Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus: N-gram Based System for Variant Document Comparison and Analysis (NGSV)

Ishii Kosei

Komazawa Junior College, Japan

Abstract:

In the case where there are a number of variant editions extant for a single document, our first task must be that of gaining an accurate grasp of the distinctive characteristics of those variant texts. Thus, in this presentation, I would like to take up the example of the variant editions of Liang Wudi's *Bodhidharma Inscription*, along with the texts are derived from this document, and analyze them through a variety of processes to clarify their genealogies. Once these kinds of processing methods are developed, they may be used broadly for purposes other than the construction of genealogy tables for variant editions.

1. Examples of the system

By virtue of electronic text, it has become extremely easy to search particular words and phrases included in a particular text. However, there has not been a system that displays an extremely easy-to-look-at list of common parts and differences in more than two documents at once. Since we are in the middle of an experimental production of such a useful system, we would like to report our findings. In this presentation, we refer to this system tentatively as “N-gram Based System for Variant Document Comparison and Analysis” (NGSV). Although NGSV is able to handle documents in any language, in this presentation I will deal documents written in Classical Chinese.

The first example is the one that processes the variant editions of each of the *Bodhidharma Inscription* (菩提達摩碑文） which has been attributed to Liang Wudi (梁武帝). This inscription is taken up as an example is because it is an important document that urges re-examination of the framework of Early Chan history, and also is due to the fact that the inscription is attracting a lot of attention among researchers in this field. To begin with, let me show you a part of the result that was processed by our system.

```
10      一眞之法      4       (寶,日,元,少,熊)
2       可禪師矣      4       (慧)
```

10	而登妙覺	4	(寶,日,元,少,熊)
8	實哉空哉	4	(寶,元,少,熊)
10	眞之法盡	4	(寶,日,元,少,熊)
2	大同二年	4	(熊)

In the list displayed above, the number in the upper-left corner indicates the number of times a particular string appeared in the texts. In other words, with respect to the first line, it shows that the phrase of "Dharma of One Truth (一眞之法)" appears a total of 10 times in all the variant editions.

The number displayed on the right of the Chinese characters shows the number of matched characters. In other words, this number shows that the string of these 4 characters appear in the texts. It is possible for you to designate this number freely so that you can designate "matched examples with less than 12 characters," etc. When handling documents in such languages as Tibetan, where sentences are divided into words, it allows you to designate in word unit rather than in character unit.

Next, characters at the upper-right corner show which variant documents include the words or phrases shown. Therefore, you see that the phrase "Dharma of One Truth" appears in many variant documents while the word "Datong 2nd year (大同二年)" only appears in one document with abbreviation of “熊."

However, the proofreading has not been completed sufficiently and our system is still at an experimental stage and it is possible that there might be errors in the results. Therefore, I would appreciate it if you see the following results of NGSV as reference examples.

Now I would like to give an overview of these variant documents.

熊	＝	熊耳山菩提達摩碑文
少	＝	少林寺菩提達磨碑文
元	＝	元符寺 (二祖寺)菩提達摩碑文
日	＝	日本天台宗最澄･ 光定所引菩提達磨碑文
寶	＝	『寶林伝』所引菩提達磨碑文
正	＝	『伝法正宗記』所引略文
駒	＝	『景德傳燈錄』 (駒澤大學圖書館所藏)所引略文
松	＝	『景德傳燈錄』 (松ヶ岡文庫所藏） 所引略文
慧	＝	『寶林傳』所引慧可碑文

First, what is referred to by the abbreviation “熊” is the text of the *Bodhidarma Inscription* of Mt. Bear-ears (熊耳山 Xionger shan) where the grave of Bodhidharma is located. This stone monument was erected either in the Ming era or Ching era.

The character “少” indicates the inscription on the stone monument of Bodhidharma erected in the famous Shaolin Temple, which was re-erected in the Yuan era.

The character “元” refers to the *Bodhidharma Inscription* of the Yuanfu Temple (元符寺) which is the location of the grave of Huike (慧可), the second patriarch of Chan School. This was written in 818 and the stone monument has only recently been discovered.

“日” refers to the *Bodhidharma Inscription* quoted by Saichō (最澄), the founder of Tendai sect in Japan (天台宗) and Kōjō (光定) who is a disciple of Saichō. This was written sometime around early-to-mid 9th century.

The character "寶" refers to the inscription quoted in the *Baolin zhuan* (寶林傳) which is an early historical document of Chan school. The *Baolin zhuan* was written in 801, but the current text was written and printed during the Jin era. Recently the *Bodhidharma Inscription* has been under suspicion for being a forgery by the author of the *Baolin zhuan*. However, this suspicion is unwarranted. In addition to the fact that the text of the *Bodhidharma Inscription* in the existing *Baolin zhuan* belongs to a fundamentally different lineage from that of other variant documents, it is full of typographies.

What is marked "正" is the part of the *Bodhidharma Inscription* that is quoted in the *Zhuanfa zhenzong ji* (傳法正宗記) which is another historiography of the Chan school written during the Song era.

In addition, there are parts of the *Bodhidharma Inscription* cited in the commentaries of the *Jingde zhuandeng lu* (景德傳燈錄). The character “駒” refers to the commentary stored in Komazawa University Library. The character “松” indicates the commentary stored in Matsugaoka Library established by Dr. Suzuki Daisetsu. These two commentaries are presumably written around the 14th century and copied around the 17th century.

Finally, the text indicated by “慧” is the inscription of the second patriarch Huike that is quoted in the *Baolin Chuan*. It is a document whose author belonged to the same lineage as that of the *Bodhidharma Inscription* and uses an extremely large number of words in common with it.

Now, let's see a little more of the result of NGSV. This is a very small part of the list which displays the part matching only 2 characters among each of the variant documents.

10	凡夫	2	(寶,日,元,少,熊)
12	妙覺	2	(慧,寶,日,元,少,熊)
2	無說	2	(寶)
4	無內	2	(寶,少)
6	無法	2	(元,少,熊)

2	唯佛	2	(慧)
10	有也	2	(寶,日,元,少,熊)
8	流法	2	(寶,元,少,熊)
14	龍珠	2	(寶,日,駒,松,元,少,熊)
2	梁大	2	(熊)
8	鱗惠	2	(寶,元,少,熊)
10	利那	2	(寶,日,元,少,熊)
8	嗚呼	2	(寶,元,少,熊)
8	嗟呼	2	(寶,元,少,熊)
2	昊天	2	(熊)
2	曰夫	2	(慧)
2	楞伽	2	(寶)

What information does this data reveal to us? The parts where matches are found in all of the variant documents are presumably in the same form as in the original text. And since there are many cases where four variant documents (寶, 元, 少, 熊) show agreement, these genealogies are presumably close. Since there are cases where each of three documents (寶, 少, 熊) match each other and there is also a case where only two documents of (寶 and 少) are in agreement with each other, these three documents (寶, 少, 熊) are much closer than others and still (寶 and 少) are furthermore closer in lineage. Moreover, the fact that there are cases where three documents of (元, 少, 熊) are in agreement with each other suggests a possibility that after the establishment of a text on which the Shaolin Temple inscription was based, the text became the standard and various variant documents were derived from it with only a little difference added to each copy in the process of copying.

When looking at it from this perspective, the above variant documents can be classified neatly into a clear-cut genealogy. This suggests the fact that there was no attempt to make a new revision by comparing editions of more than one genealogy.

Also, by comparing the frequencies of matches between a document A and a document B with the frequencies of matches between a document A and a document C, we can compare the similarity between A and B with the similarity between A and C. If we conduct such a survey on all of the documents being compared, a similarity between each of variant documents can be clarified in this way. In order to make more precise judgment a statistical examination is necessary.

I feel extremely ashamed to make the following report to you as a Japanese person. For at the beginning of Kōjō's *Denjutsu isshinkai mon* (傳述一心戒文), Kōjō rewrote the *Bodhidharma Inscription* into Saichō's accomplishment. More exactly, Kōjō made Bodhidharma's biography out to be Saichō's biography by changing necessary parts such as "Master Bodhidharma was a person born in India" into "Master Saichō was a person born in Japan," etc. and by using the rest of the *Bodhidharma Inscription*. If it were done in the present day, it would be considered to be a violation of Copyright Act. When processing texts with NGNV, we tried to restore it to the original form by comparing variant editions in advance. This was done with for the purpose of determining the genealogy of the text of the *Bodhidharma Inscription* used as a model by Kōjō. If you want to know matching between the *Denjutsu isshinkai mon* itself and the variant documents of the *Bodhidharma Inscription,* you need to input the present form of the *Denjutsu isshinkai mon* in order to make the comparison.

However, even researchers of Chan history such as Dr. Hushi and Dr. Sekiguchi Shindai have not shown interest in the similarity of these two documents. Since there is a quotation from the *Bodhidharma Inscription* in the latter half of the *Denjutsu isshinkai mon*, they directed their attention to this and did not notice the relationship between Saichō's biography and the *Bodhidharma Inscription*. I accidentally noticed this resemblance. Using our system, however, the fact that both documents bear great resemblance to each other may be noticeable by anyone even if he is not a specialist in this field. This suggests that our system is extremely useful not only for comparing variant documents but also discovering similarities and differences between any set of documents.

2. Contents and the background of the system

This system is based on a free program for statistics, “ngram,” which was prepared by Mr. Hujiwara Shigeru for the analysis of Japanese sentences, and has been placed on the Internet (http://www.jaist.ac.jp/~shigeru/ngram.html) for view by the general public. This program works on UNIX and can handle only the files written with either of two Japanese codes, JIS and EUC. Personally, I am using this program on a Windows 2000 PC with the environment of Cygwin, which creates a quasi UNIX environment. Mr. Fujiwara’s “n-gram” can handle not only Chinese characters and Japanese characters but also any languages as long as it is Romanized. The advantage of his program is that it can work at an extremely high speed with a small amount of memory. Since “ngram” is a tool for N-gram statistics, I would like to take this opportunity to explain a little bit about it.

The N-gram statistics is a theory developed by Claude E. Shannon for the purpose of linguistic analysis and extracts the strings of an arbitrary length in a text and calculates the frequency of occurrence. Let's take a famous sentence at the beginning of the *Heart Sutra*.

觀自在菩薩行深般若波羅蜜多時照見五蘊皆空度一切苦厄

In this case, if we extract using the unit of 1 gram (unigram), in other words using the unit of one character, the result is as follows.

觀 自 在 菩 薩 ……

Next, if we extract using 2 gram (bygram) or a unit of two characters, the result becomes as follows.

觀自 自在 在菩 菩薩 薩行 ……

If we extract using 3 gram (trigram) or a unit of three characters, the result becomes as follows.

觀自在 自在菩 在菩薩 薩行深 ……

In this way you can process a text by the unit of any gram (*n*-gram) you need. Mr. Fujiwara's "ngram" extracts all the strings divided by grams equal to and less than the designated number. Although it is possible to expand the gram number as large as possible, an appropriate number of gram that should be used depends on the purpose of the study. Mr. Fujiwara's "ngram" displays a list of results with the frequencies of occurrence of each string at the left side and the number of matched gram at the right side as follows (numbers are not correct):

```
4       色即是 3
6       觀自在菩薩 5
17      菩薩 2
6       在菩薩 3
2       無上呪是 4
```

By executing a "sort" on this result, we can arrange the lines in the order of the frequency of occurrence or arrange them in the order of the number of matches. Although it creates strings that do not make sense such as "在菩薩, they can serve as important information when comparing variant documents.

Since the present version of "ngram" can deal only two Japanese codes and it ignores strings that occur once in the text, we are asking Mr. Fujiwara to improve his program.

3. Application of the n-gram statistics in classics research

Although the use of the N-gram statistics for language analysis in Japan has been attempted for a long time, the attempt at using it in classics research has just begun. Mr. and Mrs. Kondo became the forerunners in this attempt. Mr. Kondo, Yasuhiro who is both a Japanese linguist and a computer expert, has applied N-gram statistics to the *Kokin-waka-shu*, a representative anthology of Japanese poems and to the famous *Tales of Genji* and compared the results. He found that a number of the words and phrases of both works are in agreement with each other, and clarified that the main text of the *Tales of Genji* skillfully uses the words and phrases of the poems in the *Kokin-waka-shu*.

On the other hand, Mrs. Kondo Miyuki applied the N-gram statistics to the poems of male and females poets in the *Kokin-waka-shu* and compared the results. She discovered that the words used in both categories of poets are different and that male poets usually used male expressions and female poets usually used female expressions, clarifying the attitudes of males and females at that time. For instance, the verb “love” is used both by males and females in the *Man’yō-shu* which is the oldest anthology in Japan while the same word is used, to our surprise, only by males in the *Kokin-waka-shu*. That is, entering the Heian era, females came to be obliged to take a passive role of “being loved." This study was epoch-making as a study of the 31-syllable Japanese poems and also as a "gender" study in Japan's classic literature.

When I heard Mr. Kondo Yasuhiro made a presentation about the application of N-gram statistics to Japanese Classics for the first time in 1999, I was very much interested in it. And I mentioned it and stressed on various possibilities of N-gram use at the Japan Association for Asian Text Processing (JAET). Thereafter, there have been an increasing number of people who are interested in N-gram including Mr. Tanimoto Sachihiro and Mr. Moro Shigeki (who is also making a presentation at this conference). Mr. Tanimoto, who is studying the relation between the Chinese poetry and 31-syllable Japanese poetry enabled the program to search for character strings that are different from the string being searched in one character by using a fuzzy search program to process the result obtained by N-gram. That is, when there is a phrase such as “観自在菩薩 (Avalokiteśvara Bodhisattva)” in one of the documents, his system searches for strings of characters that match with the phrase except one character. That is, he asked the program to automatically extract these strings as follows:

観自在菩薩 ＝ 観自由菩薩，観自音菩薩， 観世在菩薩，観自在布薩，観自在菩提

If you use “regular expressions”, needless to say, such searching is possible. But you have to designate what kind of character patterns to search for. Since Mr. Tanimoto has asked the program to carry out searching by designating all the results obtained by “ngram”, all the character strings different from the searching words or phrases by one character are also automatically searched. This is an extremely useful tool, as it allows us to search words or phrases that differ by only one character. And it allows us to find words or phrases where only one character was miscopied.

But, the attempts described so far only compare two documents at a time. Therefore, I have started an experiment where I compare multiple documents processed by “ngram” and display the result in an easy-to-read list. Then, Mr. Moro Shigeki, who is cooperating with me in the SAT and INBUDS project, took the trouble of writing an extremely convenient Perl script for Perl according to my designs. The result of this is the system that I first showed you above.

4. Use of the system for research other than variant document processing

Of course, our system can handle tasks other than variant document processing. Let me show you an example where several works of the same author are processed by this system. The following is an example where various works by Wŏnhyo, the scholar-monk from ancient Korea, are processed by NGSV and arranged. First, I will show you a table of 20 works by Wŏnhyo and their designated abbreviations.

慧	=	大慧度経宗要
法	=	法華宗要
金	=	金剛三三昧經論
無	=	兩卷無量壽経宗要
阿	=	佛說阿彌陀経疏
涅	=	涅槃宗要
彌	=	彌勒上生經宗要
海	=	海東疏 (起信論疏)
別	=	大乘起信論別記
戒	=	菩薩戒本持犯要記
中	=	中邊分別論疏
判	=	判比量論
本	=	本業經疏
花	=	花嚴経疏
二	=	二障義
梵	=	梵網經菩薩戒本私記
和	=	十門和諍論

懺　=　大乘六情懺悔 (疑）
遊　=　遊心安樂道 (疑）
發　=　發心修行章 (疑)

The last three are probably either forgeries or the work of another author. NGSV is effective even when it tackles such a problem and can be one of the materials to serve as a decision aide. When processing these various documents using NGSV without designating the upper limit of the number of matched characters, an output file is generated with the size of 1.93MB. Of the output file, the maximum number of characters that matches are as follows:

2　　若得無念者即知心相生住異滅以無念等故而實無有始覺之異以四相俱時而有皆無自立本來平等同一覺故案云　　47　　(金)
2　　滅者不異外道斷見戲論諸外道說離諸境界相續識滅相續識滅已即滅諸識大慧若相續識滅者無始世來諸識應滅　　47　　(海)

That is, of the instances where the same character string appears more than once in the works of Wŏnhyo, the maximum number of characters is 47 and the character string appears twice in the *Kŭmgang-samme-kyŏngnon* and the *Haedong-so*.

The former is the quotation from the *Awakening of Faith in Mahāyāna* and the latter is the part quoted from the *La□kāvatāra-sūtra*. In addition, of the above 20 documents, the words used in as many as 17 documents are “all (一切)” and “sentient beings (衆生)."

Furthermore, in the file of documents by Wŏnhyo, all the punctuation has been removed. There are many things that can be clarified if punctuation is included in the processing. However, its accuracy depends on the uniformity of using punctuation. Conversely, if such a processing is done, we can clarify the fact that punctuation is used in an inconsistent manner. And therefore, it becomes easier to correct them.

Let's see a little more of the results of processing done on various documents by Wŏnhyo. The number of gram that yields useful information is from 7 gram. In other words, it is at about these instances where 7 characters are matched.

4　　爲無等無倫最上 7　　(無,遊)
2　　爲欲說第一義諦 7　　(慧)
(中略）
2　　當坐道場證得無 7　　(法)
2　　當諸門由非異故 7　　(涅)
10　　當知此中道理亦 7　　(海,別,二)

2	當得往生彼佛世 7	(遊)
2	當得成阿耨菩提 7	(涅)
3	當學般若波羅蜜 7	(慧)
2	當滿地成佛菩提 7	(金)
5	當發無上菩提之 7	(無,遊)

Looking at this part, we can see such an expression as “You should know that it also holds true here (當知此中道理亦)” appears in the *Haedong-so*, the *Kisillon-pyŏlgi*, and the *Ijang-ŭi*. Moreover, since it appears as many as 10 times, we can infer that it was a favorite phrase used by Wŏnhyo. Actually, Wŏnhyo uses this expression once in the *Muryangsugyŏng-chongyo* (無量壽經宗要). Therefore, I made a rough search to see if there are other people besides Wŏnhyo who use this expression. Then, I found that Fazang (法藏) who was a famous priest of the Huayan School in the Tang period was using it in the *Wujiao zhang* (五教章) and the *Huayanjing zhigui* (華嚴經旨歸). It is well known that Fazang got philosophical influence from the works of Wŏnhyo. However, the study has not been conclusive as to which parts were actually influenced by Wŏnhyo. NGSV will be useful in such a study, too.

Furthermore, there are not many instances of a 7-character match. However, looking at the above list, we can see two phrases that occurred both in the (無) or the *Muryangsugyŏng-chongyo* and (遊) or the *Yusim-allak-to*. Actually, since the *Yusim-allak-to* was forged after the *Muryangsugyŏng-chongyo* with an addition of some mantras, it is a matter of course that there are matches found in both documents. The fact that two works resemble each other in a strange way becomes clear if you see the following part.

4	雖願	2	(**無,遊**)
4	雖見	2	(涅)
2	雖盜	2	(梵)
24	雖然	2	(法,**無**,涅,海,戒,**遊**,二)
3	雖念	2	(海)
32	雖非	2	(金,涅,海,戒,二)
(略)			
34	邊故	2	(金,**無**,涅,海,**遊,**本)
13	邊際	2	(金,**無**,**遊**,本)
2	邊大	2	(慧)
7	邊地	2	(**無,遊**)

Other than this part, there are many parts where these two documents appear together and there are too many strange coincidences where only two documents are in agreement. That is, NGSV can become an extremely effective means to clarify the similarity between documents and the original influence. For instance, if all the works of Wŏnhyo are compared with the *Mahāparinirvā□a-sūtra* (涅槃經) through NGVS's processing, all the matched strings including words and phrases that match by chance can be extracted and be clearly displayed in the arranged way even if Wŏnhyo did not say “The *Mahāparinirvā□a-sūtra* says as follows.” When I was a graduate student, the most difficult thing to prepare for the class was to find a reference that says "A Sutra says so." However, we are now able to search for such things almost perfectly in a really short amount of time. Researchers can save time and think more about the texts referring such results shown above.

Furthermore, I believe it is important to know how and why Wŏnhyo used specific expressions in order to understand Wŏnhyo's way of thinking. Let us see the use of "though (雖)" as one of the examples. This time, the result is displayed with the number of occurrence in each text as follows. Thanks to Mr. Moro's convenient script, we can choose either way of displaying.

14	雖是	2	(金,涅,遊,二)
2	雖成	2	(梵)
27	雖然	2	(法,無,涅,海,戒,遊,二,梵)
2	雖多	2	(遊)
2	雖盜	2	(梵)
2	雖得	2	(梵)
3	雖念	2	(海)
2	雖犯	2	(梵)
43	雖非	2	(金,涅,海,戒,本,二,梵)
39	雖不	2	(法,金,無,涅,海,別,戒,遊,本,二,梵)
18	雖復	2	(金,涅,海,別,二)
5	雖未	2	(金,二)
36	雖無	2	(金,涅,彌,海,懺,遊,本,二,梵)
2	雖名	2	(涅)
54	雖有	2	(金,阿,涅,海,別,戒,發,二,梵)

It is natural that phrases such as "even if he steals" and "even if he violates it" appear in (梵) or Wŏnhyo’s commentary on rules of conduct for bodhisattvas (菩薩戒). What we notice here is the fact that there are so many uses of negating a negative situation such as "even though it is not .... (雖非/雖不)” and “even without .... (雖無)." If the number of such expressions increase or decrease with time, it may be related to the change in the philosophy of Wŏnhyo. By studying such a change and by investigating situations of the Buddhist world that served as the background for the change, we can step deeper into the philosophy of Wŏnhyo.

5. Combining NGVS and XML

There may be various ways of using NGVS. However, the most important is its compatibility with XML. For example, if we mark up the text like this:

<verse>歸命盡十方　最勝業遍知</verse>

Then it is possible to extract only the parts of verses and other parts. By processing them with NGSV, we can find specific words and phrases used only in verses. If we deal analects or history documents of Chan school, we can insert tags that show the lineage of Chan masters and tags that show dialogue between Chan master and their apprentices. Then we can extract only the parts of statement of Chan masters belonging to a specific lineage and compare them with those of other lineages through NGSV.

Such an attempt has already been done before to some extent in a traditional way such as using cards. However, a more detailed study can be conducted since NGSV extracts all the character strings. Also conversely, with the advent of NGSV, it will become clear what kind of XML tag can lead to useful information. Namely, NGSV is also useful in devising ways of tagging itself.

How we advance the study by effectively using voluminous digital data will become more important from now on. NGSV should become an extremely powerful means to attempt such use. I am hoping that many people in this field will improve on our NGSV and devise various ways to accomplish fruitful results with it.