digital humanities 2012

Conference Abstracts

University of Hamburg, July 16–22

Hamburg University Press

Stylometric Analysis of Chinese Buddhist texts: Do different Chinese translations of the 'Gandhavyūha' reflect stylistic features that are typical for their age?

Bingenheimer, Marcus
m.bingenheimer@gmail.com
Temple University, USA

Hung, Jen-Jou
Jenjou.Hung@gmail.com
Dharma Drum Buddhist College, Taiwan

Hsieh, Cheng-en
chengen.xie@ddbc.edu.tw
Dharma Drum Buddhist College, Taiwan

Buddhist Hybrid Chinese is a form of Classical Chinese that is used in the translation of Buddhist scriptures from Indian languages to Chinese between the 2nd and the 11th century CE. It differs from standard Classical Chinese of the period in vocabulary (esp. the use of compounds and transcriptions of Indian terms), register (esp. the inclusion of vernacular elements), genre (esp. the use of prosimetry), and rarely even syntax (at times imitating the syntax of the Indian original). Texts in Buddhist Hybrid Chinese are central to all traditions of East Asian Buddhism, which is practiced in China, Korea, Japan and Vietnam.

No comprehensive linguistic description of Buddhist Hybrid Chinese has been attempted so far and perhaps never will, due to the great diversity between translation idioms that at times use different Chinese terms for one single Indian term, and in other cases one single Chinese term for different Indian terms. In as far as Buddhist Hybrid Chinese has been described, the research generally concentrates on grammatical particles (e.g. Yu 1993), single texts (e.g. Karashima 1994), single terms (e.g. Pelliot 1933) or even single characters (e.g. Pulleyblank 1965). The stylometric study of Buddhist Hybrid Chinese – as that of Classical Chinese in general – has only just begun. Only since 2002, when the Chinese Buddhist Electronic Text Association (CBETA) distributed the texts in XML are the canonical texts available in a reliable digital edition.[1]

The Chinese Buddhist canon was printed first in the 10th century and regarding texts before that date its contents have been relatively consistent since then. The currently most widely referenced edition (the Taishō edition, published 1924-34) is based on a Korean edition from the 14th century. It contains ca. 2200 texts from India and China. Due to insufficient and unreliable bibliographic information for texts translated before the 7th century, the attributions to individual translators – where they exist at all – are often questionable. This again has an impact on the dating of the early texts, as they are usually dated via their translator(s). Since most stylometric methods, including those for authorship attribution, were developed for European languages, they often rely on easily parsable word-boundaries, which in the case of Buddhist Hybrid Chinese do not exist. Our wider aim is therefore to develop methods to identify stylistic clues for certain eras in Chinese translations from Indian texts. Can we, based on stylometric features, find a way to date Chinese Buddhist texts or at least to meaningfully corroborate or contradict traditional attributions?

In this study we have compared three translations of the same text, i.e. the *Gandhavyūha* section (ch. Ru fajie pin 入法界品) of the *Avatamsakamsūtra* (ch. Huayan jing 華嚴經). The *Gandhavyūha,* which contains a long narrative of the quest of the young man Sudhana to visit spiritual teachers, was translated into Chinese three times:

T. 278 by Buddhabhadra 佛陀拔陀羅et. al. (Chang'an 418-20 CE)

T. 279 by Śiksānanda實叉難陀 et. al. (Chang'an 695-699 CE)

T. 293 by Prajña 般若et. al. (Chang'an 796-8 CE)

Our task in this particular case was to develop an algorithm that can demonstrate that the T.278 was translated three to four hundred years earlier than T.279 and T.293, and show which of its features can identify a translation idiom that is earlier or at least different from that of T.279 and T.293. Can it be shown that the two Tang dynasty translations (T.279 and T.293) truly are more closely related to each other than to the translation from the Eastern Jin (T.278)?

Our approach here combines a general statistical weighing of n-grams with a focus on grammatical particles (*xuci* 虛詞). A ranking of their importance for our corpus must factor in occurrence as well as variance. The algorithm must also provide for the fact that characters that function as particles can also be used in nominal or verbal compounds. These instances must be filtered out by applying a list of compounds from a large dictionary of Buddhist terms

(Soothill & Hodous 1937). The algorithm for this is developed in the first section.

The following sections describe the sampling procedure and the preparation of the corpus. Although ostensibly all versions of the same Indian text, the three translations differ greatly in length, mainly because the volume of the Indian *Gandhavyūha* expanded between the 5th and the 8th centuries. To counter this problem and to produce enough samples for our analysis, each translation will be divided into sub-divisions of equal length. Then, the frequencies of grammatical particles in these divisions will be calculated and used for defining the stylometric profile of the three translations. We will therefore deal with text clusters on which we can use Principle Component Analysis (PCA), which we have used in a previous study (Hung, Bingenheimer, Wiles 2010). Using PCA on the extracted profiles and plotting the values of first and second components in 2-d charts we are able to discern clearly that T.279 and T.293 are closer to each other and more distant/different from T.278. The two Tang dynasty translations seem indeed to differ from the Jin dynasty translation in its use of particles, and the first and second component of the PCA analysis result shows, which particles create the distinction.

Thus stylometric analysis can give us a better understanding of the translation styles of Buddhabhadra, Śiksānanda and Prajña. All translators have several other translations attributed to them and comparing their *Gandhavyūha* translation to the rest of their corpus, and then again their corpora with each other, could in the future help us to improve our algorithms that ideally would be able to describe and demarcate the work of different translators. The general aim is to get a first handle on the quantitative analysis of the corpus written in Buddhist Hybrid Chinese and extract significant features, which can then be used for a more accurate linguistic description of the idiom.

What the analysis does not account for is changes in the Indian text. The Eastern Jin translation was translated from a somewhat different version of the Indian text than the two Tang translations 300-400 years later. This does, however, not impact our analysis. It is possible to distinguish how grammatical particles were used by different translators, because they reflect different styles of Buddhist Hybrid Chinese, which is what we are looking to describe. Even taking into account that the Sanskrit text of the *Gandhavyūha* has evolved between the 5th and the 8th century, its grammar could not have changed to the degree as there are changes in the translation idiom.

References

Hung, J.-J., M. Bingenheimer, and S. Wiles (2010). Quantitative Evidence for a Hypothesis regarding the Attribution of early Buddhist Translations. *Literary and Linguistic Computing* 25(1): 119-134.

Karashima, S. 幸嶋靜志 (1994). *Chōagonkyō no gengo no kenkyū – onshago bunseki o chushin toshite* 長阿含經の原語の研究 – 音写語分析を中心として[Study of the language of the Chinese Dīrghāgama]. Tokyo: Hirakawa平河出版社.

Pelliot, P. (1933). Pāpīyān > 波旬 Po-siun. *T'oung Pao* (Sec. Series), 30 (1-2): 85-99.

Pulleyblank, E. G.. (1965). The Transcription of Sanskrit K and Kh in Chinese *Asia Major* 11 (2): 199-210.

Soothill, W. E. and L. Hodous (1937). *A Dictionary of Chinese Buddhist Terms*. London: Kegan. [Reprint Delhi: Motilal, 1994]. Digital as XML/TEI file at `http://buddhistinformatics.ddbc.edu.tw/glossaries/`.

Yu, L. 俞理明 (1993). *Fojing wenxian yuyan* 佛経文獻語言 [The Language of the Buddhist Scriptures]. Chengdu: Bashu shushe巴蜀書社.

Notes

1. The CBETA edition is an openly available digital edition of the Chinese Buddhist Canon (the texts can be downloaded in various formats at `http://www.cbeta.org/`.

A Computer-Based Approach for Predicting the Translation Time Period of Early Chinese Buddhism Translation

Hung, Jen-Jou
jenjou.hung@gmail.com
Dharma Drum Buddhist College, Taiwan

Bingenheimer, Marcus
m.bingenheimer@gmail.com
Temple University, USA

Kwok, Jieli
guo.jieli@ddbc.edu.tw
Dharma Drum Buddhist College, Taiwan

Buddhism is a world-religion which has managed to take roots in cultures vastly different from that of its origin. Its transmission from India to China between the 2nd and the 10th centuries happened against all odds. The 'Buddhist conquest of China' can be partly attributed to the successful translation of a great number of texts translated into Chinese from Indian languages. The current standard edition of the Chinese Buddhist canon (*Taishō shinshū daizōkyō* (Abbr.: T.) 大正新修大蔵經, edited 1924-1934) contains 3053 works in 85 volumes, including about 1000 texts of Indian (or alleged Indian) provenance. However, ca. 150 of these texts are marked as *shiyi* 失譯, indicating that the name(s) of the translator(s) are unknown. Furthermore, for the texts that were translated between the 2nd and the late 6th century, many attributions are uncertain, problematic or simply incorrect. The issue of doubtful and wrong attributions has been debated in the field of Buddhist studies over the last few decades, e.g., Zürcher (1991), Harrison (1993), (and) Nattier (2008).

Over the years Buddhist scholars have leveraged traditional text-critical methods to corroborate or dispute traditional attributions yet like every method philology has its limits. Faced with a large number of texts in 'Buddhist Hybrid Chinese' of unknown provenance/origin, the long-established note-taking on the usage of characters and words quickly runs into problems. As with European languages, computational linguistics might offer new avenues of data collection and verification. The corpus of Buddhist Hybrid Chinese is available in a reliable digital format (XML/TEI) since the first 55 volumes of the Taishō edition were published freely by the Chinese Buddhist Electronic Texts Association (CBETA).

We are now able to apply statistical methods and artificial intelligence algorithms to the analysis of this corpus. This enables us to obtain new evidence bearing on translatorship attribution problems. The major advantage of quantitative methods for translatorship attribution is being able to analyze large amounts of data and to discover patterns which are not evident to the human reader.

Quantitative translatorship attribution is often considered to be a classification problem, that is, a text with uncertain or problematic authorship will be analyzed and compared with a corpus of texts by possible authors and then attributed to the author which whose works the texts shares most 'characteristics.' Recent years have seen renewed interest in many issues involved in optimizing quantitative authorship attribution. One of them is the effect of the size of possible candidate authors. As Luyckx and Daelemans (2010) have shown the accuracy of authorship analysis will decrease as the number of possible authors increases. It is therefore advisable to limit the number of possible authors in order to get a high accuracy analysis result. In our case, however, many of the early Chinese Buddhist translations are only rarely mentioned in historical records and canonical catalogues, and few have attracted the attention of philologists. For these translations, it is difficult to reduce the range of possible translators.

Therefore, as part of our attempt to establish a foundation for quantitative translatorship attribution for early Chinese Buddhist translations, we propose a classification mechanism based on predicting the translation time or period of a text. The advantage of this mechanism is twofold. First, within a given time bracket for the translation, the number of possible authors is limited, thereby improving the performance of the translatorship attribution. Second, by examining the result classification mechanism, we are able to identify possible and probable stylistic features of translations for different periods.

The time periods we focus on in the present study include three early Chinese dynasties: the Eastern Han (C.E. 25-220), the Three Kingdoms (C.E. 220-280) and the Western Jin (C.E. 266-316). These three dynasties constitute the earliest phase of Buddhist translation history and most of the translations from these periods present attribution problems. In this research, we build up classification mechanisms for each of the three dynasties. These can be used to test whether the translation style of a text is similar to the one *prevalent* during a certain period. We are aware of the fact that within

Buddhist Hybrid Chinese translation styles within a given period can vary greatly.

For the Eastern Han (C.E. 25-220) and the Three Kingdoms periods we build on recent philological scholarship (Nattier 2008), which has ascertained a number of attributions for this period. For the Western Jin textual corpus, we rely on contemporary research on traditional Buddhist sūtra catalogs, from which we exclude those texts for which current scholarship has not reached a consensus (Lancaster 2008; Lü 1981; Ren 1985; Yu 1993; Xu 1987). We then adopt the Variant Length N-gram algorithm (Hung et. al. 2009) to extract the stylometric features from the three corpora of ascertained texts. Variant Length N-gram is an extended form of the traditional n-gram algorithm. In the traditional n-gram algorithm, the length of grams *n* is fixed. Although the exploitation of n-gram algorithm has great impact on the performance of following analysis, deciding the best value of *n* is not straightforward. The Variant Length N-gram algorithm generates grams of all possible lengths, then removes those which are not significant. Thus, the importance of stylometric features is measured across grams of different length. This is crucial as there are no word boundaries in Buddhist Hybrid Chinese: gram-based analysis must therefore include grams of any length.

In the final stage, we use Fisher Linear Discriminant Analysis (FLDA) to analyze the stylometric features that have been extracted from the translations and to build up the classification mechanisms. The FLDA is a well-known dimension reducing and classification algorithm. It returns a linear function that transfers the high dimension source data of different groups into one-dimension points such that the ratio of total variances of projected points to the variances between groups of projected points is maximized. Since the FLDA's transformation is based on assigning weight to n-grams, the analysis is capable of yielding distinctive features, i.e. strings of Chinese characters, that are characteristic of the dynasties in question.

According to our experiments, the classification mechanisms for the three dynasties have all reached an accuracy rate higher than 90%. Moreover, when the three classification mechanisms are combined and usedto predict the translation time of an unknown translation, we can achieve an accuracy rate and a recall rate both above 80%. Besides, we are able to identify characteristic translation terms for different time periods.

References

Harrison, P. (1993). The Earliest Chinese Translations of Mahāyāna Buddhist Sūtras: Some Notes on the Works of Lokaksema. *Buddhist Studies Review* 10(2): 135-177.

Hung, J., M. Bingenheimer, and S. Wiles (2009). Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations. *Literary and Linguistic Computing* 25(1): 119-134.

Lancaster, L. (2008). *Catalogues in the Electronic Era: CBETA and The Korean Buddhist Canon: A Descriptive Catalogue*. CBETA, Taipei, 2008 (electronic publication). Retrieved from `http://jinglu.cbeta.org/lancaster.htm`.

Lu, Cheng 呂澂 (1981). *Xinbian hanwen dazangjing mulu* 新編漢文大藏經目錄. Jinan: Jilu shushe 齊魯書社.

Luyckx, K., and W. Daelemans (2010). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*. 26(1): 35-55.

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

Ren Jiyu 任繼愈 (1985). *Zhongguo fojiao shi* 中國佛教史. Vol 1. Beijing: Zhongguo shehui kexue 中国社会科学出版社.

Xu Lihe 許理和 (1987). Zui zao de fojing yiwen zhong de donghan kouyu chengfen 最早的佛經譯文中的東漢口語成分, Yu yan xue lun cong 語言學論叢, Vol. 14. Beijing: Shangwu yinshuguan 商務印書館, pp. 197-225.

Yu Liming 俞理明 (1993). *Fojing wenxian yuyan* 佛経文献語言 [The Language of the Buddhist Scriptures]. Chengdu: Bashu shushe巴蜀書社, p. 206.

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