# Harmony in diversity: The language codes in English–Chinese poetry translation

**Xiaxing Pan**
Chinese Language and Culture College, National Huaqiao University, China

**Xinying Chen**
School of Foreign Studies, Xi'an Jiaotong University, China

**Haitao Liu**
Department of Linguistics, Zhejiang University, China; Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

## Abstract

Translating poetry is a very complex process. The paradoxical nature of untranslatability and translatability of poetry has been noticed by Hai An (The translation of poetry by the translator-cum-poet. Chinese Translator Journals 2005; 6: 27–30), citing two famous scholars who are holding totally different opinions toward poetry translation. Robert Frost purports that 'poetry is what gets lost in translation', and Susan Bassnet advocates 'poetry is what we gain in translation'. However, the common ground between these two drastic opinions is that poetry translation is no more a repetition of the original works than a reproduction. There are both similarities and discrepancies between the translated works and the original pieces, or in another word: 'harmony in diversity'. This study aims to testify the above-mentioned proposal in a clear and objective manner, by comparing the original poetry texts (twenty randomly selected poems from Shakespearean sonnets) with their translated versions (the corresponding Chinese-translated versions by four different translators) from the perspective of vocabulary, word frequency distribution and part-of-speech (POS) frequency distribution. The results have corroborated the previous proposal: first, there is no significant difference in terms of vocabulary size and the text management styles between the translated poems and the original ones. Second, there is a significant difference in the word frequency distribution and POS frequency distribution between translated poems and the original ones. Third, there are also differences in the POS frequency distribution in poems translated by different authors. Furthermore, the translation style could distinguish professional translators from professional poets.

**Correspondence:**
Haitao Liu, Department of Linguistics, Zhejiang University, No. 866 Yuhangtang Road, 310058 Hangzhou, China.
**E-mail:**
lhtzju@yeah.net

Digital Scholarship in the Humanities, Vol. 33, No. 1, 2018. 
doi:10.1093/llc/fqx001 Advance Access published on 14 February 2017

# 1 Introduction

Since the 'literary revolution' at the beginning of twentieth century in mainland China (Hsia, 2004; Tang, 1998), there have been many published books, papers, and other works studying the poetry translation, especially poetry translated from European and North American countries and areas. These efforts bring great improvement to the Chinese Mandarin new poetry creation as well as to the research of poetry translation. Nowadays, the poetry translation and the relative studies in mainland China are mainly focusing on following points:

(1) the relationship between poetry creation and poetry translation (cf. Zheng, 2001),
(2) applied translation techniques and translation principles (cf. Hai, 2005),
(3) the effects of poetry translation on the development of Chinese Mandarin new poetry (cf. Bian, 1989), and
(4) some famous Chinese poets and poetry translators, as well as their works (Huang, 1988).

Nevertheless, the similarities and discrepancies between the original poetry and translated works are the main aspects of poetry translation research. For instance, Malmkjær (2004) suggests that translation is affected by four aspects:

(1) different translators may interpret the same original text in different perspectives;
(2) translation always has a particular goal;
(3) the reason for a translation of a text differs from the reason for writing the text; and
(4) the readers of original texts and their relative translations are different.

Wang and Li (2012, p. 83) also suggest that we have to pay attention to two common problems while studying translator styles: (1) How the translator satisfies the different stylistic demands on language and culture of the original texts and their translations? (2) Is there any difference between the styles of the original writer and the translator? Furthermore, Cao and Zheng (2011, p. 113) emphasize that there would be no exact equation between the original foreign literary works and their translations. Once translated, the new creation turns out to be a completely independent literary work. In conclusion, there would be a style difference between the original writer and the translator, as well as in original works and their translations.

It is widely accepted that translation is a kind of recreation of its original work. Translators prefer putting their own ideas and opinions into the translations to repeating and imitating the original works. Some common types of the differences between original texts and their translations have been discussed, as well as the motivations of these differences (as Malmkjær, 2004 lists). However, it seems that the differences caused by different translators' identities, culture, education, social backgrounds, and target audiences have been generally ignored. Thus, in the present article, we define the discrepancies and similarities between translated works and the original works as 'harmony in diversity' and observe them through the following two questions:

(1) What are the differences in the word and part-of-speech (POS) distribution between the original works and the translated works?
(2) Are there significant differences between professional translators and part-time translators who are supposed to be professional poets? If so, what are the features of the differences?

These two questions are closely related to our hypotheses:

*$H_0(1)$: There is no significant difference between original poetry and their translations on the word distribution.*
*$H_0(2)$: There is no significant difference between original poetry and their translations on the POS distribution.*
*$H_0(3)$: There is no significant difference between the translated works of professional translators and professional poets.*

The approaches and methods we are going to apply for testifying the hypotheses are quite different from most of the previous studies which are methodologically qualitative. Although qualitative studies are good at describing and recording the performance of linguistic units in translation, we can induce and summarize the properties of poetry translation in an intuitive perspective. This

kind of study could not draw an objective and verifiable conclusion quite easily. We propose that the application of quantitative methods into poetry translation studies will be helpful in testing the validity of observations and conclusions drawn from the qualitative studies. Modern quantitative linguistics has improved lots of available quantitative methods.

Language is no doubt the basic and literal material of poetry creation, and the direct medium of poetry transmission and inheritance. Since poetry language is highly related to aesthetic and rhythmic qualities of language, e.g. meter, rhythm, phonaesthetics, etc. (Greene *et al.*, 2012, p. 1046; Masters, 1915, p. 308). We would approach the features of poetry creation and poetry translation with a quantitative linguistic perspective. To be specific, we would observe the quantitative language data to reveal some non-evident or underground language rules in poetry translation, which would bring a more comprehensive and refined analysis to poetry translation study.

Quantitative linguistics has developed many techniques to make objective observations, accurate descriptions, rational predictions, and linguistic explanations to language phenomena. The goal of quantitative linguistic is to discover the underlying language laws, and then explore the mechanisms of the language system (Altmann, 1993, p. 7, 1997, p. 13; Köhler *et al.*, 2005; Liu and Huang, 2012; Těšitelová, 1992, p. 13; etc.). In the present article, our goal is to quantitatively explore the underground language regularities in the process of English–Chinese poetry translation.

For previous quantitative poetry translation studies, Andreev (2003) adopts discriminant analysis techniques to investigate the meter, rhythm, and syntactic properties and so forth of fifteen sonnets of Keats as well as their two corresponding Russian translations. The result of this study makes it clear that the difference between original poetry and translated poetry is significant, which is mainly caused by the difference between the two languages. The result also reveals that there are some differences mainly distributed over the verse lines between the two kinds of translated poetry. Moreover, the study proposes that it is necessary to examine poetry at every language level for complexity, since both poetry strophes and verse lines have extremely strong interaction with all of the language units from every language level. Then Sorvali (2007) points out that the number of words, the frequency of POS, individual words, etc. could be used for distinguishing original poetry and translated poetry. Popescu *et al.* (2015) adopt quantitative parameters to investigate 150 poetry texts written by the Romanian poet Mihai Eminescu on two language levels, namely, phonology and vocabulary. In addition, it is a thorough study of textual quantitative features on the two language levels.

In this work, we randomly selected 20 sonnets from Shakespeare's 154 sonnets and their translations to testify our hypotheses. These texts are sonnets 5, 8, 11, 17, 28, 36, 52, 60, 65, 69, 91, 93, 104, 109, 110, 124, 133, 138, 142, and 151.We collected the respective translations of these sonnets from four Chinese translators, namely Liang Zongdai[1], Tu An,[2] Cao Minglun,[3] and Gu Zhengkun.[4] The first two are part-time translators as well as professional poets, while the other two are professional translators. These selected texts are segmented into words and annotated.[5] Accordingly, language data, such as text vocabulary, word frequencies, POS frequencies, etc., are all extracted from the texts. Then, the extracted data are to be computed and interpreted via the following indexes and methods:

(1) vocabulary-related 'a-index',
(2) word frequency-related 'Zipf's law' and 'writer's view',
(3) analysis of variance (ANOVA) test,
(4) ALSCAL (Alternative Least Square Scaling), and
(5) cluster analysis.

The indexes and data discussions are assigned into five sections accordingly, followed by a brief conclusion, Section 6.

## 2 The 'a-index' of Original Poetry and their Translations

Quantitative linguistic studies are based on objective language data. However, it must be noted that the

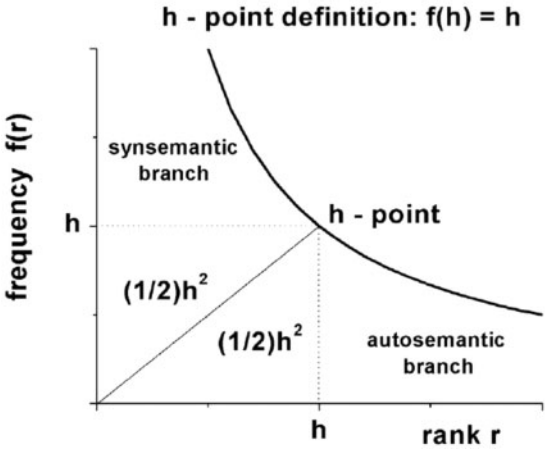


**Fig. 1** The definition of 'h-point' and its position on a rank–frequency curve (cf. Popescu and Altmann, 2006, p. 25)

data are created by means of a definition, and different researchers may define the given entities in a different way. Therefore, we must define the data specifically first and then to process, analyze, and interpret the data. In this chapter, we are going to introduce the definition and the analysis of 'a-index' in our work.

The '*a*-index' is closely related with the '*h*-point' which is proposed by Hirsch (2005), and is cited by Popescu and Altmann (2006, 2007). Popescu *et al.* (2012, p. 121) point out that these two indexes can be used to describe and distinguish the genre features of texts. In quantitative linguistics, any text with a vocabulary $V$ (i.e. the number of word types in the text is $V$) can be reconstructed as a word list, in which the words are ordered into a descending rank according to their frequency in the text. The '*h*-point' is the point whose rank number equals its frequency (Fig. 1). It is defined as:

$$h = \begin{cases} f(h), \text{ if there is an } r = h = f(h) \\ \dfrac{r_j * f(i) - r_i * f(j)}{r_j - r_i + f(i) - f(j)}, \text{ if there is no } r = h = f(h) \end{cases} \quad (1)$$

where $i < j$, $r_i < f(i)$, $r_j > f(j)$, $i$ and $j$ represent the positions of two adjacent words in the word list, and $f(i)$ and $f(j)$ are the frequencies of these two words.

**Table 1** Average values of the 'a-index' of the five poetry groups, as well as their standard deviations, minima, and maxima

| Poetry groups | Number of texts | Average | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| ENG | 20 | 6.4215 | 1.9996 | 3.9500 | 10.1100 |
| CML | 20 | 6.9515 | 2.4997 | 3.2400 | 12.2200 |
| GZK | 20 | 7.5275 | 2.9760 | 3.8100 | 14.5600 |
| LZD | 20 | 6.8325 | 2.6566 | 3.6300 | 12.1700 |
| TA | 20 | 6.2490 | 2.4546 | 3.0400 | 10.4400 |

*Note*: ENG = the group of the twenty original poems, CML = the group of the twenty translated poems by Cao Minglun, GZK = the group of the twenty translated poems by Gu Zhengkun, LZD = the group of the twenty translated poems by Liang Zongdai, TA = the group of the twenty translated poems by Tu An.

So $r$ is the rank of a given word in the list, and $f(r)$ is the respective frequency of this word.

Meanwhile, the 'a-index' is the ratio between 'h-point' and the text vocabulary $V$ (see Function (2)).

$$a = \frac{V}{h^2}. \quad (2)$$

Martináková *et al.* (2008) propose that this index can differentiate texts: the higher the value of the 'a-index', the richer the vocabulary of a text. Table 1 lists the average 'a-index' values of the five poetry groups, as well as the standard deviation, minimum, and maximum. In the table, the five average 'a-index' values are very similar. It seems that there is no vocabulary richness difference between the poems.

To testify the hypothesis, an ANOVA test is adopted here. ANOVA is a useful statistical test in verifying statistic hypotheses and has been widely adopted in linguistics (Biber, 1993). The testing result of ANOVA ($F = 0.777$, $P > 0.05$; Table 2) confirmed our preliminary observation drawn from Table1. Therefore, we can accept our hypothesis above that there is no significant difference in vocabulary richness between the original poetry and their translations. An acceptable explanation would be that the vocabulary richness of the translated poetry is probably heavily dependent on the original poetry.

**Table 2** ANOVA test on the average values of the 'a-index' of the five poetry groups

| | | | Sum of squares | *df* | Mean square | *F* | Significance |
|---|---|---|---|---|---|---|---|
| Between groups | (Combined) | | 20.001 | 4 | 5.000 | 0.777 | 0.543 |
| | Linear term | Contrast | 0.431 | 1 | 0.431 | 0.067 | 0.796 |
| | | Deviation | 19.571 | 3 | 6.524 | 1.013 | 0.390 |
| Within groups | | | 611.531 | 95 | 6.437 | | |
| Total | | | 631.532 | 99 | | | |

## 3 The 'Writer's View' and the Aesthetic Tendency

Martináková *et al.* (2008), Popescu *et al.* (2012), and Tuzzi *et al.* (2010a, b) find out that there is a concentration tendency, which is metaphorically named as 'Golden Ratio', in word frequency distributions of texts. Specifically, this tendency is reflected by a series of 'writer's view' values, and tells the authors' control of function words (e.g. words with no lexical meaning like prepositions, conjunctions, auxiliary words, interjections, etc.) and content words (e.g. words with lexical meaning like nouns, verbs, adjectives, etc.) in their production processes. Their investigations reveal that, with the increase of the size of the texts, the author is going to lose his/her control over the texts gradually, and the arrangement of words in the texts would approach to the 'Golden Ratio' law, which represents authors' pursuit of aesthetics.

Pan *et al.* (2015) investigate the 'writer's view' properties of modern Mandarin poetry, and Fig. 2 is the word rank–frequency distribution curve of the Mandarin translation of Shakespeare's 91st sonnet. Point *A* represents the 1st rank word with the highest frequency 9, point *B* represents the 77th rank word with the lowest frequency 1, and point *h* is the so-called 'h-point'. These *A*, *B*, and *h*-like points on any descending curve can form a triangle with an angle $\alpha$ ($\angle AhB$ in Fig. 2). This angle is metaphorically called 'writer's view' (Popescu *et al.*, 2007, 2009), with its cosine value (see Function (3)) converging to the golden section (≈1.618).

$$\cos\alpha = \frac{(h-1)(f(1)-h)+(h-1)(V-h)}{\sqrt{(h-1)^2+(f(1)-1)^2}*\sqrt{(h-1)^2+(V-h)^2}}. \quad (3)$$

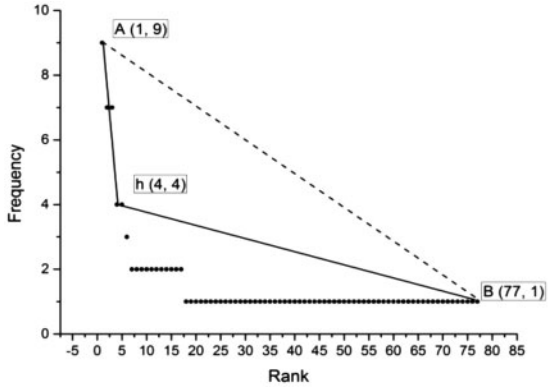


**Fig. 2** The word rank–frequency distribution curve of Shakespeare's 91st sonnet, as well as the 'h-point' and the 'writer's view'

In Function 3, *V* is the number of word types and $f(1)$ is the frequency of the word ranked 1st.

Popescu *et al.* (2012, p. 121) propose that this kind of mechanism controls the language usage, and may have different realizations in different languages. Since the means of the 'writer's view' of our five poetry groups seem similar, Fig. 3, we hypothesize that there is no difference between the original poetry and their translations toward 'writer's view'.

A further ANOVA test result ($F = 16.528$, $P < 0.05$; Table 3) denies the hypothesis and indicates a significant difference between the original poetry and their translations. This result can be interpreted, at least, in two different perspectives. First, the 'writer's view' figures out the difference on the arrangement of function words and content words, which projects the typological language difference between English and Chinese. Second, the investigation also reflects the difference of aesthetic appreciation among English and Chinese audiences.

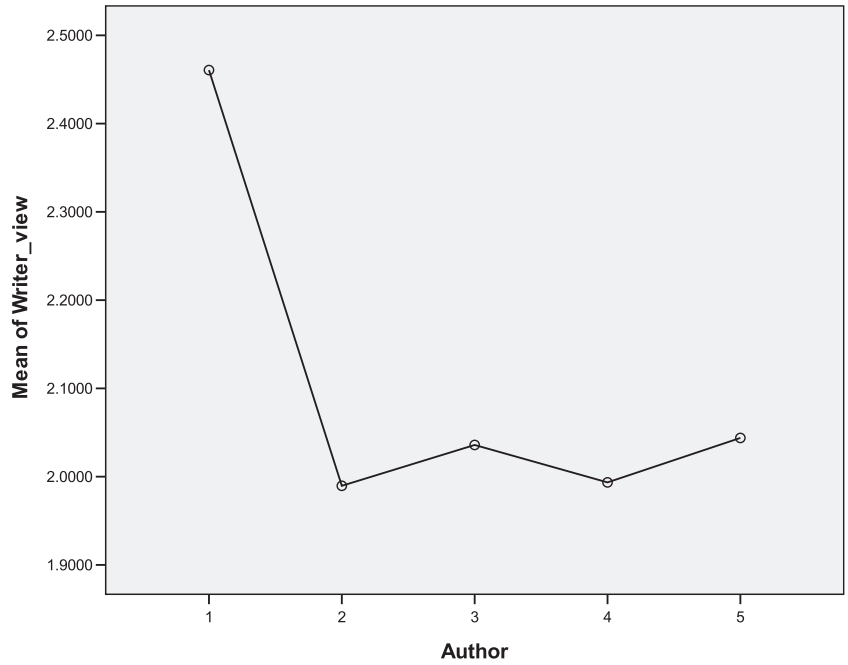


**Fig. 3** Means of the 'writer's view' of the five poetry groups

**Table 3** ANOVA test for values of the 'Writer's view' of the five poetry groups

| | | | Sum of squares | *df* | Mean square | *F* | Significance |
|---|---|---|---|---|---|---|---|
| Between groups | (Combined) | | 3.214 | 4 | 0.804 | 16.528 | 0.000 |
| | Linear term | Contrast | 1.377 | 1 | 1.377 | 28.314 | 0.000 |
| | | Deviation | 1.838 | 3 | 0.613 | 12.599 | 0.000 |
| Within groups | | | 4.619 | 95 | 0.049 | | |
| Total | | | 7.834 | 99 | | | |

## 4 'Zipf's Law' and the Parameter *b*

'Zipf's law' (Zipf, 1935) is one of the most basic and important language laws in quantitative linguistics. In 1916, the French shorthand expert J. Estoup, who concentrated on frequency dictionary study, proposed a new word list of the words for the dictionary. In the list, words' ranks are ascribed according to their frequencies. Then, based on the rank, Estoup calculated the product of the rank position $r$ of every word and every corresponding frequency $n_r$, which is a constant. In 1928, Condon (1928) said that the product of the word rank $r$ and the word probability $n_r/N$ equals $k/N$. Due to the fact that the size of any text is defined, the ratio $k/N$ is nevertheless a constant, too. Condon then defined this ratio as a constant $c$ whose value should be about 0.102. In 1935, Zipf counted the word frequencies of *Ulysses*, and demonstrated $c$ as a constant whose value ranges from 0 to 0.1. Eventually, he proposed the famous linguistic law as Function (4):

$$y = ax^b. \tag{4}$$

Where $y$ is the word's frequency in a descending order, and $x$ is the rank of the words, the previous constant $c$ is replaced by $a$ here, and $b$ is a parameter, $b < 0$. This formula has been criticized by Mandelbrot, and a slightly more complex formula

**Table 4** Fitting Zipf's law to the word frequency distribution of Shakespeare's 133rd sonnet

| Rank | Word | Frequency | Fitted frequency | Rank | Word | Frequency | Fitted frequency |
|---|---|---|---|---|---|---|---|
| 1 | my | 9 | 9.29 | 10 | Be | 3 | 2.56 |
| 2 | heart | 6 | 6.25 | ... | | | |
| 3 | and | 5 | 5.04 | 74 | Cruel | 1 | 0.83 |
| 4 | me | 5 | 4.27 | 75 | From | 1 | 0.83 |
| 5 | in | 4 | 3.76 | 76 | Must | 1 | 0.82 |
| 6 | to | 4 | 3.41 | 77 | Eye | 1 | 0.82 |
| 7 | that | 4 | 3.12 | 78 | Next | 1 | 0.81 |
| 8 | thou | 3 | 2.89 | 79 | Taken | 1 | 0.80 |
| 9 | friend | 3 | 2.72 | 80 | Hath | 1 | 0.80 |

$N = 129$, $V = 80$, $a = 9.2895$, $b = -0.5599$, $DF = 78$, $R^2 = 0.9584$.

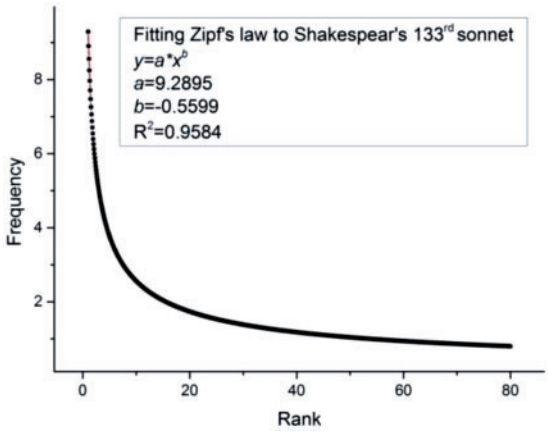


**Fig. 4** Fitting Zipf's law to the word frequency distribution of Shakespeare's 133rd sonnet

has been proposed afterward (Manderlbrot, 1953, p. 492).

An example is given below to demonstrate the Zipf's law more specifically. Table 4 records a data fitting result of Shakespeare's 133rd sonnet's word frequency distribution, where words in the text are ranked in descending order. In the table, numbers in the 'frequency' column represent the times that the words appear in the text, while numbers in the 'fitted frequency' column represent the theoretical frequency that is fitted by Zipf's law. It is clear that the word frequency distribution of Shakespeare's 133rd sonnet can be fitted by Zipf's law quite well with the goodness-of-fit coefficient $R^2 = 0.9584$.[6] Fig. 4 delivers a more intuitive understanding of this sonnet's word distribution. According to Fig. 4, the majority of the words appear rarely, except a few words that have relatively high frequencies.

This law is a part of the 'self-organization' of natural language, which represents a tendency in organizing language units by writers and readers, and it is widely considered as a universal language property (Köhler, 1993). This mechanism can therefore assure the uniformity and diversity in word usage. However, in Chinese poetry writing, especially ancient poetry writing, there is a norm that the poetry language should be succinct but implicit. Poets are more likely to use the simplest words that have most distinctive features. Therefore, the readers of ancient Chinese poetry, especially in modern times, may have completely different ideas when they are reading and translating the ancient poems. As the instances shown in Table 5 and Fig. 5, fitting Zipf's law to the word frequency distribution of Du Mu's *Qingming*[7] failed. All the word frequencies in this ancient Chinese poetry are one, which go against the 'self-organization' law. What is more, words in most of the ancient Chinese poems are deliberately selected and not repeated. Hence, there is no spontaneous 'self-organization' (Liu and Pan, 2015).

Moreover, Liu and Pan (2015) investigate poetry data, including translated poetry texts, and demonstrate the literal connections between ancient and modern Chinese poetry. However, most of the language features of modern Chinese and ancient Chinese are quite different. For instance, ancient Chinese usually treat one Chinese character (*hanzi*) as one word, but in modern Chinese, one

**Table 5** Fitting Zipf's law to the word frequency distribution of Du Mu's *Qingming*

| Rank | word | Frequency | Rank | word | Frequency |
|---|---|---|---|---|---|
| 1 | 何 | 1 | 11 | 纷纷 | 1 |
| 2 | 处 | 1 | 12 | 路 | 1 |
| 3 | 酒家 | 1 | 13 | 雨 | 1 |
| 4 | 魂 | 1 | 14 | 清明 | 1 |
| 5 | 借问 | 1 | 15 | 时节 | 1 |
| 6 | 指 | 1 | 16 | 欲 | 1 |
| 7 | 杏花村 | 1 | 17 | 断 | 1 |
| 8 | 遥 | 1 | 18 | 人 | 1 |
| 9 | 有 | 1 | 19 | 上 | 1 |
| 10 | 牧童 | 1 | 20 | 行 | 1 |

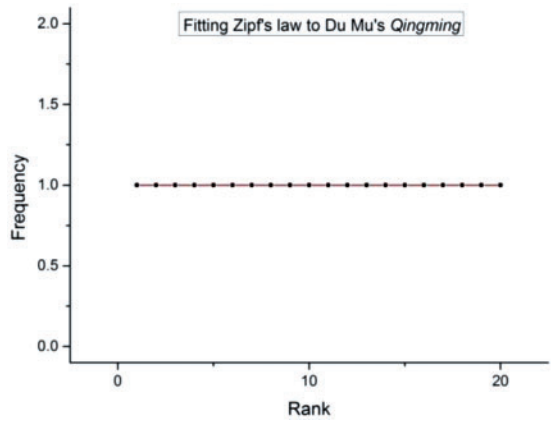


**Fig. 5** The fitting curve of fitting Zipf's law to the word frequency distribution of Du Mu's Qingming (The fitting result says that: $a$=1.0000, $b$=0.0000, $R^2 = -0.0556$)

Chinese character is usually treated as one syllable. The new poetry and translated English poetry cannot be compared with ancient Chinese poetry (or ancient Chinese style poetry) taking every character and word into consideration. Furthermore, there are more function words, such as auxiliaries, conjunctions, etc., are used in English and modern Chinese texts, which make the texts more 'self-organized' than ancient Chinese. Meanwhile, English literary writing is also quite different from Chinese. In English, poets do not need to pay too much attention to the uniformity of the verses, while the uniformity is fundamental for Chinese poets. Thus, generally speaking, English poetry works are comparatively more 'self-organized', that is, the effort of the writer is equilibrated with that of the reader.

For investigating the 'self-organized' feature of our texts, we fit Zipf's law to the word frequency distribution of all the texts in five poetry groups, and Table 6 lists all the fitting results (i.e. $R^2$). Fig. 6 is the box plot of these results listed in Table 6. It is easy to conclude from our results that the first group in Fig. 5 (i.e. ENG group) has the shortest box, with the lowest point above 0.8082 (Table 6), which states that all of the word frequency distributions of the texts in this ENG group obey 'Zipf's law' well ($R^2 > 0.8000$). Meanwhile, since the ENG group has the shortest box, it can be deduced that this group has the smallest variance of the fitting results among these five groups, which means that the texts in the ENG group share great similarities.

The four Chinese translation groups' boxes are all longer than the ENG one, which means that the similarity among the translated works is less than that in the ENG group. Additionally, there are some translated works that failed to fit to 'Zipf's law', e.g. the 60th sonnet in the GZK group (the lowest point in the 'Author = 3' box, $R^2 = 0.6937$), the 65th sonnet in the TA group (the lowest point in the 'Author = 5' box, $R^2 = 0.7449$).The fitting results of 'Zipf's law', especially the lengths of the boxes indicate that there would be some kinds of differences in organizing texts between the original poets and their translators.

For further testing the results, we use ANOVA to testify our $H_0$ hypothesis that there is no significant difference between the original English poetry and their translations on the goodness-of-fit coefficient values.

The fitting results ($F = 0.965$, $P > 0.05$; Table 7) show that we can accept our hypothesis above. They also show that the outliers in Fig. 6 are not significantly influential to our test results.

Joos (1936) figures out that the parameter $b$ in Function (4) is significantly correlated with the text size $N$. However, in our study here, the parameter $b$ seems more correlated with language types rather than with $N$ Table 8. An ANOVA test is conducted to test our hypothesis that the parameter $b$ does not indicate significant difference between the original English sonnets and their Chinese translations.

The ANOVA test results ($F = 13.456$, $P < 0.05$; Table 9) deny the hypothesis and certify a significant

**Table 6** Goodness-of-fit coefficient results $R^2$ of Zipf's law to every text in the five poetry groups

| Text | ENG | CML | GZK | LZD | TA | Text | ENG | CML | GZK | LZD | TA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.8758 | 0.7909 | 0.8008 | 0.8066 | 0.8304 | 11 | 0.9221 | 0.9102 | 0.9304 | 0.8975 | 0.9330 |
| 2 | 0.9068 | 0.9165 | 0.9295 | 0.9006 | 0.9425 | 12 | 0.9387 | 0.8899 | 0.8860 | 0.8502 | 0.9633 |
| 3 | 0.9161 | 0.9237 | 0.8485 | 0.9120 | 0.8301 | 13 | 0.9073 | 0.9124 | 0.8708 | 0.9578 | 0.9149 |
| 4 | 0.9103 | 0.9006 | 0.8961 | 0.9010 | 0.9178 | 14 | 0.9452 | 0.8971 | 0.8788 | 0.8833 | 0.8739 |
| 5 | 0.9124 | 0.9447 | 0.9298 | 0.9450 | 0.9579 | 15 | 0.8585 | 0.8659 | 0.8409 | 0.8520 | 0.9051 |
| 6 | 0.8082 | 0.8879 | 0.9415 | 0.8507 | 0.8987 | 16 | 0.8765 | 0.8426 | 0.9208 | 0.9005 | 0.9200 |
| 7 | 0.9082 | 0.8675 | 0.8431 | 0.9160 | 0.9564 | 17 | 0.9584 | 0.9666 | 0.9446 | 0.9304 | 0.9571 |
| 8 | 0.8598 | 0.7831 | 0.6937 | 0.7667 | 0.7901 | 18 | 0.9470 | 0.8843 | 0.9232 | 0.9350 | 0.9370 |
| 9 | 0.8607 | 0.8876 | 0.8199 | 0.8346 | 0.7449 | 19 | 0.9171 | 0.9436 | 0.8956 | 0.9256 | 0.9395 |
| 10 | 0.8864 | 0.8698 | 0.7995 | 0.8461 | 0.8628 | 20 | 0.8956 | 0.9406 | 0.8796 | 0.8994 | 0.9477 |

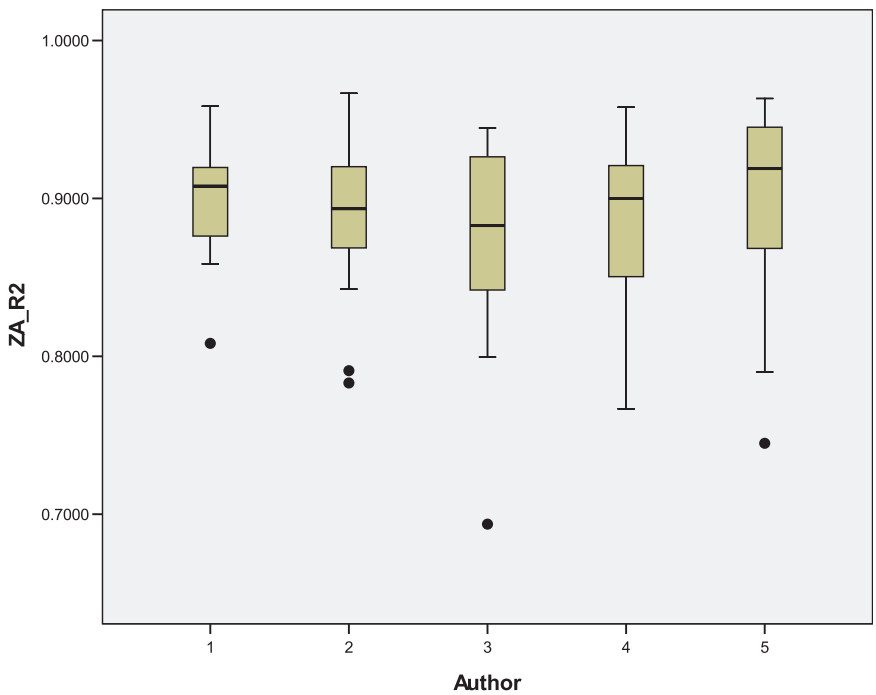


**Fig. 6** The box plot of Zip'f law fitting results to the word frequency distribution of the five poetry groups (The numbers on X-axis represent that: 1, the ENG group. 2, the CML group. 3, the GZK group. 4, the LZD group. 5, the TA group. The same for the Fig. 6. 'ZA_R2' represents the values of the goodness-of-fit coefficient of fitting Zipf's law to the word frequency distribution of the five poetry groups.)

difference between the original poetry texts and the translations on the *b* values. For investigating whether there is any intergroup difference among the four translated groups, we could conduct an Least Significant Difference (LSD) post hoc test for testifying the hypothesis that there is no significant difference among the four translated poetry groups.

The results (Table 10) accept this hypothesis, since all the significant values of among the four translated groups are more than 0.05. Bujdosó (2006, 2008) compared 21 languages and concluded that the value of the parameter *b* in Zipf's law can be used to distinguish the types of languages. Our investigation result is in agreement with this

**Table 7** ANOVA test for the fitting results of Zipf's law to the five poetry groups

| | | | Sum of squares | *df* | Mean square | *F* | Significance |
|---|---|---|---|---|---|---|---|
| Between groups | (Combined) | | 0.010 | 4 | 0.003 | 0.965 | 0.430 |
| | Linear term | Contrast | 0.000 | 1 | 0.000 | 0.015 | 0.902 |
| | | Deviation | 0.010 | 3 | 0.003 | 1.282 | 0.285 |
| Within groups | | | 0.258 | 95 | 0.003 | | |
| Total | | | 0.268 | 99 | | | |

**Table 8** Values of the parameter *b* of the Zipf's law fitting to the five poetry groups

| Text | ENG | CML | GZK | LZD | TA | Text | ENG | CML | GZK | LZD | TA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | −0.3792 | −0.5651 | −0.5142 | −0.5712 | −0.6073 | 11 | −0.6134 | −0.5560 | −0.4410 | −0.5587 | −0.5675 |
| 2 | −0.4173 | −0.5827 | −0.5094 | −0.5964 | −0.5554 | 12 | −0.6450 | −0.6210 | −0.6724 | −0.6980 | −0.6366 |
| 3 | −0.5391 | −0.5216 | −0.5725 | −0.5372 | −0.6150 | 13 | −0.4017 | −0.6686 | −0.4423 | −0.7061 | −0.5250 |
| 4 | −0.4741 | −0.6481 | −0.7932 | −0.8125 | −0.7858 | 14 | −0.5179 | −0.6777 | −0.7188 | −0.7675 | −0.7260 |
| 5 | −0.5126 | −0.5586 | −0.5997 | −0.4907 | −0.6312 | 15 | −0.4384 | −0.6802 | −0.5916 | −0.8172 | −0.6542 |
| 6 | −0.3598 | −0.6745 | −0.6777 | −0.6620 | −0.6755 | 16 | −0.4362 | −0.5848 | −0.5937 | −0.6300 | −0.6343 |
| 7 | −0.5242 | −0.7055 | −0.5539 | −0.6649 | −0.5867 | 17 | −0.5599 | −0.7893 | −0.9083 | −0.8355 | −0.7930 |
| 8 | −0.4783 | −0.5889 | −0.6674 | −0.7322 | −0.5072 | 18 | −0.5348 | −0.7220 | −0.7064 | −0.6945 | −0.7044 |
| 9 | −0.3844 | −0.6414 | −0.7076 | −0.6533 | −0.5106 | 19 | −0.4476 | −0.7289 | −0.8066 | −0.7290 | −0.6595 |
| 10 | −0.5279 | −0.8288 | −0.6562 | −0.8166 | −0.7702 | 20 | −0.4917 | −0.5931 | −0.6437 | −0.6592 | −0.6956 |

**Table 9** ANOVA test for values of the parameter *b* of the Zipf's law fitting to the five poetry groups

| | | | Sum of squares | *df* | Mean square | *F* | Significance |
|---|---|---|---|---|---|---|---|
| Between groups | (Combined) | | 0.476 | 4 | 0.119 | 13.456 | 0.000 |
| | Linear term | Contrast | 0.246 | 1 | 0.246 | 27.789 | 0.000 |
| | | Deviation | 0.230 | 3 | 0.077 | 8.678 | 0.000 |
| Within groups | | | 0.840 | 95 | 0.009 | | |
| Total | | | 1.316 | 99 | | | |

**Table 10** The value of 'sig.' in LSD post hoc test for multiple comparison

| Author | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 0.000 | 0.000 | 0.788 | 0.245 | 0.872 |
| 3 | 0.000 | 0.788 | 0.000 | 0.153 | 0.914 |
| 4 | 0.000 | 0.245 | 0.153 | 0.000 | 0.186 |
| 5 | 0.000 | 0.872 | 0.914 | 0.186 | 0.000 |

conclusion. In more detail, our study here reveals that the language differences between English and Chinese have a significant effect on poetry translation.

In a brief conclusion, the fitting results ($R^2$ in Tables 6 and 7) show that there is no significant difference among our texts, which means that 'self-organization' generally exists in the original works and their translations. On the other hand, the values of parameter *b* in Zipf's law demonstrate the variances among our texts. These two results together reveal the harmony and diversity between poetry.

Prison/v my/r heart/n in/p thy/r steel/n bosom/n 's/pos ward/n ,/w
But/c then/d my/r friend/n 's/pos heart/n let/v my/r poor/a heart/n bail/v ;/w
Whoe'er/r keeps/v me/r ,/w let/v my/r heart/n be/v his/r guard/n ;/w
Thou/r canst/v not/d then/d use/v rigor/n in/p my/r gaol/n :/w
And/c yet/d thou/r wilt/v ;/w for/p I/r ,/w being/v pent/a in/p thee/r ,/w
Perforce/d am/v thine/r ,/w and/c all/dt that/dt is/v in/p me/r ./w

你/r 那/r 使/v 我/r 心/n 呻吟/v 的/u 心/n 真是/d 该死/v ，/w
因为/p 它/r 深深/a 伤害/v 了/u 我/r 和/c 我/r 朋友/n ！/w
难道/d 折磨/v 我/r 一/m 个/q 人/n 还/d 不够/a 惬意/a ，/w
非/d 得/v 让/v 我/r 的/u 朋友/n 也/d 成为/v 阶下囚/n ？/w
你/r 冷酷/a 的/u 眼睛/n 早/a 已经/d 把/p 我/r 俘获/v ，/w
如今/n 又/d 无情/a 地/u 把/p 另/r 一/m 个/q 我/r 霸占/v 。/w
你/r 和/c 他/r 以及/c 我/r 自己/r 都/d 抛弃/v 了/u 我/r ，/w
于是/c 我/r 经受/v 着/u 三/m 三/m 九/m 重/q 的/u 苦难/n 。/w
请/v 把/p 我/r 心/n 囚/v 于/p 你/r 铁石/n 般/u 的/u 心房/n ，/w
让/v 我/r 不幸/a 的/u 心/n 去/v 把/p 他/r 的/u 心/n 保护/v ，/w
无论/c 谁/r 囚/v 我/r ，/w 让/v 我/r 心/n 为/p 他/r 筑/v 墙/n ，/w
在/p 我/r 的/u 狱中/n 你/r 就/d 不/d 能/v 让/v 他/r 受苦/v 。/w
可/v 你/r 仍/d 会/v 得逞/v ；/w 因/p 我/r 囚/v 在/p 你/r 心里/n ，/w
所以/c 我/r 心中/n 的/u 一切/r 都/d 必然/d 属于/v 你/r 。/w

**Fig. 7** POS annotation of Shakespeare's 133rd sonnet and its Chinese-translated version by Cao Minglun

## 5 POS Frequency Distributions

According to above investigations, although the differences on the vocabulary richness of original English poetry and their translations are not significantly different, the parameter *b* in Zipf's law and the values of 'writer's view' both demonstrate a sizeable divergence. So, a test for a following hypothesis $H_0$ (2) that 'there is no significant difference on the POS arrangement' is still needed.

Fig. 7 is an instance of POS annotation for Shakespeare's 133rd sonnet and its Chinese translation written by Cao Minglun. Table 11 then lists the frequencies of POS in each poetry group. For testing the hypothesis, alternative least square scaling and cluster analysis are adopted (Figs. 8 and 9).

It can be deduced in ALSCAL, when the distance between two data points is comparatively great, the difference between the two points is rather large, otherwise, it is comparatively small. Meanwhile, there are two statistical parameters (*S-stress* and *RSQ*, whose values both range from 0 to 1) are used in ALSCAL to value the analysis results. In detail, if *S-stress* is closer to 0, and *RSQ* is closer to 1 at the same time, the analysis result is better. In this study, ALSCAL would be beneficial to the intra-investigation of text differences among our five poetry groups. Cluster analysis is another common and

**Table 11** Frequencies of POS in the five poetry group

| | ENG | CML | GZK | LZD | TA | | ENG | CML | GZK | LZD | TA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n | 511 | 457 | 494 | 420 | 435 | o | 7 | 8 | 8 | 9 | 7 |
| r | 336 | 359 | 368 | 366 | 367 | u | 47 | 244 | 234 | 262 | 275 |
| v | 436 | 618 | 628 | 572 | 615 | l | 0 | 21 | 35 | 24 | 36 |
| a | 195 | 184 | 217 | 189 | 205 | q | 0 | 35 | 41 | 39 | 54 |
| d | 151 | 275 | 289 | 267 | 255 | c | 112 | 91 | 87 | 95 | 75 |
| b | 0 | 6 | 4 | 4 | 7 | i | 2 | 26 | 31 | 12 | 18 |
| rp | 2 | 0 | 0 | 0 | 0 | zx | 0 | 3 | 4 | 0 | 0 |
| m | 16 | 39 | 58 | 42 | 52 | pos | 34 | 0 | 0 | 0 | 0 |
| p | 312 | 130 | 86 | 114 | 110 | dt | 176 | 0 | 0 | 0 | 0 |
| e | 0 | 0 | 0 | 2 | 0 | w | 366 | 326 | 393 | 403 | 447 |

*Note*: n, noun; r, pronoun; v, verb; a, adjective; d, adverb; b, distinguishing word; rp, particle; m, numeral; p, preposition; e, onomatopoeia; o, interjection; u, auxiliary word; l, location word; q, classifier; c, conjunction; i, idiom; zx, suffix; pos, genitive marker; dt, determiner; and w, punctuation[8].

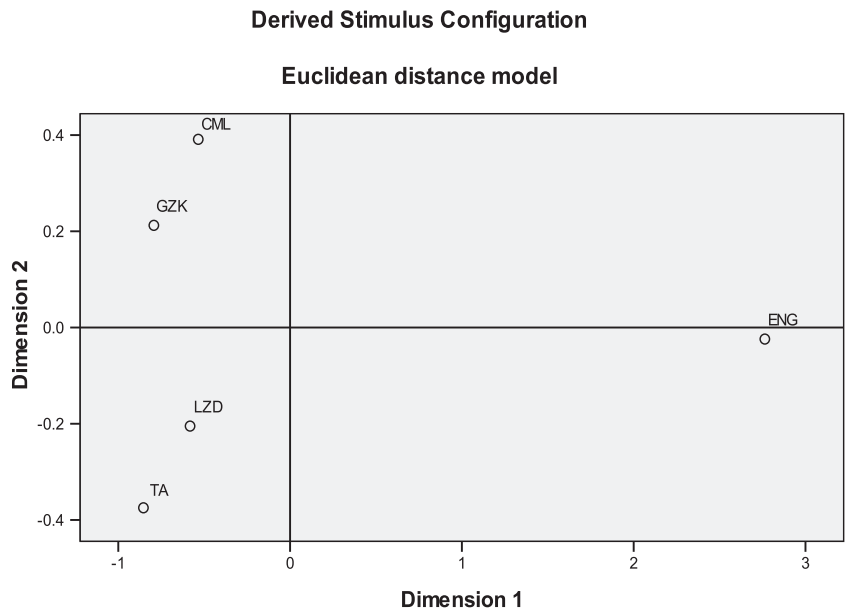


**Fig. 8** ALSCAL analysis on POS arrangement of the five poetry groups
*Note*: The English sonnets are quite different with their Chinese-translated versions, while poetry translated by Cao Minglun and Gu Zhengkun is different from the poetry translated by Liang Zongdai and Tu An.

useful sample classification method, which is good at exploring the similarities of different texts (Ji, 2013). Liu and Pan (2015) successfully adopt this method to discover the difference between ancient Chinese poetry and modern Chinese new poetry. This method should also work effectively for this study.

The two parameters *S-stress* and *RSQ* in our test are 0.01467 and 0.99949, respectively, indicating that the result is well accepted. In detail, our ALSCAL result (Fig. 8) shows that there are clear differences among Chinese and English poetry texts, since the ENG group locates on the left side and another four Chinese translation groups scatter on the right side. The *x*-axis (i.e. Dimension 1) can distinguish texts' language difference very well. Meanwhile, we find out that the *y*-axis (i.e. Dimension 2) can further cluster the four translated poetry groups into two subgroups, with CML

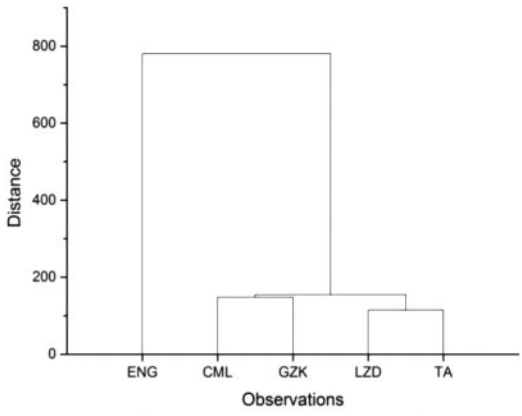


**Fig. 9** Cluster analysis on POS arrangement of the five poetry groups
*Note*: Poetry texts translated by Cao Minglun and Gu Zhengkun are clustered into one group, while the poetry texts translated by Liang Zongdai and Tu An are clustered into another group. Then the two groups are clustered into one group differing from the English group.

and GZK as one, LZD and TA as the other. Interestingly, Cao Minglun and Gu Zhengkun are both professional translators, while Liang Zongdai and Tu An are part-time translators/professional poets. Thus, the *y*-axis actually reveals a translation style difference between these two groups of authors. A further cluster analysis in Fig. 9 also yields the same result. The $H_0(2)$ is therefore rejected, as well as $H_0(3)$.

To conclude, the results of Figs. 8 and 9 demonstrate the difference between the original English poetry and their translations. Meanwhile, they reveal a difference among the translators on POS distribution. According to Malmkjær (2004) and Hai (2005), translation can be viewed as a triple with interplay of three elements, namely, author, translator, and translated work. This can be applied to explain our results here in a way that translators' different identities, the translators' different 'habitus' (Xu and Chu, 2015), or different language environments for creation would lead to text differences/style differences.

## 6 Conclusion

The present article compares twenty of Shakespeare's sonnets with four groups of corresponding Chinese-translated sonnets in a quantitative perspective. Vocabulary richness, word frequency distribution, POS frequency distribution, etc. are adopted as quantitative properties. The empirical language data extracted from objective texts have helped us to deepen and enrich our understanding on the harmonies and diversities in poetry texts. In summary,

(1) the vocabulary richness of the English sonnets is not changed significantly when they are translated into Chinese;

(2) according to the Zipf's law fitting, most of the translated poetry texts are still equilibrated as the original texts;

(3) according to the parameter *b* in Zipf's law and the values of the 'writer's view', the word distribution of English sonnets and their translations are rather different;

(4) according to the result of the ALSCAL analysis on POS, POS arrangement of English sonnets and their translations are significantly different; and

(5) additionally, the investigation further reveals a difference in POS arrangements of the translators. We assume that the reason may be that translators with different identities are generally live in indifferent language environments, and target different readers. More research on the assumption will be conducted in further studies.

## Funding

This work is funded by the National Social Science Foundation of China (11&ZD188) and the High Level Talent Project of National Huaqiao University (15SKBS313). This study was partly supported by the Fundamental Research Funds for the Central Universities (Program of Big Data PLUS Language Universals and Cognition, Zhejiang University) and the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies.

## References

**Altmann, G.** (1993). Science and linguistics. In Köhler, R. and Rieger, B. B. (eds), *Contributions to Quantitative Linguistic*. Dordrecht: Kluwer Academic Publishers, pp. 3–10.

**Altmann, G.** (1997). The art of quantitative linguistic. *Journal of Quantitative Linguistic*, **1–3**: 13–22.

**Andreev, S.** (2003). Estimation of similarity between poetic texts and their translations by means of discriminant analysis. *Journal of Quantitative Linguistic*, **2**: 159–76.

**Bian, Z. -L.** (1989). Merits and demerits of the poetry translation since 'May 4th'. *Translations*, **4**: 182–88.

**Biber, D.** (1993). The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings. *Computers and the Humanities*, **26**: 331–45.

**Bujdosó, I.** (2006, 2008). Rangado–Vortstatistikaekzamenado de la plurlingvateksto de la konstituciopropono de EŭropaUnio. http://www.oocities.org/bujdosxo/rangadoe8.htm.

**Cao, Sh. -Q. and Zheng, Y.** (2011). Translation literature and nationalism of literature. *Foreign Literature Studies*, **6**: 111–17.

**Condon, E. U.** (1928). Statistics of vocabulary. *Science*, **67**: 300.

**Greene, R., Cushman, S., Cavanagh, C., Feinsod, H., Ramazani, J., Marno, D., and Slessarev, A.** (2012). *The Princeton Encyclopedia of Poetry and Poetics*, 4th edn. Princeton; Oxford: Princeton University Press.

**Hai, A.** (2005). The translation of poetry by the translator-cum-poet. *Chinese Translator Journals*, **6**: 27–30.

**Hirsch, J. E.** (2005). An index to quantify an individual's scientific research output. *PNAS*, **46**: 16569–72.

**Hsia, C. T.** (2004). *C. T. Hsia on Chinese Literature*. New York, NY: Columbia University Press.

**Huang, W. -L.** (1988). Influences from American and British on May 4th new poetry. *Journal of Beijing University*, **5**: 25–38.

**Ji, M.** (2013). *Exploratory Statistical Techniques for the Study of Literary Translation*. Lüdenscheid: RAM-Verlag.

**Joos, M.** (1936). Review of G. K. Zipf "*The Psycho-biology of Language*". *Language*, **12**: 196–210.

**Köhler, R.** (1993). Synergetic linguistics. In Köhler, R. and Rieger, B. B. (eds), *Contributions to Quantitative linguistic: Proceedings of the First International Conference on Quantitative linguistic, QUALICO, Trier, 1991*, Dordrecht: Kluwer Academic Publishers, pp. 41–51.

**Köhler, R., Altmann, G., and Piotrowski, P.** (2005). *Quantitative Linguistik: ein Internationales Handbuch*. Berlin; New York: Mouton de Gruyter.

**Liu, H. -T. and Huang, W.** (2012). Quantitative linguistic: state of the art, theories and methods. *Journal of Zhejiang University (Humanities and Social Science)*, **2**: 178–92.

**Liu, H. -T., and Pan, X. -X.** (2015). Quantitative properties of Chinese contemporary poetry. *Journal of Shanxi University (Philosophy and Social Science)*, **2**: 40–7.

**Malmkjær, K.** (2004). Translational stylistics: Dulcken's translations of Hans Christian Andersen. *Language and Literature*, **1**: 13–24.

**Manderlbrot, B.** (1953). An informational theory of the statistical structure of language. In Jackson W. (ed.), *Communication Theory*. London: Butterworths, pp. 486–502.

**Martináková, Z., Mačutek, J., Popescu, I. -I., and Altmann, G.** (2008). Some problems of musical texts. *Glottometrics*, **16**: 80–110.

**Masters, E. L.** (1915). What is poetry. *Poetry*, **6**(6): 306–8.

**Pan, X., Qiu, H., and Liu, H.** (2015). Golden section in Chinese contemporary poetry. *Glottometrics*, **32**: 55–62.

**Popescu, I. -I. and Altmann, G.** (2006). Some aspects of word frequencies. *Glottometrics*, **13**: 23–46.

**Popescu, I. -I. and Altmann, G.** (2007). Writer's view of text generation. *Glottometrics*, **15**: 71–81.

**Popescu, I. -I., Mačutek, J., and Altmann, G.** (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

**Popescu, I. -I., Čech, R., and Altmann, G.** (2012). Some geometric properties of Slovak poetry. *Journal of Quantitative Linguistic*, **2**: 121–31.

**Popescu, I. -I., Lupea, M. L., Tatar, D., and Altmann, G.** (2015). *Quantitative Analysis of Poetic Texts*. Berlin; Boston: Walter de Gruyter.

**Sorvali, I.** (2007). Different translations of one original text in a quantitative & qualitative perspective. In Grzybek, P. and Köhler, R. (eds), *Exact Method in the Study of Language and Text*. Berlin; New York: Gruyter, pp. 611–22.

**Tang, T.** (1998). *History of Modern Chinese Literature*. Beijing: Foreign Language Press.

**Těšitelová, M.** (1992). *Quantitative Linguistic*. Amsterdam; Philadelphia: John Benjamins Publishing Company.

**Tuzzi, A., Popescu, I. -I., and Altmann, G.** (2010a). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

**Tuzzi, A., Popescu, I. -I., and Altmann, G.** (2010b). The golden section in texts. In Wilson, A. (ed.), *Emperical Text and Culture Research 4: Dedicated to Quantitative*

*Emperical Studies of Culture*. Lüdenscheid: RAM-Verlag, pp. 30–41.

**Wang, Q. -Q. and Li, D.** (2012). Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses. *Literary and Linguistic Computing*, **27**: 81–93.

**Xu, M. and Chu, C. -Y.** (2015). Translator's professional habitus and the adjacent discipline: the case of Edgar Snoco. *Target*, **27**(2): 173–91.

**Zheng, H. -L.** (2001). On dissimilation and optimization of target language. *Chinese Translator Journals*, **3**: 3–7.

**Zipf, G. K.** (1935). *The Psychobiology of Language*. Boston: Houghton-Miflin.

## Notes

1 Liang Zongdai (梁宗岱), 1903–83, is a famous modern Chinese poet and translator.

2 Tu An (屠岸), 1923-, is a famous modern Chinese poet and translator.

3 Cao Minglun (曹明伦),1953-, is a translation major professor in Sichuan International Studies University, Chongqing, China.

4 GuZhengkun (辜正坤), 1952-, is a translation major professor in Beijing University, Beijing, China.

5 The word segment and annotate tools are 'Stanford Log-linear Part-Of-Speech Tagger' for English word segmentation and annotation, and 'Segtag' for Chinese word segmentation and annotation.

6 The goodness-of-fit coefficient $R^2$ is always applied in result testing of fitting experiments. Usually, if the value of $R^2$ is more than 0.7500, the fitting result is accepted, to be rejected otherwise.

7 Du Mu, (1803–52), a very famous Chinese poet in late Tang Dynasty. The poetry *Qingming* says: 清明qīngmíng/ 时节shíjié/雨yǔ/ 纷纷fēnfēn, 路lù/ 上shàng/ 行xíng/ 人rén/ 欲yù/ 断duàn/ 魂hún. 借问jièwèn/ 酒家jiǔjiā/ 何hé/ 处chù/ 有yǒu, 牧童mùtóng/ 遥yáo/ 指zhǐ/ 杏花村xìnghuācūn (A drizzling rain falls like tears on the mourning day. The mourner's heart is going to break on his way. Where can a wine shop be found to drawn his sad hours? A cowherd points to a cot mid apricot flowers—translated by Xu Yuanchong许渊冲, 1994).

8 Note: We have paid attention to the effects from the number of punctuation work to the cluster result. The four authors were then clustered into a little different result. We then noted that the punctuation may be one of the aspects which can tell the difference of authors.