Quantitative evidence for a hypothesis regarding the attribution of early Buddhist translations

Jen-Jou Hung, Marcus Bingenheimer and Simon Wiles
Library and Information Center, Dharma Drum Buddhist College, Taiwan, ROC

Abstract

This article provides quantitative evidence for a hypothesis concerning fourth-century translations of Indian Buddhist texts from Prakrit and Sanskrit into Chinese. Using a Variable Length *n*-Gram Feature Extraction Algorithm, principal component analysis and average linkage clustering we are able to show that 24 sutras, attributed by the tradition to different translators, were in fact translated by the same translator or group of translators. Since part of our method 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 translator(s). This is the first time that these techniques have successfully been applied to medieval Chinese texts. The results of this study open up a number of new directions for the lexicographic and syntactic study of early Chinese translations of Buddhist texts.

Correspondence:
Jen-Jou Hung,
Assistant Professor,
Dharma Drum Buddhist College,
No. 2-6 Xishihu,
Jinshan 20842,
Taipei County,
Taiwan, ROC.
E-mail:
jenjou.hung@ddbc.edu.tw

1. Introduction

For several decades now the problem of computational authorship attribution for works written in European languages has led to the emergence of a variety of methods and to some useful results. However, the absence of word separation in both modern and classical Chinese and the grammatical and semantic polyvalence of Chinese characters are just two ways in which authorship attribution poses special problems when it comes to Chinese texts. There have been number of studies focused on developing effective authorship identification approaches for modern Chinese texts (Peng *et al.*, 2003; Liu *et al.*, 2007). However, research on quantitative techniques for authorship attribution in classical Chinese is extremely scarce.

In the specific field of classical Buddhist Chinese texts of Indian origin, computational methods for authorship attribution have almost never been investigated.[1] The sizable corpus of Buddhist texts that were translated from Indic languages into Chinese from the second through the eleventh-centuries poses formidable challenges to analysis. First, the texts were translated over a long period of time and thus bear witness to changes in the target as well as in the source language. Secondly, the production of these translations has in turn influenced the development of Chinese vocabulary. Thirdly, the translation efforts were not—as for

Literary and Linguistic Computing, Vol. 25, No. 1, 2010.
doi:10.1093/llc/fqp036 Advance Access published on 16 November 2009

example was the case in Tibet—based on a common glossary, and the transcription of Indian terms into Chinese has never been standardized outside of translation workshops (on these, see Bingenheimer, 2009). Current algorithms for authorship attribution are geared to identify the style of authors not translators.

We usually know little or nothing about the authors of Indian Buddhist texts, many of which do not in fact have a single author at all, and were transmitted orally for several centuries. In many cases a Chinese translation is the first available witness of a text. Medieval catalogers and editors in China made attempts to record what had been translated by whom, but for a number of reasons, which go beyond the scope of this article, scholars of Buddhism are faced with the fact that more than half of the attributions of pre-eighth century Buddhist translations are known to be inaccurate (see note 4).

In what follows we will discuss one case where 24 sutras that are attributed to different translators were in fact translated by the same persons or group. The Madhyama Āgama (MĀ), in English sometimes translated as the 'Collection of Middle-length Sayings', is an important collection of 222 early sutras. It exists only in Chinese (*Zhong Ahan Jing* 中阿含經), although we have single sutras or fragments of sutras from this collection in Sanskrit and Prakrit.[2] The collection was translated into Chinese twice: the first translation was completed in 384 by Dharmanandin (Tanmonanti 曇摩難提) (fl. 384–391) and Zhu Fonian 竺佛念 (fl. 365–410). In all probability it was Zhu Fonian, a gifted linguist who later became one of the leading translators of his time, who bore the main responsibility for the actual translation of the text into Chinese. Dharmanandin probably merely read and interpreted the text for him. However, this first attempt came to be considered unsatisfactory and only fifteen years later, in 397–398, a retranslation of the text began under Saṅghadeva (Sengqietipo 僧伽提婆) (fl. 383–398). This second translation was finalized in 401. The first translation was soon lost and the second came to be included in the various canonical editions of the Buddhist canon.

As with all Āgama collections, however, translations exist of single sutras from the MĀ which stand on their own in the Canon. These sutras were generally translated independently from the collection as a whole, at different times and by various translators. In the case of the MĀ, the authoritative edition of the Buddhist canon[3] contains 72 such single sutras. In 1969, the eminent scholar of Buddhism Mizuno Kōgen proposed that 24 of these—all previously attributed to very different translators—were in fact translated by Zhu Fonian and Dharmanandin and belong together as remnants of the first translation of the MĀ. Mizuno based this judgment on his consummate knowledge of Buddhist Chinese, Pāli and Sanskrit. He was able to distinguish that 24 of these 72 belonged together simply by reading through them—that is to say, his judgment was qualitative. We have taken it as our task to attempt to verify his hypothesis by providing more substantial quantitative evidence.

The question could be framed as a typical author/translator verification problem. Unfortunately, however, there are still too many open questions regarding the stylistic features of Zhu Fonian's translation corpus to provide us with a reliable sample set against which to compare the 24 single sutras. Moreover, before attempting to identify the translator of these 24 documents, we must begin by examining the first part of Mizuno's hypothesis: do these 24 documents really constitute one set?—were they translated by the same translators? To test this premise we employ a technique called principal component analysis (PCA) to examine the variation between documents.

PCA is a statistical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. With a small number of components, it is easier to quantify the variations between documents. Furthermore, PCA does not require any prespecified class assumption: if the 24 translations identified by Mizuno can be shown to manifest specific and significant trends in the results of the PCA, this may be considered to have demonstrated Mizuno's hypothesis. We will further examine portions of the principal components to discover

characteristic translation usages. This will provide scholars of Buddhism and linguists with valuable material for further research.

The contribution made in this article is threefold: (1) we have, for the first time, developed a systematic approach to analyzing translation style for Chinese Buddhist texts of Indian origin; (2) we have provided quantitative evidence supporting Mizuno's hypothesis; (3) we have discovered term usage characteristics for the first Chinese translation of MĀ. In Section 2, we provide a comprehensive review of previous research. In Section 3, we describe the details of the quantitative authorship attribution approach and report the result of the analysis. Based on these results from Section 3, we propose in Section 4 an enhanced Feature Extraction Algorithm, which improves analytic quality. Section 5 illustrates the result of applying the enhanced algorithm. Finally, Section 6 summarizes the conclusions and outlines the directions of future research.

2. Previous Research

Most existing authorship attribution approaches consists of two steps: first establish the stylometry, extracting quantifiable stylistic features from documents; then proceed to classifying documents on the basis of these features.

The choice of stylometrics employed dramatically influences the results of the final analysis. In the pursuit of higher quality results, various definitions of stylometry have been adopted for use in attribution algorithms. One commonly used type of stylometry is based on textual measurement, such as frequency of function words (Mosteller and Wallace, 1984; Koppel and Schler, 2004; Zhao and Zobel, 2007), frequency of word collocations (Hoover, 2002), vocabulary richness (Grieve, 2007) and average sentence length (Bozkurt *et al.*, 2007). The major advantage of textual measurement techniques is the extremely low computing load. Though computationally simple, these features still prove useful for authorship attribution and have achieved high accuracy rates in many experiments (Grieve, 2007).

Other approaches utilize non-trivial NLP software for extracting contextual grammatical descriptors from documents, e.g. part-of-speech (POS) tags, or morphological characteristics (Stamatatos *et al.*, 2001; Zhang and Lee, 2006). The features extractable by NLP software often allow more precise description of an author's writing style, which may enhance the effectiveness of attribution analysis. In Zhao and Zobel (2007) and Zhao *et al.* (2006) the authors propose several high-precision authorship attribution approaches by applying entropy analysis models on POS tags. However, these types of feature extracting algorithms may encounter practical difficulties when dealing with documents written in Asian languages such as Chinese and Japanese, where determining the word boundaries is not straightforward. To overcome this difficulty, several attribution approaches have adopted *n*-gram extracting algorithms (Kjell, 1994; Peng *et al.*, 2003, 2004; Houvardars and Stamatatos, 2006; Liu *et al.*, 2007) to retrieve features from documents. *n*-Gram extracting algorithms ignore the meaning of text and instead cut the text into arbitrary chunks called grams. Stylometric definitions can then be calculated based on statistical analysis of the grams.

The second step in authorship identification requires the adoption of a suitable statistical classification method or artificial intelligence algorithm with which to classify documents of unknown authorship into one of several possible document classes for which authorship is undisputed. Commonly used statistical classification methods include: the chi-square distribution (Grieve, 2007); Naive-Bayes classifiers (Peng *et al.*, 2003; Bozkurt *et al.*, 2007); multivariable analysis (Stamatatos *et al.*, 1999); PCA (Burrows, 1992; Holmes and Forsyth, 1995; Labbé, 2007); maximal likelihood estimator (Liu *et al.*, 2007); and discriminant analysis (Stamatatos *et al.*, 2001; Tambouratzis and Vassiliou, 2007). In addition, machine learning methods, such as support vector machines (SVM) (Koppel and Schler, 2004; Houvardars and Stamatatos, 2006; Zhang and Lee, 2006; Bozkurt *et al.*, 2007 Automatic Clustering (Bozkurt *et al.*, 2007; Labbe, 2007), Neural Netwroks (Kjell, 1994;

Menevitz and Yousef, 2007) and Data Mining approaches (lqba *et al.*, 2008) are also widely used for this task.

A major challenge for authorship verification is determining whether a document belongs to a specified class given only positive cases. This is what is known as a one-class classification problem. Specialized one-class classification methods exist, and these are often applied to the problem of authorship verification, e.g. one-class SVM (Koppel and Schler, 2004), one-class neural network (Manevitz and Yousef, 2007).

3. Quantitative Analysis

3.1 Corpus

To verify Mizuno's hypothesis, we conducted a quantitative analysis of the translations of the 72 individual sutras which derive from the MĀ (T.27–T.98). We used the digital texts of these 72 sutras, as provided by the Chinese Buddhist Electronic Tripiṭaka Text Collection, version 2008 (CBETA, 2008), which are well marked up according to the TEI XML standard (TEI P5). To retrieve necessary parts from the XML documents, we wrote an XSLT script to strip the XML tags and elements that are not required for analysis, such as headings, revisions, descriptions and commentaries. The resulting files were then stored in plain text format. If Mizuno is correct, the 24 sutras of Table 1 should show common stylistic features in some way distinct from the remaining 48 texts which comprise the other group.

As Mizuno had surmised and we are going to show below, it is likely that all these attributions are wrong. On the basis of the present research we are not able to prove that these 24 texts were translated, as Mizuno holds, by Dharmanandin and Zhu Fonian specifically, but we can demonstrate that the same translation team is responsible for the whole group. In the next subsection, we describe the algorithm for extracting style features from the texts.

3.2 *n*-Gram Feature Extraction

To extract stylistic features from the 72 sutra translations, we adopted an *n*-gram extraction algorithm to tokenize the texts into grams, and then calculated the style features based on these grams. There are two reasons to use the *n*-gram extraction model. First, to the best of our knowledge, there is not yet (and may never be) any reliable method for establishing word boundaries for classical Chinese. We cannot, therefore, implement any style extracting strategies which operate at the word level, let alone apply advanced NLP analysis functions. Secondly, there is no punctuation in the original texts—in fact, all the punctuation in the current editions has been added in the twentieth-century. Many commonly used indicators which rely on the structure of the text, e.g. number of punctuations in a sentence, average length of sentences, etc., are therefore also not available to us. For these reasons it is clear that the *n*-gram extraction model is the best choice for extracting features from classical Chinese texts.

After all modern punctuation and tags are removed from the documents, the text becomes one long string. The string is then cut into grams of fixed length *n*. After the grams are generated, significant grams are identified and selected into the feature set which will be used to evaluate the writing style of documents. In order to avoid selecting content-dependent grams into feature set, which may bias analysis towards content classification rather than authorship classification, we define an arbitrary number[4] of documents in which a gram must appear as a threshold to inclusion in the feature set. Since the contents of the 72 sutras are in fact quite disparate, the occurrence of a gram in several different sutras can be considered an indication that it is a widely-used but *content-independent* term, and is therefore suitable to include in the feature set. Letting S denote the set of all possible grams, and $D(s_x)$ denotes the number of documents in which the occurrence of gram s_x is larger than 0, the feature set 'FS' can be defined as follows:

$$\mathrm{FS} = \left\{ s_x \in S \middle| D(s_x) > |D| \times Th \right\} \qquad (1)$$

where $|D|$ is the number of total documents, and Th is a threshold between 0 and 1.[5] Let fs_x denote a gram in FS and let $F(d,fs_x)$ denote the frequency

Table 1 The 24 sutras Mizuno considers to have been translated by Dharmanandin and Zhu Fonian, together with their traditional attribution according to the various sutra catalogs[6]

Taishō No.	Title	Traditional date (CE)	Traditionally attributed translator	Length (in Chinese Characters)
T47	離睡經 (*li shui jing*)	265–316	Dharmarakṣa (竺法護)	1080
T49	求欲經 (*qiu yu jing*)	265–316	Faju (法炬)	4601
T50	受歲經 (*shou sui jing*)	265–316	Dharmarakṣa	1914
T51	梵志計水淨經(*fan zhi ji shui jing jing*)	317–420	Unknown	789
T53	苦陰經 (*ku ying jing*)	25–220	Unknown	2236
T55	苦陰因事經 (*ku yin yin shi jing*)	265–316	Faju	2258
T56	樂想經 (*le xiang jing*)	265–316	Dharmarakṣa	418
T58	阿耨風經 (*a nou feng jing*)	317–420	Zhu Tanwulan (竺曇無蘭)	2521
T60	瞿曇彌記果經 (*ju tan mi ji guo jing*)	420–479	Huijian (慧簡)	2938
T64	瞻婆比丘經 (*zhan po bi qiu jing*)	265–316	Faju	1323
T65	伏婬經 (*fu yin ching*)	265–316	Faju	1106
T66	魔嬈亂經 (*mo rao luan jing*)	25–220	Unknown	3381
T70	數經 (*shu jing*)	265–316	Faju	2045
T73	須達經 (*xu da jing*)	479–502	Guṇavṛddhi (求那毘地)	1286
T75	佛為黃竹園老婆羅門說學經 (*fo wei huang zhu yuan lao po luo mean shuo xue jing*)	420–479	Unknown	1601
T77	尊上經 (*zun shang jing*)	265–316	Dharmarakṣa	1384
T79	鸚鵡經 (*ying wu jing*)	420–479	Guṇabhadra (求那跋陀羅)	3826
T82	意經 (*yi jing*)	265–316	Dharmarakṣa	1035
T83	應法經 (*ying fa jing*)	265–316	Dharmarakṣa	1659
T90	鞞摩肅經 (*bing mo su jing*)	420–479	Guṇabhadra	1757
T91	婆羅門子命終愛念不離經 (*po luo men zi ming zhong ai nian bu li jing*)	25–220	An Shigao (安世高)	1571
T92	十支居士八城人經 (*shin zhi jus hi ba cheng ren jing*)	25–220	An Shigao	1185
T93	邪見經 (*xie jian jing*)	420–479	Unknown	460
T94	箭喻經 (*jian yu jing*)	317–420	Unknown	1478

of fs_x as it appears in d. $F(d,fs_x)$ can then be calculated by formula (2) as shown in following:

$$F(d,fs_x) = \frac{C(d,fs_x)}{L_d} \quad (2)$$

L_d is the length of document d, $C(d,fs_x)$ is the number of occurrences of fs_x in document d.

3.3 PCA

To produce quantitative information about the translation idiom of the 72 single MĀ sutras, we use PCA techniques. In this section, we report the analysis results and discuss several related issues. Hereafter we use '*TD group*' to refer to the 24 sutras that Mizuno considers as translated by Dharmanandin and Zhu Fonian, and refer to the other 48 sutras as the '*NTD group*'.

3.3.1 *PCA analysis result with bi-gram features*

In first analysis, the *n*-gram length was set to 2; 1003 bi-grams were selected into the feature set, and the frequency of these 1003 features in 72 documents are calculated according to formula 2. Then the PCA analysis was performed on the 72 texts. Figure 1 shows the scatter plot of first and second principle components generated by the PCA analysis. In Fig. 1, the 24 translations in the TD group are plotted as red squares, and the 48 translations of the NTD group are represented by blue diamonds.

In Fig. 1, it is obvious that the red points are separable from the blue points. In fact, all red points have larger first component than any of the blue points. More precisely, the line described by '*First Component = 0.0053*' perfectly divides red

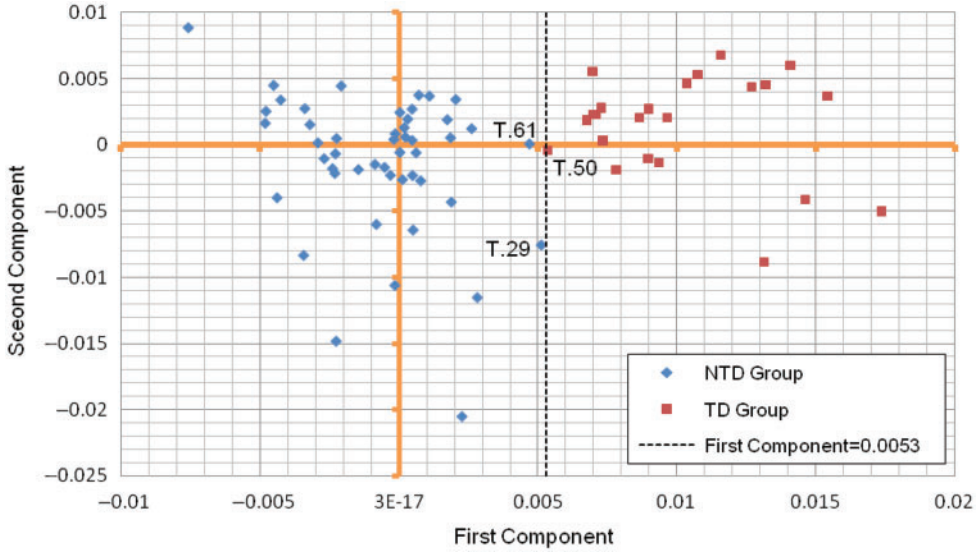

Fig. 1 Scatter plot of the first and second components resulting from a PCA analysis with bi-gram features

points and blue points into two disjoint sets. According to this result, it may be safely concluded that the idiom of the documents in TD group differs from that of the documents in the NTD group.

In order to further specify this difference, we examined the constitution of the first component. Since the documents in the TD group have larger first component values, the 'heaviest' bi-grams of the first component are the most significant, and therefore distinctive, features of the TD group. As the corollary, the 'lightest' bi-grams of the first component are the more significant features of the NTD group. Table 2 shows the 30 bi-grams with the largest and smallest weight in the first component.

In Table 2, 'weight' indicates the weight of the gram in first component and the 'NTD' and 'TD' columns record the number of documents in which the gram is found, from the NTD or TD groups, respectively. From Table 2 we can observe that the bi-grams 喜而, 而樂, 世尊, 是說, 尊所, 佛如, 婆伽, 伽婆, 時婆, 婆在, 聞世, 說歡, 園彼 and 所說 are very commonly used in sutras from the TD group, but only rarely appear in the translations belonging to the NTD group. This result accords with Mizuno's observation (Mizuno, 1969) that most of the 24 sutras in the TD group begin with:- 如是,一時婆伽婆在舍衛國祇樹給孤獨園,彼時世尊 ('Thus have I heard, once the Buddha stayed in Śrāvastī, in the Jeta grove of the Anāthapiṇḍika park, and at that time...') and end with 佛如是說,...聞世尊所說 ,歡喜而樂 ('This is what the Buddha said, (...) having heard the Buddha, [the listeners] were pleased and delighted'). In addition, we found that the terms 却坐, 一面, 面已, 白世, 尊曰, which are substrings of the phrases 却坐一面(已) ('to sit to one side') and 白世尊曰 ('to address the Buddha') are also common only for the TD group. The very same phrases are common in other Buddhist sutras: however, documents in the NTD group do not contain these patterns.

The significant bi-grams of the NTD group, on the other hand, are not part of longer distinct phrases. This suggests that the phraseology is inconsistent, and therefore makes it likely that the texts in the NTD group were not translated by the same group. However, these significant terms still provide valuable information for discriminating the translation styles characteristic of the TD and NTD groups. For example, 佛言 and 佛告 are commonly used in

Table 2 Top 30 bi-grams with largest and smallest weight in the first component[7]

Bi-gram	Weight	NTD (48 Total)	TD (24 Total)	Bi-gram	Weight	NTD (48 Total)	TD (24 Total)
喜而	0.1040	1	21	佛言	−0.07617	25	0
而樂	0.1028	1	19	言人	−0.06835	8	0
世尊	0.0975	25	24	言我	−0.06799	20	3
面已	0.0972	0	16	善惡	−0.06704	14	0
到已	0.0964	4	18	第三	−0.0633	14	0
是說	0.0946	7	22	下人	−0.06315	8	0
一面	0.0920	14	17	如人	−0.06279	8	0
却坐	0.0898	2	12	第四	−0.06256	12	0
尊所	0.0894	3	19	佛告	−0.06242	31	1
佛如	0.0881	1	19	其人	−0.06186	8	0
彼時	0.0876	1	22	第二	−0.05773	14	0
婆伽	0.0862	2	20	時佛	−0.05763	38	1
伽婆	0.0857	2	20	道中	−0.05718	8	0
時婆	0.0851	4	22	天下	−0.05679	16	1
婆在	0.0843	3	22	第五	−0.05679	8	0
已白	0.0840	1	13	為佛	−0.05473	11	0
行至	0.0826	2	12	相見	−0.0536	9	0
所到	0.0824	7	17	當自	−0.05347	8	1
聞世	0.0798	0	18	者佛	−0.0532	9	0
說歡	0.0797	5	22	人皆	−0.0530	8	0
於是	0.0753	13	16	上天	−0.0530	9	0
是尊	0.0747	3	9	人復	−0.0529	7	1
所說	0.0718	26	24	有四	−0.0525	14	2
至世	0.0691	2	13	佛所	−0.0518	28	1
彷徉	0.0686	0	9	門道	−0.0517	8	0
門瞿	0.0685	5	9	佛作	−0.0515	10	1
不久	0.0678	1	7	以不	−0.0514	10	0
園彼	0.0677	0	10	自思	−0.0509	13	1
尊曰	0.0670	4	14	為人	−0.0507	12	0
白世	0.0662	6	14	道人	−0.0504	10	0

the NTD group for 'the Buddha said', but 佛言 is only found in one document from the TD group, and 佛告 not at all.[8] Another interesting observation is that the ordinal numbers that appear as 第二, 第三, 第四 in NTD are entirely absent from TD. Since ordinal numbers are basic features of any language and are commonly used in other Buddhist texts, their absence could be considered a significant characteristic translation pattern of the TD group.

In the center of Fig. 1, there are three documents from different groups which have very similar first component values. Those texts are: T.61 *Shou xin sui jing* 授新歲經; T.50 *Shou sui jing* 受歲經; and T.29 *Xian shui yu jing* 鹹水喻經. The high proximity of T.50 and T.61 stems straightforwardly from the fact that they are translations of the same Indian text. Moreover, the phraseology of T.50 (TD) exhibits various small differences in comparison with the other texts in TD while T.61 (NTD) and T.29 (NTD) contain several translation terms commonly found in the TD group. For example, in contrast to other sutras of the TD group, T.50 does not include the common phrase describing a monk approaching the Buddha, nor does it use 白世尊曰 for 'he addressed the Buddha'. On the other hand, T.29 (NTD) begins with 聞如是，一時婆伽婆在舍衛城祇樹給孤獨園, which is characteristic of TD sutras. T.61 also contains several others terms characteristic of the TD group,

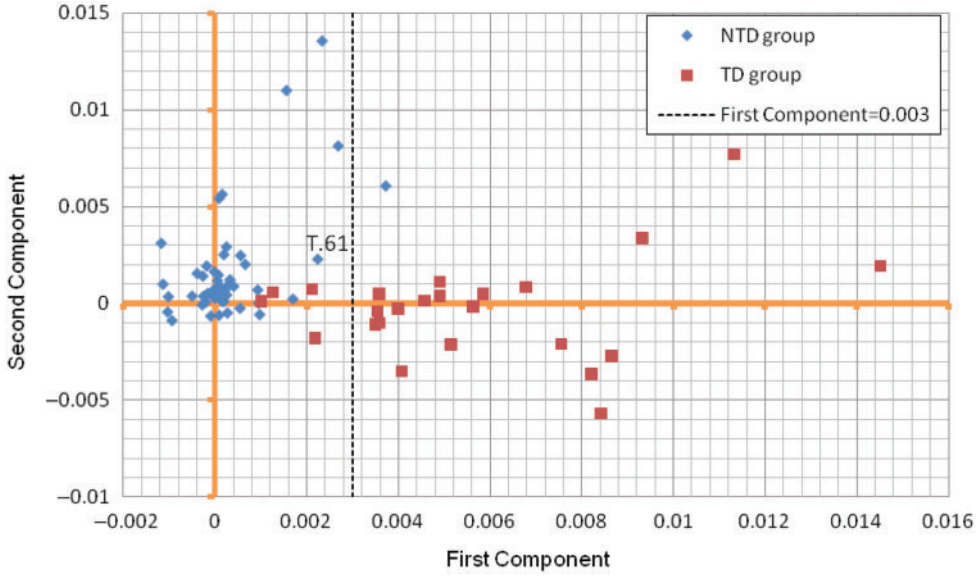

Fig. 2 Scatter plot of the first and second components resulting from a PCA analysis with tri-gram features

including 至世尊所 and 白世尊, which make T.61 a possible candidate for the TD group.

In the results of this analysis we have discovered solid quantitative evidence to support Mizuno's hypothesis. However, since bi-grams are only two characters, the characteristic terms resulting from the analysis are not always semantically meaningful. Perhaps more significantly, due to their shortness and the nature of the Chinese language bi-grams may sometimes be semantically ambivalent, and this ambivalence cannot be accounted for in the analysis. Therefore, in order to enhance the quality of the results by generating more semantically meaningful terms, we conducted a further analysis based on tri-grams.

3.3.2 *PCA analysis result with tri-gram features*

In this analysis, we set the length of grams to 3: i.e. documents are cut into 3-character strings. After applying the *n*-gram feature extraction algorithm described in Section 3.2, the grams appearing in less than 7.2 documents are filtered out, and the remaining 192 significant and commonly used tri-grams constitute the feature set. As is to be expected, as the length of the gram increases, so the cardinality of the feature set decreases, as the number of statistically significant grams is smaller. Figure 2 shows the values of the first and second principle components of the 72 texts under this analysis.

In Fig. 2 it can be seen that the two groups still differ significantly in their first component values, although the division is less clear-cut. Most of the documents in the TD group have a larger first component than documents in the NTD group. Based on the line described by '*First Component* = *0.003*', we can classify the 72 documents into two sets, with a low rate of discrepancy against the expected result. The set described by '*First Component* > *0.003*' consists of 21 texts from the TD group and only one text from the NTD group, giving a misclassification rate of 1/22=0.045. The set described by '*First Component* < *0.003*' contains three texts which come from the TD group and 47 from the NTD group; here the misclassification rate is 3/50 =0.06. The top 20 tri-grams having the largest and smallest first component weights are listed in Table 3.

In comparison with the results from the bi-gram analysis the boundary between the two groups is less

Table 3 Top 20 tri-grams with largest and smallest weight in the first component[9]

Tri-gram	Weight	NTD	TD	Tri-gram	Weight	NTD	TD
世尊所	0.164921	2	19	我聞一	−0.06789	15	0
尊所說	0.157548	0	18	聞一時	−0.06524	14	0
聞世尊	0.156512	0	18	是我聞	−0.06524	14	0
佛如是	0.156292	0	19	一時佛	−0.06465	36	0
歡喜而	0.151162	0	21	白佛言	−0.05598	17	0
喜而樂	0.148293	0	19	時佛在	−0.05527	29	0
如是說	0.147355	3	22	時佛告	−0.05522	10	1
一時婆	0.146176	3	22	詣佛所	−0.04989	10	0
婆伽婆	0.145225	2	20	復如是	−0.04741	12	0
時婆伽	0.145225	2	20	佛告諸	−0.04726	15	1
伽婆在	0.145225	2	20	苾芻眾	−0.0466	9	0
獨園彼	0.135526	0	9	亦復如	−0.04628	11	0
園彼時	0.134716	0	10	諸苾芻	−0.04511	10	0
一面已	0.133612	0	16	佛作禮	−0.04406	9	0
尊曰唯	0.132367	0	9	到佛所	−0.04379	8	0
尊所到	0.131977	1	14	世間人	−0.04376	14	0
已白世	0.13173	0	12	佛言我	−0.04361	11	0
已禮世	0.130612	0	9	為佛作	−0.04258	9	0

distinct, and the texts in the NTD group are more closely grouped and located closer to the origin of the axes of the graph. This is explained by the fact that as the length of gram increases, the number of grams in the feature set will decrease. Moreover, most of grams in the tri-gram feature set are in fact sub-strings of longer translation phrases. As can be seen from Table 3, fully two-thirds of the heaviest tri-grams are from the sentences: 一時婆伽婆在舍衛國祇樹給孤獨園,彼時世-... and 佛如是說,...聞世尊所說,歡喜而樂, and several terms with negative weights are apparently from the pattern 如是我聞,一時佛在. Since the same feature is counted multiple times, the analysis overemphasizes the long translation patterns and reduces the effectiveness of the analysis.

Another interesting observation again concerns T.61, the *Shou xin sui jing* 授新歲經, which belongs to the NTD group. T.61 contains a number of instances of 世尊所 and 白世尊 which are frequently used in TD group and for this reason T.61 is very close to TD group in Fig. 2. However, the term 世尊所 in T.61 is from 至世尊所, describing a monk approaching the Buddha, while in most TD group documents, 世尊所 actually comes from the phrase 世尊所說, introducing a passage spoken by the Buddha. 白世尊 in T.61 is part of the phrase 白世尊言 where the common pattern found in the TD group is 白世尊曰. That both phrases have similar meanings (both indicating the words of the Buddha), and perhaps translate the same Indic construction, only reinforces the significance of the difference in idiom. However the discernment of such differences is beyond the reach of a tri-gram-based approach.

In a bi-gram analysis, even in classical Chinese, semantically meaningful translation idioms will frequently be truncated.[10] This is undesirable for two main reasons: (1) whilst over a large enough sample the effect would be statistically negligible, with smaller data sets there is the potential for significant distinguishing features to be broken into ambiguous sub-strings, as in the example above; (2) the resulting feature set and their corresponding component weights are far less interesting (and convincing) to scholars who work directly with the texts. Using tri-grams mitigates these difficulties in certain respects, and exacerbates them in others. It also yields a feature set containing fewer significant grams, which results in a less distinct analysis as the capacity to differentiate documents is reduced.

For these reasons, a superior approach will abandon reliance on fixed-length grams, and instead employ a variable length feature extraction

Table 4 Count before and after 如是我聞一時佛在 is selected into the feature set

Before Adjustment			After Adjustment		
Gram	**Text1**	**Text2**	**Gram**	**Text1**	**Text2**
如是我聞一時佛在	1	0	如是我聞一時佛在	1	0
如是我聞一時佛	1	0	如是我聞一時佛	0	0
如是我聞一時	1	0	如是我聞一時	0	0
如是我聞一	1	0	如是我聞一	0	0
如是我聞	2	1	如是我聞	1	1
如是我	2	1	如是我	1	1
如是	3	4	如是	2	4

algorithm whilst at the same time seeking to avoid overemphasizing especially long features. The next section proposes just such an algorithm, which is capable of generating improved feature sets for quantitative analysis.

4. Variable Length *n*-Gram Feature Extraction

In order to generate better feature sets for analysis, 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—a string comprising the whole text. Then we remove all non-significant grams from the feature set. Since long grams will contain numerous sub-grams, the feature selection algorithm proposed in Section 3, whereby child grams are selected into the feature set in addition to their parent grams, serves to exacerbate the overemphasis problem. To counter this difficulty, we propose a new variable length *n*-gram counting procedure, in which the principle is that duplicate counts for the same long gram should be eliminated.

Table 4 provides an example to illustrate this idea. The table on the left shows the counts of 如是我聞一時佛在 and its sub-grams for a hypothetical document, as calculated according to the original feature extraction algorithm. Since our long gram appears once, the algorithm counts one appearance for the whole string, and one for each sub-string.[11] If 如是我聞一時佛在 is selected into the feature set, and its count is thereby considered in the analysis, the occurrences of sub-strings which come from the longer gram should not be considered separately. The right side of Table 4 shows the counts after 如是我聞一時佛在 is chosen into the feature set. Notice that since Text2 does not contain any instance of 如是我聞一時佛在, the counts in both tables are the same.

Figure 3 shows pseudo code for the new counting procedure. In Fig. 3, the sub-procedure documentCount, lines 1–7, returns the number of documents containing s_x. Lines 9–23 are a procedure called determineFS, used to determine which grams should be selected into the feature set FS. From lines 10 to 14, we first calculate the number of occurrences of all possible grams across all documents. $C(d,s_x)$ is used to denote the number of times a gram s_x appears in document d. In lines 15–23, we inspect each gram in turn, from long to short, to see whether the gram s_x appears in more than $|D| \times Th$ documents. If the return value of documentCount is larger than $|D| \times Th$, then s_x will be added to FS. In this case, for all sub-grams of s_x, denoted by s_z, we subtract the value of $C(d,s_x)$ from $C(d,s_z)$ to reflect its true significance. The loop in lines 15–23 will be executed until all grams are processed, at which time the procedure is completed.

5. Quantitative Analysis II

5.1 Analysis with Variable Length *n*-Grams

In this section, we apply the variable length *n*-gram extraction procedure proposed in Section 4 to our

```
1  Function documentCount(sx)
2   Count=0;
3   for all document d
4     if C(d,sx)>0 then Count = Count + 1
5   end for
6   return Count
7  end Function
8
9   Sub determineFS()
10   for all possible gram sx
11    for all possible documents d
12      set C(d,sx) to number of occurrence of sx within document d
13    end for
14   end for
15   for all sx sorted by length of S_{x} in descending order
16    if documentCount(sx) > |D| × Th then
17      add sx to FS
18       for all sub-gram s23sx
19          for all possible document d
19            C(d,s2) = C(d,s2) - C(d,sx)
20          end for
21       end for
22    end if
23   end for
24  end Sub
```

Fig. 3 Variable length *n*-gram count calculation procedure

analysis, such that one analysis takes account of all different length *n*-grams.

Figure 4 plots the values of first and second components for the 72 individual sutra translations from the Madhyama Āgama collection. It is immediately apparent that the TD instances are more densely clustered than the NTD instances, a result which is more consistent with Mizuno's hypothesis than any previous analysis result. Moreover, the NTD instances are no longer predominantly grouped near the origin, indicating that this analysis has a higher discriminatory efficacy than our previous analyses. In fact, consideration of the first component alone permits a significant division between the TD and the NTD group. The points belonging to the TD group and the non-TD group can be perfectly separated into two distinct sets by the line described by '*First Component = 0.0035*'.

The feature set created for this analysis consists of 996 variable length *n*-grams. The grams with the 40 highest and lowest weights in the first component are listed in Table 5. In sharp contrast to Tables 2 and 3, only one term is represented more than once (世尊 is also to be found in three of the longer grams), and following the method described above its significance in the feature set is calculated accordingly.

From Table 5 it can immediately be observed that most of the significant features are bi- and tri-grams, but some among the heavy group are longer. Since greater weight for the first component is the factor which most distinguishes one group from the other, the heaviest grams in this

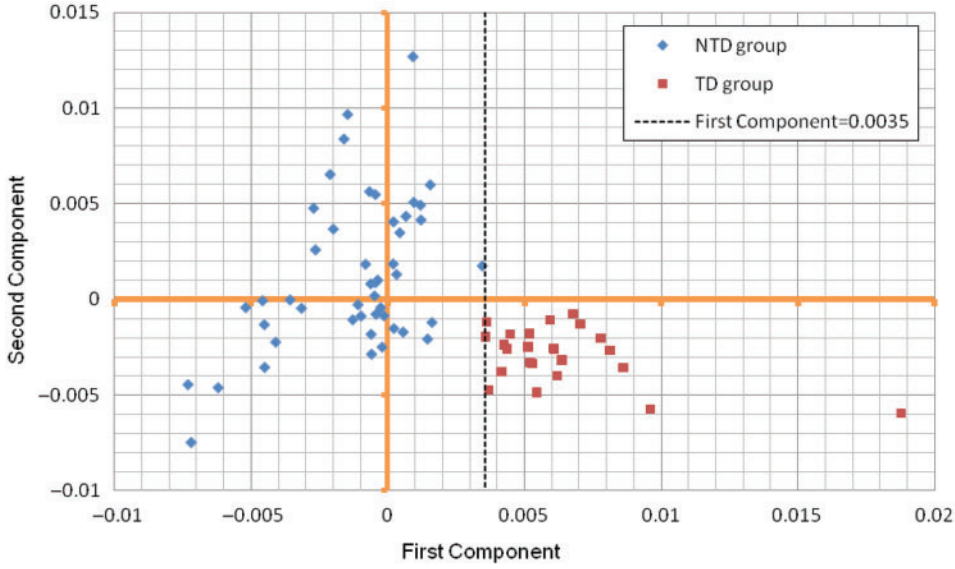

Fig. 4 The PCA analysis result, where the document features are extracted by the variable-length *n*-gram feature extraction algorithm proposed in Section 4

component are the ones most responsible for the neat separation of the groups (i.e. they move the plot-point in Fig. 4 further to the right, for documents in which they are found). They may therefore be said to most characterize the TD group in contrast with the NTD group. The average length of these 40 grams is 3.125, while the average length of the 40 with the lowest weights is just 2.05—only marginally above the minimum gram size. This suggests that the fact that the documents in the TD group bear more similarity to one another than is the case in the NTD group, or across the entire set, is a semantic phenomenon and not simply a statistical co-incidence.

5.2 Hierarchical Clustering

To provide more grouping information regarding the 72 single sutra translations of the MĀ, we apply Average Linkage Clustering, a commonly used hierarchical clustering algorithm, to the PCA analysis result from Section 5.1. In the Average Linkage Clustering algorithm, every document is initialized as the smallest possible group, consisting only of itself, and then the distance between any pair of groups is measured. The function used to calculate the distance between groups is defined as the average Euclidean distance between any pair of members from different groups. More specifically, if $d(A,B)$ denotes the distance between Cluster A and Cluster B, then $d(A,B)$ can be calculated as below:

$$d(A,B) = \frac{1}{n_A n_B}\left(\sum_{p\in A, q\in B}\sqrt{(p.x-q.x)^2+(p.y-q.y)^2}\right) \quad (3)$$

In equation 3, n_A and n_B are the numbers of items in group A and group B, respectively. In each iteration, the pair with shortest distance will be clustered into a new group. The iteration proceeds until all the points are finally clustered into one group. Figure 5 shows the dendrogram of the clustering analysis result.

In Fig. 5, the values on the X-axis are the 72 single-sutra translations from the MĀ, the '(D)' symbol next to a Taishō number indicates the document is one of the documents from the TD group. The value on the Y-axis indicates the distance between two groups when they are combined. Figure 5 shows

Table 5 Top 40 variable length *n*-grams with highest and lowest weight in the first component[12]

Gram	Weight	NTD	TD	Gram	Weight	NTD	TD
世尊	0.08785	18	18	言我	−0.09121	14	1
行至	0.08282	2	12	言人	−0.09121	8	0
却坐一面已	0.08139	0	10	下人	−0.08856	8	0
彷徉	0.07350	0	9	其人	−0.08739	8	0
所到已	0.07152	3	9	如人	−0.08702	8	0
沙門瞿曇	0.07001	5	9	善惡	−0.08589	14	0
聞世尊所說歡喜而樂	0.06776	0	15	佛言	−0.08471	21	0
不久	0.06680	1	7	第四	−0.07933	12	0
有此	0.06674	3	11	第三	−0.07777	14	0
今日始	0.06606	1	7	人皆	−0.07597	8	0
却坐	0.06486	2	6	門道	−0.07515	8	0
如來無	0.06464	3	7	第五	−0.0748	8	0
猶若	0.06460	2	17	人復	−0.07337	7	1
法及	0.06389	3	7	道中	−0.07322	8	0
無所著	0.06340	3	8	以不	−0.07321	10	0
來無所	0.06324	1	7	當自	−0.07317	8	1
是說	0.06288	7	6	上天	−0.07264	9	0
佛如是說	0.06219	0	10	父母不	−0.07234	8	1
於是	0.06187	12	14	第二	−0.07113	14	0
一面已白世尊曰	0.06183	0	10	為人	−0.07064	12	0
涅槃	0.06166	6	9	相見	−0.0692	9	0
見如	0.06144	3	6	道人	−0.06918	10	0
曰此	0.06118	3	9	人從	−0.06904	8	1
及比丘僧	0.06104	2	8	天下	−0.06775	16	1
彼時	0.06054	0	17	何以	−0.06774	7	1
命終	0.05952	8	10	見之	−0.06604	10	1
邪見	0.05946	6	5	者皆	−0.06489	11	3
足却坐一面	0.05926	0	8	惡道	−0.06461	8	0
我所	0.05819	15	11	在世	−0.06451	11	2
世尊所到已	0.05799	1	7	人生	−0.06371	9	2
尊者阿難	0.05762	5	4	其中	−0.06345	8	2
所問	0.05730	4	4	者佛	−0.06341	9	0
知此	0.05723	8	10	人身	−0.06311	7	3
優婆塞	0.05675	2	7	各自	−0.0631	9	1
此瞿曇	0.05632	0	8	生天上	−0.06303	11	1
坐一面	0.05620	4	6	中有	−0.0628	19	4
見此	0.05590	6	5	復還	−0.06235	8	1
有知	0.05579	1	8	所從	−0.06223	6	2
何以故	0.05573	10	16	人言	−0.06123	7	1
聞如是一時婆伽婆在	0.05529	2	12	所作	−0.061	14	4

that the documents in TD group, with the exception of T.93 (discussed below), will cluster into one group, which contains only TD documents. Figure 6 overlays the clustering result on the scatter plot, where the maximum cluster distance is set to 0.007 (as indicated by the dashed line in Fig. 5).

As can be seen in Fig. 6, sutra T.93 is quite distant from the other TD documents. T.93 (the 'Sutra On Heterodox Views' *Xiejian Jing* 邪見經) only consists of 460 characters, making it among the shortest of sutras, and since the subject of T.93 is 'Heterodox Views' (*xiejian* 邪見), it contains many occurrences of this term. Since 邪見 is a gram with a high positive weight, the result is an exaggerated first component value for T.93. Nevertheless, it is clear that T.93 belongs with the texts of the TD group rather than with those of the NTD group.

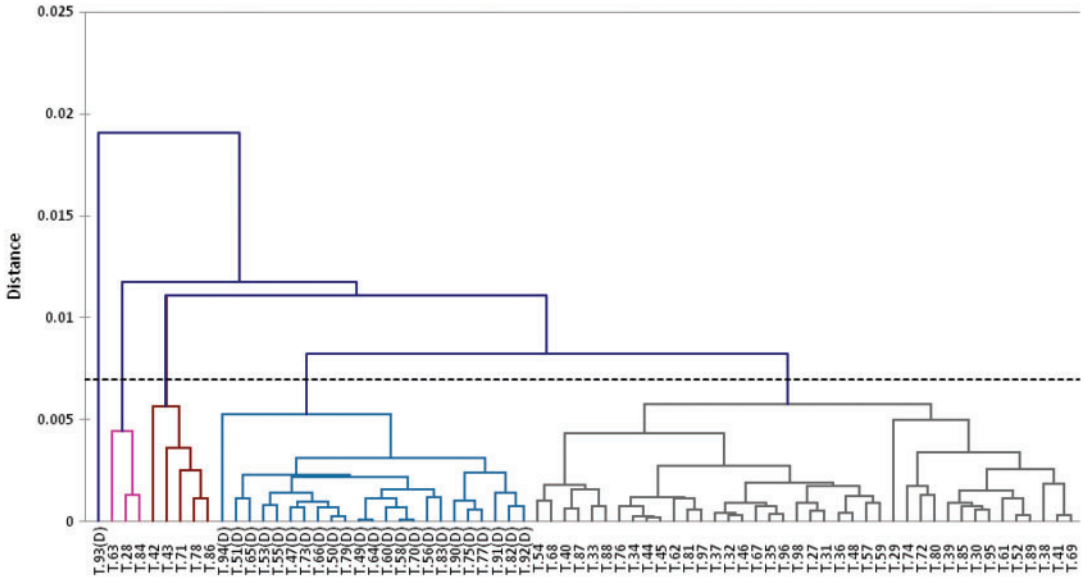

Fig. 5 Dendrogram of clustering analysis result

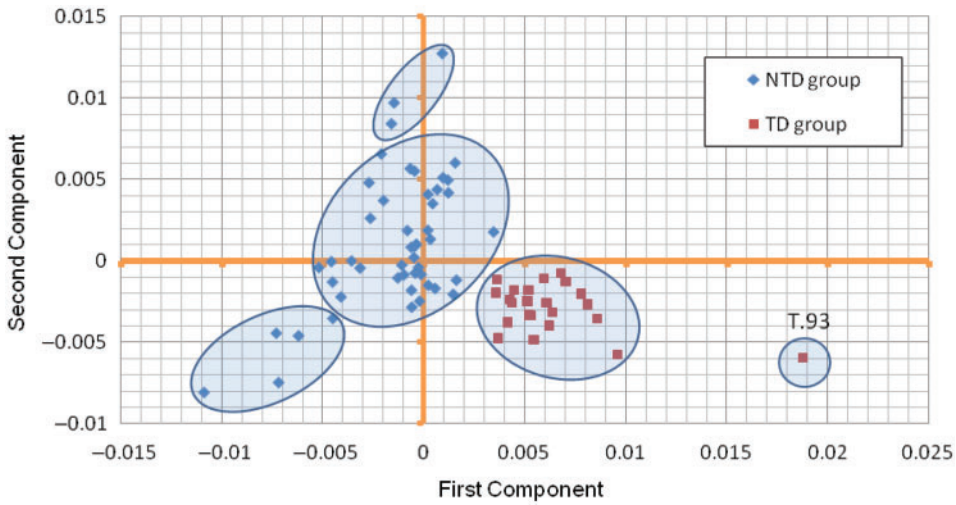

Fig. 6 Clusters of PCA analysis result by using average linkage clustering with the maximum distance between two groups set to 0.007

6. Conclusions and Future Research

In this research, we have analyzed the 72 individual sutra translations from the MĀ in various ways to verify the hypothesis proposed by Mizuno (1969) that 24 of these 72 translations were translated by the same translation team. We employed an *n*-gram feature extraction algorithm for extracting feature sets from the texts, and then performed PCA analysis on these sets. The results provide strong evidence in support of Mizuno's hypothesis.

Furthermore, by examining the components produced by the PCA analysis we discovered significant grams for the TD and NTD groups, which is valuable material for further authorship identification and attribution. In addition, in order to enhance the effectiveness of our analysis, we proposed a new variable length *n*-gram feature extraction algorithm, which considers grams of every possible length for each text. According to our results, the proposed variable length *n*-gram extraction algorithm can generate a superior feature set, which avoids overemphasizing long translation patterns and enhances the quality of the analysis.

From here, we plan to continue our quantitative analysis in order to be able to identify the translator of the 24 documents in TD group.

References

Bingenheimer, M. (forthcoming 2009). Problems and Prospects of Collaborative Edition and Translation Projects in the Era of Digital Text. In Meisig, K. (ed.), *Proceedings of Translating Buddhist Chinese: Problems and Prospects – An International Workshop*. Gutenberg Universität, Mainz 2008.

Bozkurt, I.N., Bağhoğlu, Ö., and Uyar, E. (2007). Authorship Attribution: Performance of Various Features and Classification Methods. In *Proceedings of the 22nd International Symposium on Computer and Information Sciences*. Ankara, Turkey, IEEE press, pp. 1–5.

Burrows, J. (1992). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing*, **7**(2): 91–109.

CBETA (2008). *Chinese Electronic Tripiṭaka Collection CD Version 2008.* CBETA, Taipei.

Grieve, J. (2007). Quantitative Authorship Attribution, an Evaluation of Techniques. *Literary and Linguistic Computing*, **22**(3): 251–270.

Holmes, D. and Forsyth, R. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, **10**(2): 111–127.

Hoover, D.L. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing*, **17**(2): 157–179.

Houvardars, J. and Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. In *Proceedings of 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications,* Varna, Bulgaria.

Iqbal, F., Hadjidj, R., Fung, B., and Debbabi, M. (2008). A Novel Approach of Mining Write-prints for Authorship Attribution in E-mail Forensics. *Digital Investigation*, **5**: S42–S51.

Ishii, Kōsei 石井公成 (2003). 「『大乗起信論』の用語と語法の向 NGSM による比較分析」 The Trend of Terms and Phrasings in the Awakening of Faith: comparison and analysis through NGSM. *Indogaku Bukkyōgaku Kenkyō* 『印度学仏教学研究』 103 (52–1): 202–208.

Kjell, B. (1994). Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing*, **9**(2): 119–124.

Koppel, M. and Schler, J. (2004). Authorship Verification as a One-Class Classification Problem. In *Proceedings of the 21st International Conference on Machine Learning,* Banff, Canada, ACM press, pp. 62.

Labbé, D. (2007). Experiment on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, **14**(1): 33–80.

Liu, W., Allison, B., Guthrie, D., and Guthrie, L. (2007). Chinese Text Classification without Automatic Word Segmentation. In *Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology*, IEEE press, pp. 45–50.

Manevitz, L. and Yousef, M. (2007). One-class Document Classification via Neural Networks. *Neurocomputing*, **70**(7–9): 1466–1481.

Mizuno, K. 水野弘元 (1969). Chū agon kyō kaidai 中阿含經解題. Kokuyakuissai-gyō 國譯一切經, Agon bu 阿含部 **6** Revised Ed.: 403–411.

Mosteller, F. and Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer Verlag, New York.

Onō, G. 小野玄妙 (1936). *Bussho kaisetsu daijiten – Bekkan* 仏教解說大辞典 別卷: *Bukkyō kyōten sōron* 仏教經典總論 [A general study of Buddhist Canonical Literature]. Daitō Shuppansha, Tokyō.

Peng, F., Schuurmans, D., and Wang, S. (2004). Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, **7**(3–4): 317–345.

Peng, F., Schuurmans, D., Keselj, V., and Wang, S. (2003). Language Independent Authorship Attribution

using Character Level Language Models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. ACM press, pp. 267–274.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (1999). Automatic Authorship Attribution. In *Proceedings of EACL '99: Automatic Authorship Attribution*, Bergen, Norway. ACM press, pp. 158–164.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Computer-Based Authorship Attribution without Lexical Measures. *Computers and Humanities*, **35**(2): 193–214.

Tambouratzis, G. and Vassiliou, M. (2007). Employing Thematic Variables for Enhancing Classification Accuracy within Author Discrimination Experiments. *Literary and Linguistic Computing*, **22**(2): 207–255.

Zhang, D. and Lee, W.S. (2006). Extracting Key-substring Group Features for Text Classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining,* Philadelphia, USA, ACM press, pp. 474–483.

Zhao, Y. and Zobel, J. (2007). Searching with Style: Authorship Attribution in Classic Literature. In *Proceedings of the 13th Australasian conference on Computer science*. Ballarat, Australia, ACM press, pp. 59–68.

Zhao, Y., Zobel, J., and Vines, P. (2006). Using Relative Entropy for Authorship Attribution. In *Proceedings of the Third AIRS Asian Information Retrieval Symposium*, Lecture Notes in Computer Science (LNCS) 4182, pp. 92–105, Springer Verlag, 2006.

Notes

1 Ishii (2003) has used an *n*-gram based approach to inquire into the authorship of the *Daqixinlun*.

2 The 'Middle Length Sayings' also exists—under the same name but with slightly different content—in Pāli.

3 The authoritative edition of the Buddhist canon is the Taishō edition (T.), Taishō shinshū daizōkyō 大正新修大藏經, created 1924–1932 under the leadership of Takakusu Junjirō 高楠順次郎 and Watanabe Kaikoku 渡辺海旭. This edition has been improved on and developed digitally by the Chinese Buddhist Electronic Text Association CBETA and all texts are available at http://www.cbeta.org.

4 For the purposes this study, we set this parameter to 10% of total number of documents—i.e. $Th = 0.1$ (see below).

5 In following analysis, the document count threshold (*Th*) is set to 10%, which means that only grams which appear in more than $72 \times 10\% = 7.2$ documents will be used to generate the style features.

6 These and many other inaccurate attributions can mostly be traced to Fei Changfang's 費長房 early catalog *Lidai sanbaoji* 歷代三寶紀 (T.2034) dated 597. Fei attributed many translations of previously unknown provenance. These spurious attributions made their way into the *Kaiyuan Lu* 開元錄 (T.2154), the authoritative catalog for the Tang dynasty Buddhist canon, and from there on down the ages into modern times. Though the problem has been known since the tenth-century, it was never adequately addressed, probably since it would mean doing away entirely with the majority of the attributions for pre-seventh century translations. As a result, most of the attributions found in the Taishō edition for the period before the Sui dynasty should be considered incorrect (Onō 1936, p. 4).

7 A table listing all bi-grams in the feature set can be accessed at http://dev.ddbc.edu.tw/~joey/all_bigrams.pdf.

8 One reason for this is that the texts of the TD group usually refer to the Buddha as 世尊 ('World-honored One') instead of simply as 佛.

9 A table listing all tri-grams in the feature set can be accessed at http://dev.ddbc.edu.tw/~joey/all_trigrams.pdf.

10 Whilst in classical Chinese the semantic unit is usually very small, in Buddhist Chinese, almost a dialect of its own, it is frequently longer on account of such phenomena as the frequent use of transcribed technical terms.

11 For the sake of example, the string 如是我聞 is assumed to occur once in Text1 outside the long gram, and 如是 a further time outside of that, and only the sub-strings which have the same start point are shown in the table.

12 A table listing all *n*-grams in the feature set can be accessed at http://dev.ddbc.edu.tw/~joey/all_ngrams.pdf.