Finding related passages in the Zen Knowledgebase – A status report

Christian Wittern

Introduction

The Zen Knowledgebase is a long-term project that has gone through several iterations. It basically consists of the texts in the Zen/Chan section of the Kanripo with some information added to it.

Texts in the Zen tradition have a high degree of intertextuality, referring to things mentioned in earlier texts, quoting and so on. Here, the aim is to find passages similar to a passage that is currently in focus, for example in an interactive application.

However, since the identification of parallel passages, even in a medium sized text corpus, requires time-consuming calculations, the passage will be calculated beforehand and saved for later use. Part of the work described here thus focuses on a description format for these linked passages.

Most research on text reuse, text parallels and so forth have been done by considering mainly the sequence of characters, ignoring higher level textual units. This study will take a different approach, by using a phrase as the unit of comparison.

The research reported here is in a very early stage, and should be considered more of a status report, not a polished paper.

What is a related passage?

Zen texts are (or pretend to be) to a considerable degree records of oral interactions. In these interactions, there are a small number of questions that come up at many different places, some of them are standard interactions like greetings etc. So although the character content is the same, for the purpose of this study, such cases will not be considered related.

On the other hand, some records of interactions seem to have been made independently or have been edited considerably, which makes the characters different, although a passage should be considered related.

In this study, passages will be considered related, if at least two adjacent (or almost adjacent) phrases match. We will also consider phrases related if they are occurring only very rarely.

What is a phrase?

Most texts can be quite naturally decomposed into nesting structural units, such as

- section (章)
- paragraph (段)
- sentence (文)
- phrase (句)

In such an analysis, the phrase is the smallest unit. In texts with punctuation, this will be typically the chunks of text between punctuation marks. Of course, the exact placement of punctuation marks frequently is open to debate and has several plausible solutions. A method using phrases as analytical unit therefore needs to be somewhat flexible and can not simply look for more or less exact matches. For all these reasons, the strategy for finding a matching phrase becomes rather complicated. However, using the structural information that is available in texts with markup in addition to simply using the character content should allow for findings that can not be achieved otherwise.

What is not a related phrase?

Table 1 shows the most frequent phrases in the text corpus of the pilot study. All of these are questions, but the answers are in most cases quite different. All these phrases therefore occur in contexts that are not considered related for the purpose of this study. Some effort has to be made to make sure, that these phrases are excluded most of the times. They should be found however in cases, where the answers also show some similarity, since that most likely indicates some sort of relation.

Table 1: Most frequent phrases in the pilot corpus

No	Phrase	Score
1	問如何是佛	005254
2	如何是佛法	005036
3	問如何是道	005008
4	曰如何是佛	004971
5	云如何是佛	004941
6	如何是和尚	004917
7	如何是佛性	004882
8	如何是佛身	004875
9	如何是佛心	004872
10	如何是佛語	004871
11	如何是佛光	004870
12	如何是佛地	004869
13	如何是佛理	004868
14	問如何是本	004867
15	問如何是禪	004848
16	問如何是玄	004843
17	問如何是教	004832
18	問如何是定	004821
19	問如何是頓	004819
20	問如何是細	004817
21	問如何是易	004815
22	曰如何是祖	004790
23	曰如何是道	004725
24	云如何是道	004696
25	如何是祖意	004691
26	如何是祖令	004688
27	如何是道人	004635
28	曰如何是無	004633
29	如何是道者	004628
30	曰如何是法	004612
31	云如何是法	004583
32	畢竟如何是	004577
33	曰如何是賓	004566
34	曰如何是禪	004565
35	如何是法身	004564
36	如何是道如何是禪	004554

No	Phrase	Score
37	曰如何是此	004553
38	如何是諸佛	004547
39	曰如何是末	004546
40	曰如何是異	004545
41	如何是無為	004543
42	曰如何是僧	004542
43	曰如何是有	004541
44	曰如何是目	004540
45	如何是一句	004538
46	曰如何是九	004537
47	如何是衲僧	004536
48	曰如何是短	004535
49	如何是大道	004533
50	如何是不顧	004532
51	如何是不道	004531
52	如何是大乘	004529
53	如何是一喝	004527
54	如何是一法	004526
55	如何是古佛	004525
56	如何是大千	004522
57	如何是一燈	004521
58	如何是一劍	004520
59	如何是大覺	004519
60	如何是法眼	004516

Strategies for finding related passages

Some earlier research (ウィッテルン 2019) suggested that using an n-gram based approach should provide usable results. The method used there has been refined and now proceeds as follows.

1. All phrases with at least 3 characters will be used, shorter phrases ignored.
2. For every text, every selected phrase will be assigned a serial number, which indicates the position in the text.
3. Every phrase is then split into ngrams of 3 characters.
4. The ngrams are collected, ordered and assigned to buckets by frequency.
5. For every ngram in every phrase, other phrases that contain some or all of the ngrams in this phrase are collected
6. Three scores are assigned to the phrase based on the matching ngrams:
    1. @score : the number of ngrams matching in the other phrase, divided by the number of existing ngrams
    2. @tscore : the number of ngrams matching from the other phrase, divided by the number of ngrams in the target phrase
    3. @w (weight) : an indicator of the rareness of the phrase, calculated from the frequency buckets of the ngrams in the phrase (a higher bucket index is a less frequent ngram)
7. Target matches with a low @score will be discarded immediately
8. The results of this process will be analysed:
    1. For paragraphs, the matches recorded for the containing phrases will be clustered by

aligning phrases that are (near) neighbors in both the source and the target text.

2. In addition, the potential matches are separated bidirectional matches and those that only match either in the ’outgoing’ or the ’incoming’ phrase.

9. Finally, the resulting clusters with at least two phrases will be confirmed as candidate matches and made available for further analysis. For the confirmation, they need to fulfil the following conditions:
    - the source and target sequences have the same order
    - no big jumps in either sequence
10. Single matches with a high @w (weight) score will also be kept

The following section will illustrate how this works out in practice.

A pilot study to develop the method

Texts used for this study

An idea such as the one proposed here has to be tested and fleshed out sufficiently, before it can be generalized. A pilot study has therefore been started for this purpose.

Table 2: The 16 texts used in the pilot study, by creation date

ID	Title	Date	Characters	Phrases	Avg phrase len	Comment
B25n0144	祖堂集	0954	186964	38773	4.82	Lamp history
T51n2076	景德傳燈錄	1004	326859	53487	6.11	Lamp history
KR6q0003	景德傳燈錄	1004	372161	82671	4.50	Lamp history
X78n1553	天聖廣燈錄	1036	185002	34005	5.44	Lamp history
X78n1556	建中靖國續燈錄	1101	245936	53785	4.57	Lamp history
T48n2003	佛果圜悟禪師碧巖錄	1125	110408	19553	5.65	Koan collection
X79n1557	聯燈會要	1125	318834	71375	4.47	Merger of two Lamp histories
X79n1559	嘉泰普燈錄	1147	255218	51423	4.96	Lamp history
X68n1315	古尊宿語錄	1150~	442305	94520	4.68	Anthology of Recorded Sayings

ID	Title	Date	Characters	Phrases	Avg phrase len	Comment
T47n1998B	大慧普覺禪師宗門武庫	1163+	19898	3372	5.90	Koans and other anecdotes.
T47n1998A	大慧普覺禪師語錄	1180~	181536	30383	5.97	Recorded Sayings of Dahui Zonggao
T48n2004	頌古從容庵錄	1223	87113	15844	5.50	Koan collection
KR6q0610	眞字正法眼藏	1227	25119	5762	4.36	Koan collection by Dogen
T48n2005	無門關	1228	7865	1485	5.30	Koan collection
X80n1565	五燈會元	1253	569109	124638	4.57	Compilation of Lamp histories
X80n1566	五燈會元續略	1644	119009	18982	6.27	Continuation of 五燈會元

For this pilot study, a subset of 16 texts with approximately 2 million characters has been selected from the Zen Knowledgebase. These texts vary in genre and composition time, the details can be seen in Table 2. Among these is one text, the 景德傳燈錄 from 1004, which is used in two different versions - they differ somewhat, especially in markup, but represent still essentially the same content. Including both will allow specifically to understand how slight differences in markup will influence the outcome.

Text format

The texts are in XML format using the markup conventions by the TEI consortium. Basic textual units are marked as is appropriate for the text type, e.g. Koan collections have a Koan as the smallest structural division (div element), while lamp histories, which are texts that record sayings and doings by Zen masters, have the entry of one Zen master as smallest div (exception to this is KR6q0003, which has one encounter dialog etc. as the smallest unit. Below the div is most of the times a paragraph (p element), sometimes, e.g. for verses, also a line group (lg element) and (l element) for the lines. Sentences are not marked in most texts. In all cases the text is wrapped in a seg element for each phrase, and all seg elements have been given a unique identifier, so that it can be recognized and addressed. This is the same format as in the TLS database, as shown in Listing 1.

Listing 1: A short excerpt from one of the texts used

```
<div type="other">
  <p xml:id="pT51p0290a1801">
    <seg xml:id="T51n2076_CBETA_012-0290a1801.s1">鎮州臨濟義玄禪師<c n="。"/>
    </seg>
    <seg xml:id="T51n2076_CBETA_012-0290a1801.s2">曹州南華人也<c n="。"/>
    </seg>
    <seg xml:id="T51n2076_CBETA_012-0290a1801.s3">姓邢氏<c n="。"/>
    </seg>
    <seg xml:id="T51n2076_CBETA_012-0290a1801.s4">
      <lb n="0290a19" ed="T"/>幼負出塵之志<c n="。"/>
    </seg>
    <seg xml:id="T51n2076_CBETA_012-0290a1801.s5">及落髮進具便慕禪宗<c n="。"/>
    </seg>
    [... many more seg elements omitted ...]
  </p>
</div>
```

Data preparation

The first step in preparing the texts for analysis is to produce the ngram files. The example from Listing 1, transformed to the intermediate ngram format is shown in Listing 2.

Listing 2: Phrases with added ngrams.

```
<txt xmlns="http://hxwd.org/ns/1.0/reuse">
  <seg sid="T51n2076_CBETA_012-0290a1801.s1" sn="14682" s="鎮州臨濟義玄禪師">
    <g org="鎮州臨"/>
    <g org="州臨濟"/>
    <g org="臨濟義"/>
    <g org="濟義玄"/>
    <g org="義玄禪"/>
    <g org="玄禪師"/>
  </seg>
  <seg sid="T51n2076_CBETA_012-0290a1801.s2" sn="14683" s="曹州南華人也">
    <g org="曹州南"/>
    <g org="州南華"/>
    <g org="南華人"/>
    <g org="華人也"/>
  </seg>
  <seg sid="T51n2076_CBETA_012-0290a1801.s3" sn="14684" s="姓邢氏">
    <g org="姓邢氏"/>
  </seg>
  <seg sid="T51n2076_CBETA_012-0290a1801.s4" sn="14685" s="幼負出塵之志">
    <g org="幼負出"/>
    <g org="負出塵"/>
    <g org="出塵之"/>
    <g org="塵之志"/>
  </seg>
  <seg sid="T51n2076_CBETA_012-0290a1801.s5" sn="14686" s="及落髮進具便慕禪宗">
    <g org="及落髮"/>
    <g org="落髮進"/>
    <g org="髮進具"/>
    <g org="進具便"/>
    <g org="具便慕"/>
    <g org="便慕禪"/>
    <g org="慕禪宗"/>
  </seg>
</s>
```

This is a slightly simplified version of what will actually be used in the first step of the processing.

To facilitate inspection along the processing pipeline, the plain text of the phrase is retained in the @s attribute; also shown is the sequential number of the phrase has been added in the @sn attribute. The actual ngram that is matched against is in the attribute @org, since at some point a @nor attribute might be introduced to apply some kind of normalization to reduce complexity and give more opportunity for matches.

From the ngram files, now the frequency distribution for the sample can be calculated. Table 3 shows the details. For this table, all ngrams occurring more than 10 times have been lumped together in one entry; for the less frequent ones, they go in separate buckets. Both the number of items and the percentage is given.

Table 3: Ngrams buckets and associated information

Occurrences (buckets)	No of ngrams in group	% of ngrams	No of ngram matches	% of ngram matches	Samples
+10	22083	2.88	640252	31.18	如何是.作麼生.時如何
10	3440	0.45	34400	1.68	曹溪來.心如工.意如和
9	4440	0.58	39960	1.95	無人承.山前檀.前檀越
8	5993	0.78	47944	2.34	曾到西.須道取.取一半
7	7788	1.01	54516	2.66	莫曾到.工伎兒.更無消
6	12028	1.57	72168	3.51	西天亦.天亦無.即有也
5	17963	2.34	89815	4.37	弘濟禪.非但曹.子莫曾
4	34775	4.53	139100	6.77	到即有.山斷際.因参馬
3	69431	9.05	208293	10.14	觀音導.音導利.導利興
2	137260	17.89	274520	13.37	眼藏上.原山弘.嗣大鑑
1	452252	58.93	452252	22.03	山弘濟.曾問石.全靠希

What can be seen here is that the most frequent 2.88% of ngrams account for 31.18 of all matches. On the other side, 58.93%, that is much more than half of all ngrams, occur only one time, altogether accounting for 22.03% of all ngram matches. Now, this means that all these ngrams in the last bucket will never match any other ngram in the text corpus. The first row, with matches of more than 10 items on the other hand are so frequent that they are too unfocused to be useful for identifying related passages.

Some metrics

In order to develop the algorithm to produce useful results, - finding all correct matches, but no incorrect ones - a few parameters need adjustment through practical testing. These include

- length of ngram: 1, 2 or 3
- calculation method and use of the score parameters
- weighting of matches
- filters in the post processing of matches

The weight score @w

TFIDF (term frequency / inverse document frequency) is a frequently used metric to gauge the rareness of a term: frequent here but not frequent across all documents.

While we can not use this exact formula here directly, we still need a metric that tells us how frequent a phrase is compared to all phrases in the corpus. This is what is called the weight score here.

If we look at the frequency distribution of the ngrams in the test set (Table 3), we can take the row it occurs in as a way to sub-dividing the ngrams by frequency, using that as a way to assign a weight to this ngram. And for the phrase, the cumulative weight can quite easily taken as the average weight of the individual ngrams. In Listing 3 the weights and counts have been added on the ngram **g** for illumination purpose, for the algorithm there is no need to retain this information.

Listing 3: The ngram table with @w scores

```
<txt xmlns="http://hxwd.org/ns/1.0/reuse">
  <seg w="0.00" sid="T51n2076_CBETA_012-0290a1801.s1" sn="14682" s="鎮州臨濟義玄禪師">
    <g w="0" cnt="15" org="鎮州臨"/>
    <g w="0" cnt="17" org="州臨濟"/>
    <g w="0" cnt="12" org="臨濟義"/>
    <g w="0" cnt="12" org="濟義玄"/>
    <g w="0" cnt="11" org="義玄禪"/>
    <g w="0" cnt="32" org="玄禪師"/>
  </seg>
  <seg w="1.75" sid="T51n2076_CBETA_012-0290a1801.s2" sn="14683" s="曹州南華人也">
    <g w="2" cnt="9" org="曹州南"/>
    <g w="0" cnt="18" org="州南華"/>
    <g w="4" cnt="7" org="南華人"/>
    <g w="1" cnt="10" org="華人也"/>
  </seg>
  <seg w="5.00" sid="T51n2076_CBETA_012-0290a1801.s3" sn="14684" s="姓邢氏">
    <g w="5" cnt="6" org="姓邢氏"/>
  </seg>
  <seg w="5.00" sid="T51n2076_CBETA_012-0290a1801.s4" sn="14685" s="幼負出塵之志">
    <g w="6" cnt="5" org="幼負出"/>
    <g w="6" cnt="5" org="負出塵"/>
    <g w="2" cnt="9" org="出塵之"/>
    <g w="6" cnt="5" org="塵之志"/>
  </seg>
  <seg w="4.71" sid="T51n2076_CBETA_012-0290a1801.s5" sn="14686" s="及落髮進具便慕
```

```
禪宗">
    <g w="4" cnt="7" org="及落髮"/>
    <g w="3" cnt="8" org="落髮進"/>
    <g w="1" cnt="10" org="髮進具"/>
    <g w="8" cnt="3" org="進具便"/>
    <g w="8" cnt="3" org="具便慕"/>
    <g w="6" cnt="5" org="便慕禪"/>
    <g w="3" cnt="8" org="慕禪宗"/>
  </seg>
  <seg w="5.17" sid="T51n2076_CBETA_012-0290a1801.s6" sn="14687" s="初在黃蘗隨眾參
侍">
    <g w="4" cnt="7" org="初在黃"/>
    <g w="2" cnt="9" org="在黃蘗"/>
    <g w="9" cnt="2" org="黃蘗隨"/>
    <g w="9" cnt="2" org="蘗隨眾"/>
    <g w="0" cnt="20" org="隨眾參"/>
    <g w="7" cnt="4" org="眾參侍"/>
  </seg>
</s>
```

Is the weight a good metric to use in place of TFIDF? How should we treat these values? To help answer this question, it might be useful to look at the distribution across the whole set of phrases.

Table 4: Distribution of @w in 10 buckets

No	Percent	Range of @w
1	19.21%	0 - 0
2	3.14%	0.06 - 1.21
3	5.15%	1.22 - 2.42
4	7.68%	2.43 - 3.64
5	8.85%	3.65 - 4.85
6	11.53%	4.86 - 6.06
7	10.24%	6.07 - 7.27
8	11.24%	7.29 - 8.48
9	12.93%	8.5 - 9.7
10	10.02%	9.71 - 10

For Table 4, the values of @w have been calculated in a way that they should preferencially fall in equal partitions of 10%. However, due to the distribution, there are some distortions, most notably the first row, which has almost 20% of the phrases consisting entirely of frequent ngrams. However, considering that the percentage for the most frequent ngrams was more than 30%, this is already an improvement. Also interesting is that the maximum value observed here seems to be 10.

The match score @score

For the initial test run, the cut-off value for @score has been set to 0.6. This resulted in more than 1.7 million potential matches (for this first run, no @w based pruning has been applied yet). The distribution is as follows, because of the extreme distortion, only 7 buckets have been generated. There are about 10% just above the cut-off value, and the remaining matches are almost entirely nearly exact matches.

Table 5: Distribution of @score in 7 buckets

No	Percent	Range of @score
1	10.3455%	0.61 - 0.8
2	0.3478%	0.81 - 0.83
3	0.2271%	0.84 - 0.86
4	0.0307%	0.88 - 0.89
5	0.0030%	0.9 - 0.93
6	0.0011%	0.94 - 0.95
7	89.0447%	1 - 2

The target match score @tscore

Table 6 shows the @tscore, which is the same dataset and conditions as for @score, except that the match direction is reversed, here we calculated the score for incoming matches, from the direction of the target. In this case there is no cut-off value, so the range of values is slightly larger. However, the distribution is quite different and only about 43% of values are above 0.6.

Table 6: Distribution of @tscore in 8 buckets

No	Percent	Range of @tscore
1	10.1997%	0.05 - 0.2
2	13.1972%	0.21 - 0.33
3	3.8099%	0.35 - 0.46
4	10.9404%	0.47 - 0.6
5	3.7887%	0.61 - 0.73
6	1.8589%	0.74 - 0.86

7	0.0336%	0.88 - 0.95
8	56.1717%	1 - 2

Some preliminary results

The method is still under heavy development, so the results here should be considered only preliminary. As a result of the matching process, pairs of aligned matches with only small gaps have been generated. Such sequences will be called ’windows of related phrases’, or just ’windows’. They vary greatly in length, but still give some kind of rough handle on the ’relatedness’ of two texts. These windows will be ultimately the ‘related passages’ this research attempts to identify, but a cursory inspection shows that they need to be further scrutinized, so the actual numbers in Table A1 will probably change.

The Appendix has the full matrix table for all 16 texts. The matches are directional and differ slightly by direction, but are not fundamentally different. There can and are also self-matches to other sections in the same text, which is why there are also numbers recorded. Indicated is the number of matching ‘windows’ for all text pairs.

Example of aligned window

Finally, Table 7 gives a first preview of how the alignment of related passages could be shown. In this case, a section from the biography of the Buddha Śākyamuni in three different texts, the 五燈會元 (1253) compared with the 景德傳燈錄 (1004) and the 聯燈會要 (1125). The third and fifth column indicate a match (=) or a non-match (x) for this specific line. The display algorithm will display the whole text of this passage between the first and the last entry of the ‘window’. Sometimes, as in row 15 or 21, the text is inserted even if no match has been registered. This demonstrates that the alignment algorithm developed here can indeed cope with differences in the texts both on the level of characters, as well as on the level of phrase segmentation. This also shows that the counting of matching phrases is not solely a good metric for evaluation; the number of related passages, even if they vary by length, might be a better choice.

Table 7: Matches between 五燈會元 (X80n1565) and two selected other texts

No	X80n1565	m	T51n2076	m	X79n1557
1	姓刹利	=	釋迦牟尼佛(賢劫第四尊)姓刹利。	x	
2	父淨飯天	=	父淨飯天。	x	
3	母大清淨妙位	=	母大清淨妙。	x	
4	登補處	=	位登補處生兜率天上。	x	
5	生兜率天上	x		x	

6	名曰勝善天人	=	名曰勝善天人。	x	
7	亦名護明大 士	=	亦名護明大士。	x	
8	度諸天眾	=	度諸天眾說補處行。	x	
9	說補處行	x		x	
10	於十方界中	=	亦於十方界中現身說法。	x	
11	現身說法	x	普耀經云。	x	
12	普曜 經云	x		x	
13	佛初生刹利王家	=	佛初生刹利王家。	=	佛初生刹利王家。
14	放大智光明	=	放大智光明照十方世界。	=	放大智光 明。
15	照十方世界	x		x	照十方法界。 地湧金蓮華。
16	地 涌金蓮華	=	地涌金蓮華自然捧雙足。	x	
17	自然捧雙足	x		=	自然捧雙足。
18	東西及南北	=	東西及南北各行於七步。	=	東西及南 北。
19	各行於七步	x		=	各行於七步。
20	分 手指天地	=	分手指天地作師子吼聲。	=	分手指天地。
21	作師子吼聲	x		x	作大師子吼。
22	上下及四維	=	上下及四維無能尊我者。	=	上下及四 維。
23	無能尊我者	x		=	無能尊我者。
24	即 周昭王二十四年甲寅歲四月八日也	=	即周昭王二十四年甲寅歲四月八日也。	x	
25	至四十二年 二	=	至四十二年二月八日。	x	

	月八日				
26	年十九	=	年十九欲求出家。	x	
27	欲求出家而自念言	x	而自念言。	x	
28	當復何遇	=	當復何遇。	x	
29	即 於四門遊觀	=	即於四門遊觀見四等事。	x	
30	見四等事	x		x	

Conclusions

A method has been developed to find related passages, even if there are comparatively few exact character matches, based on structural characteristics of the documents in addition to character based matches. Based on a preliminary qualitative evaluation, the method is promising, but there is still a lot of room for improvement. It will need a proper quantitative process to study the influence of parameter variation and settle on the most performant set.

If extended to the whole Zen Knowledgebase and beyond that to Kanripo and the TLS database, this method will provide a useful way to study intertextuality and discover which text passages have been most influential and how this changed over time.

In an earlier study (Wittern 2020), *KanripoX*, a tag set for recording textual variation had been developed. This operates on character tokens in a document and records nexus points and sequences thereof to make this information available to applications in an interchangeable way. It would be useful to further develop this tag set, or provide a similar one for the parallels found in this study, to make them available and interchangeable.

References

ウィッテルン・クリスティアン (2019), 《景德傳燈錄》から《五燈會元》へ　ー　禅宗の変遷と燈史の編集, 東方学報（京都）　第９４冊.

Wittern, Christian (2021), KanripoX: A tagset for connecting digital texts, 東洋学へのコンピュータ利用第 33 回研究セミナー. (http://kanji.zinbun.kyoto-u.ac.jp/seminars/oricom/PDFs/2021PDFs/2wittern.pdf)

Appendix A

Table A1 part 1: Number of ‘related windows’ by text

Text	(1) B25n0144	(2) T51n2076	(3) KR6q0003	(4) X78n1553	(5) X78n1556	(6) T48n2003	(7) X79n1557	(8) X79n1559
B25n0144 祖堂集	4135	2653	3020	664	217	214	1590	116
T51n2076 景德傳燈錄	1748	15206	31573	4100	1459	597	5275	594
KR6q0003 景德傳燈錄	1412	19469	1262	2557	301	369	3914	203
X78n1553 天聖廣燈錄	376	1920	2046	4919	1885	161	2585	469
X78n1556 建中靖國續燈錄	119	248	279	1787	10724	275	1700	522
T48n2003 佛果圜悟禪師碧巖錄	309	919	954	599	617	4432	1695	432
X79n1557 聯燈會要	1989	7923	8830	5565	3136	1996	20813	3778
X79n1559 嘉泰普燈錄	192	1369	1278	2563	3792	506	3975	4602
X68n1315 古尊宿語錄	1984	12933	12076	21309	14529	2815	16069	5043
T47n1998B 大慧普覺禪師宗門武庫		19	13	9	46	6	155	136
T47n1998A 大慧普覺禪師語錄	191	1118	1064	783	852	635	2049	1142
T48n2004 頌古從容庵錄	132	622	717	195	276	704	1166	168
KR6q0610 眞字 正法眼藏	218	634	661	219	186	253	1169	119
T48n2005 無門關	14	63	59	31	24	29	155	30
X80n1565 五燈會元	4016	33637	36036	17363	17622	2952	32117	24674
X80n1566 五燈會元續略	18	389	77	365	279	53	303	278

Table A1 part 2: Number of ‘related windows’ by text

Text	(9) X68n1315	(10) T47n1998B	(11) T47n1998A	(12) T48n2004	(13) KR6q0610	(14) T48n2005	(15) X80n1565	(16) X80n1566
B25n0144 祖堂集	1150		301	119	175	35	2416	39
T51n2076 景德傳燈錄	5418	16	718	363	515	36	17829	464
KR6q0003 景德傳燈錄	1305		490	246	152	41	11742	73
X78n1553 天聖廣燈錄	6294	6	221	97	97	22	3270	19
X78n1556 建中靖國續燈錄	3655	12	303	175	151	31	3636	37
T48n2003 佛果圜悟禪師碧巖錄	2089	9	659	719	306	74	1726	98
X79n1557 聯燈會要	13063	257	2886	1713	1294	272	20692	393
X79n1559 嘉泰普燈錄	5340	207	1761	218	122	54	22942	629
X68n1315 古尊宿語錄	38655	111	3523	1157	1453	238	20617	973
T47n1998B 大慧普覺禪師宗門武庫	33	47	17	10	3		303	7
T47n1998A 大慧普覺禪師語錄	2203	32	4665	332	311	90	2885	186
T48n2004 頌古從容庵錄	732	7	309	1674	247	32	1208	163
KR6q0610 眞字正法眼藏	950		351	364	343	80	1209	32
T48n2005 無門關	135	3	86	53	52	45	134	15
X80n1565 五燈會元	24623	461	5465	1901	1992	226	35672	2124
X80n1566 五燈會元續略	526	9	197	116	12		895	657