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

Marcus BINGENHEIMER[*], Jen-Jou HUNG[**], Cheng-en HSIEH[**]

Abstract

Below we develop a method to determine whether the use of grammatical particles in Chinese Buddhist scriptures is characteristic for the period of their translation. The corpus consists of three different Chinese translations of an early Indian Mahāyāna text from two different periods. We use the results of Principal Component Analysis (PCA) to discern if the samples of texts from different periods cluster together.

We found that PCA used on common grammatical particles that exhibit great variance between texts, but occur evenly distributed within the texts, does in this case indeed yield distinctive patterns that distinguish the translations from different time periods. This is relevant for historical Chinese linguistics and Buddhist studies. It allows us to identify grammatical particles the use of which differs between translations from different periods. This in turn is an important basis for further research into Buddhist Hybrid Chinese translation idioms and the better attribution and dating of Chinese Buddhist texts.

1. The Texts

The *Gaṇḍavyūha*[1] is an important Mahāyāna text which relates the journey of the merchant's son Sudhana, who visits 52 teachers in his quest for enlightenment.[2] As usual for the genre we do not have a date or single author for the *Gaṇḍavyūha*, which started to evolve in Sanskrit probably in Central Asia around the 3rd century CE and was translated

[*] Temple University, Philadelphia, USA
[**] Dharma Drum Institute of Liberal Arts, Taiwan

1 Chinese in most versions *Rufajie* 入法界 (**Dharmadhātu-praveśana*). On the name *Gaṇḍavyūha* and related terms, see Osto (2008). On the history of the text and its study, in the context of the *Buddhāvataṃsaka sūtra* see the contributions in Hamar (2007).

2 For an English translation of the G-Tang-1 version see Cleary (1993). Verses from the *Gaṇḍavyūha* have also been translated in careful comparison with the Sanskrit and Tibetan versions in Gomez (1967).

into Chinese and Tibetan.[3] There are three complete translations of the *Gaṇḍavyūha* into Chinese:

- As part of the *Buddhāvataṃsaka* (T. 278) translated by Buddhabhadra 佛陀拔陀羅 et al. in Chang'an between 418 and 420 CE in the final years of the Eastern Jin dynasty (265-420). (Siglum below: **G-Jin**)
- As part of the *Buddhāvataṃsaka* (T. 279) translated by Śikṣānanda 實叉難陀 et al. in Chang'an between 695 and 699 CE at the time of the Tang dynasty (618-907).[4] (**G-Tang-1**)
- A later, enlarged, Sanskrit version of the *Gaṇḍavyūha* section was transmitted to China in the 8th century and translated as an independent text (T. 293) by Prajña 般若 et al. again in Chang'an between 796 and 798 CE. (**G-Tang-2**)

The three full translations represent different stages in the development of the Sanskrit text and differ in length, but agree in general as to their content and structure.

There is also an earlier, shorter translation of the *Gaṇḍavyūha* (**G-Qin**, T. 294) by the monk Shengjian 聖堅 (fl. 385-422). According to the traditional sutra catalogs Shengjian translated G-Qin between 385-388 in the North-West, during the Western Qin.[5] We will bring G-Qin, which is close in time to G-Jin, into play only later, once we have trained our algorithm on the three full translations.[6] If our method would show that G-Qin and G-Jin are related, it would be an indication that it does distinguish features due to the time of translation.

2. Research Question and Approach

Grammatical information in Chinese is expressed through syntax and with the help of a limited number of grammatical particles (*xuzi* 虛字), the use of which changed considerably over time.[7] Our question is: Is there stylometric evidence that the use of grammatical particles in the two Tang versions (G-Tang-1, G-Tang-2) differs from that of the G-Jin version and can the algorithm help to identify which particles make a

[3] The main witnesses for the Sanskrit text we have today are based on Nepalese Manuscripts and might thus show the text at a later stage than the Chinese or even the Tibetan versions. The Tibetan (Derge No.44) was translated in the early 9th century. Two short (Buddhist-) Sanskrit fragments that were found in Central Asia, were published in Hori (2002).

[4] Strictly speaking, this translation was done under empress Wu Zetian's short-lived Zhou dynasty, but we will forgo this distinction here.

[5] See T.55.2146.119c14 and T.55.2149.254c11.

[6] See Gómez (1967: xxiv) for some differences between the abridged early version and the later translation. Another *Gaṇḍavyūha* text in the canon (T.295) is merely a fragment and cannot be used here.

[7] See e.g. the three volumes by Dobson (1959, 1962, 1964) that demonstrate the transition from Early Archaic, to Late Archaic and on to Late Han Chinese by analyzing the changing use of the characters used as particles. For the issue of two-character particles see below.

difference? Can it be shown that the two Tang dynasty translations are stylistically more closely related to each other than to G-Jin? And if yes, which particles are responsible for this difference?

Apart from being influenced by the period in which they were produced, the wording of translations obviously depends on the individual style of the translator or the translation team. Traditionally, the language of G-Jin is considered elegant and relatively accessible, and because it represents an early form of the text, it is shorter than the Tang versions. G-Tang-1 is supposedly less easy to read as it retains in places the syntax of its Sanskrit original (Fang 1981: 10 *f*). The style of the longest version of the *Gaṇḍavyūha*, G-Tang-2, is considered more fluent (Wen 2000: 19). Content-wise, G-Jin and G-Tang-1 are very close, while G-Tang-2 was translated from a considerably expanded Sanskrit text (Gómez 1967: xxvi-xxviii).

Traditionally, researchers have studied different translation idioms through close, comparative reading of the texts. For this they either have to limit themselves to a small number of texts or compare only a small number of features.[8] Stylometric analysis can attempt to address longer texts, without the need to sample. Crucially, it is now possible to test a large amount of features, orders of magnitude beyond the reach of the unaided human reader.[9] Two factors must be decided beforehand. First, one has to settle on what type of style markers or features to use; second, which method of comparison to apply. Regarding the latter, current approaches typically combine statistical analysis with machine learning techniques. Regarding the former, generally we try to chose features that appear frequently ("most common words" (Burrows 1992), or "very frequent words" (Hoover 2002)). This works well in many cases, but unfortunately the structure of literary Chinese, whether in its Buddhist or more "classical" variants, is ill suited for a choice of "very frequent words." In East Asian writing words were traditionally not separated by spaces.[10] Chinese word-segmentation is a difficult subject even for modern Chinese with its better understood dictionary and grammar.[11] The difficulty to disambiguate phrase level compounds and words has been called "one of the most vexing problems in modern Chinese grammar" (Norman 1988: 156).

---

8 This form of research has produced excellent results for a limited number on texts. Nattier (2008), for instance, gives an overview of the work on the translatorship (de-)attribution for c.50 texts that were translated into Chinese before 280 CE. The arguments have to rely heavily on the catalogue tradition as well as on the comparative use of a relative small number of phrases.

9 The possibilities of stylometric analysis for new discoveries in the field of Buddhist textual scholarship have attracted attention already more than ten years ago when Wan Jinchuan (2002: 73 ff) expressed his hope that stylometrics can decide the doubtful attribution of some Chinese texts to Nāgārjuna.

10 For Chinese and Japanese this is still the case. Vietnamese started to put spaces after each syllable after it adopted the Latin script. Only modern Korean writing, and only since about the seventies, has adopted word segmentation by spaces.

11 This impacts other operations such as PoS parsing or even relatively basic Named Entity Recognition (Bingenheimer 2015).

It has therefore long been held that non-contextual words or phrases are better features for stylometric analysis than words or phrases that human readers tend to focus on. Style is encoded in the use of features such as grammatical particles, basic verbs or even punctuation rather than in the literary use of rare, but "stylish", nouns and adjectives.[12] Our approach here focuses on grammatical particles (*xuzi* 虛字), which, next to syntax, are the main carriers of grammatical information in classical Chinese.

Apart from deciding on which features to process, stylometrics needs a comparison method that allows to decide on and measure the similarity of features in a multi-dimensional vector space. This type of clustering or classifying problem is currently most often approached with machine learning algorithms, which are either supervised or unsupervised (Stammatos 2009). In supervised methods, such as Bayesian classification (Bozkurt et al. 2007), Support Vector Machines (Diederich et al. 2003), Neural Networks (Tearleet et al. 2007) and others, the algorithm is trained on a set of already marked or classified data. With unsupervised methods, such as Principal Component Analysis (PCA) (Binongo and Smith 1999) or other forms of cluster analysis (Labbé and Labbé 2001), there is no training dataset with pre-classified samples. If the results of the analysis turn out to confirm a hypothesis of how the texts would cluster these types of analysis amount to strong evidence for the hypothesis, because the clustering has not been influenced by pre-conceived categories in a set of training data.

In this study we are using Principal Component Analysis to analyze and compare the translation style of the three *Gaṇḍavyūha* translations based on their use of grammatical particles. Do we see a difference in the use of particles between the G-Jin, G-Tang-1 and G-Tang-2? And do these differences align with the translation date? PCA is often used for exploratory data analysis. It reduces the complexity of a multivariate dataset, allowing an informative lower-dimensional view of the data emphasizing variance in the data.[13] Based on the variables from the original data PCA determines a number of components, which are sorted in order of variance, the first component being the one which expresses maximal variance between all variables. Plotting the values of first and second components in 2D charts we are able to visualize the distance between different translations and see whether G-Tang-1 and G-Tang-2 are closer to each other and further from G-Jin, as the dating suggests.

To produce enough units for statistical analysis, each translation is divided into units of equal length and the frequencies of grammatical particles in these divisions are calculated relative to text-length. Sub-dividing the texts we can use PCA to explore their relative distance of the various units according to weighted features.[14] Since characters

[12] Mosteller and Wallace 1984; Zhao and Zobel 2007.

[13] For a comprehensive description of PCA see Jolliffe (2010).

[14] We have applied PCA before for a different, but related problem in the same domain (Hung, Bingenheimer, Wiles 2010).

that serve as grammatical particles can also appear as part of names or doctrinal terms, such cases must be filtered out as far as possible based on a domain-specific dictionary.

3. Dataset description

For our stylometric analysis on the three different translations of the *Gaṇḍavyūha* (G-Jin, G-Tang-1, G-Tang-2), we use the digital text as contained in the CBETA (ver. 2011) corpus.[15] We remove all markup, punctuation, front- and back-matter, as well as all headings. Table 1 shows the character count for the texts after clean-up.

Table 1 Character count for G-Jin, G-Tang-1 and G-Tang-2

	Number of fascicles (juan 卷)	Number of characters
G-Jin	17	143,957
G-Tang-1	21	169,122
G-Tang-2	40	250,452

4. Text Analysis Procedure

4.1 Procedure: Step 1 – Reducing the list of particles

Chinese characters can often serve in various grammatical functions as verb, noun, grammatical particle, part of a compound (term with two or more characters) etc. This is especially true for Classical Chinese, a written idiom, where compounds are much rarer than in the vernacular. In Buddhist Hybrid Chinese, on the other hand, compounds of two or more characters are quite common, especially for names and doctrinal terms. The grammatical particles that we are interested in can also appear as part of such compounds which would distort the analysis. We therefore filter out all instances where particles appear in the compounds that appear as entry in the *Dictionary of Chinese Buddhist Terms* (Soothill and Hodous 1937 [1994]). I.e, for instance, that the character 如 will not be counted as particle when it appears as part of 如來 (for Skr. *tathāgata*), or 如意 (meaning "at will" etc.).

The initial list of 228 grammatical particles we start with is adapted from Wang (2007). Running PCA with all 228 particles would give too many variables for effective PCA weighing. Also, many particles from a generic list are redundant in cases that compare only a small number of texts. Most are not relevant because they simply do not

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

appear in the texts at all, others appear but not as particles. However, the list of possible candidates can be refined even further. Since we are looking for *variance* in particle use, the best list of particles contains all those particles that are used relatively frequently across all texts, and at the same time have maximum variance between texts. The algorithm must therefore select *the most frequent particles, with the highest (inter-textual) variance*. After applying the dictionary filter mentioned above, we use the following formula to optimize our list of particles.

Let F_j^ibe the frequency with which a particle j appears in a text S_i (disregarding those instances filtered out with the help of the dictionary). F_j^i can be expressed as:

$$F_j^i = \frac{WC_j^i}{L_i} \quad (1)$$

Where WC_j^i (word count) is the number of instances a particle j appears in text S_i, and L_i the total number of characters in S_i. From this follows that the average frequency$\overline{F_j}$across the set of texts *S* can be expressed thus:

$$\overline{F_j} = \frac{1}{n}\sum_{i=1}^{n} F_j^i \quad (2)$$

Lastly we can calculate the *inter-textual variance* $diff_j$ for a particle j based on its frequencyF_j^i and average frequency$\overline{F_j}$as below:

$$diff_j = \sum_{i=1}^{n}\left|F_j^i - \overline{F_j}\right| \quad (3)$$

We start our analysis by selecting the 30 particles with the highest $diff_j$ value. From this algorithmically generated list we eliminate certain characters which in our case are not used as particles.[16] On the other hand we also decided to add some synonyms which were known to appear in the text as synonyms to characters in the list, but did have a lower variance value.[17] Finally we arrive at a list of 33 characters, which appear as particles in all texts and display a high variance between texts.

Table 2 Frequency, average frequency and variance value of each character

Particle 虛字	G-Jin(A)	G-Tang-1(B)	G-Tang-2(C)	$\overline{F_j}$ (D)	$diff_j$(E)

16 This is an intervention based on domain knowledge. Originally the list contained, for instance, 見 which in some periods was used as passive marker, or 能 which in combination with other characters sometimes is used as particle (e.g. as auxiliary of number s. Dobson (1974, sub voc.)), rather than part of the verb phrase. In the G translations both are simply used as verbs.

17 For example we included 說 and 言 which can mark direct speech similar to 曰 and 謂, which were on the original list.

其	0.3536%	0.7302%	0.5873%	0.5571%	0.4069%
彼	0.5703%	0.3583%	0.3498%	0.4261%	0.2884%
諸	1.3678%	1.1879%	1.1619%	1.2392%	0.2572%
所	0.5300%	0.7030%	0.7063%	0.6465%	0.2329%
於	0.6891%	0.8692%	0.8345%	0.7976%	0.2170%
而	0.2577%	0.4547%	0.3669%	0.3598%	0.2041%
如	0.3348%	0.4665%	0.4835%	0.4283%	0.1869%
或	0.2272%	0.3707%	0.3182%	0.3054%	0.1564%
以	0.3175%	0.4322%	0.4272%	0.3923%	0.1497%
無	0.2452%	0.2874%	0.3773%	0.3033%	0.1480%
者	0.4946%	0.3631%	0.4504%	0.4360%	0.1459%
說	0.3480%	0.4565%	0.3526%	0.3857%	0.1416%
爾	0.2383%	0.1537%	0.1282%	0.1734%	0.1298%
故	0.5724%	0.6640%	0.6612%	0.6325%	0.1203%
時	0.5210%	0.4512%	0.4109%	0.4610%	0.1200%
曰	0.1292%	0.0509%	0.0299%	0.0700%	0.1184%
是	0.1702%	0.2495%	0.2452%	0.2216%	0.1029%
此	0.4481%	0.5268%	0.4564%	0.4771%	0.0995%
一	0.2515%	0.1839%	0.1944%	0.2099%	0.0831%
為	0.6676%	0.7385%	0.7199%	0.7087%	0.0822%
常	0.1014%	0.1443%	0.1745%	0.1401%	0.0773%
已	0.2126%	0.2667%	0.2104%	0.2299%	0.0736%
亦	0.1952%	0.2560%	0.2380%	0.2297%	0.0691%
當	0.0820%	0.1431%	0.1206%	0.1152%	0.0665%
謂	0.0507%	0.0857%	0.1150%	0.0838%	0.0662%
猶	0.0493%	0.0982%	0.0986%	0.0820%	0.0654%
不	0.4585%	0.4914%	0.5211%	0.4903%	0.0637%

之	0.2855%	0.3447%	0.3162%	0.3155%	0.0600%
言	0.1917%	0.2371%	0.2352%	0.2213%	0.0592%
及	0.1473%	0.1809%	0.2012%	0.1765%	0.0584%
由	0.0104%	0.0313%	0.0643%	0.0353%	0.0579%
若	0.1306%	0.1803%	0.1657%	0.1589%	0.0566%
又	0.1368%	0.1206%	0.0894%	0.1156%	0.0524%

The first three columns A, B, C in Table 2 list the frequency of the character in the text. D is the average of A, B, and C. The table is sorted according to the variance value in column E which is calculated as *abs(A-D)+abs(B-D)+abs(C-D).*

4.2 Procedure: Step 2 – Dividing the texts

In order for the statistical analysis to work, we have to divide each of the three translations in smaller units to be able to use a larger number of samples for PCA. PCA is sensitive to the scaling of the units and the number of units is critical for the quality of the results. As of yet we have no exact method to determine the optimal number of units and therefore have to experiment in order to see which amount of units yields the clearest results.

4.3 Procedure: Step 3 – First Principal Component Analysis (PCA) results

We are looking for a set of variables that causes units from the same texts to cluster together. This would indicate that there is a characteristic difference in the occurrence of grammatical particles between the texts. If there is no clear distinction in the way particles are used in the Jin vs. the Tang dynasty one would expect that no set of particles would result in a clearly discernible difference between texts. For the first round of PCA we run the analysis with the full list of 33 particles for three different numbers of units for each text, splitting the three translations in 40, 20 or 10 units each. Plotting the first and second component of the result we arrive at the following (Figs. 1, 2, 3).[18]

18 The variances on the x and y axes are simply the output of the PCA eigenvector transformation, they do not have a semantic unit beyond their numeric value.

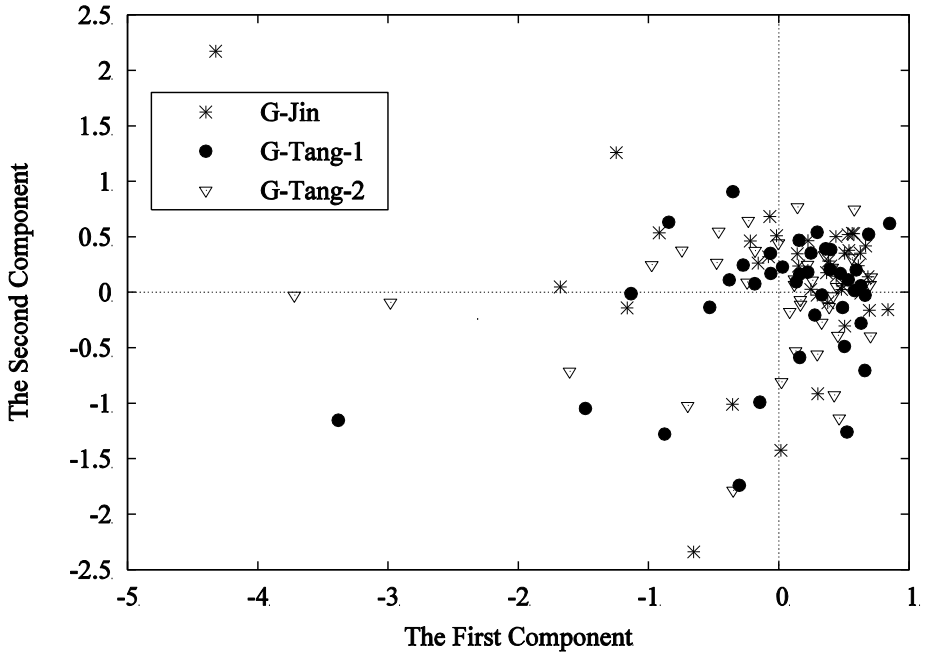

Fig. 1 PCA result for all 33 particles of Table 2, 40 units per text (= total 120)

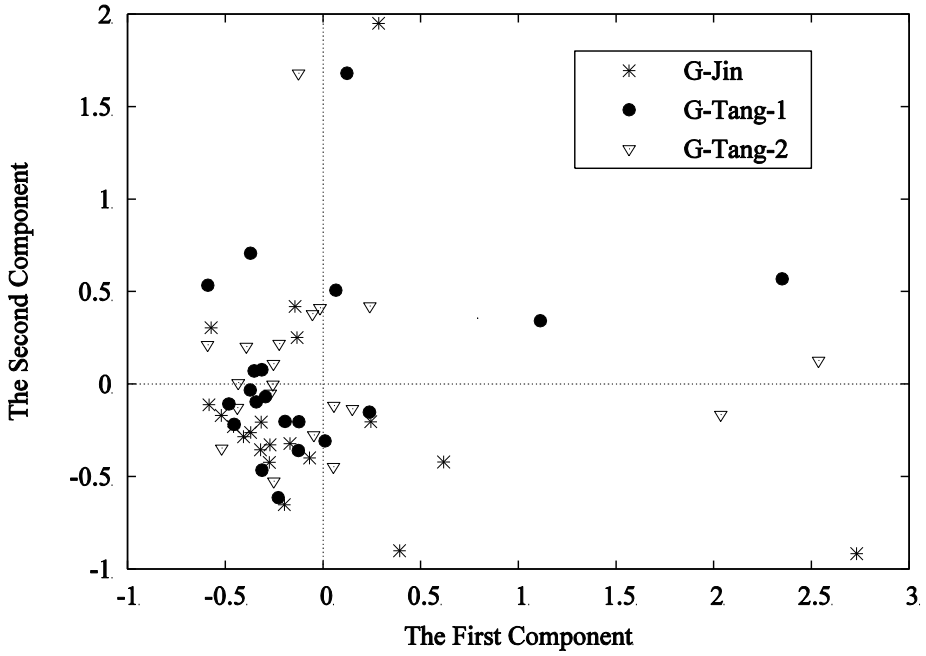

Fig. 2 PCA result for all 33 particles of Table 2, 20 units per text (= total 60)

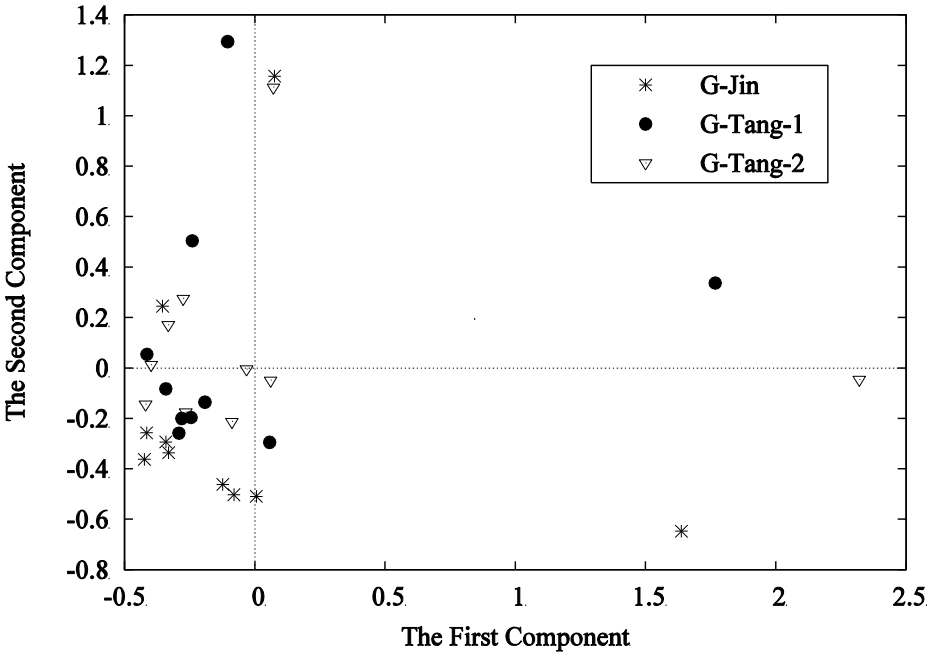

Fig. 3 PCA result for all 33 particles of Table 2, 10 units per text (= total 30)

None of the three experiments results in a distinct clustering of the translations, but we can discern certain trends:

1. The result of the first experiment (40 units per text) (Fig. 1) shows all units clustered together in a way that suggests that PCA cannot cope with the number of units. 120 appears to be above the limit that allows the analysis to result in a distinct clustering and dividing the texts in 40 units each is therefore not going to work.

2. In Experiment 2 (Fig. 2) each text is divided in 20 units and G-Jin seems to start forming an independent cluster but the clustering is still indistinct.

3. Only in Experiment 3 (Fig. 3) the PCA clusters the majority of G-Jin units (7 out of 10) in the same region of the plotted graph (lower left). Dividing the texts in 10 units each seems to yield clearer clusters and this number is used for the following steps.

Although Experiment 3 appears to cluster G-Jin, giving an indication that its use of the 33 particles differs from the two Tang dynasty translations, the majority of units are still close together, the clustering is indistinct, and there are two sets of outliers that distort the result. Having tuned the number of units to work with, the next step is to test what set of *high frequency / high (inter-textual) variance* particles yields the greatest possible differentiation between the texts.

4.4 Procedure: Step 4 – Optimizing the particle list I

In order to better understand the influence of the length and the composition of the particle list on the analysis we start from a shorter list of 10 particles (see Table 2) (still sorted for variance value) and progressively increase the length.

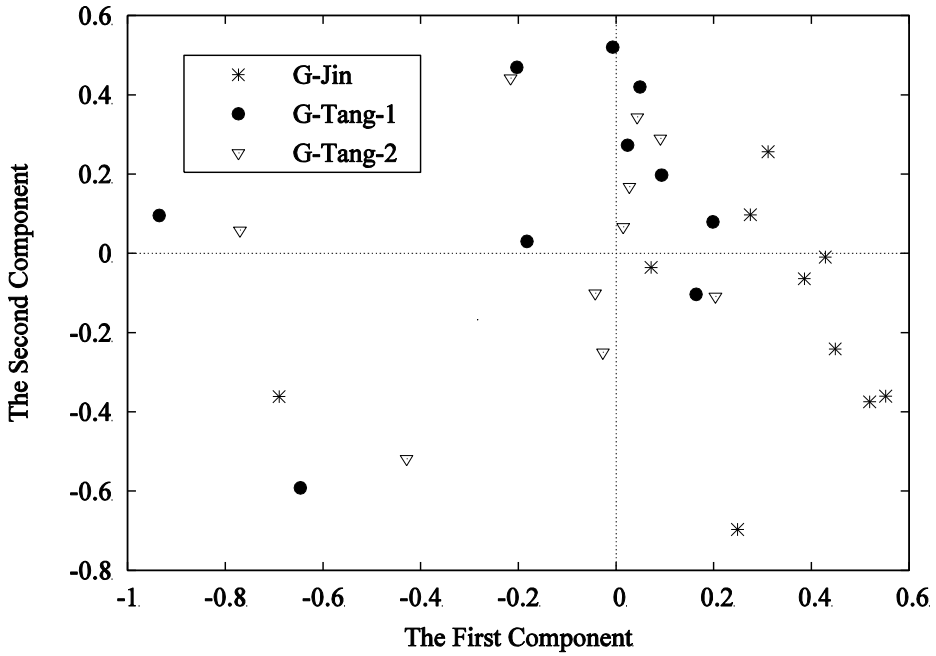

Fig. 4 PCA result for the first 10 particles (其彼諸所於而如或以無), 10 units per text

Fig. 4 shows that decreasing the number of particles used in the analysis results in a much clearer distinction between G-Jin on the one hand (right side of the graph), and G-Tang-1 and G-Tang-2 on the other. Also, as one would expect if the particle use was characteristic for the period, the samples of the two Tang versions are closely mingled and do not show independent clustering that would be indicative of variance in particle use.

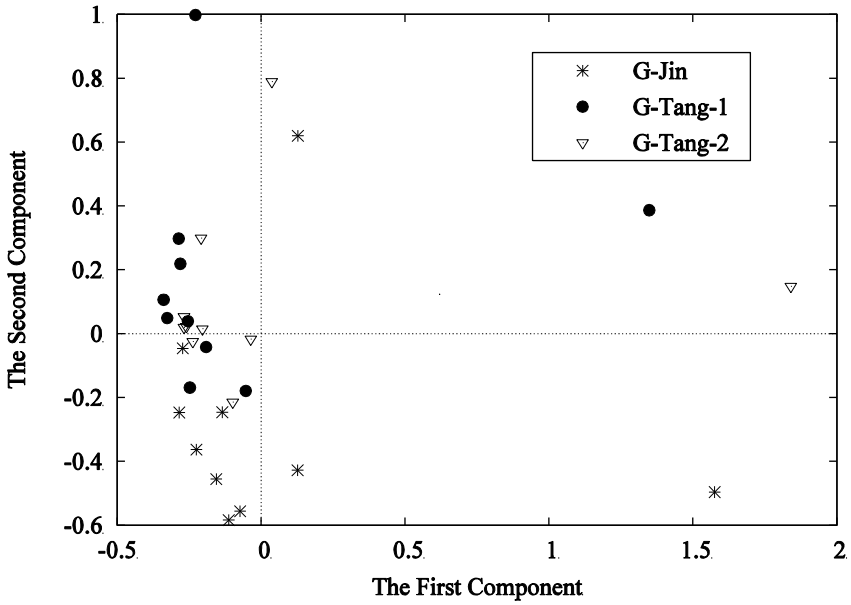

Fig. 5 PCA result for the first 11 particles of Table 2 (其彼諸所於而如或以無者), 10 units per text

In experiment 5 the addition of the 11th particle '者' results in unexpected changes. In Fig. 5, Distances between units of all translations have decreased though the clustering of G-Jin is still discernible (lower left). It seems that the addition of 者 has strengthened the two sets of outliers and distorts rather than helps the analysis (we will see how and why below).

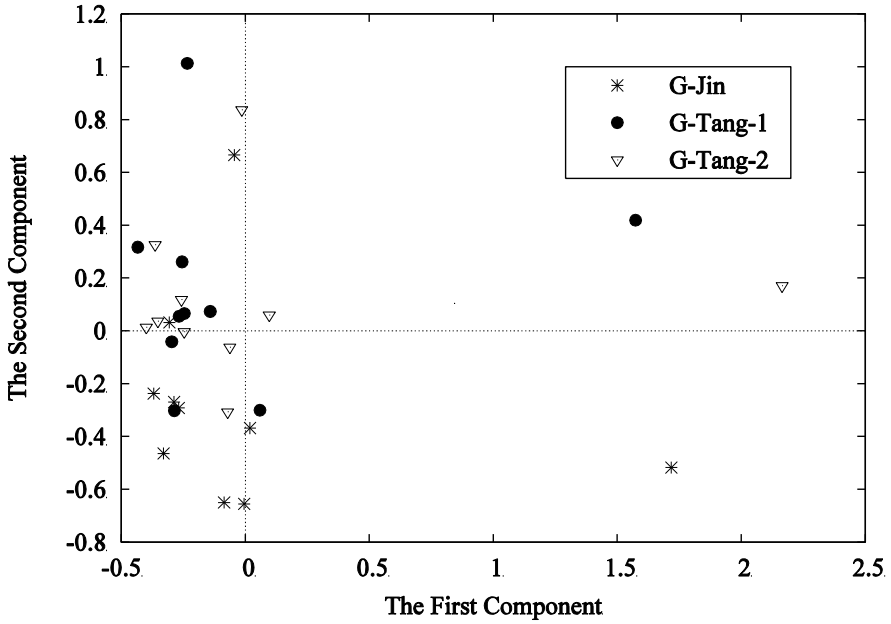

Fig. 6 PCA result for the first 14 particles of Table 2 (其彼諸所於而如或以無者爾故時), 10 units per text

Fig. 6 shows that the addition of 爾, 故, and 時 does not change the power of the outliers, if anything, the clustering has become more indistinct.

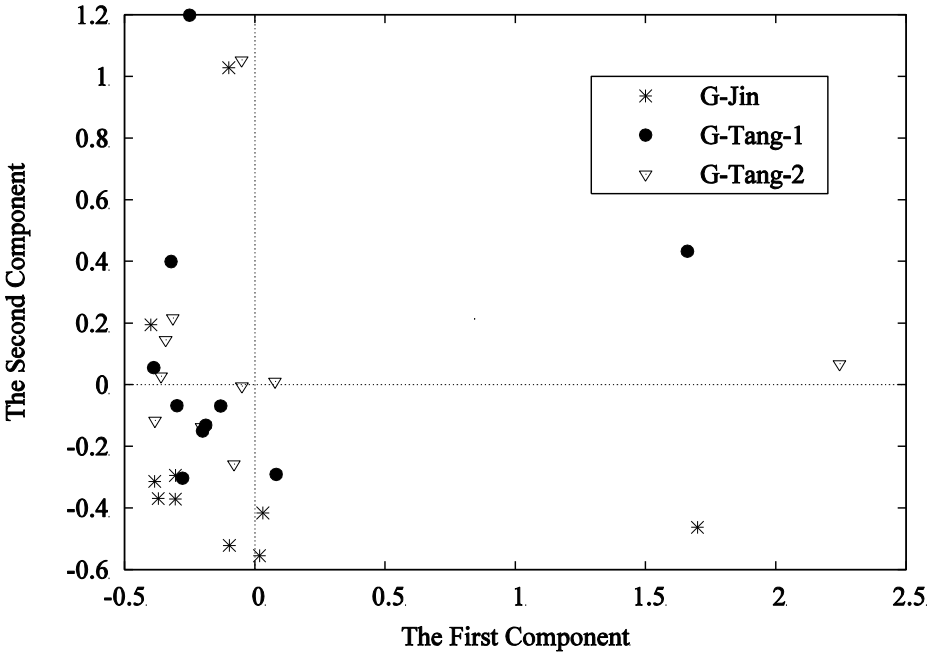

Fig. 7 PCA result for the first 25 particles of Table 2 (其彼諸所於而如或以無者爾故時曰是此一為常已亦當謂猶), 10 units per text

In Experiment 7 the number of particles is raised to 25 but the result does not become clearer. In Fig. 7, although most G-Jin units cluster clearly in the lower left corner, the two outlier sets still determine the distribution.

This picture does not change significantly if the final 7 particles are added and the analysis is done for all 33 particles on the list (i.e.: 其彼諸所於而如或以無者爾故時曰是此一為常已亦當謂猶不之言及由若又說) (see Fig. 3).

We will now further investigate why the addition of the 11th particle on the list has reduced the distinctiveness of the analysis. As it turns out, of all particles *zhe* 者 has the *highest internal variance* among the 10 units, i.e. it appears disproportionally more often in some units. This leads us to suspect that the usage of 者 is determined not mainly by syntactic "style", but by semantics, i.e. it is used more frequently in some sections of the sutra because there are content-related reasons to do so.

Table 3 lists the frequency of particle 者 in different units. We can easily identify in which unit of the three translations 者 appears disproportionally often.

Table 3 Frequency of particle 者 in different units

	G-Jin	G-Tang-1	G-Tang-2
Unit 1	0.1459%	0.1360%	0.0958%
Unit 2	0.2570%	0.1892%	0.2196%
Unit 3	0.2570%	0.2188%	0.2196%
Unit 4	0.3682%	0.2661%	0.3434%
Unit 5	0.2987%	0.3370%	0.2436%
Unit 6	0.6947%	0.3311%	0.2995%
Unit 7	0.3334%	0.3370%	0.2476%
Unit 8	0.2987%	0.1951%	0.2915%
Unit 9	1.9799%	1.4073%	2.0882%
Unit 10	0.3125%	0.2128%	0.4551%

On further investigation it quickly becomes obvious why the particle 者 appears so often in Unit 9. Table 4 lists sample excerpts illustrating the use of 者 in Unit 9. It shows that 者 is used frequently in the compound *putixinzhe* 菩提心**者** ("one who has engendered the wish for enlightenment (*bodhi-citta*)"), a term that is repeated more than hundred times in one single fascicle in all translations.[19]

Table 4 Sample excerpts illustrating the use of 者 in a compound in Unit 9

G-Jin	「汝為法器，善根潤澤，長清白法，淨勝欲性，為善知識之所總攝，諸佛護念。何以故？**菩提心者**，則為一切諸佛種子，能生一切諸佛法故；**菩提心者**，則為良田，長養眾

[19] 118 occurrences in G-Jin (Fasc. 59), 121 occurrences in G-Tang-1 (Fasc. 78), and 125 occurrences in G-Tang-2 (Fasc. 35).

	生白淨法故；**菩提心者**，則為大地，能持一切諸世間故；**菩提心者**，則為淨水，洗濯一切煩惱垢故；**菩提心者**，則為大風，一切世間無障礙故；**菩提心者**，則為盛火，能燒一切邪見愛故；**菩提心者**，則為淨日，普照一切眾生類故；**菩提心者**，則為明月，諸白淨法悉圓滿故；**菩提心者**，則為淨燈…..
G-Tang-1	汝身是善器，為諸善根之所潤澤。汝為白法之所資持，所有解欲悉已清淨，已為諸佛共所護念，已為善友共所攝受。何以故？善男子！**菩提心者**，猶如種子，能生一切諸佛法故；**菩提心者**，猶如良田，能長眾生白淨法故；**菩提心者**，猶如大地，能持一切諸世間故；**菩提心者**，猶如淨水，能洗一切煩惱垢故；**菩提心者**，猶如大風，普於世間無所礙故；**菩提心者**，猶如盛火，能燒一切諸見薪故；**菩提心者**，猶如淨日，普照一切諸世間故；**菩提心者**，猶如盛月，諸白淨法悉圓滿故；**菩提心者**，猶如明燈….
G-Tang-2	汝身即是真善法器，為諸善根之所潤澤。汝為白法之所資持，信樂廣大，慧解清淨，已得諸佛之所護念，已為善友共所攝受。何以故？謂能發大菩提心故。善男子！**菩提心者**，猶如種子，能生一切諸佛法故；**菩提心者**，猶如良田，能長眾生白淨法故；**菩提心者**，猶如大地，能持一切諸世間故；**菩提心者**，猶如大水，能滌一切煩惱垢故；**菩提心者**，猶如大風，普行世間無所礙故；**菩提心者**，猶如大火，能燒一切諸見薪故；**菩提心者**，猶如淨日，普照一切諸世間故；**菩提心者**，猶如盛月，普能圓滿白淨法故；**菩提心者**，猶如明燈…..

The use of 者 here does not reflect preferences in a certain translation style, but is an artifact of the content of the sutra, which in this fascicle contains a description of those "who have engendered *bodhi-citta* 菩提心者." Terms like these should in principle be filtered out by the dictionary before the analysis, but of course no dictionary is complete. The character 者 must in this case be excluded from the list on account of its high *internal variance*. The reason for highlighting this example here in detail is, because it drew our attention to the issue of 'internal' variance in frequency between different units. Are there perhaps other characters that evince an uneven distribution between units and therefore are distorting the PCA analysis by creating outliers? Taking the particles with the highest internal variance off the list should reduce distortion.

4.5 Procedure: Step 5– Optimizing the particle list II

In order to do so we have to re-order our particle list. By $F_j^{(i,k)}$ we express the frequency of a particle j in the unit k of text s_i. We use the following formula to express the *internal (intra-textual) variance* $diff_j^i$ of a particle *j* between units of a text s_i.

$$diff_j^i = \sum_{k=1}^{10} \left| F_j^{(i,k)} - F_j^i \right| \quad (4)$$

Here, like in previous sections, F_j^i is the frequency with which a particle *j* appears in a text s_i (see Formula 1). Table 5 lists the particles sorted according to their internal variance. Columns A, B, and C list the internal variance for a particle between units of one text, column D is the sum of their absolute values. 者, not surprisingly, tops the list, but other particles such as 故，為，or 諸 also display high internal variance.

Table 5 Particles sorted according to intra-textual variance

Particle	G-Jin internal variance (A)	G-Tang-1 internal variance (B)	G-Tang-2 internal variance (C)	Combined internal variance ($\sum_i diff_j^i$) (D)
者	3.3707%	2.0885%	3.2852%	8.7444%
故	2.4510%	2.3747%	3.2837%	8.1094%
為	2.0781%	2.4420%	2.0667%	6.5867%
諸	2.4238%	1.7111%	1.7248%	5.8597%
或	1.2485%	2.5648%	1.9556%	5.7689%
如	1.2858%	2.1427%	1.7536%	5.1821%
不	1.5279%	1.7702%	1.6490%	4.9471%
彼	2.3815%	1.3411%	1.0653%	4.7879%
時	1.2645%	1.6379%	1.5301%	4.4325%
亦	1.7610%	1.3138%	1.1674%	4.2422%
其	1.0213%	1.5907%	1.4494%	4.0614%
此	1.2922%	1.3789%	1.3137%	3.9848%
說	1.2450%	1.3694%	0.7786%	3.3929%
所	0.9693%	1.0689%	1.2338%	3.2720%
於	0.8228%	1.0051%	1.2938%	3.1216%

之	0.6181%	1.1009%	1.0077%	2.7267%
言	0.8420%	0.8941%	0.8569%	2.5929%
一	0.5306%	0.6954%	1.1867%	2.4126%
而	0.8279%	0.9508%	0.4424%	2.2211%
無	0.7251%	0.7497%	0.6348%	2.1097%
是	0.6876%	0.5960%	0.7970%	2.0807%
當	0.3890%	0.8917%	0.7059%	1.9867%
已	0.6031%	0.7273%	0.6508%	1.9812%
若	0.6750%	0.6504%	0.6189%	1.9443%
以	0.6559%	0.6564%	0.5830%	1.8952%
猶	0.2015%	0.7474%	0.6197%	1.5686%
及	0.5556%	0.4683%	0.4951%	1.5190%
爾	0.5808%	0.4612%	0.3634%	1.4054%
又	0.4612%	0.4139%	0.3594%	1.2345%
常	0.2723%	0.4612%	0.4296%	1.1632%
謂	0.1542%	0.4789%	0.5079%	1.1410%
曰	0.5558%	0.2294%	0.1677%	0.9529%
由	0.0486%	0.2732%	0.3881%	0.7099%

In order to arrive at a list of those particles which are most expressive for differences in the translation idiom we must look for relatively *frequent particles* (to allow PCA to work), and which have *high variance between translations* (to arrive at distinctive usage), but at the same time *low internal variance* between units of a translation (to avoid content related distortion). The first two conditions have yielded the set of 33 particles, and Table 5 presents the set sorted for the third condition. Once the particles with the highest internal variance value have been identified they can be excluded from the PCA.

4.6 Procedure: Step 6 – Excluding Particles with High Internal Variance

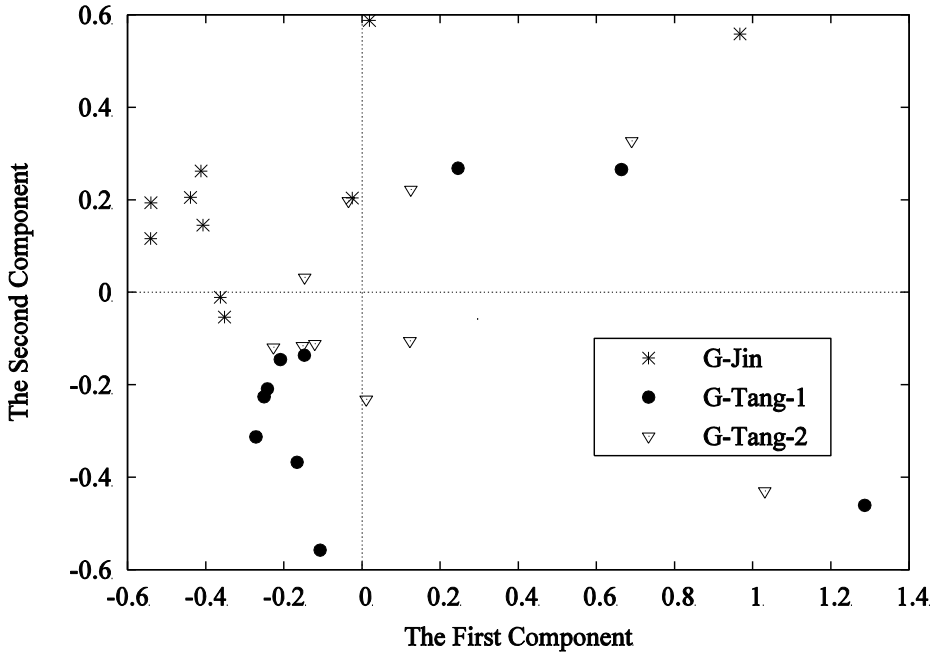

Fig. 8 PCA result for 28 particles (33 minus the top 5 of Table 5 (the five particles with the highest internal variance) (者故為諸或)), 10 units per text

Compared to Experiment 3 (all 33 particles) the clustering in the PCA result of Experiment 9 (Fig. 8) is far more distinct. Based on the use of these 28 particles G-Jin clearly forms a discreet cluster, and even the units of G-Tang-1 and G-Tang-2 seem to divide in two different clusters. This affirms that the exclusion of particles with high internal variance is conductive for this method of analysis. But how many particles should be excluded? What should the threshold for internal variance be, above which to exclude a character? In the absence of data of comparable studies we must continue to experiment. The following figures (Figs. 9-14) repeat Experiment 9, for 27 to 22 particles, successively excluding the particle with highest internal variance.

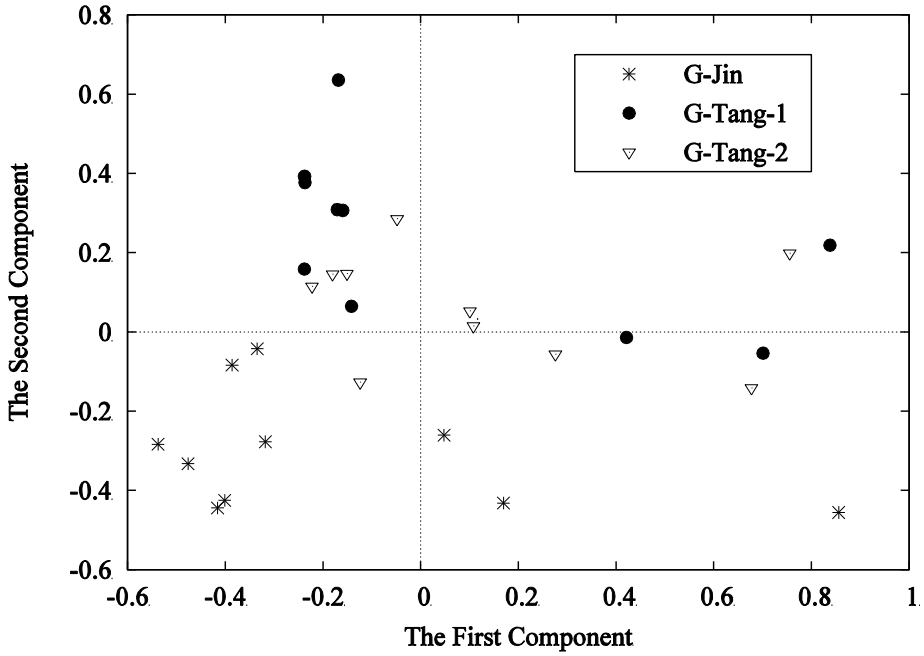

Fig. 9 PCA for 27 particles (33 minus the top 6 of Table 5 (者故為諸或如)), 10 units per text

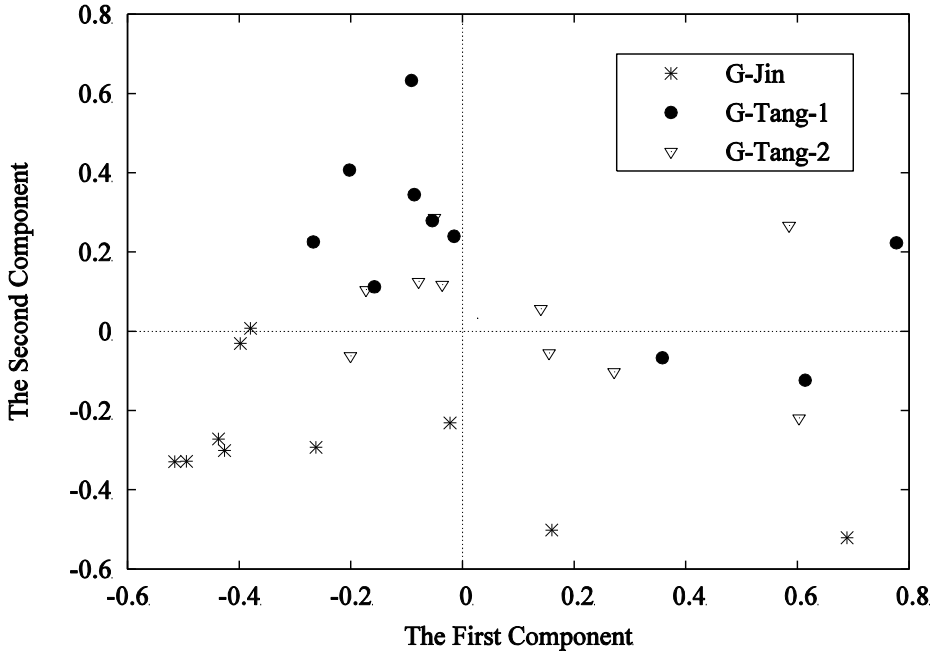

Fig. 10 PCA for 26 particles (33 minus the top 7 of Table 5 (者故為諸或如不)), 10 units per text

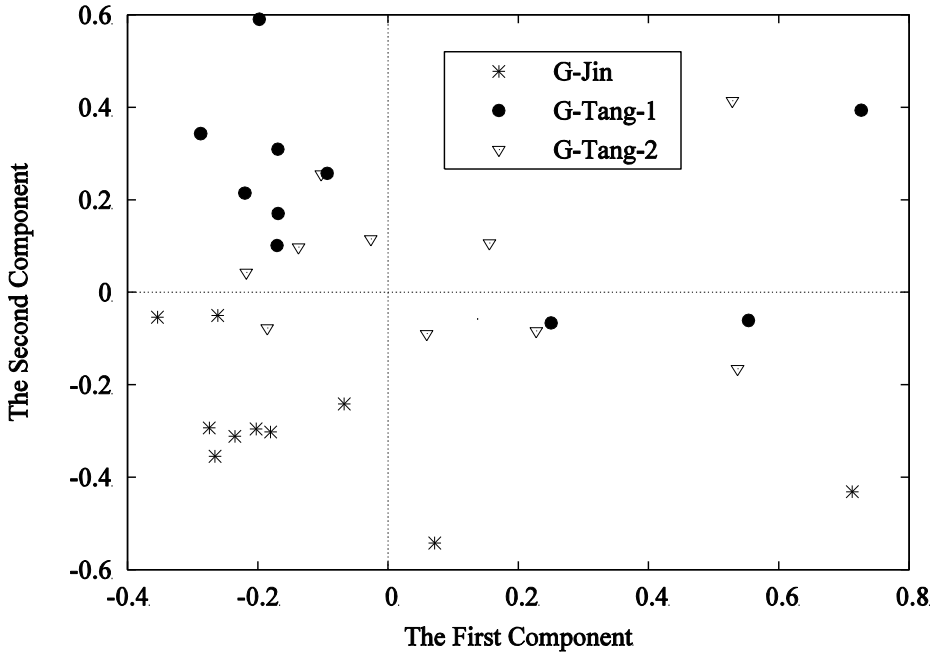

Fig. 11 PCA for 25 particles (33 minus the top 8 of Table 5 (者故為諸或如不彼)), 10 units per text

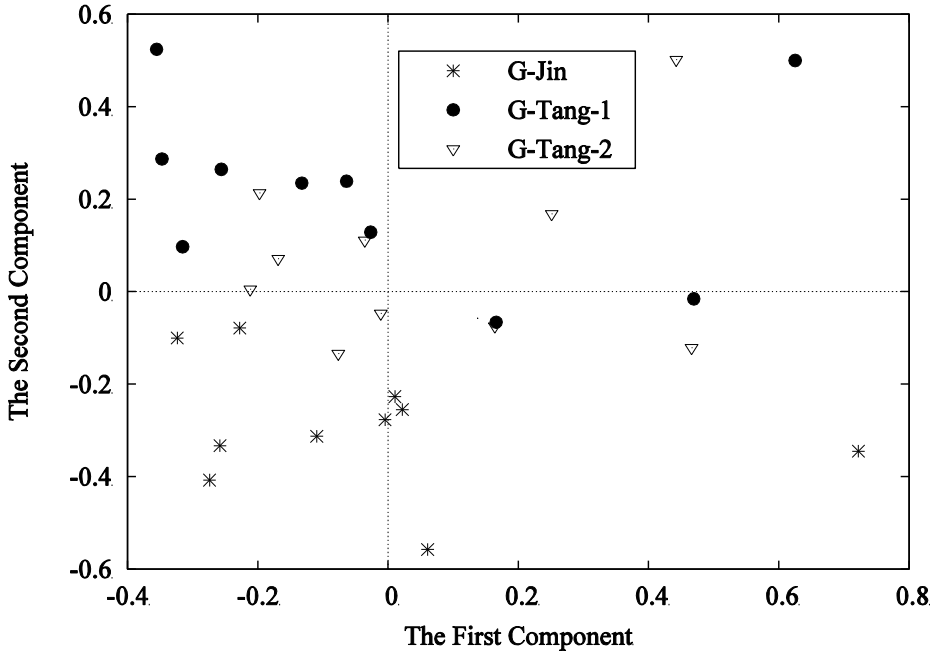

Fig. 12 PCA for 24 particles (33 minus the top 9 of Table 5 (者故為諸或如不彼時)), 10 units per text

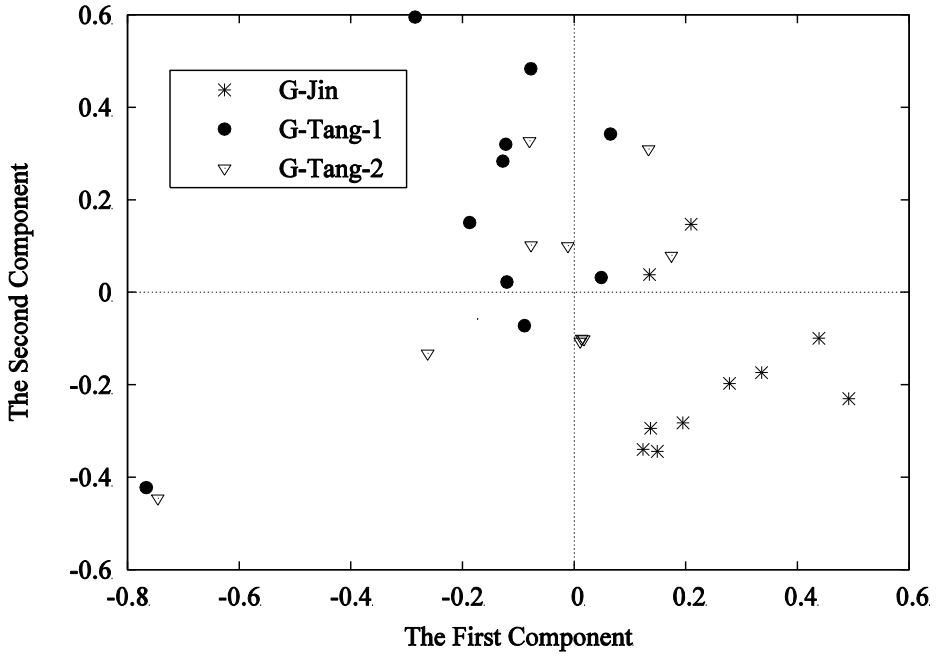

Fig. 13 PCA for 23 particles (33 minus the top 10 of Table 5 (者故為諸或如不彼時亦)), 10 units per text

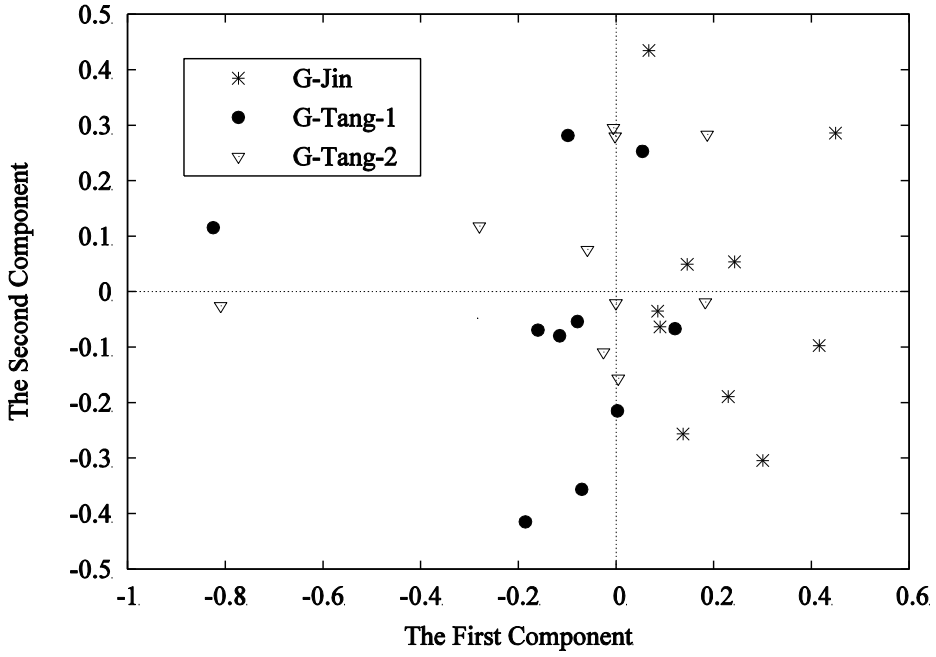

Fig. 14 PCA for 22 particles (33 minus top 11 of Table 5 (者故為諸或如不彼時亦其)), 10 units per text

The first five results in this series (Figs. 9-12) are encouraging. G-Jin units cluster consistently and distinctively and G-Tang-1 and G-Tang-2 are closer to each other than

to G-Jin. Excluding the 10 particles with the highest internal variance factor, as shown in Fig. 12, even the two persistent outliers come into the fold, and, while still not quite part of the main cluster of G-Jin, they are not distorting the output anymore. This suggests that the usage of the remaining 23 particles (33-10) is characteristic of G-Jin when compared to the two Tang versions. With the exclusion of 其 (Fig. 13), however, a threshold seems to be reached and the clustering becomes less distinct again. We have seen in Table 1 that this character has a particular high external variance value as well as high frequency and its disappearance obviously impacts the efficiency of the PCA analysis.

5. Testing: Adding another text (G-Qin)

The *Luomaqiejing* 羅摩伽經 (G-Qin, T. 294) is a partial translation of the *Gaṇḍavyūha*, which is traditionally ascribed to Shengjian 聖堅. G-Qin is nearly contemporaneous with G-Jin, however, it is less than 25% of its size. Table 6 shows the character counts for all four texts.

Table 6 Character counts for G-Jin, G-Tang-1, G-Tang-2 and G-Qin.

	Number of Fascicles	Number of Characters
G-Jin	17	143,957
G-Tang-1	21	169,122
G-Tang-2	40	250,452
G-Qin	3	33,212

We run the PCA with the particle sets that have proven effective in the last series of experiments: the 33 particles of Table 5 minus the 8-10 particles with highest internal variance. Figs. 15-17 demonstrate the PCA result of the experiments that include G-Qin.

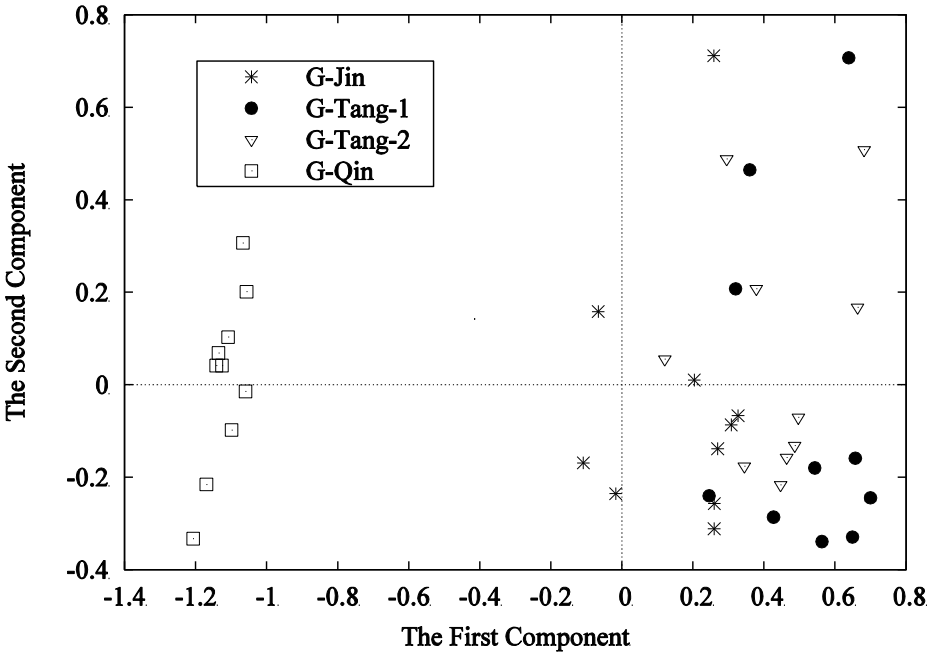

Fig. 15 PCA with G-Qin added to the test set. Using 25 particles (33 minus the top 8 of Table 5 (者故為諸或如不彼)), 10 units per text

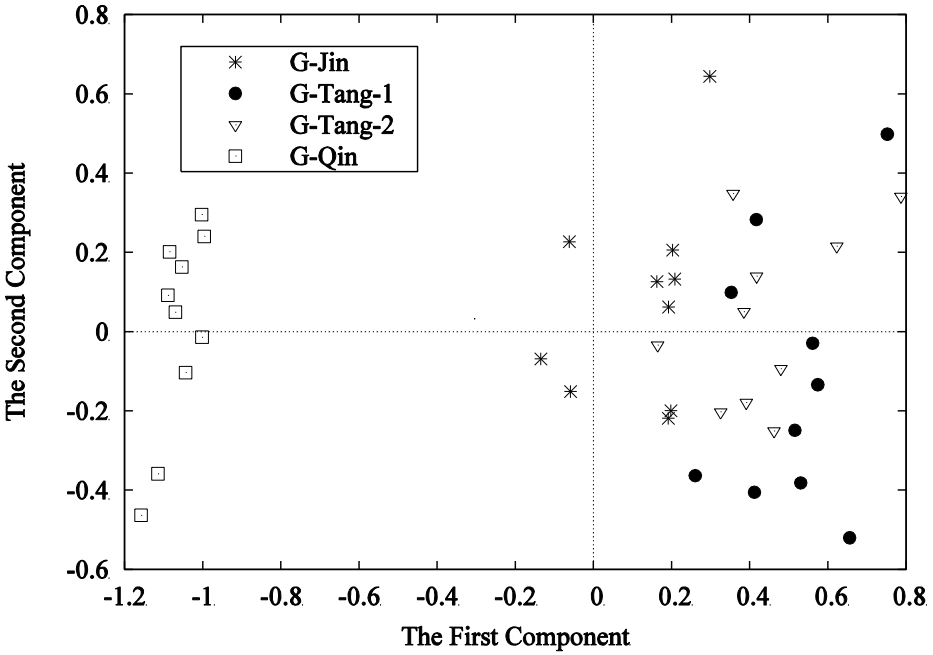

Fig. 16 PCA with G-Qin added to the test set. Using 24 particles (33 minus the top 9 of Table 5 (者故為諸或如不彼時)), 10 units per text

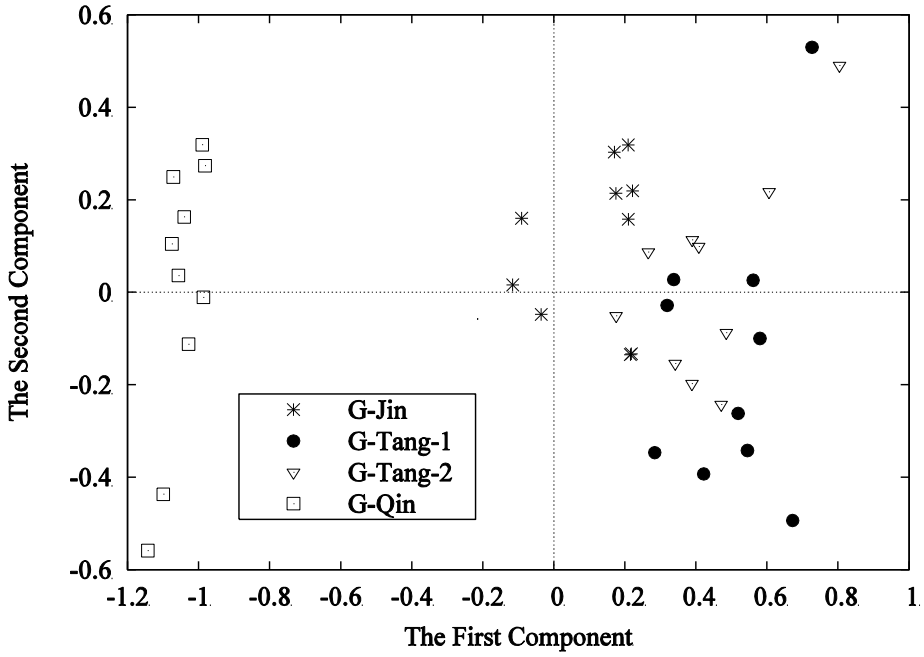

Fig. 17 PCA with G-Qin added to the test set. Using 23 particles (33 minus the top 10 of Table 5 (者故為諸或如不彼時亦)), 10 units per text

The results of the experiments shown in Figs. 15-17 indicate that the PCA weighing of grammatical particles does indeed distinguish how they are used in texts from different periods. The clear distance between the units of G-Qin and that of the other three texts, however, is mainly due to the significant different size of the units, which single them out as a distinctive group. The interesting part is that in Fig. 16 and Fig. 17 the units of G-Jin are clustering and moving toward G-Qin indicating that the use of the 24 or 23 particles tested is indeed characteristic for the difference between the two earlier texts (G-Jin and G-Qin) on the one hand, and the two Tang dynasty versions (G-Tang-1, G-Tang-2) on the other. This is an indication that the difference in particle use is not only due to the idiosyncrasy of the translations, but also to the fact that they were produced at different times.

6. Conclusion

The method outlined here uses PCA with a relatively small set of grammatical particles that fulfill the triple condition of "high occurrence / high inter-textual variance / low intra-textual variance." It has successfully distinguished translations of the same text from different periods. These findings open up a number of venues for future research. Grammatical particles could be used with PCA to compare Buddhist texts from the same periods. This would test if the stylistic features are typical for the periods in general or specifically apply to the *Gaṇḍavyūha* translations. The analysis could then be extended to non-Buddhist texts and it would be interesting to see if the same results can be achieved

with texts from different genres. If the method works for these scenarios as well and we are confident that the clustering distinguishes texts of different periods, we can use it to assess texts of hitherto unknown translation periods. It could be a breakthrough in the field of Buddhist studies, if we can suggest probable translation dates on the basis of stylometric analysis. We hope the method outlined here is a first step in this direction.

Stylometric analysis has identified 23 particles, the use of which distinguishes our three (or, including G-Qin, four) translations. These are the high-frequency particles listed in Table 5, without the top ten which are used unevenly across the units. The set that remains: 其，此，說，所，於，之，言，一，而，無，是，當，已，若，以，猶，及，爾，又，常，謂，曰，由. Though our main concern here was to develop a methodology for computational analysis, this list of distinctive particles can in turn provide a starting point for philological research. Philologists can now analyze and describe how these particles are used differently in the four translations, and compare their use to other texts of the period. We now also could attempt to use those particles which indicate a diachronic difference to describe the different translation styles of Buddhabhadra, Śikṣānanda and Prajña (and their teams). All three translators have other translations attributed to them and comparing their *Gandhavyūha* translation to the rest of their corpus (and then their corpora with each other) will help us to improve our algorithms. Ideally, algorithms would be able to describe and help to demarcate the work of different translators or translation teams. With better quantitative and qualitative descriptions it might be possible to identify wrong attributions and to improve the dating of anonymous translations of Buddhist scriptures into Chinese.

References

Bingenheimer, M. (2015). "The Digital Archive of Buddhist Temple Gazetteers and Named Entity Recognition (NER) in Classical Chinese." *Lingua Sinica* 1:8 (2015), pp. 1-19.

Binongo, J. N.; Smith, M. W. A. (1999). "The application of principal component analysis to stylometry." *Literary and Linguistic Computing*, 14(4): 445-465.

Bozkurt, I. N.; Bağhoğlu, O.; Uyar, E. (2007). "Authorship Attribution: Performance of Various Features and Classification Methods." In *Proceedings of the 22nd InternationalSymposium on Computer and Information Sciences.* Ankara,Turkey, 2007, 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.

Cleary, T. (1993). *The Flower Ornament Scripture: A Translation of the Avatamsaka Sutra*. Boston: Shambala.

Diederich; J., Kindermann; J., Leopold, E.; Paass, G. (2003). "Authorship Attribution with Support Vector Machines." *Applied Intelligence*, 19 (1-2): 109-123.

Dobson, William A. C. H. (1959). *Late Archaic Chinese – A Grammatical Study*. Toronto: University of Toronto Press.

Dobson, William A. C. H. (1962). *Early Archaic Chinese – A Descriptive Grammar*. Toronto: University of Toronto Press.

Dobson, William A. C. H. (1964). *Late Han Chinese – A Study of the Archaic-Han Shift.* Toronto: University of Toronto Press.

Dobson, William A. C. H. (1974). *A Dictionary of the Chinese Particles*. Toronto: Toronto University Press.

Fang, Dongmei 方東美. (1981). *Huayanzongzhexue 華嚴宗哲學*. Taipei: Liming wenhua 黎明文化.

Gómez, L. O. (1967). "Selected Verses from the Gaṇḍavyūha: Text, Critical Apparatus and Translation," PhD Dissertation, Yale University, 1967.

Hamar, I. (2007). *Reflecting Mirrors: Perspectives on Huayan Buddhism*. Wiesbaden: Harrassowitz.

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

Hori, Shin'ichiro. (2002). "Gandavyuha-Fragmente der Turfan-Sammlung" *Journal of the International College for Advanced Buddhist Studies* 5 (2002): 118-99. (Shigeo Kamata Memorial Volume).

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

Jolliffe, I.T. (2010). *Principal Component Analysis*. 2nd Edition. New York, Springer.

Labbé, C.; Labbé, D. (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière." *Journal of Quantitative Linguistics*, 8(3): 213-231.

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.; 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.

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

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: International Research Institute for Advanced Buddhology, Soka University.

Osto, Douglas. **(**2008). "The Supreme Array Scripture: A New Interpretation of the Title Gaṇḍavyūha-sūtra*." Journal of Indian Philosophy*, ISSN 0022-1791, 06/2009, Volume 37-3: 273-290.

Peng, F.; Schuurmans, D.; Keselj, V.; 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.

Soothill, W. E.; Hodous, Ls. (1937). *A Dictionary of Chinese Buddhist Terms.* London: Kegan, 1937 [Delhi: Motilal, 1994].

Stamatatos, E. (2009). "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology*. 60(3): 538–556.

Tearle, M.; Taylor, K.; Demuth, H. (2007). "An Algorithm for automated authorship attribution using neural networks." *Literary and LinguisticComputing*, 23(4): 425-442.

Wan, J. 萬金川. (2002). "Zongjiao zhuanbo yu yuwen bianqian – hanyi fodian yanjiu de yuyanxue zhuanxian suoxianshi de yiyi 宗教傳播與語文變遷:漢譯佛典研究的語言學轉向所顯示的意義(之二) – Part 2." *Zhengguan zazhi* 正觀雜誌, 20: 6-82.

Wang, S.王叔岷. (2007). *Guji xuzi guangyi 古籍虛字廣義*. Taipei: Zhonghua shuju 中華書局 (2nd Edition).

Wen, M. 溫美惠. (2000). "'Huayanjing – Rufajie pin zhi' zhi wenxue tezhi yanjiu《華嚴經・入法界品》之文學特質研究." Unpublished MA thesis, Zhengzhi University, Taipei 國立政治大學碩士論文.

Zhao, Y.; Zobel, J. (2007). "Searching with Style: Authorship Attribution in Classic Literature." *Proceedings of the 13th Australasian conference on Computer science*. 2007. pp. 59-68.

Table and Figure captions list

Fig. 14 PCA for 22 particles (33 minus top 11 of Table 5 (者故為諸或如不彼時亦其)), 10 units per text

Table 6 Character counts for G-Jin, G-Tang-1, G-Tang-2 and G-Qin.

Fig. 15 PCA with G-Qin added to the test set. Using 25 particles (33 minus the top 8 of Table 5 (者故為諸或如不彼)), 10 units per text

Fig. 16 PCA with G-Qin added to the test set. Using 24 particles (33 minus the top 9 of Table 5 (者故為諸或如不彼時)), 10 units per text

Fig. 17 PCA with G-Qin added to the test set. Using 23 particles (33 minus the top 10 of Table 5 (者故為諸或如不彼時亦)), 10 units per text