ΜΕΤΑ

Quantifying Phraseological Style in Two Modern Chinese Versions of *Don Quijote*

Meng Ji

Cite this note

Article abstract

Quantifying style, or stylometry, has always been one of the oldest traditions in Western literary studies. It seems, however, that such a well-explored and long-standing scientific methodology has been rarely applied to translations, as opposed to original literary texts. The present paper, which focuses on the stylistic use of phraseology in two contemporary Chinese versions of Cervantes' *Don Quijote*, shall endeavour to address the two current problems in corpus-based translation stylistics, i.e., the lack of debate on the question of semantically-rich linguistic units in quantifying style of translations, and the need for testing the use of methods and techniques adapted from corpus statistics in detecting stylistic traits in translations. It is hoped that this study, which aims at expanding the current methodological framework for translation stylistics, will help in the development of this growing area of research in Translation Studies.

érudit

BLOC-NOTES

Quantifying Phraseological Style in Two Modern Chinese Versions of *Don Quijote*

RÉSUMÉ

L'évaluation de style, ou stylométrie, est depuis toujours l'une des traditions les plus anciennes des études littéraires occidentales. Il semble, toutefois, qu'une telle méthodologie scientifique, aussi connue et de longue date soit-elle, n'ait été que rarement appliquée au domaine de la traduction, contrairement à son application aux textes littéraires originaux. Cette étude, qui porte sur l'emploi stylistique de la phraséologie de deux versions contemporaines chinoises de *Don Quichotte* de Cervantes, se propose d'aborder deux problèmes actuels en matière de stylistique comparée de traduction fondée à partir de corpus, à savoir : l'absence de débat sur la question des unités linguistiques riches sémantiquement dans le cadre de l'évaluation des styles de traduction, ainsi que la nécessité d'expérimenter l'emploi de méthodes et de techniques adaptées à la statistique des corpus et ce, afin de détecter les traits stylistiques en traduction. Il est à espérer que cette étude, qui vise à élargir la structure méthodologique actuelle de la stylistique de traduction, contribuera au développement du domaine de recherche grandissant de la traductologie.

ABSTRACT

Quantifying style, or stylometry, has always been one of the oldest traditions in Western literary studies. It seems, however, that such a well-explored and long-standing scientific methodology has been rarely applied to translations, as opposed to original literary texts. The present paper, which focuses on the stylistic use of phraseology in two contemporary Chinese versions of Cervantes' *Don Quijote,* shall endeavour to address the two current problems in corpus-based translation stylistics, i.e., the lack of debate on the question of semantically-rich linguistic units in quantifying style of translations, and the need for testing the use of methods and techniques adapted from corpus statistics in detecting stylistic traits in translations. It is hoped that this study, which aims at expanding the current methodological framework for translation stylistics, will help in the development of this growing area of research in Translation Studies.

MOTS-CLÉS/KEYWORDS

translation studies, corpus linguistics, literary stylistics, Don Quijote, Chinese four-character expressions

1. Quantifying style in corpus-based Translation Studies

Quantifying style, or stylometry, has always been one of the oldest traditions in Western literary studies. It seems, however, that such a well-explored and long-standing scientific methodology has been rarely applied to translations, as opposed to original literary texts. As a result, most of the past works dealing with an individual translator's style has largely remained at the level of scholars' impressionistic evaluations or based on the assessment of a handful of textual excerpts selected from the parallel source/ target texts.

In recent times, as a result of the introduction of corpus linguistic methodologies to the discipline of descriptive translation studies, more attention has been given to the generation and codification of quantitative linguistic data from electronically-stored translation material, since it is believed that such a corpus-driven or corpus-based approach to translation studies will yield results based on a better empirical grounding than does the traditional line of inquiry into the stylistic nature of translations.

From the initial discussions on the ontological status of a translator's style (Mikhailov and Villikka 2001), through to the outline of a potentially prolific research methodological framework for the study of stylistic patterns in translational texts (Baker 2001), and finally to the actual practice of quantifying highlighted linguistic features of translations in search of possible stylistic traits (Saldahan 2005; Winter 2005), it seems that such a novel research agenda when exploring the visibility of translators in their works is steadily gathering momentum in reshaping the way in which we see translation as an act of re-creation in its own right, and the status of translators, as legitimate creative writers assimilating foreign elements into their own cultures, is thereby considerably enhanced.

However, it should be pointed out that despite the huge research potential for translation studies that the quantification of style has shown, as well as the promising results so far obtained in this regard, the kind of quantitative analysis pursued in these early works shows clearly the limited versatility and lack of sophistication in their arguments. For example, preferred choices of linguistic features by translation stylisticians are, among others, function words, punctuation marks, syntactic or part-of-speech (POS) information, etc., while rarely touching upon semantically-rich textual elements, such as lexis or phraseology.

This is supposed to have been caused by substantial technical problems arising from a corpus-based quantitative analysis of relevant lexical or phrasal categories in the attribution to a translator of a particular style. Nevertheless, as will be shown in the present study, this apparent hurdle when trying to quantify a translator's style is far from being unsolvable given the developing quality of certain corpus linguistic tools used in translation studies; rather, it is argued that such a situation has actually more to do with the collective research habits of corpus-oriented translation stylistic, which in turn has largely reduced the variety of approaches in translation studies.

Secondly, an important notion that needs to be clarified in an effort to improve and advance the current methodology designed for corpus-based translation stylistics is that an integral procedure for the quantification of a writing style, as established in authorship attribution works (Hoover 2001; 2003), entails a spectrum of computational-tool-assisted analyses which may be divided into two major types, i.e., the provision of descriptive data and the generation of inferential statistics (Oakes 1998: 1).

An account of descriptive data involves the extraction and summarization of countable linguistic events within corpus texts, from which one might infer the characteristic use of language of a particular author. In fact, most corpus-based translation stylistic studies so far have remained at this level. The stylistic remarks thus made by translators are based on *observed* patterns of a limited number of descriptive data, rather than on a very wide base of general distribution of quantitative linguistic features throughout the corpus texts of the appropriate corpus.

However, the methodological limitations of such an approach become more than apparent when translation researchers are faced with the processing of linguistic data in hundreds or thousands, instead of tens. In such situation, it is necessary to resort to certain statistical techniques to be able to uncover underlying linguistic patterns in translations as may be hidden behind an agglomeration of corpus data.

Such corpus processing procedure actually entails what Oakes has called as inferential statistics, which answer questions, formulated as hypotheses regarding whether one author is different from another. Despite its wide application in corpus linguistics and other related disciplines of social science, it seems that such methodology has yet to be tested in Translation Studies, especially in corpus-based translation stylistics.

In this part of the present thesis, which focuses on the stylistic use of four-character expressions (FCEXs) in two contemporary Chinese versions of Cervantes' *Don Quijote,* I shall endeavour to address the two aforementioned problems in corpus-based translation stylistics, namely, the lack of debate on the question of semantically-rich linguistic units in quantifying style of translations, and the need for testing the use of methods and techniques adapted from corpus statistics in detecting stylistic traits in translations. It is hoped that this study, which aims at expanding the current methodological framework for translation stylistics, will help in the development of this growing area of research in Translation Studies.

2. Sampling

Faced with a very large number of four character segments detected in Liu's and Yang's translations, a total account of all the instances extracted from the parallel corpus of *Don Quijote* and its two modern Chinese versions, also known as CSCHDQ, would appear to be impractical; moreover, the manual assignment of each four-character item to the relevant four-character category would also turn out to be extremely time-consuming and error-prone. The quantitative nature of the corpus data thus gathered requires a probability-based statistical approach to the identification of linguistic patterns which may underlie the use of four-character expressions (FCEXs) in the two Chinese translations. In other words, since it is not practical to study all the examples of four-character expressions retrieved from CSCHDQ, we would need to select a fraction of the database which will be then used to represent the entire corpus as a whole.

The significance of sampling for corpus-based textual analyses consists in that it may provide important information regarding the characteristics of a large population, the internal complexity of which may not be at all susceptible to human inspection and manual analysis. To ensure that samples are chosen at random and hence hold certain representativeness in describing the distribution of FCEX across the two target texts, an online random number generator has been employed,[1] which helps extract on a computer-driven basis samples from the two translations. An underlying characteristic of this kind of computer assignment technique is that the numbers generated by automatic randomizers, once computed, will be soon discarded by the program and the same data would never be used again. This is to secure the genuineness of the sequences of digits thus produced.

Due to the very large size of FCEXs retrieved from each translation, only three hundred four-character expressions have been selected randomly from each translator's work, which roughly represent some fifteen per cent of the entire database of

FCEXs in each translation. All the FCEX instances included in the test sample sets are then classified according to their structural or semantic properties and incorporated into a cross-tabulation, which exhibits the distribution of FCEXs across the three sample sets in Liu's and Yang's works, respectively.

To detect whether there are any significant differences between the two translations in terms of the distribution of main four-character expression types, we shall conduct the statistical procedure, χ2 (chi-square) test, which is widely used in corpus linguistics or social sciences to evaluate statistically significant differences between proportions for two groups in a dataset. It is a non-parametric test and has the great advantage of not depending on the population being normally distributed. However, it should be noted that the use of χ2 test in quantitative analyses will be inappropriate, if any expected frequency is below one or if the expected frequency is less than five in more than twenty per cent of the contingency table (Dawson and Trapp 2004: 153-154). The expected frequency can be found using the following formula:

Formula I Expected frequency calculation in χ2 test

$$\text{Expected value} = \frac{\text{Row total x Column total}}{\text{Grand total of items}}$$

The computed expected values for the observed frequencies have also been incorporated into Table I. As we can see, in none of the cells has the expected value turned out to be less than five, which suggests the χ2 test may well be deployed with the data presented in Table I. Now, we may proceed to compute the chi-square value for the contingency table, by using the formula shown below:

Formula II Chi-square test

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In the formula, O stands for the observed frequency and E, the expected value. The interpretation of the calculated χ2 value needs the contrast of the actual result with a set of predefined critical values at different levels (Oakes 1998: 27). This is because the chi-square test, just as many other statistical tests, is performed under the null hypothesis that there is no significant difference between the proportions for the two groups. In order to prove that the initial assumption does not apply with the current study, the computed χ2 value has to be larger than the threshold value, normally set at five per cent. The computation of the data shown in Table I gives a χ2 value of 61.164 (Preacher 2001)[2], which is much larger than the critical value 26.12 at the 0.001 level[3], suggesting that the differences between the two translations are indeed statistically significant.

3. Pattern recognition

Table I shows the distribution of the main types of FCEXs in the testing sample sets. It gives us a first impression of the numerical contrast between Liu and Yang regarding their different phraseological profiles. Noting that the reason why we have used the raw frequency, instead of normalized proportions to plot the histogram, is due to the same size of the sample set selected from each translation. The figures shown in the cross-tabulation are used to generate a linear graph, which helps visualize

TABLE I

**Distribution of FCEX types in Liu's and Yang's work**

No	Code	Liu		Yang		
		Observed	Expected	Observed	Expected	Row Total
1	SP	19	18.5	18	18.5	37
2	I-FIG	15	10.5	6	10.5	21
3	SS	16	15.5	15	15.5	31
4	SSY	13	16.5	20	16.5	33
5	MP	23	42.5	62	42.5	85
6	SB	27	45	63	45	90
7	I-AI	30	18.5	7	18.5	37
8	FIG	75	66	57	66	132
9	AI	82	67	52	67	134
Column Total	FCEX	300	300	300	300	600

N.B. MP= morphologically patterned FCEXs; SS= syntactically schematic FCEXs; SSY= structurally symmetrical FCEXs; SB= semantically bipartite FCEXs; SP= shortened phrases; AI= conventionalized archaic idioms; I-AI= instantiated archaic idioms; FIG= conventionalized figurative idioms; I-FIG= instantiated figurative idioms.

the trends hidden behind the abstract data. As can be seen from the graph displayed above, a number of important and consistent patterns underlying Liu's and Yang's use of FCEXs in their translations of *Don Quijote* begin to emerge.

Diagram I

Identification of general contrastive patterns between Liu's and Yang's work
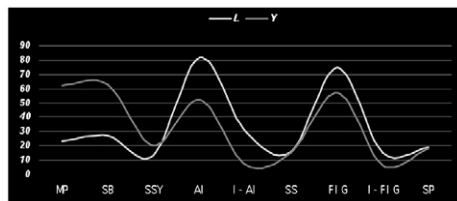
The major difference between the paired linear trends occurs when the pink line (Yang) runs much higher than its green counterpart (Liu) across the initial three categories, i.e., morphologically patterned FCEXs (MPs), semantically bipartite FCEXs (SBs) and structurally symmetrical FCEXs (SSYs). Such finding shows an important aspect of the phraseological profile of Yang's work. The relatively higher use of these three FCEX categories by Yang in her first direct translation of *Don Quijote* into Chinese share a common feature of modifying the phraseological patterns at a structural rather than semantic level, which in turn signifies an initial and original attempt at assimilating the source text to the linguistic characteristics of the target language. In this section, we shall concentrate on this particular profile of Yang's work, by studying in detail Yang's use of MPs, SBs and SSYs in relation to the original, and also in comparison with Liu's choice of other Chinese phraseological categories.

For example, metaphor, which is an important rhetorical trope in figurative speech, is ubiquitous in Chinese figurative idioms. *Don Quijote*, as an early seventeenth-century Castilian masterpiece, is well-known for its abundance of metaphors, which have been attributed to Cervantes' very special language style. (Martín 1991: 79-81; Mariscal 1994: 213-30; Raffel 1993: 5-30). Unfortunately, more often than not, the metaphorical expressions used in *Don Quijote* are unfamiliar to the target audience, which is quite understandable taking into account the huge gap (cultural, geographical, and diachronic) between Cervantes' contemporaries and his modern Chinese readership. To solve this problem, Liu has substituted the original metaphors with conceptually analogous metaphors in Chinese, thus creating a reading experience of *Don Quijote* with which the target readership may feel more at ease. As a result, when

compared with Yang's work, which shows an obvious lack of conceptual or cultural assimilation of the original into Chinese, Liu's language, from the viewpoint of a modern Chinese reader, is more natural and spontaneous, with greater levels of idiomaticity.

ES1 como son mujercillas de poco más a menos, pajecillos y truhanes de pocos años y de poca experiencia, que, a la más necesaria ocasión y cuando es menester dar una traza que importe, se les yelan las migas entre la boca y la mano, y no saben cuál es su mano derecha.
Liu: 像 那些 平淡无奇 的 娘儿们，乳臭未干、涉世不深的 毛 孩子 和 无赖，关键 时刻 需要 他们 拿主意 的时候，他们 却 举棋不定，手足无措 .
Yang: 在 紧要 关头，必须 有 急 智 的时候，这些 人 往往 拿着 面包 不会 往 嘴边 送，自己的 左右手 都 分辨 不出 。

In translating the figurative expressions *se les yelan las migas entre la boca y la mano* (holding the bread between the mouth and the hand) and *no saben cuál es su mano derecha* (not know which is the right hand), whereas Yang's approach is largely literal, Liu has used two well-known Chinese figurative idioms 举棋不定 (be hesitant in making a move in chess) and 手足无措 (not know where to put one's hands and feet; be all in a fluster).

ES2 ¿Qué me ha de suceder -respondió Sancho-, sino el haber perdido de una mano a otra, en un estante, tres pollinos, que cada uno era como un castillo?
Liu: "怎么回事 ？" 桑 乔 说，"转眼之间 我 就 丢了 三头 驴。每头 驴 都 价值连城 。
Yang: 桑 丘 说：" 出了 什么事 吗 ？ 我 一 换 手、一眨眼 的 功夫，丢失 了 三 匹 驴 驹 子，每 一 匹 都 抵 得 一座 大 房 子 呢

Similarly, in translating the figurative expression *era como un castillo* (was worth a castle), Yang has rendered the source text as 抵 得 一座 大房子 (be equal to a big house), while Liu has again opted for using a typical Chinese figurative idiom 价值连城 (be worth several cities).

ES3 Era el espejo en que se miraban, el báculo de su vejez, y el sujeto a quien encaminaban, midiéndolos con el cielo, todos sus deseos; de los cuales, por ser ellos tan buenos, los míos no salían un punto.
Liu: 他们 对 我 奉若神明，把 我 当成 他们 老 年 的 依靠，凡事 都 同 我 商量，从 我 的 需要 出发，我 总是 能 随心所欲 。
Yang: 我 是 他们 照 鉴 自己 的 镜子，是 他们 老来 的 拐杖 。他们 所有 的 愿望，只要 上 天 容许，都 以 我 为主，而且 都 是非 常 好的，和 我 本人 的 愿望 没 一点 参差 。

ES3 is another good illustration of the very different approaches adopted by Liu and Yang in assimilating the source linguistic elements into Chinese. In translating *(Dorotea) era el espejo en que se miraban* (I was the mirror in which they looked at themselves), Yang has essentially retained the original metaphor, while Liu has substituted it with the Chinese figurative idiom 奉若神明 (worship somebody as a god) to emphasize the affection that Dorotea's parents had towards their only daughter.

4. Conclusion

The main issue discussed in the present study is the quantification of style in translations. As shown above, the methods used in the quantification of phraseological style in the target texts has been successful in that it helps substantiate the researcher's intuition regarding the phraseological style of each Chinese translator in translating *Don Quijote*. In her first direct translation of *Don Quijote* from Spanish into Chinese – all previous versions have been via English –, Yang has been in favour of a plain and explicit language style, assuming a faithful attitude in cross-cultural transmissions and communications. Two decades later, Liu, in preparing his popular modern version of Cervantes' masterpiece, has been more spontaneous in bringing in cultural-linguistic elements as very typical of the target language, i.e., his palpably higher use of the unique four-character expressions of Chinese.

Such a stylistic shift from Yang to Liu may be explainable in Liu's improved consciousness of the importance of translating with a view to accommodate the literary expectation of the target readership, rather than seeing himself as heavily loaded with the responsibility of translating faithfully, just as Yang did as a state-commissioned translator of *Don Quijote*. However, it is easy to understand that in a publishing market with an every-increasing competitive urge from his peer translators, Liu has forged a language style of translation with enhanced idiomaticity as embodied in his higher use of four-character expressions, and all these efforts aim at making his translation more appealing to a wider target readership among his contemporaries.

Meng Ji
Imperial College London, London, UK
m.ji05@imperial.ac.uk

NOTES

1. See <http://www.randomizer.org/>.
2. Available from <http://www.quantpsy.org>.
3. The appendix of chi-square critical values used may be found in Oakes (1998: 266).

REFERENCES

Baker, M. (2001): "Towards a methodology for investigating the style of a literary translator," *Target* 12-2, pp. 241-266.

Baroni, M. and S. Bernardini (2006): "A new approach to the study of translationese: machine-learning the difference between original and translated text," *Literary and Linguistic Computing* 21-3, pp. 259-274.

Dawson, B. and R.G. Trapp (2004): *Basic and Clinical Statistics*, New York, Lange Medical Books/McGraw Hill.

Hoover, D.L. (2001): "Statistical Stylistics and Authorship Attribution: an Empirical Investigation," *Literary and Linguist Computing* 16, pp. 421-444.

Hoover, D.L. (2003): "Multivariate analysis and the study of style variation," *Literary and Linguist Computing* 18, pp. 341-360.

Mariscal, G. (1994): "Woman and Other Metaphors in Cervantes's 'Comedia famosa de la entretenida'," in *Theatre Journal* 46-2, pp. 213-230.

Martín, A. L. (1991): *Cervantes and the Burlesque Sonnet*, Berkeley and Los Angeles, University of California Press.

Mikhailov, M. and M. Villikka (2001): "Is there such a thing as a translator's style?," in *Proceedings of Corpus Linguistics 2001*, Birmingham, pp. 378-385.

Milic, L. (1991): "Progress in stylistics: theory, statistics, computers," *Computers and Humanities* 25, pp. 393-400.

Oakes, M. P. (1998): *Statistics for Corpus Linguistics*, Edinburgh, Edinburgh University Press, pp. 1-2.

Oppenheim, R. (1988): "The mathematical analysis of style: A correlation-based approach," *Computers and Humanities* 22-4, pp. 241-252.

Preacher, K. J. (2001): *Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence* [Computer software].

Raffel, B. (1993): "Translating Cervantes: *Una vez más*," in *Cervantes: Bulletin of the Cervantes Society of America* 13-1, pp. 5-30.

Xu, S. H. (2006): "Proximity and complementation: studies on the formation mechanism of Chinese four-character idioms from cognitive point of view," *Journal of Si Chuan International Studies University* 22-2, pp. 107-111.