xiao, 2005). Unlike its correspondent LCMC, CCTF was, due to various present limitations, only roughly grammatically tagged and segmented at sentence level to barely satisfy our present research purpose. The author did not mark up the collected texts to such a deeper degree as is in LCMC but provided some basic extralinguistic information of the texts and had them POS-tagged using ICTCLAS 1.0 (a free version of the software). It can not be denied that the accuracy of POS tagging needs to be improved largely, though. An excerpt of our corpus is as follows:

<s id=116> 结果/n 手/n 的/u 麻痹/vn 就/d 和/c 蛇/n 的/u 幻觉/n 联系/v 起来/v 了/y )/w ./w </s>

<s id=117> 等/u 蛇/n 不见/v 之后/f ,/w 她/r 惊魂未定/i 地/u 想/v 要/v 祈祷/vn ,/w 却/d 又/d 在/p 语言/n 上/f 遇到/v 了/u 麻烦/an -/w 她/r 找/v 不/d 到/v 自己/r 能/v 讲/v 的/u 语言/n 了/y ,/w 直到/v 最后/f 她/r 忽然/d 想到/v 几句/q 英语/nz 的/u 童谣/n ,/w 于是/c 她/r 发现/v 自己/r 只能/v 用/p 这/r 门/q 语言/n 思考/vn 和/c 祷告/v 了/y ./w </s>

Our corpora consist of Chinese translations from English fiction, collected mainly from world wide webs and published e-books on CD-ROM; they constitute a broad sample of parallel but comparable texts. Specific techniques of analysis are adapted from the literature, and where appropriate, new techniques are devised. Wordsmith (versions 5) and the free linguistic tool ACWT (An integrated linguistic tool by Hongyin Tao ) and Antconc (version 3.2.2w) will be our primary tools used for corpus analysis. We hope to testify Baker's hypothesis by our empirical evidence gathered in the present research: whether these four features universally exist in Chinese translated fiction or not, and if they do, what their patterns are.

To summarize, we have designed an extract, synchronic, mixed-terminological written corpus with translations that have been published by some major publishers and presses and

produced by some experienced translators providing some guarantee of quality.

## 4. Discoveries and Discussion

Once the corpora have been compiled as described in the previous section, we are ready to launch our qualitative analysis. First we used Wordsmith 5 to do the basic statistics of CCTF and LCMC(K-P), finding out that due to a strategy of retaining balance and representativeness of the corpora CCTF is relatively larger in size than LCMC(K-P). As you may notice from the following two graphs, the overall number of tokens in CCTF is almost twice as that of LCMC(K-P), which seems to some extent to question its legality as a comparable corpus, and a standardized comparison is thus required in the study whereafter:

| N | Overall | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| text file | Overall | cctf_k.txt | cctf_l.txt | cctf_m.txt | cctf_n.txt | cctf_p.txt |
| file size | 10,489,196 | 1,830,036 | 1,722,115 | 1,718,900 | 3,158,610 | 2,059,535 |
| tokens (running words) in text | 2,180,121 | 380,158 | 357,559 | 353,793 | 654,553 | 434,058 |
| tokens used for word list | 2,180,121 | 380,158 | 357,559 | 353,793 | 654,553 | 434,058 |
| types (distinct words) | 37,784 | 17,365 | 15,909 | 14,710 | 18,429 | 15,402 |
| type/token ratio (TTR) | 1.73 | 4.57 | 4.45 | 4.16 | 2.82 | 3.55 |
| standardised TTR | 28.18 | 28.97 | 28.47 | 27.69 | 27.99 | 27.92 |
| standardised TTR std.dev. | 71.33 | 70.05 | 71.08 | 71.20 | 72.81 | 71.70 |
| standardised TTR basis | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| mean word length (in characters) | 1.37 | 1.37 | 1.35 | 1.37 | 1.38 | 1.34 |
| word length std.dev. | 0.71 | 0.69 | 0.69 | 0.70 | 0.75 | 0.69 |
| sentences | 59,057 | 9,989 | 9,030 | 9,972 | 17,867 | 12,199 |
| mean (in words) | 36.91 | 38.05 | 39.60 | 35.48 | 36.61 | 35.58 |
| std.dev. | 34.42 | 38.64 | 45.86 | 37.37 | 25.24 | 29.56 |
| paragraphs | 5 | 1 | 1 | 1 | 1 | 1 |
| mean (in words) | 436,024.19 | 380,158.00 | 357,559.00 | 353,793.00 | 654,553.00 | 434,058.00 |
| std.dev. | 126,291.82 | | | | | |

Graph1. Basic Information of CCTF

| N | Overall | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| text file | Overall | lcmc_k.xml | lcmc_l.xml | lcmc_m.xml | lcmc_n.xml | lcmc_p.xml |
| file size | 5,051,148 | 1,253,100 | 1,042,611 | 260,822 | 1,243,164 | 1,251,451 |
| tokens (running words) in text | 218,120 | 55,108 | 44,883 | 11,294 | 52,735 | 54,100 |
| tokens used for word list | 218,120 | 55,108 | 44,883 | 11,294 | 52,735 | 54,100 |
| types (distinct words) | 19,986 | 8,666 | 7,708 | 2,864 | 8,427 | 8,348 |
| type/token ratio (TTR) | 9.16 | 15.73 | 17.17 | 25.36 | 15.98 | 15.43 |
| standardised TTR | 45.25 | 44.27 | 46.43 | 44.61 | 46.01 | 44.69 |
| standardised TTR std.dev. | 54.60 | 54.60 | 50.93 | 50.89 | 53.05 | 53.89 |
| standardised TTR basis | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| mean word length (in characters) | 1.59 | 1.57 | 1.63 | 1.67 | 1.54 | 1.59 |
| word length std.dev. | 0.86 | 0.87 | 0.86 | 0.91 | 0.84 | 0.87 |
| sentences | 12,119 | 3,287 | 2,434 | 528 | 2,738 | 3,132 |
| mean (in words) | 17.96 | 16.74 | 18.40 | 21.21 | 19.22 | 17.24 |
| std.dev. | 11.92 | 11.06 | 11.99 | 12.37 | 13.16 | 11.29 |
| paragraphs | 5,258 | 1,218 | 1,053 | 251 | 1,551 | 1,185 |
| mean (in words) | 41.39 | 45.17 | 42.53 | 44.61 | 33.94 | 45.57 |
| std.dev. | 49.63 | 55.49 | 50.87 | 30.19 | 35.28 | 59.52 |
| headings | 0 | 0 | 0 | 0 | 0 | 0 |
| mean (in words) | | | | | | |
| std.dev. | | | | | | |
| sections | 5 | 1 | 1 | 1 | 1 | 1 |
| mean (in words) | 43,624.00 | 55,108.00 | 44,883.00 | 11,294.00 | 52,735.00 | 54,100.00 |

Graph 2. Basic Information of LCMC(K-P)

However, we hold that LCMC(K-P) is still the best choice for the time being if there is no other better alternative to take its place, and we can minimize this scientific faults by taking these elements such as the smaller size of LCMC(K-P) and its inconsistent size of sub-corpora into consideration when a quantitative conclusion is drawn. Here we also want to point out that CCTF was not tagged at a paragraph level, so careful readers my notice the number of paragraphs of CCTF is almost equal to its number of sections of LCMC(K-P), which is rather unbelievable intuitionally, and largely due to the computer's inability to distinguish them without knowledge of boundaries of sentences and sections provided by man.

As mentioned above, our primary objective of the present study is to compare Chinese translated fiction with non-translated fiction, identifying features that may be considered

qualitatively distinctive to translated texts. As is discussed earlier, three hypothesized "universals of translation", namely normalization, simplification, and explicitation have been investigated in the foregoing studies carried out by our forerunners. To keep up with the methodology and goals of the present research, we borrowed the methods applied in the previous investigations in order to make our research close and comparable to the previous ones. In what follows in the passage, our research and findings are described, and our interpretation of the results elaborated, for the four individual universals.

**4.1 Investigation of Normalization**

As proposed in the hypotheses section, normalization is the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them. We deem that any texts demonstrating conservativeness embody the feature of normalization. To learn if a text carries such a feature, we need to manifest whether fewer instances of unattested or "abnormal" usage, less foreignness, and lower frequencies of function words occur in translated texts. In other words, we need see if translated texts are lexically normalized.

Laviosa (1998:8) advanced and testified four patterns of lexical use in comparable corpus of English narrative prose: The translational component of the comparable corpus of narrative texts has a lower lexical density and mean sentence length than the non-translated corpora; the translational component of the comparable corpus of narrative texts contains a higher proportion of high frequency words and its list head covers a greater percentage of text with fewer lemmas than the non-translational component. Do we have the same findings?

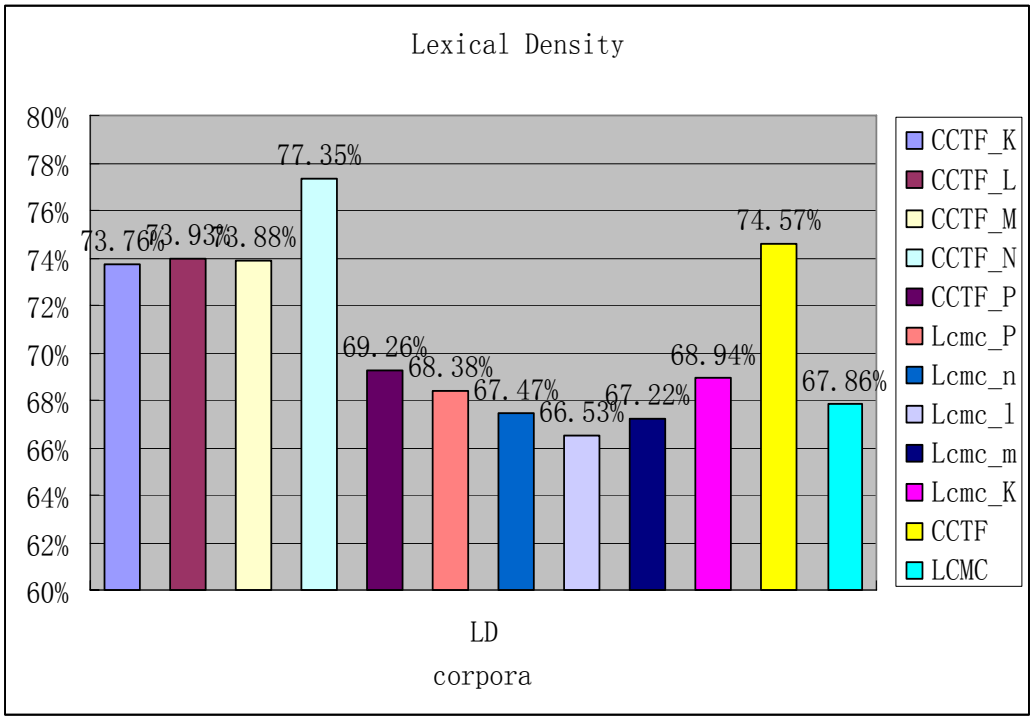
***4.1.1 Lexical Density***

There are at least two different ways to measure Lexical Density (hereinafter LD). According to UsingEnglish.Com, Lexical Density is calculated in the formula of "LD=(Number of different words / Total number of words) x 100".UE.COM claims that, as a guide, any lexically dense text

has a lexical density of around 60-70% and those which are not dense have a lower lexical density measuring around 40-50%. J. Ure (1971) and Michael Stubbs (1986), however, propose the following formula for LD: (Content Word Forms /number of Running Words) x 100. We took the second way to calculate LD, in which content words refer to nouns, verbs, adjectives, adverbials, pronouns, quantifiers, and numerals as well, opposite to function words which functions grammatically and possess no fixed meanings like prepositions, connectives, articles, auxiliaries, etc.. We use the free concordance program Antconc to count all the content words and calculate them in the total number of words in both CCTF and LCMC(K-P). Contrary to our presupposition is that neither the separate LD of individual translation corpus nor the LD of overall translation corpora is lower than that of the corpora's in LCMC and LCMC's, perhaps this is mainly attributed to the fact that most translators are experienced and skilled and they produced translations as though they were writing in Chinese, and, that is to say, the lexical usage of translated texts in CCTF is in a tendency of being normalized. To some extend, this tendency is more or less overemphasized that this exaggeration resulted in an average high performance in pursuit of lexical variety, as can be seen from the graph that follows. The average lexical density of CCTF is almost 7% higher than that of its comparable corpora LCMC(K-P). Our finding in regard of lexical density thus doesn't support Laviosa's but validate our hypothesis that translations tend to be normalized as and even conscientiously more natural than non-translated texts in order to achieve higher popularity and acceptance among readers.

Meanwhile, with such high content words to running words ratios, this finding further explains why translated texts have a relatively lower frequency of function words, which will enable the texts to be more parataxis but hypotaxis (in the sense translations follows strictly to the original by means of connectives and any other grammatical function words) (see also Hu, 2006:118), and of course makes translations not a bit foreign. In view of two language systems,

Chinese is more a parataxis language than a hypostasis language in the sense it depends less on grammatical function words like connectives, prepositions and other types of empty words to convey the meaning, which, nevertheless, is contained in the larger context of words and clauses that entail an implication of grammatical meaning and logical relationship.




Graph 3 Lexical Density of LCMC(K-P) &CCTF (K-P)

As a result of it, low frequency of grammatical function words (empty words) and high frequency of content words is a symbol of natural non-translated Chinese fiction. From this point of view, we can safely draw the conclusion that CCTF shares a feature of being target language oriented, or normalization.

### *4.1.2 Lemma Words and Frequency*

In fact, the term “lemma” affects no Chinese since every Chinese word at the same time is its lemma word. But lemma words in a corpus do reflect the overall trend of the word choices as pointed out by Laviosa. Here again, we will review and compare the lemma words list of CCTF

and LCMC(K-P) and LCMC to see if there is anything in common or significant enough for our attention. First, we used Antconc (Version 3.2.2w) and wordsmith tools 5 to make two separate lists of lemma words and calculate out their normalized frequencies in the respective corpora. We found that lemma words in the wordlists of LCMC(K-P) and CCTF vary little within a range of the top 270 words in the list as is shown in table 1 below of the top 30 words in two wordlists，but one point deserves everyone's attention is that their normalized frequencies (item's occurrence in a corpus per 1000 words, here counted in the formula "normalized frequency =item frequency*1000/number of running words in a corpus) in CCTF are much lower than them in LCMC. Although the corpora sizes are different, normalized frequency happen to suit the needs of a scientific measurement of words frequencies in different corpora. From table 1, we can clearly notice that those high frequency words in LCMC(K-P) non-translated fiction are used also the most frequently but relatively lower in CCTF, which to some extent reveals the truth that translations tend to use "normal" language as non-translations, but sometimes this tendency is often simplified since we can find out that the normalized frequencies of those frequently-used word are commparatively lower in CCTF.

| N | Word | Nor.Freq in LCMC | Word | Nor. Freq in CCTF |
|---|---|---|---|---|
| 1 | 的 | 44.7 | 的 | 26.4 |
| 2 | 了 | 21.4 | 我 | 10.4 |
| 3 | 是 | 13.0 | 他 | 9.7 |
| 4 | 一 | 12.7 | 了 | 9.3 |
| 5 | 我 | 12.1 | 是 | 6.8 |
| 6 | 他 | 11.8 | 在 | 6.5 |
| 7 | 在 | 10.6 | 你 | 5.1 |
| 8 | 不 | 8.4 | 她 | 5.0 |
| 9 | 她 | 7.9 | 不 | 4.4 |
| 10 | 你 | 7.9 | 说 | 3.4 |
| 11 | 着 | 7.4 | 着 | 3.3 |
| 12 | 说 | 7.3 | 这 | 3.0 |
| 13 | 这 | 6.3 | 和 | 2.5 |
| 14 | 人 | 5.9 | 有 | 2.4 |

| | | | | |
|---|---|---|---|---|
| 15 | 地 | 5.8 | 就 | 2.4 |
| 16 | 有 | 5.6 | 人 | 2.4 |
| 17 | 也 | 5.3 | 地 | 2.3 |
| 18 | 就 | 5.3 | 上 | 2.2 |
| 19 | 上 | 4.6 | 也 | 2.2 |
| 20 | 那 | 4.2 | 他们 | 2.1 |
| 21 | 到 | 3.8 | 我们 | 2.1 |
| 22 | 又 | 3.7 | 到 | 2.0 |
| 23 | 一个 | 3.7 | 会 | 1.9 |
| 24 | 和 | 3.5 | 要 | 1.8 |
| 25 | 来 | 3.4 | 都 | 1.7 |
| 26 | 个 | 3.4 | 那 | 1.7 |
| 27 | 得 | 3.3 | 对 | 1.7 |
| 28 | 去 | 3.2 | 把 | 1.7 |
| 29 | 都 | 3.2 | 里 | 1.6 |
| 30 | 把 | 2.7 | 来 | 1.5 |

Table 1 The Top 30 Most Frequently-used Words in CCTF and LCMC

**4.1.3 Attested Use of Words**

Unlike English, Chinese doesn't have compounding words that can illustrate the writers' or translators' creativity; On the other hand, CCTF is only roughly tagged that we could not search and observe those creative usages of words in the translations. However, we can compare the normalized frequency of idioms, as we all know, which to some extent can best represent the idiomatic degree of the language. Higher frequency of idioms can be viewed as a consequent of fewer instances of unattested usages.

By virtue of Antconc, we listed out all the idioms in Both CCTF and LCMC(K-P) and fathomed out the respective normalized frequency in the two corpora. We found the normalized frequency of idioms in CCTF is around 4.96 per 1000 words and in LCMC(K-P) is 6.80 per 1000 words. Though idioms in CCTF are less frequent than them in LCMC (K-P), we can still safely infer that the language in translation corpora CCTF makes for employing as many idiomatic expressions as possible to make translations closer to the target language readers' expectations and

gain more popularity.

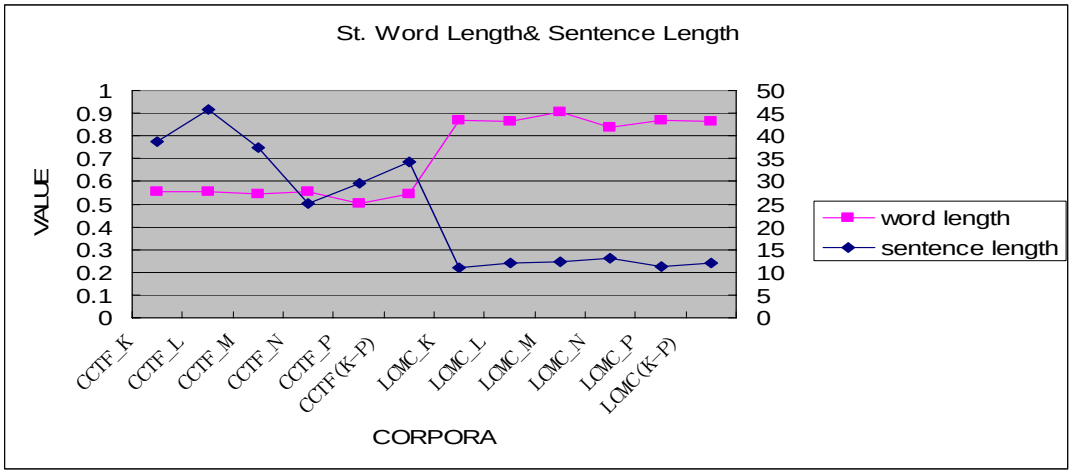
From above analysis, we can detect a kind of conservativeness of translated texts, i.e. a kind of fidelity to the target language in our translation corpora CCTF. We call this quality of translations normalization.

### 4.2 Investigation of Simplification

Put forward beforehand, simplification of translation is judged by the shorter type-token ratio, lower proportions of content words to running words, and shorter word length and sentence length, of which lower proportions of content words to running words does not seem to hold water since in section 4.1.1 we have proved that the lexical density of CCTF (note that we took the J.Ure way and included adverbials and idioms as content words) is much higher than that of LCMC(K-P). But, as far as type-token ratio and sentence length are concerned, the two aspects deserve our digging up.

#### 4.2.1 Standardized word and Sentence Length

From Graph1 and Graph 2 we can draw a graph of Sentence lengths in characters of each corpus (note here we took the standardized deviation of word length and sentence length to minimize possible deviating influences caused by different sizes of corpora):




Graph 4 Standardized Sentence Length of Corpora in CCTF and LCMC(K-P)

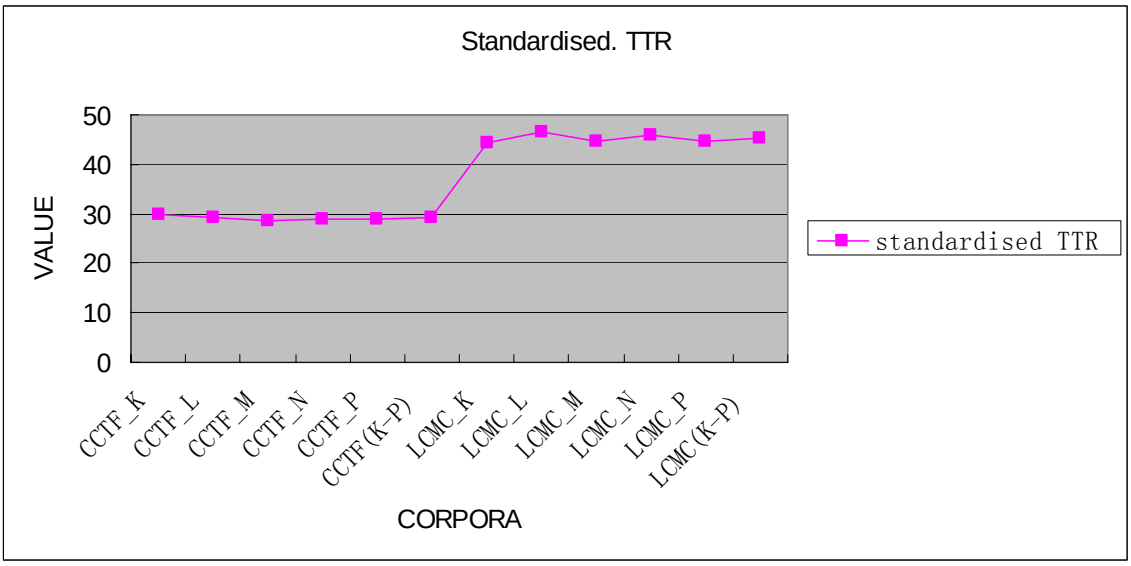
It was obvious from the above graph that the average St. sentence length of CCTF is 34.42,

which is remarkably higer than 11.94 of the LCMC(K-P). this interesting phenomenon seems to contradict our presupposition of a shorter sentence length. However, as far as the word length is concerned, a mean value of 0.54in CCTF is apparently lower than 0.86 in LCMC. We hold that this paradox , nevethelss, best explains, on the one hand, the feature of simplification in translations as illustrated by the short word length, and on the other hand the feature of explicitation. Translations resort to longer sentences to make explicit the same meaning or certain words and expressions in the source texts, which, according to our findings, are generally spread out throught translated texts.

### *4.2.2 Type-Token Ratio*

Similarly, we can also make use of the basic information made available in section 4.1.1 to count the type-token ratio of each corpus and see if it is really the case translation corpora have lower type-token ratio. It is generally believed that breadth of vocabulary can be measured in terms of type-token ratio, which is a ratio of word forms(types) to running words(tokens). Here again we took normalized or standardized type-token ratio deviations as our new measurement to compare LCMC(K-P) and CCTF because it can minimize the difference caused by corpora size.

Using data from Graph 1 and Graph2, we made a graph of standardised TTR of CCTF and LCMC but didn‘t include punctuations, symbols and numbers as tokens. we can read from graph 5 below that CCTF does have a lower type-token ratio compared with corpora in LCMC(K-P). The overall normalized type-token ratio of CCTF’s is 28.18, which is 17.07 lower than that of LCMC(K-P)’s. Nevertheless, it’s noticeable that in CCTF corpus of general fiction has the highest type-token ration while in LCMC the corpus of mystery and detective fiction does, and the reasons, however, remain unknown.




Graph 5 Type-token Ratio of CCTF and LCMC(K-P) standardized at1000 words

To sum up, our findings seem to contradict our hypothesis concerning content words to running words ratio and sentence length but are in favor of our hypotheses about word length and type-token ratio.

### 4.3 Investigation of Explicitation

Explicitation as a proposed universal of translation is a parallel to simplification. In section 4.2, we have demonstrated that in the corpora of CCTF translators are inclined to apply longer sentences, which is expected to hold true, and coincides with the third point in the theory of explicitation in section 2. And yet CCTF is only roughly tagged that we can not examine annotations in the process of translation adopted by translators except the most common strategy of annotating in brackects, so the two practical aspects left for our exploration are the explanatory markers like "*huan ju hua shuo*", "*ji*", and "*zhi*", etc. and the annotation in brackets. By using Regex to count brackets, "*huan ju hua shuo*", "*ji*", and "*zhi*" in both CCTF and LCMC, we found "*ji*" and "*zhi*" used rarely in LCMC to further explain something, and in CCTF only 2 of "*ji*" is located too. For "*huan ju hua shuo*", it is identified 7 times in CCTF and 2 in LCMC; for "*ye jiu shi shuo*", 5 times in LCMC and

17 times in CCTF. When it comes to annotating brackets in texts, we found 88 in CCTF and only 5 in LCMC. Obviously, CCTF is not 16 times larger than LCMC(K-P). This unnatural high frequent usage of annotations in brackets serves only one purpose, that is to say, to make the texts more explicit and easier for readers to understand.

It seems that this investigation of explicitation has an inborn fault and is criticism-provoking – we do not have a parallel corpus to by comparison scientifically find out what is being explicitized and in what ways, for instance, to tell if there was an increase of number of sentences in translation corpora, compared with corpora of original texts, if certain target units in original texts were rendered in a spread-out way embracing any additional elements. Limited by the time and lack of a well-annotated English and Chinese parallel corpora, we did not penetrate deeply into this problem. However, our findings that translations tend to use annotations in brackets and employ more frequent explanatory markers like "*ye jiu shi shuo*" and "*huan ju hua* ", which, to some degree, are good illustrations of explicitation in translations.

**4.4 Investigation of Leveling- out**

We will examine corpora in CCTF to see if they share a kind of homogeneity so far as type-token ratio, readability, sentence length and lexical density ratio are concerned.

Our specific hypothesis in section 2 is that translated texts will generate more harmonious sets of scores and show a central tendency in a continuum of measurement. In other words, compared to non-translated texts, translated texts will generate a narrower range of scores; their scores will have a lower standard deviation, indicating greater closeness. This time we introduced the term standard deviation to measure whether a set of scores are homogeneous or kind of distantly dispersed.

The following table seems only to partially support our hypothesis since only standard deviations of sentence length and type-token ratio are higher than them in CCTF but the standard

deviation of lexical density in LCMC is lower than that in CCTF. This central tendency of lexical use in LCMC(K-P) perhaps can be attributed to the consistent variety of lexical usages by originals writers and different personal tastes of translators when producing works.

| corpora | sent. Length | Lexical density | St. TTR |
|---|---|---|---|
| CCTF-K | 38.05 | 73.76 | 28.97 |
| CCTF-L | 39.60 | 73.93 | 28.47 |
| CCTF-M | 35.48 | 73.88 | 27.69 |
| CCTF-N | 36.61 | 77.35 | 27.99 |
| CCTF-P | 35.58 | 69.26 | 27.92 |
| St. deviation | 1.76 | 2.88 | 0.51 |
| LCMC-K | 16.74 | 68.94 | 44.27 |
| LCMC-L | 18.40 | 67.32 | 46.43 |
| LCMC-M | 21.21 | 66.53 | 44.61 |
| LCMC_N | 19.22 | 64.74 | 46.01 |
| LCMC-P | 17.24 | 68.38 | 44.69 |
| St. deviation | 1.77 | 1.65 | 0.95 |

Table 2 Standard Deviation of Sentence Length, Lexical Density, and Type-token ratio

Above table only tells us translated texts showing homogeneity in case of sentence length and typo-token ratio but a more dispersed manner in lexical density.

Another criterion is readability. Readability indices satisfy Shlesinger's (1989:96-97) precondition that the "equalizing effect" of translation should be measured using a generally recognized, "pre-established" continuum and also make it possible to follow Baker's (1996:184) suggestion that leveling-out should be measured with sets of numerical values, such as those

generated by readability indices. We believe in that translated texts show a similar degree of readability. Now we will examine this point from aspects of Flesch and Lix indexes. Both Flesch and Formula were designed to measure the readability of English texts, here we borrowed them into our study of the readability of corpora and have them adapted to a corpus-based study of Chinese.

The Flesch Reading Ease formula assigns scores on a scale of 0 to 100. The higher the score, the more readable the text is. The designated standard level of reading difficulty is a score of 60 to70. Texts with scores dropping below 60 are considered more difficult to read; those with scores above 70 are deemed easier to read. Both Flesch and Lix formula were calculated on a basis of selected 100 words per text. Thus, in order to conform the way developing the formulas, samples of 10 lines about 100 words are selected, at evenly-spaced intervals of every other 1000 lines in CCTF and every other 500 lines in LCMC considering their sizes, throughout the corpus and the average number of syllables per word(Chinese characters are typically uni-syllabical) and average number of words per sentence are calculated. The Flesch Reading Ease score is calculated in the formula (Flesch 1948:221-233):

*Reading Ease = 206.835 - (1.015*ASL) – (84.6*ASW)*

*Where:*

*ASL = average sentence length (the number of words divided by the number of sentences)*

*ASW = average number of syllables per word (the number of syllables divided by the number of words) (see alsoWilliams 2005:167)*

The Lix readability formula is a useful addition to Flesch index, and is quite simple:

*Lix = Lo + Ml*

*Where:*

*Lo= the number of long words (containing six or more letters)*

*Ml = the arithmetic mean of the sentence lengths*

Lix scores are ranged from a lowest of 20 points to the highest score of around 55 points. However, to avoid the enormous of manual labor for taking 100-word samples, this formula is modified (Williams 2005:171) as:

Lix = ASL+ 100*(Number of long (above 6 letters) words/ Number of words)

We therefore calculated Flesch readability index and Lix Readability index for all the sub-corpora in CCTF and LCMC(K-P) by using the basic information retrieved from Wordsmith 5 and annotations in corpora. See the table below:

| corpora | syllables(per 100) | total words | total sentences | Flesch Score | St. Dev. |
|---|---|---|---|---|---|
| CCTF_K | 100.00 | 4005.00 | 112.00 | 85.94 | 1.27 |
| CCTF_L | 100.00 | 3404.00 | 100.00 | 87.68 | |
| CCTF_M | 100.00 | 4261.00 | 129.00 | 88.71 | |
| CCTF_N | 100.00 | 3605.00 | 100.00 | 85.64 | |
| CCTF_P | 100.00 | 4638.00 | 135.00 | 87.36 | |
| Lcmc_P | 100.00 | 1947.00 | 101.00 | 102.67 | 2.33 |
| Lcmc_n | 100.00 | 1927.00 | 99.00 | 102.48 | |
| Lcmc_l | 100.00 | 1323.00 | 90.00 | 107.31 | |
| Lcmc_m | 100.00 | 582.00 | 32.00 | 103.77 | |
| Lcmc_K | 100.00 | 1491.00 | 99.00 | 106.95 | |
| corpora | number of long words (above 6) | total words | total sentences | Adapted Lix | St. Dev. |
| CCTF_K | 144.00 | 380158.00 | 9989.00 | 38.10 | 1.76 |
| CCTF_L | 167.00 | 357559.00 | 9030.00 | 39.64 | |
| CCTF_M | 121.00 | 353793.00 | 9972.00 | 35.51 | |
| CCTF_N | 531.00 | 654553.00 | 17867.00 | 36.72 | |
| CCTF_P | 61.00 | 434058.00 | 12199.00 | 35.60 | |
| Lcmc_P | 57.00 | 54100.00 | 3132.00 | 17.38 | 1.96 |
| Lcmc_n | 57.00 | 52735.00 | 2738.00 | 19.37 | |
| Lcmc_l | 54.00 | 44883.00 | 2434.00 | 18.56 | |
| Lcmc_m | 49.00 | 11294.00 | 528.00 | 21.82 | |
| Lcmc_K | 53.00 | 55108.00 | 3287.00 | 16.86 | |

Table 3 Flesch Scores and Lix Indexes of CCTF and LCMC (K-P)

From the standard deviation of the Flesch scores and Adapted Lix Readability indexes below, we know that translation corpora CCTF's readability vary little, compared with LCMC(K-P). Both the lower standard deviation of Flesch scores and Lix indexes indicate the comparatively homogeneity of CCTF. This readability ease further explains why we think translations tend to be simplified. However, in so far as difficulty is concerned, translated fiction tend to be more readable as we can read higher scores of Flesch indexes and lower Lix indexes from the above table.

In conclusion, translated Chinese fiction texts show a central tendency in sentence length and type-token ratio but not in lexical density, as illustrated in CCTF. Therefore, the feature of leveling-out is only relatively valid just as we have presupposed in the second section.

## 5. Conclusion

In the present study, we have been concentrated on the investigation of all four of the "universals of translation" originally proposed by Baker. Our present study has been based upon previous studies, working particularly with translated Chinese fiction, and carried further into the study of leveling-out, a fourth recurrent feature having not yet been explored systematically and in a corpus-based manner. Hereinafter, we will give a summary of what we have found and interpret them to the best of our knowledge and finally discuss the outlook of the future study.

We took three measures to testify normalization in translation corpora CCTF. Our findings relating to the lexical density, Lemma words and attested use of words appear to support our hypothesis that translations embodying a strong tendency to use more content words, and adopt idiomatic expressions to achieve, we think, as much as necessary the equivalent effect to the original, which is normalization, sometimes to a extent of exaggeration. This normalization,

perhaps, is due to another reason that all the translations are carefully chosen from works by some renowned translators who are either experienced or formally trained and believe in a normalized or target-language oriented translation gains more popularity and wider readership.

Content words to running words ratio, together with standardized word and sentence length as well as type-token ratio, is employed to measure simplification of a translation. However, they do not seem to provide a consistent evidence to support the hypothesis of simplification as the sentence lengths and lexical densities are unexpectedly higher than that in LCMC(K-P) . The results appear to depend on the vocabulary and grammar of the particular language involved, and not on the translated or non-translated status of a corpus. These results suggest that even though simplification, as we have supposed, is a recurrent feature of translation, it is maybe not limited to richness of vocabulary, lower content words to running words ratio, shorter as well as simpler sentence structures.

The measures applied to investigate explicitation frankly can only offer some very superficial evidence in support of the hypothesis of this feature. It's pitiful that we do not have a corresponding parallel corpus to CCTF in which we could make use of specially annotated information to examine what linguistic phenomena are made explicit and spelled out in texts.

Along with that of sentence length, and type-token ratio, lower standard deviation of Readability indexes of CCTF obviously supports our hypothesis of leveling-out. Their lower standard deviations of readability indexes show a greater homogeneity in a continuum of these measures. However, a 1.23 higher standard deviation of lexical density of CCTF reveals the truth that leveling-out may exist in many characteristic ways, including but may not be restricted to above mentioned features. What we are supposed to do is to find appropriate ones that can be quantified and of quality value distinguishable.

In the future studies, based on carefully and scientifically designed corpora, more detailed

study of either normative or creative expressions from a diachronic or a comparative perspective would be rather applicable; With a viable parallel corpus researchers can also work on certain parts of an utterance in translated texts to compare them with their original forms in non-translated texts so that explicitation is better examined and we can acquire better knowledge of how explicitation is formed and processed in the process of translation; Alternatively, people can examine the specific instances of simplification in translations to describe by analogy their patterns from a macro perspective to a micro perspective; For Chinese, translated texts embracing many other features of leveling-out are observable and worth digging up. For instance, the frequency of various "*Bei*" structures (a kind of passive voice structure, say, "*bei+verb*", "*wei…suo*"," *jiao*", "*gei*", "*rang*") in both translated and non-translated texts and their semantic prosody and distribution in different genres and registers (see McEnery and Xiao, 2005) sometimes can be a measure of leveling-out.

To conclude, we have demonstrated the general hypothesis about recurrent features in translations advanced in section 2 and proved that our hypotheses concerning specific universal features of translations are relatively true under the circumstances provided by this research design, except for some unexpected findings making some particular hypotheses null, say, our finding about lexical density in translation corpora.

**Reference**

Baker, M. (1993): "Corpus Linguistics and Translation Studies: Implications and Applications"[A], Text and Technology: In Honour of John Sinclair, Baker, Francis and Tognini-Bonelli (Eds), Amsterdam/ Philadelphia, John Benjamins, pp. 233-250.

_____. (1995): "Corpora in Translation Studies: An Overview and Suggestions for Future Research"[J], Target 7 (2), pp. 223-243.

_____. (1996). "Corpus-based Translation Studies: The Challenges That Lie Ahead." In Somers, ed., pp. 175—186.

_____. (2000): "Towards a methodology for Investigating the Style of a Literary Translator" [J],Target 12 (2), 241-266.

Ding, S.D. (2001) : A Study of Western Translational English Corpus [[J].Journal of

Foreign Languages，2001(5), pp 61-66.

Toury,G. (1995): Descriptive Translation Studies and Beyond, Amsterdam/Philadelphia, John Benjamins.

Kenny, D. (2001). Lexis and Creativity in Translation: A Corpus-based Study. Manchester: St. Jerome.

_____. (1999b). Norms and Creativity: Lexis in Translated Text [D]. Manchester: Centre for Translation and Intercultural Studies LJMIST. Ph.D Thesis.

Liao, Q.Y. (2000): Corpora and Translation Studies[J]. Foreign Language Teaching and Research Press. 2000, 32(5), pp 380-384.

Laviosa, S. (1998a) : "The Corpus-based Approach:a New Paradigm in Translation Studies" [J], Meta 43(4), pp. 473-479.

_____. (1998b) : "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose" [J], Meta, 43(4), pp. 557–570.

_____. (1998c): "The English Comparable Corpus: a Resource and a Methodology"[A], Bowker, Cronin, Kenny and Pearson (Eds.), Unity in Diversity? Current Trends in Translation Studies, Manchester, St. Jerome Publishing.

_____. (1997): "Investigating Simplification in an English Comparable Corpus of Newspaper Articles"[A], Klaudy and Kohn (Eds), Transferre Necesse Est, Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting 5-7 September, 1996, Budapest, Hungary, Scholastica, pp. 531-540.

_____. (1996): "Comparable Corpora: Towards a Corpus Linguistic Methodology for the Empirical Study of Translation"[A], Thelen and Lewandowska-Tomaszczyk (Eds), Translation and Meaning. Part 3, Proceedings of the Maastricht Session of the 2nd International Maastricht-L ó dz Duo Colloquium on "Translation and Meaning", Maastricht, The Netherlands, 19-22 April 1995, Maastricht, Hogeschool Maastricht School of Translation and Interpreting, pp. 153-163.

McEnery, A. & Z. Xiao. (2005) Passive constructions in English and Chinese: a corpus-based contrastive study. [Powerpoint slides] Proceedings of Corpus Linguistics 2005. Birmingham University, 14-17 July, 2005. Available online: http://www.lancs.ac.uk/postgrad/xiaoz/publications.htm, Last visited June 30th, 2008.

_____. (2004). The Lancaster Corpus of Mandarin Chinese (LCMC)[OL],Retrieved from < *http://www.lancs.ac.uk/fass/projects/corpus/LCMC/*> on May 24th, 2008.

ØVERÅS, L. (1998): "In Search of the Third Code. An Investigation of Norms in Literary Translation"[J], Meta 43(4), pp. 571-588.

Olohan, M., and Baker, M. (2000): "Reporting that in Translated English: Evidence for Subconscious Processes of Explicitation?"[J] , Across Languages and Cultures 1(2), pp. 141—158.

_____. (2001): "Spelling Out the Optionals in Translation: A Corpus Study." UCREL Technical Papers Volume 13, pp. 423-432. Special Issue: Proceedings of the Corpus Linguistics 2001 Conference. Lancaster, UCREL, University Centre for Computer Corpus Research on Language Technical Papers.

_____. (2004). Introducing Corpora in Translation Studies. Chapter 7: Features of translation, Routledge, pp. 90-144.

Qian, H.W. (2004): On syntactic foreignization and domestication in translation [[J]. Foreign Language Teaching and Research Press，32(5), pp 368-373.

Shlesinger, M. (1989). Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts on the Oral-literate Continuum. Tel Aviv University: M A. Thesis.

Kennedy, Graeme. (2000) : An Introduction to Corpus Linguistics[M], Beijing: Foreign Language Teaching and Research Press, pp 60-70.

Williams, O. (2005): "Recurrent Features of Translation in Canada: A Corpus-Based Study" [D]. University of Ottawa: School of Translation and Interpretation. Ph.D. Thesis.

Xiao, R. (2005): "All You Want to Know about LCMC"[OL], Retrieved from , <http://www.corpus4u.org/showthread.php?t=692 /> on May 26th, 2008.

Zhang, M.F. (2002): Using Corpus for Investigating the Style of a Literary Translator -Introducing and commenting on Baker s new research method [J]. Journal of PLA Foreign Languages University，25 (3), pp 54-57.