TACL Methods Guide

Michael Radich

The aim of this document is to describe Buddhological-philological methods for the careful, rigorous application of TACL[1] to some representative research problems in the study of Chinese Buddhism. My main aim is to suggest how users might think their way from research questions to an effective application of the tool to find possible evidence.

This document is intended to complement other documents describing the functions of TACL itself, or the TACL GUI, and how to operate those tools. Readers should therefore read this document in conjunction with the TACL GUI User's Manual and the TACL (command line) User's Guide. This document will assume knowledge of basic TACL operations and terms (such as "corpus", "database", "catalogue", "Intersect", "Difference", "Supplied Intersect", TACL "results" or "Filter/Rationalize", and such terms as "n-gram", "string", etc.), and we will not repeat explanations of such matters here. We will also assume here that readers have learned from those other documents how to view and sort results output by TACL in Excel or an equivalent tool.

Readers should also note that additional examples of simple TACL tests are laid out in the TACL GUI User's Manual, using cases known from previous scholarship: (re)"discovering" the overlap between T150A and T735C; "discovering" the copying from T453 in **Ekottarikāgama* 48.3; finding differences between two versions of the *Ugra-paripṛcchā*, T322 by An Xuan and Yan Fotiao, and T323 by Dharmarakṣa; finding stylistic differences between Zhu Fonian and Saṅghadeva, and using them to assess the question of which is more likely to have been the (main) translator behind the **Ekottarikāgama* T125. These examples also illustrate some of the methodological principles we aim to describe here.

The abstract considerations laid out in this document are often illustrated in concrete detail, more fully than is possible here, by prior research publications applying TACL to a range of problems. Those studies were conceptualised in part as methodological pilot studies. Readers might benefit from consulting those studies, which are listed here, and at the end of the TACL GUI User's Manual. Most of the studies in question are available for download at academia.edu.

[1] The name "TACL" is a "backronym" motivated mainly by a love of puns, and it doesn't really matter what it stands for. Interested users can contact us with suggestions for fun interpretations; users who prefer to have answers can think of it as meaning "Text And Corpus Lab".

Finding sources of a text

The two core functions of TACL are to find contiguous strings in either 1) the intersection or 2) the difference between two (or more) texts, or sets of text. We deal first with a typical application of the Intersect test, to discover the sources of a text. TACL has proven worth in application to such questions. Most typically in work to date, such tests have been used to determine whether texts were composed in China (i.e. whether they are "apocrypha"), or what works a Chinese author knows or cites (especially when those citations are not explicit, or source texts are not named by the author).

In order to run a TACL Intersect test designed to find sources of a given text, then, the user first writes a *catalogue* with *labels* that place in one group the text that is the target of the test, and in a second group, all other texts that might conceivably be sources (for our present purposes, we will use as the second group all other texts presented in the Taishō as translations, about 1700 or so texts altogether).

For example, let us imagine that we are looking for sources of the *Mahāyāna Awakening of Faith* T1666 in the rest of the Taishō. A catalogue for an Intersect test for this purpose would read as follows:

```
T1666 AF

T0001 T-trans
T0002 T-trans
T0003 T-trans
T0004 T-trans
T0005 T-trans
 ...
```

(...and so on, for the entire Taishō translation corpus in the "T-trans" label.)

Now, one of the most important (Buddhological-)methodological principles in using TACL is that in order for one's results to actually capture what one is after, it is important to think through carefully what is known about the contents of the texts or corpora that one is comparing, to avoid misleading or false results. In this example, it would be sensible to exclude from the contrast corpus the supposed "second translation" of the *Awakening of Faith* by Śikṣānanda, T1667, which, as (at the very best) a revision of T1666,[2] is likely to contain significant verbatim matches with it, but is certainly not one of its sources; and similarly, to exclude the 釋摩訶衍論 T1668,[3] which, as a commentary on T1666, also cannot be among its sources. Thus, the stretch of our "catalogue" around those works will look like this:

[2] Common consensus is that the earlier version of the *Awakening of Faith,* T1666, was composed in China. This makes the existence of a second "translation" peculiar, and suspect—it is most likely that Śikṣānanda merely prettied up the older text, and expanded it on certain doctrinal points that were thought important, and by his issue of a "new translation", gave the older text the appearance of authentic Indic provenance, and his patron the merit of having sponsored a pious work.

[3] Putative Indic original ascribed to Nāgārjuna; "translation" ascribed to the shadowy *Vṛddhimata(?), 筏提摩多; but commonly regarded as "apocryphal", and probably composed as late as the ninth century.

```
    ...
    T1663 T-trans
    T1664 T-trans
    T1665 T-trans
    T1669 T-trans
    T1670 T-trans
    T1671 T-trans
    T1672 T-trans
    ...
```

(i.e. leaving T1667 and T1668 out altogether—recall that T1666 itself is already given, with a different "label", at the top of the catalogue file.) Alternatively, the same result can also be achieved by formatting the catalogue file as follows, for this section:

```
    ...
    T1663 T-trans
    T1664 T-trans
    T1665 T-trans
    T1667
    T1668
    T1669 T-trans
    T1670 T-trans
    T1671 T-trans
    T1672 T-trans
...
```

That is to say, we leave the works we want to exclude from the test without a "label", and TACL ignores them.

This Intersect test will yield a set of results giving n-grams occurring at least once each in both T1666, and at least one other text (in the second corpus defined by the catalogue). The total quantity of raw results is likely to be large, and to overwhelm the human analyst. We therefore use the functions of TACL "results" (a subset of which functions are called "Filter/Rationalize" in the TACL GUI) to filter the results by such parameters as n-gram size and n-gram count. All other things being equal, a direct borrowing from one text to another, such as we are looking for here, is more likely to be evidenced by rare or unique strings than by common ones; and we will regard as evidentially more significant long matches, rather than short ones. We can thus filter, for example, to keep in our results only strings that occur a few times (say, maximum four times), and only strings above a certain size (say, minimum six characters long). Typical protocols for such "sort" procedures are described in a little more detail in the TACL GUI User's Manual.

Obviously, if we want to use the results of such a test as evidence that one text is a possible source of another—that the later text borrowed from the earlier—we must be careful, and apply critical philological judgement. For example, 阿耨多羅三藐三菩提 (_*anuttarasaṃyaksambodhi_) is relatively long, but it is too common to serve as evidence of direct borrowing from any one text to another. Similarly, the string 生恭敬心若 is very rare—it occurs only once each in T158, T1246, and T1521—but it is not likely to indicate any direct relationship of borrowing between any of the texts that contain it.

This leads directly to another important point of method in use of TACL results: in most cases, the human philologist must interpret the results, and weigh up their likely, qualitative evidential significance. We will return to this point below.

Style, as evidence of authorship or translatorship

TACL can also be used to discover text-internal, stylistic evidence bearing upon questions of authorship or translatorship.[4] For examples of such studies, see Radich and Anālayo (2017), Radich (2017a, 2017b).

Exact methods for application of TACL to such problems depends very much upon multiple features of the particular problem. For example:

— For some problems, external evidence might suffice to confine us to only two realistic candidates for authorship/translatorship of a given text. This is the case, for example, for the **Ekottarikāgama* (Radich and Anālayo 2017, Radich 2017a), and the **Mahāmegha* 大方等無想經 (Radich 2017b).

— For some candidate translators or authors (e.g. Zhu Fonian), we may have a solid corpus of texts for which we regard the ascription as reliable, which can be used as a benchmark in determining typical style. For others (e.g. Saṅghadeva) we may have only a single benchmark text (once again, see Radich and Anālayo 2017, Radich 2017a). In extreme cases, we may even have reason to believe that *none* of our received canonical ascriptions is reliable, so that we have no benchmark—but we may nonetheless wish to investigate the probability that that figure was in fact the main person responsible for some of our canonical texts. (I believe Baoyun 寶雲 is such a case.)

— For many problems, by contrast to those above, we may have no idea where to start looking for the true author or translator of a text. Such cases can be further divided into various groups. Such a text might be canonically ascribed to a given historical figure, in which case, we might regard it as progress if we can come up with solid evidence to *dissociate* the text from that ostensible translator or author, even if we cannot go so far as to find the true author instead. In other cases, a text might be canonically regarded as anonymous, so that we do not even have this option.

[4] Complex problems obviously surround the fact that especially in the Chinese tradition, translation was often a collective endeavour. However, I believe that there is ample evidence that empirically speaking, translation groups still evince styles coherent and distinctive enough, against meaningful points of comparison, that we are warranted in treating them as consistent stylistic actors. This means that for many purposes, it is possible to treat the name of a "translator" like "Paramārtha" as a convenient label for the corporate entity (group, workshop, etc.) that produced a body of texts, and proceed with our analyses "as if" we are dealing with individuals. Due to limitations of space, I cannot substantiate this claim here.

— In some cases, external evidence might at least suffice to show that a text must belong to a given historical period (or range), for example, that it was extant by a certain date. In other cases, we may also be confronted by a very wide range of possible dates.

Our methods, and the assumptions upon which they are founded, must be carefully adapted to these considerations and others like them, as they apply to our particular case. I will therefore discuss more than one strategy.

1. Where we have reasonable grounds to consider only a limited number of candidate translators/authors (minimally, two), we can ask the following question equally of each of the two (or more) candidates:

> What stylistic features regularly recur in [CANDIDATE] and in [TARGET TEXT] but not in [OTHER CANDIDATE(S)]? In other words: What *is in A and B, but not in C*?

To give a concrete example, when Ven. Anālayo and I attempted to see whether the **Ekottarikāgama* T125 was by Zhu Fonian or *Saṅghadeva, we looked for stylistic features (i.e., in TACL, strings, because that is all that TACL can find) that appear in T125 and Zhu Fonian, but not in Saṅghadeva; and then we also attempted to find strings that appeared in T125 and Saṅghadeva, but not in Zhu Fonian. We must remember that this test will only be even-handed, and therefore rigorous, if it is applied equally in all directions, i.e. to all candidates.

In TACL terms, as described in more detail in the TACL GUI User's Manual, this is a Supplied Intersect test, "concatenating" or chaining together the results of two previous tests: a Difference, and an Intersect.

We then post-process the results, using "TACL results" (at the command line) or "Filter/Rationalize" (in the GUI), to limit the raw results to items occurring repeatedly, and relatively short items that are likely to be recurring terms or turns of phrase. Typical protocols for "TACL results" or Filter/Rationalize operations for such cases are described in the TACL GUI User's Guide. We can further increase our efficiency in zooming in on likely traits of style by adopting appropriate sort protocols in Excel (or an equivalent tool), as also described in the TACL GUI User's Guide.

Readers should note that at the first step of this Supplied Intersect, we ask, effectively: What is characteristic of A, but not of B?—in this case: What is characteristic of Zhu Fonian, but not Saṅghadeva (or vice versa). An extremely important point here is that what we get out of this test is only as good as what we put in. The benchmark corpus that we use to find or define the style of a given figure or group must be rigorously constructed, on the basis of informed, conservative criteria. Otherwise, we risk missing information vital to consideration of our question, or discovering so-called evidence that in fact does not characterise the figure or group in question. We will return to this question of rigour in the construction of benchmark corpora below.

2. A text might carry a traditional ascription, which we regard as dubious, but we might not immediately know where else we might look for a more likely translator/author. For example, we might doubt that the *Fo yin sanmei jing* 佛印三昧經 T621 was really translated by An Shigao,[5] as the tradition claims—but we might not have any other obvious alternate candidate for the authorship or translatorship of the text. In such a case, we can ask the following questions.

> What stylistic features (strings) recur in [TARGET TEXT] but never in [RECEIVED TRANSLATOR/AUTHOR]? Where else do those strings most frequently occur?

For example, the corpus of Faju 法炬 is riddled with serious problems of attribution. As Zürcher notes,[6] Dao'an ascribes only four works to Faju (sometimes in cooperation with the even more shadowy Fali 法立); of those works, only three are still extant today: T23, T211, and T683.[7] Most of the received ascriptions to him first appear in the *Lidai sanbao ji* 歷代三寶紀 T2034, as part of a wide-ranging pattern of problems in that catalogue in the presentation of hitherto unknown ascriptions.[8] Thus, we might take one of the 24 other works ascribed to Faju in the Taishō—say, the *Aṅgulimāla-sūtra* 鴦崛髻經 T119—and test it against these three texts, as the most plausible benchmark corpus for Faju's style.[9] In TACL terms, we run an "asymmetric Difference" test, to discover all n-grams that distinguish T119 from the three "benchmark" Faju texts listed above. Because we are looking for style, we filter to keep only items that recur (so at least min-count 2). We can then look for the items thrown up by this test elsewhere in the canon.

Even on preliminary examination, the results yielded by such a test are potentially quite interesting. For example, the string that most frequently recurs in T119, but not in our benchmark Faju corpus, is -城乞-. Examination of this string in context shows that in T119, it always occurs in the longer string 舍衛城乞食 ("beg for food [in] Śrāvastī"). This longer string, however, is quite restricted in its distribution: it occurs numerous times in the *Saṃyuktāgama* T99 ascribed to Guṇabhadra, the **Ekottarikāgama* T125 of Zhu Fonian, the Dharmaguptaka Vinaya T1428 ascribed to Zhu Fonian and

[5] See Nattier (2008): 15 n. 26.

[6] Zürcher (1959/2007): 70, 345 n. 254.

[7] T2145 (LV) 9c19-10a3.

[8] Radich, Michael. "Fei Changfang's Treatment of Sengyou's Anonymous Texts." *Journal of the American Oriental Society* 139.4 (2019): 819-841.

[9] A significant number of the texts ascribed to Faju are included in a group that Mizuno identified as probably having originally formed part of an alternate translation of the **Madhyamāgama,* which was then broken up and its parts canonised separately; Mizuno (1989). There is a high likelihood that these texts may have been composed in part on the basis of our extant **Madhyamāgama* T26 (or *vice versa*), and I have therefore avoided them for the purposes of this example. A number of others (e.g. T33, T34) are very short, and likely to provide us with slender handholds at best, and I have therefore avoided them too.

Buddhabhadra, and the Sarvāstivāda Vinaya T1435 ascribed to *Puṇyayaśas and Kumārajīva.[10] The same phrase also occurs less copiously—only a handful of times each—in the *Dīrghāgama* T1 of Zhu Fonian, the anonymous *Saṃyuktāgama* T100, Zhu Fonian's *Udānavarga* T212, the *Vimaladattaparipṛcchā* 無垢施菩薩應辯會 T310(33) ascribed to Nie Daozhen 聶道真, the Mahāsāṅghika Vinaya T1425, the **Sarvāstivāda-vinayamātṛkā* T1441, the *Fenbie gongde lun* 分別功德論 T1507, the **Mahāprajñāpāramitopadeśa* T1509, the *Vibhāṣā* T1546 ascribed to Buddhavarman 浮陀跋摩 and Daotai 道泰, and the *Śataka-śāstra* T1569 ascribed to Kumārajīva. This distribution is notable, first, for the way it clusters tightly in time in a few decades around 400 CE, a century later than Faju; and second, for the striking prominence of works associated with Zhu Fonian.

However, the above operation in fact yields a list of 80 n-grams in total, and it would be quite time-consuming to investigate individually each of these n-grams in detail. As a first approximation, it might be useful for us to know where, if anywhere, those 80 n-grams appear in greatest number in the translation corpus. For this purpose, we can run TACL Search on that list of n-grams. (For more details on TACL Search, see the TACL GUI User's Manual.) This allows us to see, crudely (without allowing for such factors as text length) where these 80 n-grams appear in greatest quantity in the canon—in other words, to find possible "hotspots" where the distinctive style of T119, against the benchmark Faju texts, clusters.

When we do run a TACL Search on the 80 n-grams identified by the above difference test, we find that apart from T119 itself, the results suggest possible confirmation of the pattern we began to glimpse with the single string 舍衛城乞食. Setting aside some noise,[11] 40 out of the 80 n-grams appear in Zhu Fonian's **Ekottarikāgama* T125; 29 appear in Guṇabhadra's *Saṃyuktāgama* T99; 29 (a different set) also appear in the Dharmaguptaka Vinaya T1428; 27-29[12] appear in T212; 28 appear in Yijing's T1442;[13] 26-27 in the *Madhyamāgama* T26; and 25 in the *Dīrghāgama*. Concentration of these n-grams in the works of Zhu Fonian is noticeable, and appears (in this crude overview) possibly to be too widespread to be accounted for by textual parallels to T119 alone.

We are here only exploring T119 as an example of methods for the application of TACL to one type of question, but were we investigating this text seriously, it would be worth considering, at this juncture, the hypothesis that the text is in fact by Zhu Fonian, not Faju. We might investigate that

[10] An additional methodological caution: Some of these results may be explicable because T99 and T125, at least, contain parallel texts to T119, and might in these portions overlap in content and wording for that reason; were we investigating this problem seriously, we would need to check this possibility carefully.

[11] E.g. T2121, as just mentioned; also T2122, T310.

[12] Variation in counts depends upon the textual witness.

[13] When considering material shared like this between multiple large Vinaya texts, we have to consider the fact that later Vinaya translators appear to have lent heavily upon the work of their predecessors. Anecdotal observation suggests, for instance, that Yijing's massive Mūlasarvāstivāda Vinaya in particular sometimes contains at least one instance of nearly every n-gram under the sun.

possibility further by examining the n-grams found by our difference test in their original contexts; by running tests comparing the style of T119 to a benchmark Zhu Fonian corpus; and so on.

Other examples

In Radich (2014), I used TACL to discover evidence with which I argued the following main points (confining myself to claims relevant for the present purpose of exemplifying TACL method):

1) Chapters of the *Suvarṇabhāsottama* T664 ascribed to Paramārtha were in fact probably composed in China;
2) Even setting aside passages in which we can see debts to earlier Chinese sources, these chapters also display a large number of recurring stylistic features atypical of Paramārtha, but typical of Sui translators, that suggest that the chapters may have been produced or at least revised closer to the Sui context than to Paramārtha's own group.

In preparing that study, I used TACL as follows:

1) In looking for Chinese sources, I used TACL Intersect as described above, to look broadly for unique matches between Paramārtha's chapters of T664 and any other single "translation" text in the Taishō.
2) In investigating the style of the chapters, I used roughly the same method described above (for Zhu Fonian versus *Saṅghadeva as translator of the **Ekottarikāgama*) to look for n-grams found in those chapters and in Sui translators, but not in a benchmark corpus of texts reliably ascribed to Paramārtha (or *vice versa*) (again, the key pattern is *in A and B, but not in C*, for which we use TACL Supplied Intersect).

In another study (2019, 2020a), focusing on the *Mahāparinirvāṇa-sūtra* T7, I had discovered that three texts were particularly closely related to one another on the basis of internal evidence, even though traditional ascriptions would suggest no special relation: the *Mahāparinirvāṇa-sūtra* 大般涅槃經 T7 is attributed to Faxian 法顯, the *Guoqu xianzai yinguo jing* 過去現在因果經 T189 attributed to Guṇabhadra 求那跋陀羅, and the **Mahāmāyā-sūtra* 摩訶摩耶經 T383 ascribed to Tanjing 曇景. In a follow-up study (2018), I wanted to see what could be learnt about the probable authorship/translatorship of these texts, or other aspects of the context in which they were produced. I used TACL Intersect to find rare, relatively long verbatim matches between each of these texts and other canonical translation texts. This did not enable me to pin down the translator or author of texts in this triad, but it did show that repeatedly, for a range of longer phrases usually expressing formulaic notions that recur in many Buddhist texts, this triad shared extremely specific wording with texts in a delimited historical and geographic context—the first part of the fifth century, in the South of China. On the basis of this evidence, I argued not only that these texts were products of that milieu, but also that we can thereby glimpse otherwise obscure dynamics of textual circulation and reception in that milieu—it shows us what was in the "library" (including the heads) of the people who produced these three texts, and, moreover, how they absorbed, and themselves used, the contents of the texts they knew.

Rigour in the construction of benchmark corpora, and "grey zones" of uncertainty

As mentioned above, when choosing texts or defining corpora as benchmarks or points of comparison, it is vital that we do so as rigorously and conservatively as possible. We must scrutinise our assumptions thoroughly and critically, and know as much as possible about the nature and content of the texts.

In particular, many corpora ascribed to major translators comprise numerous ascriptions that are problematic or plain wrong—in the case of Zhi Qian, for example, more than half the corpus ascribed in the Taishō is probably not really his;[14] conversely, in the case of Zhu Fonian, perhaps as many as half of his authentic texts do not carry his name in the received canon.

If we attempt to define a benchmark corpus, and thereby, find traits of style, for a translator (or group) like "Zhi Qian", but we follow an incorrect ascription and incorporate a text actually by another figure (or group) into our benchmark corpus, the "signal" that we pick up could be seriously garbled. We can see how this would work by considering two scenarios for such errors:

1) Text A is actually by our author, but wrongly ascribed to someone else in tradition (e.g. the traditional ascription of the *Sukhāvatīvyūha* T362 to *Lokakṣema instead of Zhi Qian[15]). We mistakenly put Text A into the contrast corpus, for style that contrasts with that of our author, and run a Difference test. Every n-gram that is in fact unique to our author, but which occurs in Text A, will be excluded from our evidence. We will possibly miss huge quantities of relevant evidence.
2) Text A is actually by someone else, but wrongly ascribed to our author in tradition (e.g. the ascription of the *Sūtra of Humane Kings* 仁王般若波羅蜜經 T245 to Kumārajīva).[16] We mistakenly put Text A into the benchmark corpus for our author. Every n-gram that appears in Text A, but not in the contrast corpus, will wrongly end up in our results as a supposed trait of the style of our author, even though many of these n-grams may in fact be characteristic of some other, completely different figure.

The corresponding need to exercise rigour in the construction of benchmark corpora cannot be emphasised strongly enough. It is far better to err on the side of excluding authentic texts from a benchmark corpus, and thereby to reduce the information available to us as part of our baseline, than to define too liberally a baseline (or contrast) corpus that turns out to contain junk. It is therefore vital that benchmarks be defined with extreme conservatism.

At the risk of belabouring the obvious, note, once more, that it is equally necessary to apply conservatism on the "other side" of any comparison, for the contrast corpus.

This need for rigour means that very often, in setting up two-way comparisons, it is important that we think rigorously and systematically about a *grey zone* between the options at issue, where we place

[14] Using as criterion the assessments in Nattier (2008).

[15] https://dazangthings.nz/cbc/text/1286/

[16] https://dazangthings.nz/cbc/text/189/

in limbo all items about which we might not be sure. If we were investigating the Zhi Qian corpus, once more, we might for a start conservatively place outside *both sides* of the initial comparison all texts in the half of the traditional Zhi Qian corpus treated as problematic by Nattier (2008) (our most informed assessment of the external evidence).[17]

Careful definition of the "text(s)"

It is also important that we flexibly define the "text", as a unit of analysis, in a rigorous manner that actually matches the purpose of our analysis, rather than passively accepting the units in which supposed "texts" are packaged by the Taishō (and therefore by CBETA). For example, a large text like the **Mahāsaṃnipāta* 大方等大集經 T397 actually includes multiple texts, ascribed variously to at least four translators or groups (and if ascriptions were corrected, may in fact include even more diversity than this indicates). For many purposes, then, it obviously makes no sense to treat this large collection as a single "text". At the same time, if we omit the material contained in this collection from any study of a figure like *Dharmakṣema or Narendrayaśas (to each of whom nearly half the entire collection is ascribed), we will miss a very significant source of information.

Another example, on a different level of scale, may be found in Zhi Qian's *Aṣṭasāhasrikā prajñāpāramitā* 大明度經 T225, which Nattier has shown may be divided into three heterogeneous parts: the first chapter, in which we must further distinguish between root text and an interlinear commentary; and subsequent chapters.[18] For purposes of stylistic analysis, these three different layers of material must be treated separately. Further examples of this problem are legion.

This consideration lies behind the construction of the modified "Radich" Taishō corpus, for use with TACL, which we have made available for download at Zenodo (download here; corpus described here).

A cautionary tale about mistaken assumptions

Failure to appreciate these various points could potentially lead to abuses and misapplication of the tools, and egregious error. An example of such dangers can be drawn from my experience in preparing Radich (2014), in which, as I mentioned above, I argued that four chapters of the *Suvarṇabhāsottama* ascribed to Paramārtha were in fact composed in China.

I was initially led to undertake that study when I observed in passing, in the course of other analyses of works ascribed to *Dharmakṣema, that markers typical of *Dharmakṣema but atypical of Paramārtha

[17] Nattier (2008) is our best single source of summary information about the state of critical ascription studies not only for Zhi Qian, but for all texts ascribed to figures in the period prior to 280 CE. For a far less complete or systematic source of information about other periods, researchers will hopefully sometimes find it useful to consult our "CBC@" database at http://dazangthings.nz/cbc/. It is to be hoped that over time, and with contributions from the scholarly community, this resource will gradually become more complete, and help scholars keep abreast of existing studies critically assessing traditional attributions for all of our texts.

[18] Nattier (2008[2010]).

seemed to occur repeatedly in these chapters. However, the hypothesis that I initially formed, and spent the best part of three months investigating, turned out to be completely wrong. My mistake was caused principally by one apparently well-grounded assumption—which also turned out to be wrong—abetted by a misdirected inference on the basis of one misleading circumstantial fact.

The circumstantial fact that served as springboard to launch my misguided hypothesis was that the first translation of the *Suvarṇabhāsottama* was by *Dharmakṣema (T663)—though that translation is said not to have included equivalents to the chapters ascribed to Paramārtha. The ill-fated hypothesis I formed on that basis was this: Unbeknownst to the bibliographic tradition, the chapters in question had in fact been translated by *Dharmakṣema, and either Paramārtha's translation had been a revision on the basis of that earlier translation, or the ascription of those chapters to Paramārtha was downright wrong. Note that this was a *false* hypothesis: it turned out my hunch was woefully mistaken.

The false assumption that propelled me in the direction of this hypothesis was that these chapters must be genuine translations. This assumption was based upon unusually strong external evidence[19]—especially the fact that at least one Tibetan version of the text, incorporating the same chapters, is held by Tibetan tradition to have been translated from Sanskrit, which would ordinarily indicate that these chapters indeed once existed in India.[20] In light of this last fact in particular, it simply never dawned on me that these chapters could have been composed in China—not, at least, until very late in the process of my investigations, well after I had first built a gigantic castle in the air (空中樓閣) and then had it crash down around my ears.[21]

As I argued in my eventual paper, I now believe that the real explanation for the presence of these "*Dharmakṣema-like" markers in "Paramārtha's" text was not that those chapters were originally translated by *Dharmakṣema—nor, indeed, that the composers of "Paramārtha's" text were drawing upon work by *Dharmakṣema (no works by him were among the Chinese sources I identified for the chapters). Rather, I think that those markers were part of a pattern of stylistic evidence that associates "Paramārtha's" chapters with the Sui context.[22] That is to say, it was true that these items of terminology or phraseology were more typical of *Dharmakṣema than Paramārtha—but they were also typical of the Sui translators. It seems that in many respects, the influence of *Dharmakṣema's idiom had bypassed Paramārtha, but worked powerfully upon his Sui successors.

This cautionary tale illustrates several key points of difficulty in applying TACL with rigour:

First, as already mentioned above, markers often only serve as evidence of relations or contrast within particular contexts. So long as *Dharmakṣema and Paramārtha were the only two candidates for

[19] Radich (2014): 210-211.

[20] I attempt to provide an alternate explanation for this Tibetan evidence in Radich (2015).

[21] I was very fortunate to be saved at the eleventh hour from attempting to publish an article arguing for my wrong-headed hypothesis by the cogent criticisms of Prof. Funayama Tōru, and I am very grateful to him for it.

[22] Radich (2014): 227-233.

translatorship/authorship of the chapters in question, phraseology (relatively) *more characteristic* of *Dharmakṣema *than Paramārtha* might indeed have constituted evidence in favour of the possibility that *Dharmakṣema had something to do with their production. But the restriction of the question to the framework of that two-way comparison between *Dharmakṣema and Paramārtha was based upon an assumption—and it turned out that assumption was false: I needed to expand my range of comparison to include the Sui translators as well.

Second, my travails illustrate how great the difficulty can be, at times, in being sure of our ground in assessing ascriptions on the basis of external evidence, and therefore, in rigorously defining benchmark corpora. As mentioned earlier, the external evidence in favour of Paramārtha's translatorship (not authorship!) of these chapters was extremely strong, and indeed, I was prepared to take them as part of an absolute gold standard for Paramārtha's style. But had I done so, it turns out, I would have introduced a great deal of extrinsic noise into the signal for Paramārtha's group—not just stylistic features derived from the text's earlier Chinese sources, but also features more characteristic of the Sui milieu.[23]

The human partner in "cyborg" work: human philological analysis of raw results in context

As has been implicit at several points in the discussion above, TACL results by themselves provide no answers to our research questions. Rather, even if they have been rigorously and intelligently matched to the nature of the research question and the texts or corpora they address, TACL tests at best merely provide *potential* evidence: sets of data which are likely to contain n-grams that can be used as evidence in constructing such answers. The results always need to be checked and interpreted by a human researcher with Buddhological expertise.

It is often a key part of such "checking" to return to the texts (e.g. via the CBReader) and see how the n-grams isolated by a test fit into their contexts, and what they mean there. Although TACL greatly boosts our power to address text-historical questions, it is no magic wand, that we wave to do our work without knowing how it happens; nor is it a Buddhological house-elf, that does our work for us while we watch the Quidditch. Using TACL is still hard slog, and we need to keep our wits about us. It is also important that in such work, we remember the limitations of TACL: that it only finds literal, exact matches; that it only finds contiguous strings; and that the underlying algorithms, especially for TACL Difference, sometimes find things slightly different from what interests the human reader. Here are a few examples of things we need to be alert to.

TACL cannot "read Chinese", and has no idea what things "mean". This means that anything that is represented by the same literal code in a digitised text counts as the same, for TACL purposes. Human analysis, however, may need to distinguish. For example, if we see in isolation the string 佛語, we are

[23] For another example of a problem in which the stylistic "signal" of a text turns out to be surprisingly mixed, and possibly to betray greater complexity in the history of the text than we normally entertain in analysing such questions, see He [Radich] (2019b) on T474.

most likely to read it as *fóyǔ,* meaning "the word(s) of the Buddha". In Lokakṣema's *Kāśyapa-parivarta,* however, every instance of this string actually appears in contexts like this: 爾時佛語摩訶迦葉比丘言... Thus, we see that in fact, in this context, the string is *Fó yù...,* and means "the Buddha said to [X]..." It is easy to imagine questions of style for which such a difference, between *fóyǔ* and *Fó yù,* might be significant evidence. But this is a difference that only the human can see. Here, TACL is blind.

Similarly, a striking difference between Dharmarakṣa and earlier texts is that in Dharmarakṣa's texts, we sometimes encounter 文殊 alone for Mañjuśrī; in reliable ascriptions before Dharmarakṣa, this usage only ever appears once, in T626. However, in the earlier texts, what we find instead is 文殊師利 or, rarely, 文殊尸利. Because 文殊 is part of 文殊師利 and 文殊尸利, a TACL Difference test of Dharmarakṣa against his predecessors will not discover the real fact that 文殊 *alone* is a distinctive feature of Dharmarakṣa against his predecessors—the literal string 文殊 does, in fact, appear on both sides of such a comparison. Once again, this is a real stylistic difference, to which TACL is blind.

Sometimes human analysis shows us that the strings we regard as significant differences, for purposes of analysis, are slightly longer than the strings returned by the TACL Difference test (for example, TACL returns a 2-gram, but we decide that the real difference is a 3-gram containing that 2-gram). To give a rather weird, artificial example: 云何為 is a recurring feature of the style of Xuanzang, but never occurs in Lokakṣema. If, for some reason, we ran a Difference test to find stylistic differences between core Xuanzang and Lokakṣema texts, we would therefore expect 云何為 to be among the results. In fact, however, such a difference test returns only 何為, not 云何為, and the reason is that 云何 does occur in Lokakṣema. TACL therefore returns the "real" difference, that is, the 2-gram that never appears in Lokakṣema, but not the 3-gram containing that 2-gram and another 2-gram that is itself not a "real" difference (in TACL terms).

The reason that TACL does not return the 3-gram here is technically a bit complicated, but in essence, has to do with the algorithm used for TACL Difference. Experience has shown that strictly returning every longer difference occurring in a text results in overwhelming quantities of redundant noise, of very little evidential value; it would also multiply runtimes unhelpfully. The code therefore has to strike a balance that keeps quantities of results to manageable proportions, but still returns enough information that the human analyst is prompted to notice and find the "real" difference, for evidential purposes, when examining the results back in context. This means that TACL methods rely upon human philological analysis to read results in context, and check whether the "real" difference of interest is slightly longer than the string returned by TACL.

Another example of this phenomenon, from work by Anālayo and Radich on the **Ekottarikāgama,* is that 苦出要諦 is a difference between the style of Zhu Fonian and the *Madhyamāgama*. However, a Difference test between Zhu Fonian and the *Madhyamāgama* only finds 苦出要. We rely on human analysis to see, when examining the contexts in which this 3-gram occurs, that the term at issue is in fact 苦出要諦.

Because TACL only finds literal strings, it is also usually up to the human philological analyst to discover whether or not "the same" item of evidence can be written more than one way. For example, 燕坐 and 晏坐 basically mean the same thing in early translations, and one often appears as a variant reading for the other (we can often pick up on such multiple forms by paying attention to the Taishō apparatus). However, for TACL, 燕坐 and 晏坐 are as different as chalk and cheese, and it will never discover a relation between the two. If the human analysis fails to notice that this same term can appear in two different forms, then, we will miss a part of the actual pattern of distribution of the term, which could be significant as evidence of a distinctive style.[24]

Such graphic variation can sometimes be so extreme that the real pattern of distribution of something that human readers would regard as a single item can be very thoroughly hidden from TACL. For example, one word that appears in early texts can be written at least twelve ways: 校露, 交路, 交露, 交絡, 交珞, 交結, 挍露, 挍珞, 挍絡, 珓路, 珓珞, or 絞絡. During human analysis, then, it pays to keep an eye on variant readings, and check whether they might change the overall pattern of distribution for an item that we are tempted to consider as evidence for our problem.

Hint: Often, the human analyst gets a clue that an n-gram might be susceptible to variant readings from the fact that counts for the n-gram vary between witnesses of the same work.

Next, because TACL only discovers contiguous strings, it also misses cases where, from a human perspective, the "real" or "actual" item of evidence is a pattern with some variation in the middle. For example, in the few instances in which it occurs at all, the usual order of the list of the cardinal directions, in reliable ascriptions before Dharmarakṣa, is 東西南北. By contrast, 東南西北 only occurs in one text, T313. However, a TACL Difference test between Dharmarakṣa and his predecessors will not discover the string in this full form, everywhere that it occurs, because it is very common that a point is first made for "East", with intervening text, and only then does the text say, "[the same is also true for] South, West and North"; for example, 人在大海中央，不見東方山樹木之際，亦不見南、西、北方樹木之際. Thus, a human reader can see that the "real" item of evidential interest here is 東...南西北, but a TACL difference test would only find 南西北, because the intervening text blinds it to the connection with 東. This connection must be made by the human.

Finally, Difference tests in particular tend to return numerous strings which appear to a human eye to have no semantic integrity as a unit. The evidential significance of such strings is uncertain at this stage of our research. For example, a real difference between the (expanded and corrected) Zhu Fonian corpus and some other points of comparison is the appearance of the 2-gram 繫云. However, when we look in context, most instances tend to be across phrase or sentence boundaries, as in this example: 是謂身中二俱繫。云何身中二俱不繫耶? We thus must examine such items carefully in context, in

[24] In either form, this word only every appears twice before Dharmarakṣa, in a single text (T6). In Dharmarakṣa, it is regular.

order to determine whether or not we judge them evidentially significant for the purposes of our research question.

I am very keen to help potential users apply TACL to their research problems, and also to know who is using it, and how. If readers have questions, or are willing to keep me posted about experience with TACL and any results derived using it, I would be grateful if they would please email me.