A Quantitative Textual Analysis of the Translation Idiom of the *Madhyama-āgama* and the *Ekottarika-āgama*

Jen-jou Hung (洪振洲)
Dharma Drum Institute of Liberal Arts
&
Bhikkhu Anālayo
University of Hamburg

Abstract

The attribution of the Chinese translations of the *Madhyama-āgama* (中阿含經, T 26) and the *Ekottarika-āgama* (增一阿含經, T 125) is debated, with uncertainty as to whether the translatorship of the *Ekottarika-āgama* should also be credited to Gautama Saṅghadeva, the translator of the *Madhyama-āgama*. The present article offers a quantitative textual analysis of these two collections, to complement the picture that emerges from traditional philological research.

We took the digitised text of the *Madhyama-āgama* and the *Ekottarika-āgama* from the CBETA corpus, removed all punctuation marks, and tokenized the texts into grams with the help of an n-gram extraction algorithm. We then selected the grams appearing in a significant number of documents and calculated the frequency of these grams to identify variations between T 26 and T 125. This involves PCA (Principal Component Analysis), 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.

The PCA results already convey a fairly strong impression that the translation style of the *Madhyama-āgama* and the *Ekottarika-āgama* are quite different from each other. To provide fuller evidence in support of this conclusion, the results of the PCA were further examined with a view to identifying the key phrases that cause the *Madhyama-āgama* and the *Ekottarika-āgama* to behave so differently. A comparison of these key phrases indicates that these do reflect different translation styles; the variations do not seem to be merely due to differences of content. Therefore, it seems justified to draw the conclusion that the translations of the *Madhyama-āgama* and the *Ekottarika-āgama* do not stem from the same translator, but are the products of different translators at work.

Contents

Introduction

The present article offers a quantitative textual analysis of the Chinese translations of the *Madhyama-āgama* (中阿含經, T 26) and the *Ekottarika-āgama* (增一阿含經, T 125). As discussed in more detail in Radich and Anālayo's contribution (2017), the translatorship attribution in the case of these two collections is debated, with uncertainty as to whether the translation of the *Ekottarika-āgama* should also be credited to Gautama Saṅghadeva, the translator of the *Madhyama-āgama*.[1]

I. Quantitative Analysis Procedure

To test the translatorship attribution, the digitized text of the *Madhyama-āgama* (T 26) and the *Ekottarika-āgama* (T 125) as found in the 2014 version of the CBETA corpus in TEI/XML format were transformed into plain text, the appendices and footnotes were removed, and the following procedure was applied to prepare the data for analysis.[2]

1. For performing the statistical analysis, fascicles were used as the basic unit. In this way, each fascicle in T 26 and T 125 was treated as an independent document, as a result of which the T 26 group consists of 60 samples, whereas the T 125 group consists of 51 samples.
2. All punctuation marks were removed, whereby the text became one long string of (Chinese) characters.

[1] As discussed by Anālayo in Radich and Anālayo 2017: 218, in the case of the *Madhyama-āgama* it seems safe to conclude that the one in the translation team responsible for the choice of translation terminology would have been Saṅghadeva himself.

[2] A count of the text file transformed from the XML source files in CBETA 2014 DVD results in 518,058 characters (without punctuation) in T 26 and 364,092 characters (without punctuation) in T 125.

3. With the help of an n-gram extraction algorithm the texts were tokenized into grams.[3] These grams then were the basis for calculating style features.
4. In order to generate better feature sets for analysis, at first all possible grams from the texts were generated (instead of using fixed-length grams), i.e., all bi-grams, tri-grams, quad-grams and so on up to the longest possible n-gram. Then all non-significant grams were removed from the feature set. The significance of a gram is based on deciding on a specific number of documents,[4] referred to as 'D', within which a gram must appear as a threshold to merit inclusion in the feature set.

II. Principal Component Analysis

Once the feature set had been generated, the frequency of the grams of the feature set in the altogether 111 fascicles (60 fascicles of T 26 plus 51 fascicles of T 125) could be calculated and further examined to identify variations between T 26 and T 125. This involves PCA (Principal Component Analysis), 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.

It is important to keep in mind that the PCA analysis is based on what could be called 'unsupervised learning', in that we do not instruct the program as to which characteristics we are looking for or

[3] Here a 'gram' indicates a sequence of consecutive Chinese characters, for example: 如是 is a 2-gram, and 一時佛在 is a 4-gram. A gram does not always have a complete meaning; in some cases it could just be part of a meaningful word.

[4] Here a document means a single fascicle.

expecting to be singled out. Instead, the program itself will offer 'random' results, relations between items with certain features represented on a bi-dimensional diagram. Such a relational model of analysis does end up highlighting the relationship between points, but this is not due to an input on our sides regarding what we expect to find. In short, the procedure is not deduction-based and is un-directed.

II.1 PCA Results and Discussions

This section presents the PCA analysis results of the 111 fascicles in the *Madhyama-āgama* and *Ekottarika-āgama* groups. To obtain best results, analyses with different values of D were performed. This serves to avoid using highly content-dependent grams as stylistic measurements in the analysis.[5] As the value of D increases, the algorithm will choose only those grams that appear in a large number of different documents for stylistic measurements. This will reduce the probability of including content-dependent grams in the feature set. However, a problem here is that the increase of D also raises the possibility of excluding some important stylistic features that appear only in a relatively small number of documents from the entire feature set. In order to avoid unduly influencing the results through a particular setting of D, a progressive analysis series with different settings of D seems an ideal solution. Thus, to begin with, D was set at a value of 20, about 18% of the total number of documents. Then the value of D was increased in steps of 20 until it reached 100, with a final analysis done with D set at the maximum of 111, corresponding to the total number of fascicles of the two texts compared.

[5] For an illustration and discussion of the problem that can arise because of the influence of content-related grams cf. Hung 2014.

II.1.1 PCA Analysis with D Set to 20

Figures 1 illustrates the first and second principle components generated by the PCA analysis with D set to 20. In this chart, black triangular symbols (▲) represent the documents from the *Madhyama-āgama* group (T 26) and hollow circles (○) represent the documents from the *Ekottarika-āgama* group (T 125).

Figure 1. PCA Result with D = 20

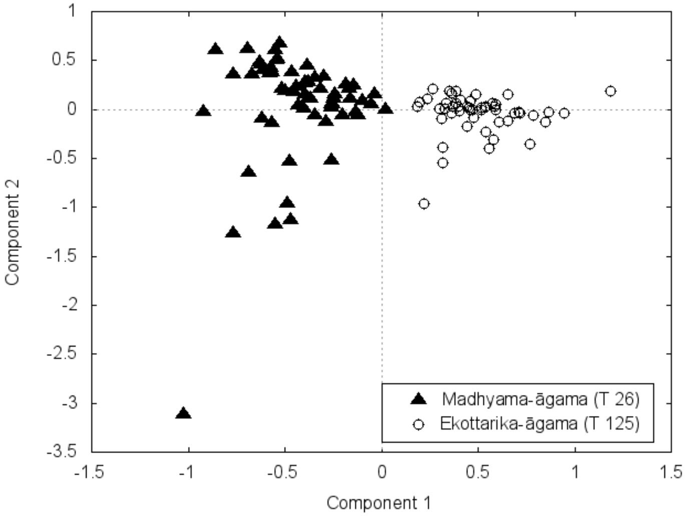


Figure 1 shows the analysis result with the value of D set to 20, which means that the set of stylometric measurements only contains grams that appear in more than 20 documents. The figure clearly shows that the two groups are in very different places compared to each other. Most of the *Madhyama-āgama* points, with only one exception, are located to the left side of the origin, whereas all of the *Ekottarika-āgama* points lie to the right side of the origin. Moreover, the two groups do not overlap on the component 1.

Already this first analysis conveys a fairly strong impression that the translation style of the *Madhyama-āgama* and the *Ekottarika-āgama* are quite different from each other.

II.1.2 PCA Analyses with D Set to 40 and 60

The next step involved raising the value of D to 40 and 60, in order to observe whether this results in a different behaviour of the researched texts. Figures 2 and 3 show that the results of PCA analyses with D set to 40 and 60 exhibit the same trends as shown in figure 1 when the value of D was set to 20. The points of the *Madhyama-āgama* and the *Ekottarika-āgama* continue to be located on different sides of the x-axis in distinct clusters. In this way, their grouping behaviour continues to be very clear even when the value of D is raised from 20 to 40 and 60.

This further confirms the impression that the translation styles of the two collections is different.

Figure 2. PCA Result with D = 40

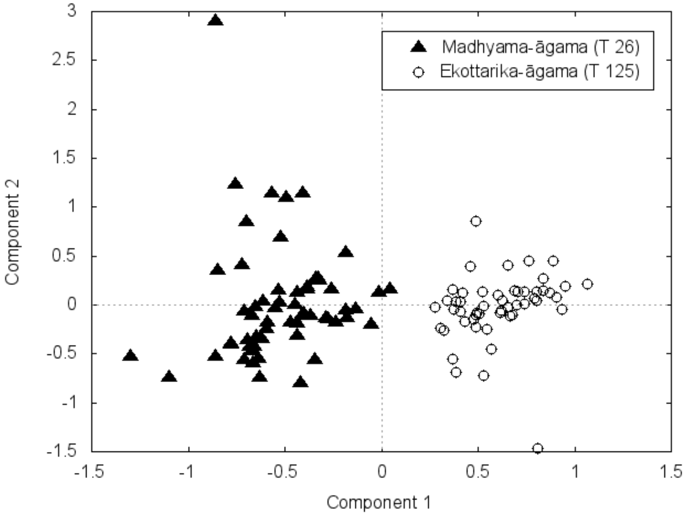
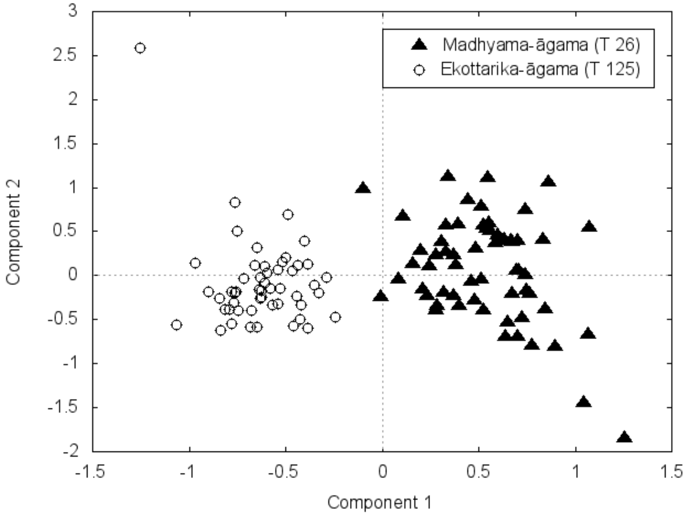


Figure 3. PCA Result with D = 60



II.1.3 PCA Analyses with D Set to 80, 100 and 111

To check the analysis result when the document threshold is set to a very high value, the value of D was increased to 80, then to 100 and finally to 111. The following figures 4, 5 and 6 show the results of the corresponding PCA analyses. Compared to the results evident in figures 1 to 3, it is noticeable that the points of the *Madhyama-āgama* texts move from the right-hand side of the coordinate plane to the top part; whereas the points of the *Ekottarika-āgama* texts move from the left-hand side to the bottom. The reason for this change would be due to the fact that the values of D in the three analyses are larger than the number of actual documents in each of the two groups: the *Madhyama-āgama* has 60 and the *Ekottarika-āgama* 51 documents (= fascicles). Therefore grams that are only used in one group but never occur in another group will no longer function as stylometric measurements, as they do not reach the threshold of D.

As a result, the difference between the *Madhyama-āgama* and the *Ekottarika-āgama* texts will inevitably be reduced and the location of points can also be subject to change.

As evident in figures 4, 5 and 6, the expectation that due to the increase in D the distance between the two groups decreases is confirmed. Nevertheless, even when the threshold is set at such a high value, still the *Madhyama-āgama* and the *Ekottarika-āgama* texts are grouped in different locations in the coordinate plane. Even when the value of D is raised to 100, only few overlaps occur. Moreover, when D is raised to the absolute possible maximum of 111, which means that only those grams that occur in every single fascicles of the *Madhyama-āgama* and the *Ekottarika-āgama* will be used, still these two groups do not show much overlap with each other.

This clearly confirms that the translation phrases employed in the *Madhyama-āgama* and the *Ekottarika-āgama* are very different from each other.

Figure 4. PCA Result with D = 80

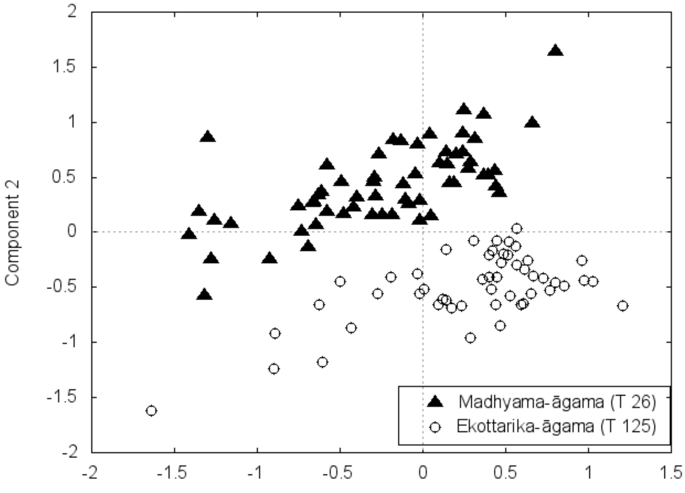


Figure 5. PCA Result with D = 100

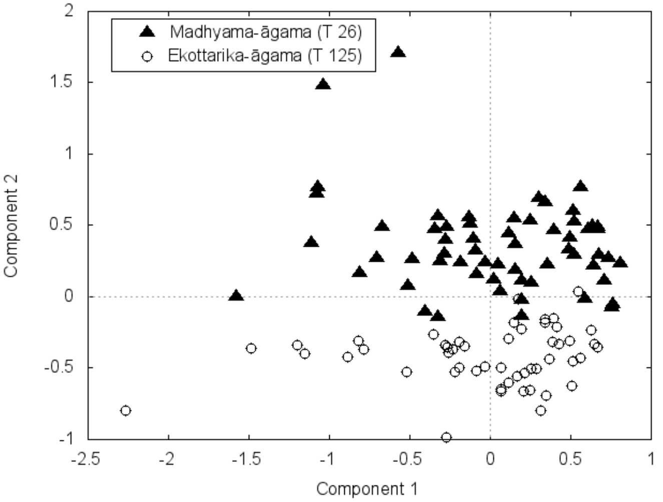
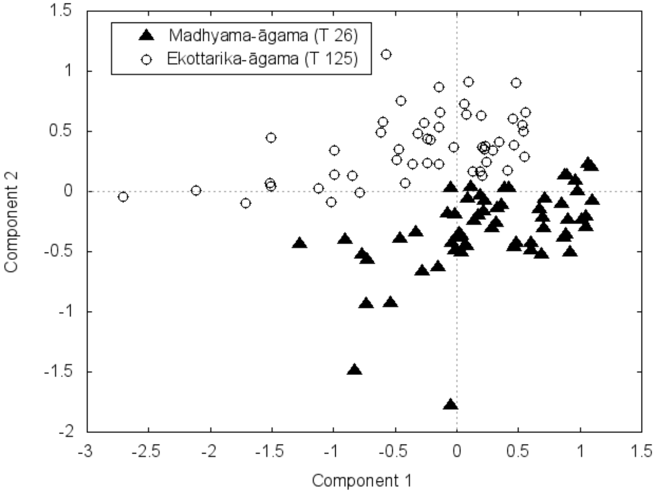


Figure 6. PCA Result with D = 111



II.2 Summary of PCA Analysis

From the above analyses it can confidently be concluded that the translation styles of the *Madhyama-āgama* and the *Ekottarika-āgama* are substantially different from each other, although both clearly belong to the same genre of scripture and their Indic originals can safely be assumed to have had considerable terminological overlap. It seems highly improbable that the *Madhyama-āgama* and the *Ekottarika-āgama* could be from the same translator.

In order to provide the full evidence to support this conclusion, in what follows the result of the PCA analysis are further examined with a view to identifying the key phrases that cause the *Madhyama-āgama* and the *Ekottarika-āgama* to behave so differently.

II.3 Gram Analysis of the PCA Results

By way of further defining the difference between the *Ekottarika-āgama* group and the *Madhyama-āgama* group, we examined the grams that are only used in one of the two collections. As the above analysis results show, right from the outset with D set to 20 the points corresponding to the *Madhyama-āgama* and the *Ekottarika-āgama* respectively are located differently. This is already significant, since with D set to a lower value more grams will be selected, which means that more information will be processed compared to when D is set to a higher value. Therefore, in what follows we take the PCA result with D set to 20 as the basis for examining the significant and distinctive features of the *Ekottarika-āgama* group and of the *Madhyama-āgama* groups.

Table 1 lists the grams only found in the *Ekottarika-āgama* group. The first two columns give the gram, followed by the number of matches of the gram in the *Ekottarika-āgama* groups. The next column, ‘Average Matches per Fascicles in EĀ’, presents the results of dividing the number of matches given in the previous column by the number of fascicles in the whole *Ekottarika-āgama*, which is 51. The last column, 'Average Matches per Characters in EĀ', then presents the results of dividing the same number of matches by the number of characters in the whole *Ekottarika-āgama*, which is 364,092.

Table 2 proceeds in the same way, this time instead taking up the grams only found in the *Madhyama-āgama* group. In this case, then, 'Average Matches per Fascicle in MĀ' are the results of dividing the number of matches by the number of fascicles in the whole *Madhyama-āgama*, which in this case is 60, and 'Average Matches per Character in EĀ' gives the results of dividing the same number of matches by the number of characters in the whole *Madhyama-āgama*, which is 518,058.

Table 1. Grams that Appear in More than 20 Fascicles of the *Ekottarika-āgama* but Are Never Used in the *Madhyama-āgama*

Phrase	Matches in EĀ	Average Matches per Fascicle in EĀ	Average Matches per Character in EĀ
所以然者	483	9.47	0.0013
聞如是一時佛在	441	8.65	0.0012
舍利弗	411	8.06	0.0011
祇樹給孤獨園	377	7.39	0.0010
白佛言	241	4.73	0.0007
白世尊言	191	3.75	0.0005
在一面坐	174	3.41	0.0005
便說此偈	168	3.29	0.0005
釋提桓因	150	2.94	0.0004
更不	130	2.55	0.0004
之類	125	2.45	0.0003

Phrase	Matches in EĀ	Average Matches per Fascicle in EĀ	Average Matches per Character in EĀ
沙門婆羅門	112	2.20	0.0003
羅閱城	108	2.12	0.0003
是故諸比丘當	105	2.06	0.0003
亂想	98	1.92	0.0003
出現於世	96	1.88	0.0003
當求方便	94	1.84	0.0003
彼云何名為	91	1.78	0.0002
諸比丘對曰	91	1.78	0.0002
之是時	89	1.75	0.0002
所在	85	1.67	0.0002
如實知之	83	1.63	0.0002
退而去	83	1.63	0.0002
生此念	82	1.61	0.0002
四部之眾	80	1.57	0.0002
是謂名為	79	1.55	0.0002
由旬	78	1.53	0.0002
便退而去	76	0.92	0.0001
如來至真等正覺	74	1.45	0.0002
阿須倫	73	1.43	0.0002
是時諸	70	1.37	0.0002
人民之	65	1.27	0.0002
臥具病瘦醫藥	63	1.24	0.0002
鬼神	63	1.24	0.0002
歡喜踊躍不能自勝	63	1.24	0.0002
得法眼淨	62	1.22	0.0002

Phrase	Matches in EĀ	Average Matches per Fascicle in EĀ	Average Matches per Character in EĀ
釋迦文	62	1.22	0.0002
以此因緣	58	1.14	0.0002
如汝所言	54	1.06	0.0001
之想	54	1.06	0.0001
在虛空	54	1.06	0.0001
生死已盡	54	1.06	0.0001
狐疑	53	1.04	0.0001
世之	51	1.00	0.0001
思惟此	50	0.98	0.0001
三惡趣	48	0.94	0.0001
爾時王	47	0.92	0.0001
之行	46	0.90	0.0001
設當	46	0.90	0.0001
到時	44	0.86	0.0001
靡不	44	0.86	0.0001
遊化	42	0.82	0.0001
成阿羅漢	41	0.80	0.0001
諸塵垢盡	40	0.78	0.0001
眾生之類	40	0.78	0.0001
愚惑	39	0.76	0.0001
三法衣	36	0.71	< 0.0001
眾僧	36	0.71	< 0.0001
在閑靜之處	35	0.69	< 0.0001
迦蘭陀竹園	35	0.69	< 0.0001
過去久遠	34	0.67	< 0.0001

Phrase	Matches in EĀ	Average Matches per Fascicle in EĀ	Average Matches per Character in EĀ
比丘從佛受教	33	0.65	< 0.0001
繫念在前	33	0.65	< 0.0001

Table 2. Grams that Appear in More than 20 Fascicles of the *Madhyama-āgama* but Are Never Used in the *Ekottarika-āgama*

Phrase	Matches in MĀ	Average Matches per Fascicle in MĀ	Average Matches per Character in MĀ
成就遊	481	8.02	0.0009
白曰世尊	248	4.13	0.0005
我聞如是一時佛遊	216	3.60	0.0004
多聞聖弟子	206	3.43	0.0004
佛說如是	201	3.35	0.0004
至信	182	3.03	0.0004
捨家無家	180	3.00	0.0003
於是世尊	165	2.75	0.0003
坐一面	163	2.72	0.0003
如意足	162	2.70	0.0003
滅道	155	2.58	0.0003
安隱快樂	155	2.58	0.0003
彼一切	147	2.45	0.0003
於是尊者	145	2.42	0.0003
著袈裟衣	142	2.37	0.0003
一向	138	2.30	0.0003
成就歡喜	138	2.30	0.0003

Phrase	Matches in MĀ	Average Matches per Fascicle in MĀ	Average Matches per Character in MĀ
正念正智	136	2.27	0.0003
世尊答曰	136	2.27	0.0003
如是知	134	2.23	0.0003
妙行	132	2.20	0.0003
燕坐	132	2.20	0.0003
說法勸發渴仰	128	2.13	0.0002
無量善	124	2.07	0.0002
梵志居士	124	2.07	0.0002
往詣佛	123	2.05	0.0002
行精勤	115	1.92	0.0002
稽首佛足	115	1.92	0.0002
調御	113	1.88	0.0002
生已盡	112	1.87	0.0002
不更受有	112	1.87	0.0002
無結無怨無恚無諍	111	1.85	0.0002
獨住	107	1.78	0.0002
自知自覺	102	1.70	0.0002
彼諸比丘聞佛所說	101	1.68	0.0002
因此故	99	1.65	0.0002
至惡處生地獄中	93	1.55	0.0002
無事處	89	1.48	0.0002
極廣甚大	89	1.48	0.0002
我今寧可	85	1.42	0.0002
叉手向佛	84	1.40	0.0002
自作證成就遊	82	1.37	0.0002

Phrase	Matches in MĀ	Average Matches per Fascicle in MĀ	Average Matches per Character in MĀ
生喜樂	81	1.35	0.0002
村邑	81	1.35	0.0002
正盡	79	1.32	0.0002
平旦	79	1.32	0.0002
詣佛所稽首	75	1.25	0.0001
離惡不善之法	75	1.25	0.0001
宴坐	73	1.22	0.0001
繞三匝而去	56	0.93	0.0001
無量方便	54	0.90	0.0001
偏袒著衣	54	0.52	0.0001
世尊聞已	53	0.88	0.0001
求安隱快樂	53	0.88	0.0001
我寧可	51	0.85	< 0.0001
愍傷	49	0.82	< 0.0001
求義及饒益	48	0.80	< 0.0001
過夜平旦	46	0.77	< 0.0001
苦滅道	45	0.75	< 0.0001
佛法及比丘眾	43	0.72	< 0.0001
斷疑	43	0.72	< 0.0001
蘭哆園	41	0.68	< 0.0001
於晡時從	40	0.67	< 0.0001
自歸乃至命盡	40	0.67	< 0.0001
敷尼師檀	40	0.67	< 0.0001
天及魔梵	39	0.65	< 0.0001
從今日始終身	38	0.63	< 0.0001

Phrase	Matches in MĀ	Average Matches per Fascicle in MĀ	Average Matches per Character in MĀ
受我為優婆塞	37	0.62	< 0.0001
至得第四禪	37	0.62	< 0.0001
受教而聽	36	0.60	< 0.0001
苦如真	33	0.55	< 0.0001
善受善持	31	0.52	< 0.0001

A comparison of the expressions in the above two tables gives the impression that these reflect different translation styles, the variations found do not seem to be merely due to differences of content.

Conclusion

The above results make it safe to conclude that the indications already evident from the figures representing the PCA analysis find confirmation on closer inspection of the grams on which they are based. It seems therefore justified to draw the conclusion that the Chinese translations of the *Madhyama-āgama* (T 26) and the *Ekottarika-āgama* (T 125) do not stem from the same translator, but are the products of different translators at work.

Abbreviations

CBETA	Chinese Buddhist Electronic Text Association
D	Document threshold
EĀ	*Ekottarika-āgama* (T 125)
MĀ	*Madhyama-āgama* (T 26)
PCA	Principal Component Analysis
T	Taishō edition (CBETA, 2014)

References

Hung, Jen-jou 2014: “A Textual Analysis of the Last Discourse in the Chinese Dīrgha-āgama Based on a Translatorship Attribution Algorithm”, in Dhammadinnā (ed.), *Research on the Dīrgha-āgama*, Taipei: Dharma Drum Publishing Corporation, 167–198.

Radich, Michael and Bhikkhu Anālayo 2017: “Were the Ekottarika-āgama and the Madhyama-āgama Translated by the Same Person? An Assessment on the Basis of Translation Style”, in Dhammadinnā (ed.), *Research on the Madhyama-āgama*, Taipei: Dharma Drum Publishing Corporation, 209–237.