**HUNG — Applying**

<9 figures>

**Applying a Combined Quantitative and Qualitative Analysis Method to Evaluate the Translatorship of the Fourth Division of the Chinese Translation of the Dīrgha-āgama**

Hung, J.

The issue of doubtful and wrong attributions of the great number of Chinese canonical texts that were translated from Indian originals during the Eastern Han dynasty (25–211 C.E.) to the Tang dynasty (618–907 C.E.) has been much debated in the field of Sinological and Buddhist studies over the last few decades, e.g., by Zürcher (1991), Harrison (1993), and Nattier (2008). Scholarship has increasingly become aware of the fact that the authorship—or, more precisely, translatorship—of the early Chinese translations recorded in the canonical catalogues and in a number of historical records is quite often unreliable.

Although not strictly speaking a case of suspicious translatorship attribution, the content of the fourth division of the Chinese translation of the Indian Dīrgha-āgama (長阿含經, Taishō 1) appears at odds with the rest of the Dīrgha-āgama collection, and scholars of early Buddhist texts have reasons to suspect that this section may have been attached to the rest of the collection at a later stage. The Dīrgha-āgama is a collection of early Buddhist discourses, preserved, in addition to the Chinese version transmitted by the Dharmaguptaka school, in Pali and Sanskrit. The Pali and Sanskrit recensions were transmitted, respectively, by the Tharavāda and (Mūla-)Sarvāstivāda schools. In terms of content, Taishō 1 closely corresponds to its Pali counterpart, the Dīgha-nikāya, and the extant Sanskrit version. There is, however, a major structural discrepancy between the Chinese translation and its parallels, in that the Chinese translation has four main divisions, instead of three. The fourth division is actually a single discourse, DĀ 30 (世記經), in five fascicles, which in terms of content is also out of place in a collection of early Buddhist discourses. In view of the activity record of the main translator of Taishō 1, Zhu Fonian (竺佛念) (Nattier, 2010; Anālayo, 2013; Hung, 2013), it seemed worthy investigating whether the redactor responsible for the addi­tion was Zhu Fonian.

To investigate this hypothesis, we apply quantitative translatorship analysis methods to analyze the text. The main idea of methods based on quantitative approaches to translatorship analysis is to check whether different texts have similar (translation) style, i.e., similar vocabulary usage pattern(s). After quantitative analysis, we study the translation phrases that have been identified by the algorithm in detail. This qualitative form of analysis is necessary to seek the confirmation of our judgment based on the quantitative testing result. Thus our analysis combines quantitative testing with a qualitative approach to the data.

**Research Method**

The digital text of Taishō 1 is taken from the CBETA 2011 corpus.1 The Taishō 1 collection consists of 22 fascicles and con­tains 30 differ­ent discourses. The discourse in question is the last discourse, DĀ 30, which, as mentioned above, consists of the last five fascicles of Taishō 1. For our analy­sis, we remove from the available digital text the entire markup, front and back matter, as well as all the headings added by the editors.

In order to extract significant and meaningful units, which are then used as stylometric measurements, the Chinese sentences are first tokenized with Variable Length N-gram algorithm (abbreviated as VL-ngram, cf. Hung et al., 2009). VL-ngram is an enhanced form of traditional n-gram extraction algorithms. In the VL-ngram algorithm, instead of using fixed-length grams, we first generate all possible grams from our texts—i.e., all bi-grams, tri-grams, quad-grams, and so on—up to the longest possible n-gram. Then we remove all non-significant grams from the features set. To do so, we define an arbitrary num­ber of documents, referred to as the ‘document threshold’, within which a gram must appear as a thresh­old to merit inclusion in the features set. With the proper setting of the document threshold, we are able to avoid the problem of selecting too many content-related words for inclusion in the features set, which would bias the analysis toward content analysis rather style analysis.

Afterward, we use Principle Component Analysis (PCA) on the extracted profiles. The PCA is a statistical procedure that transforms a number of possi­bly correlated variables into a smaller number of uncorrelated variables called *principal compo­nents*. We then plot the values of first and second components in 2-d charts and check whether the translation style of the disputed part of, for example, a scriptural collection or a single work is con­sistent with that of other parts in the same collection or work.

**Analysis Results**

In the discussion below, ‘F17’ refers to the first 17 fascicles of T 1, and ‘L5’ to the last five fascicles of T 1. Figure 1 shows the analysis result with D, the document threshold, set to 4. In Figure 1, all of the five points in L5 lie in the fourth quadrant. In contrast, the points in the F17 group are spread across the remaining three quadrants. This result shows that the five fascicles in L5 have a similar style but behave very differently compared to the fascicles in the F17 group.



Figure 1. PCA result of T 1 with D = 4.

Figures 2 and 3 illustrate the PCA result with D set to 6 and 8, respectively. The results of these two experiments are very similar to those shown in Figure 1.



Figure 2. PCA Result of T 1 with D = 6.



Figure 3. PCA Result of T 1 with D = 8.

In the subsequent analyses, we raise the value of D to 10, 12, and 14 and perform PCA analysis again. The following Figures 4–6 illustrate the result of PCA analyses with three different settings of D.



Figure 4. PCA Result of T 1 with D = 10.



Figure 5. PCA Result of T 1 with D = 12.



Figure 6. PCA Result of T 1 with D = 14.

To our surprise, from Figures 4, 5, and 6, we notice that the difference between L5 and F17 starts to decrease. When D is set to 14, all points in L5 are enclosed by points in the F17 group, and there is no longer a clear boundary between the L5 and F17 groups. In fact, the results show that the L5 and F17 groups share a very similar pattern of usage of translation phrases, as if they choose the translation phrases from the same pool.

Because the five fascicles of DĀ 30 belong to the same narrative and are very similar in content, part of our analysis is biased by the high content similarity of DĀ 30. That is, the difference in the preliminary result is very possibly from difference in content, rather than from diverse transla­tion styles. To remove content-dependent conditions of analysis, we examine the constitu­tion of the first and second components, and identify a number of most distinctive grams of DĀ 30. We find out that many of the resulting grams are topic-related. We remove these grams from the features set, and perform the analysis once again. The analysis results are illustrated in Figures 7 and 8 below.





Figure 7. PCA Result with D = 6, with removing grams that are highly related to the topic of DĀ 30.

Figure 8. PCA Result with D = 8, with removing grams that are highly related to the topic of DĀ 30.

In Figures 7 and 8, the points of F17 and L5 are actually clustering so as to be­long to a single group. This evidence con­firms the previous observation that F17 and L5 are very different in con­tent but very similar in style. At this point, the results become consistent and show no difference between DĀ 30 and the remaining 17 fascicles. This means that in terms of translation style, no significant differences between the first 17 and the last five fascicles have been detected.

More evidence to support this observation comes from the constitution of the first and second components. Then we notice that there are many long strings that appear both in DĀ 30 and in the remaining 17 fascicles. Notably, as shown in Table 1, some of these strings are very rarely or not at all found outside of T 1. Further, most of them are not constrained to a specified topic. Such a situation is likely to occur when the texts were translated by the same translator or translation team. Also, most of the matched long strings are used in narrative descriptions and clichés that commonly occur in many other early Buddhist discourses. However, these strings are largely only used in T 1. We therefore believe this demonstrates that F17 and L5 are indeed by the same translator, Zhu Fonian.

|  |  |  |  |
| --- | --- | --- | --- |
| **Text String** | **Fas­c. in F17** | **Fasc. in L5** | **Oc­curr. (excl. T 1)** |
| 所以者何？我今如來．至真．等正覺，亦說此 | 13 | 19, 22 | 0 |
| 邪婬、兩舌、惡口、妄言、綺語、貪取、嫉妬、邪見 | 7, 15 | 18 | 0 |
| 生死已盡，梵行已立，所作已辦，更不受有 | 5, 9 | 22 | 4 (T 190: 1; T 212: 3) |
| 盡壽不殺、不盜、不婬、不欺、不飲酒 | 3, 7, 11, 13, 15, 17 | 20 | 1 (T 125) |
| 遠塵離垢，得法眼淨，猶如淨潔 | 13, 15 | 20 | 0 |
| 剃除鬚髮，服三法衣，出家修道 | 1, 6, 11, 13, 14, 15, 16, 17 | 19 | 0 |
| 身行不善，口行不善，意行不善 | 6, 8 | 22 | 2 (T 1548: 2) |
| 苦聖諦，苦集、苦滅、苦出要諦 | 2, 3 | 20 | 0 |

Table 1. Significant long strings used in both the F17 and L5 groups.

**Note**

1. CBETA (Chinese Buddhist Electronic Text Association) is a mature, curated digital edition of the widely used Taishō edition of the Buddhist canon.

**References**

**Anālayo.** (2013). Two Versions of the Mahādeva Tale in the Eko­­tt­a­rika-āgama: A Study in the Development of Taishō No. 125. In Dha­mma­­dinnā (ed.), *Research on the Ekottarika-āgama (Taishō 125).* Taipei: Dharma Drum Publish­ing Corpo­ra­tion, pp. 1–70.

**Harrison, P.** (1993). The Earliest Chinese Translations of Mahāyāna Buddhist Sūtras: Some Notes on the Works of Lokakṣema. *Buddhist Studies Review,* **10**(2): 135–77.

**Hung, J., Bingenheimer, M. and Wiles, S.** (2009). Quantitative Evidence for a Hypothesis Regarding the Attribution of Early Buddhist Translations. *Literary and Linguistic Computing,* **25**(1): 119–34.

**Hung, J.** (2013). The Second Version of the Mahādeva Tale in the Ekottarika-āgama: Quantitative Text Analysis and Translatorship Attribution. In Dha­mma­­dinnā (ed.), *Research on the Ekottarika-āgama (Taishō 125)*. Taipei: Dharma Drum Publish­ing Corpo­ra­tion, pp. 107–32.

**Nattier, J.** (2008). *A Guide to the Earliest Chinese Buddhist Translations: Texts from the Eastern Han*東漢*and Three Kingdoms*三國*Periods.* International Research Institute for Advanced Buddhology, Soka University.

**Nattier, J.** (2010). Re-Evaluating Zhu Fonian’s Shizu duanjie jing (T309): Translation or Forgery? *Annual Report of the In­ter­national Research Institute for Advanced Buddhology at Soka University,* **13**: 231–58.

**Zürcher, E.** (1991). A New Look at the Earliest Chinese Buddhist Texts. In Koichi Shinohara et al. (eds), *From Benares to Beijing: Essays on Buddhism and Chinese Religion in Honour of Prof. Jan Yün-hua*.Mosaic Press, pp. 277–304.