Using TACL for research problems in Chinese Buddhism

Michael Radich[1]

LAST UPDATED AUGUST 2019.

QUESTIONS AND SUGGESTIONS FOR IMPROVEMENT TO THIS DOCUMENT WELCOME AT

michael.radich@hcts.uni-heidelberg.de

Introduction

[Readers who feel they already know what TACL does and how it works, and want to get on with installing it and getting it running, may want to jump directly to the next section on **Installation and use**, p. 2.]

TACL ("Textual Analysis for Corpus Linguistics") is a tool for the large-scale comparative analysis of strings contained in two or more bodies of digitised text. It was created by Michael Radich and Jamie Norrish. It is programmed in Python, and the code is always up-to-date and freely available on GitHub.[2] It was initially created for use with the texts of the Chinese Buddhist canon (and related collections), as digitised in the XML files of the Chinese Buddhist Electronic Text Association (CBETA).[3] In principle, however, it can be adapted for use with any Unicode corpus, and in collaboration with other colleagues, we have also conducted limited experiments (unpublished) in its application to corpora in Tibetan, Pali and Latin.

At its core, TACL is conceptually very simple. It allows the comparison of two or more user-defined texts or text groups ("A", "B", "C"...), of any size up to the entire canon, to find contiguous strings of user-defined size (n-grams) matching one of two patterns of distribution:

1) occurring in both/all of A and B (C, etc...);
2) occurring only in A, but not B (C, etc...).

That is to say, it finds either the **intersection**, or the **difference**, between the sets of n-grams contained in the user-defined texts/text groups under comparison.

Further functions in the toolkit allow the manipulation or first steps in the analysis of data resulting from intersect or difference tests. The most important of these functions are:

---

[1] michael.radich@hcts.uni-heidelberg.de

[2] https://github.com/ajenhl/tacl
Documentation at http://pythonhosted.org/tacl/

[3] http://www.cbeta.org/

— **filtering** of raw results according to various criteria (maximum/minimum number of occurrences for an n-gram; maximum/minimum number of texts in which an n-gram occurs; maximum/minimum length of n-grams) (examples below);
— **concatenation** of tests (feeding the results of one intersect or difference test into a second or subsequent test) (examples below);
— **alignment** (in an HTML display) of overlapping sequences in two texts for which an intersect test has found overlapping strings;
— **highlighted** display of one text with matches in an intersect test displayed in different colours;
— simultaneous **search** for all n-grams in a user-supplied list (which in practice is often a list of n-grams found by a prior TACL test) in texts in a user-defined corpus, yielding output that includes a count of the number of n-grams from the list occurring in each text (examples below).

Another important basic capability of TACL is that its analysis of CBETA/Taishō texts catches variant readings in alternate witnesses of the texts, as those witnesses are documented in the Taishō apparatus (that is to say, those witnesses indicated in the apparatus by sigla such as 宋, 元, 明, 聖, 宮 etc.). Although, in technical terms, it is slightly misleading to put it this way, it is a useful approximate characterisation of this functionality to say that TACL "searches the Taishō footnotes" (as they were entered into CBETA), as well as the base text.

These capabilities make TACL a powerful tool for the discovery of potential evidence concerning such problems as ascription, dating, textual history, earlier sources, later reception and impact, textual circulation, and other aspects of intertextual relations. It derives its power from the combination of two factors: 1) Scope—it can work its way through vast quantities of text (e.g. the entire canon) in a matter of minutes or hours, where it would be difficult for a human to do the same in a lifetime; 2) Blind accuracy—it finds all strings (n-grams) matching the user-defined pattern of distribution, regardless of whether they "look interesting" (in human terms), or not.

Interested readers may wish to consult some of the research publications by Radich and Funayama listed at the end of this document for examples of the application of TACL to research problems, and some discussion of methods. This document is intended as a users' guide. I will briefly describe basic aspects of the installation and operation of TACL, and then describe the ways it can be applied to the discovery of evidence for some example problems among the types listed above.

Installation and use

[Readers who already have TACL installed and running, or who are confident that they know their way around the command line well enough to do it without additional explanations, might like to skip to the next section on **Research methods**, p. 5. Readers who are not yet sure if TACL will be useful to them, or if they want to invest the time in installation, might also first read that section to get a better sense of what TACL can do, and then return here if necessary.]

This brief discussion of installation and use is designed to give readers who are unfamiliar with the command-line environment (as I was myself when I first began working with TACL) a little

supplementary information that may help them in following the instructions given in the TACL `documentation`[4] and the `helps` (see below).

This brief discussion will assume familiarity with some computing terms and methods. Terms so assumed will be presented in `consolas font` to allow easy identification. Users unfamiliar with those terms and methods may need a little additional background learning to make sense of the use of those terms in this description (Googling will usually suffice).[5]

Installation of TACL differs to some degree, depending upon the user's operating system. Instructions for installation using `pip` may be found at https://github.com/ajenhl/tacl. On Windows (and perhaps MAC), experience has shown that `pip` often does not work so well, in which case users can follow the instructions for their respective OS at https://github.com/ajenhl/tacl/wiki/Installation.

Once installed, TACL is run from the `command line` or `shell`. Users then can call up instructions for use, i.e. `help`, which specifyies and exemplifyies the correct way to format commands for the various TACL functions. For example, if the user types at the `command prompt`:

```
tacl –h
```

...this calls up a menu listing subcommands available within TACL (each achieving one of the types of functionality described above): `align, catalogue, counts, diff` etc.

If the user then types the name of one of those functions followed by `–h`, for example:

```
tacl align –h
```

...this calls up a different `help` menu specific to that subcommand. Under `usage`, a typical `help` gives, in a strictly required sequence, required `parameters` (in lowercase font) and `arguments` (in CAPS), and optional `parameters` and `arguments` (enclosed in [square brackets]). This is followed by text that explains the operation of the subcommand, and some explanation of each of the `parameters` and their `arguments`.

In order to make TACL execute a given subcommand, it is necessary to type in the command, with its `parameters` and their `arguments`, exactly right, and in the right sequence.

From this point, this document will assume that the user has TACL correctly installed, and can use the `helps` to correctly input commands for the various TACL subcommands. We will also assume that users know how to the `pipe` results of commands to files on the hard drive.[6] It will also

[4] Once more, at http://pythonhosted.org/tacl/

[5] This page may also be useful: https://en.wikipedia.org/wiki/Command-line_interface

[6] Basically, piping is achieved by typing after the correctly formatted TACL command (including `parameters` and `arguments`) a right arrow, and then the `(relative) file path` and file name of the file to which the results should be piped. For example:

```
> "tests\Zhi Qian\results 1.csv"
```

which, after a correctly formatted TACL command, would like something like this:

be useful for users to know (if they do not already) that `verbose output`, as specified by the `-v` argument, which can optionally be passed as part of any TACL command, instructs the code to output (to the command `line/shell window` from which the code is run) a running commentary on progress as the code is implemented.

Preparation of the corpus and database

As mentioned above, TACL (when used for Chinese texts) is adapted for application to the digitised Taishō, as made available by CBETA in XML format. Users should therefore create, in a location handy for use from the main directory containing TACL, a folder containing these XML files (in TACL terms, the *corpus*). Before performing any intersect or difference tests, or further manipulations and analyses of their results, it is then necessary to prepare the CBETA XML files in three steps. (Once more, details on how to perform these steps can be discovered by calling up the `help` pertaining to each subcommand.)

First, one must run `tacl prepare` on the folder of CBETA XML files. Details of what is achieved at this stage need not concern users, unless they are interested in the workings of the code itself—in which case they can probably find out what they want to know by directly examining the code.

Next, one must run `tacl strip` on the folder of files resulting from the previous `tacl prepare` operation. This step removes all TEI/XML `markup` (`tags`) from the CBETA files, transforming them into plain `text files`. (This is also the step at which TACL handles variant readings in other witnesses, indicated in the Taishō apparatus by 宋, 元, 明, 聖, 宮 etc.; for each witness, TACL replaces corresponding material in the Taishō base text with the reading indicated in the Taishō/CBETA footnote, and thereby reconstitutes a full text-file equivalent of the original witness, as it is described by the Taishō apparatus.)

Finally, one must run `tacl ngrams` on the *corpus* produced by `tacl strip`, in order to build a database.

Here, it is useful for users to know that TACL tests (`tacl diff, tacl intersect` and associated functions) do not operate by directly searching in the text-only versions of the CBETA corpus created by `tacl strip`. Rather, with `tacl ngrams`, TACL builds a database which lists all n-grams (of user-defined length) in every text in the corpus, and the count (number of instances) for every n-gram in each text. Subsequent TACL operations then work by `querying` that database.

Users should be aware that if one builds a database for a large corpus (e.g. the entire CBETA corpus, or the entirety of the Taishō or the *Zokuzōkyō*), the resulting database can be very large, and the `tacl ngrams` operation that builds the database is likely to be the single TACL operation that takes the longest time and the greatest amount of processing memory (RAM). For example, a database for the entire CBETA corpus, for 2-10-grams (on n-gram length in the database, see immediately below), is a little over 300 GB in size. I myself use a souped-up computer, custom-built

---

```
tacl diff "full db.db" "corpora/xml stripped" "tests\Zhi
Qian\catalogue.txt" > "tests\Zhi Qian\results 1.csv"
```

(Red is used here only to draw attention to part of the command, and is not part of the actual formatting.)

for use with TACL, which has 64 GB of RAM. Running at around 28-30 GB of RAM, `tacl ngrams` takes a little over 24 hours to build that database. Work is currently in progress to try and improve these speeds.

This is one occasion where it can be especially reassuring to run a command with the abovementioned optional `-v` argument for `verbose output`; this allows the user to keep tabs on what the computer is doing, whereas otherwise, if the computer sits silent for a day or more, it can be easy to get worried than nothing is happening, or something has gone wrong.

Because this process is so memory-hungry, users should also note that `tacl ngrams` has an optional `parameter -r`, which allows the user to specify a maximum quantity of RAM to be used by the process. If one does not thus specify how much RAM the process should use, it usually results in an `out of memory error` (the process crashes, and you have to start again).

However, it is also possible to build a database for any corpus of any size. Thus, users who are certain that they want to use TACL for comparisons only within a smaller set of texts can custom-build a smaller database just for the relevant corpus. If the corpus and resulting database are small enough, this will avoid the challenges described above.

Users should also note that the `tacl ngrams` command requires users to pass `arguments` specifying the longest and shortest n-grams to be indexed. I myself have built, and use for all tests, a general-purpose database, which includes 2- to 8-grams. However, users should also note that (as will be discussed below) for some types of research question (e.g. questions devolving upon distinctive styles), shorter n-grams *tend to* be more significant, whereas for other types of question (e.g. questions of highly specific or unique intertextual relations, longer n-grams *tend to* be more significant (though these are tendencies only). In conjunction with the `extend` and `reduce` functions of `tacl results` (also discussed below), these tendencies may mean that if users are only interested in using TACL to examine particular questions, they might be able to afford to build a custom database with a narrower range of n-gram lengths.

Once you have a database, you are ready to apply TACL to research problems.

Research methods

In this section—the core of this document—I aim to describe methods for the careful, rigorous application of TACL to some representative research problems. 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.

Along the way, especially in the course of explaining the first problem (Finding sources), I will also walk readers through aspects of the operation of the software, in ways that hopefully supplement (not substitute for) the TACL `help` and `documentation`. As above, I will show TACL commands in `Courier New` font. I will introduce other TACL-related terms in *italics*.

Finding sources of a text

As mentioned above, 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. These operations are achieved by `tacl intersect` and `tacl diff`.

One type of research problem for which TACL has proven worth is the discovery of sources of a text. Such tests could be useful in determining whether a text presented as a translation was composed in China ("apocryphal"), 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). For examples of studies of the former type, see Radich (2014) and Funayama (2016).

The most useful tool in looking for evidence of this type is `tacl intersect`. The basic method is simple: One runs a `tacl intersect` test with the text under scrutiny on one side of the comparison, and all possible source texts on the other side.

In describing how this might work, I will first make a moderately lengthy detour to describe some additional basic features of how users operate TACL in greater detail than the documentation.

Users tell TACL about the groups into which they want to organise texts, for purposes of comparison, by means of a *catalogue.* For TACL purposes, a *catalogue* is a `text file` (saved with the suffix .txt) which lists texts according to their *identifiers* (for TACL studies of the CBETA corpus, Taishō numbers), and assigns each text to a group, for the purposes of the analysis, using a *label*. *Identifiers* are set properties of the *corpus,* that is to say, catalogue files must identify texts by the exact filenames that identify them in the folder of stripped CBETA files that was used to build the database. Generally, these names will take the form of the CBETA siglum for the collection that includes them (e.g. T for Taishō, X for *Zokuzōkyō*), plus a four digit number (numbers of less than four digits are filled out with zeroes; see examples below). *Labels* are arbitrarily defined by users at the point at which they create a catalogue file.

For example, a catalogue file for comparison between the **Madhyamāgama* T26 and a small corpus of texts by Zhu Fonian (including the **Ekottarikāgama* T125; cf. Radich and Anālayo 2017, Radich 2017a), might look like this:

```
T0026 MA

T0001 ZFn
T0125 ZFn
T0212 ZFn
T0309 ZFn
T0656 ZFn
T1428 ZFn
T1464 ZFn
```

The user must then save this file in plain text format (.txt), and wherever the syntax of a TACL command requires a catalogue, input the name of the catalogue file as an `argument` to the `parameter -c` (for *catalogue*).

Obviously, this aspect of TACL usage immediately raises the problem of how to create catalogue files for very large corpora. For example, one might wish to compare a single text against the rest of the Taishō, but nobody wants to sit and type up a list of identifiers and labels for the entire Taishō. For this purpose, TACL includes a subcommand `tacl catalogue`, which automatically generates a catalogue file listing all texts in a given directory (folder). Running this command over a corpus provides a useful base catalogue, and it can be convenient to construct catalogues by further editing such a base (e.g. by deleting unwanted files) using a `text editor`.

In order to run an 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 place in a second group all other texts that might conceivably be sources (for our present purposes, we will say, all other texts presented in the Taishō as translations). For example, a catalogue for a test to look for sources of the *Mahāyāna Awakening of Faith* T1666 in the rest of the Taishō would begin 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.)

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 carefully think through 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, 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,[7] which, as a commentary on T1666, also cannot be among its sources.

The basic intersect test described above will yield a set of results giving n-grams occurring at least once in both T1666 and at least one other text (in the second corpus defined by the catalogue). It is useful, when inputting the `tacl intersect` command to run the test, to `pipe` the results to

[7] 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.

a file, which can then be further manipulated using other TACL functions, or software such as Excel (as described below). For such purposes, I find it useful to save piped results in `comma separated values` format (`.csv` file suffix).

Now, at this point in the process, potential problems arise for human users from the format in which raw TACL results are output:

- As mentioned above, when the text-only corpus is created with `tacl strip`, the software reconstitutes text-only versions of all the witness texts documented in the Taishō apparatus. The database built with `tacl ngrams` then actually contains a separate row giving information about n-grams and counts for each one of those witnesses. As a result, any `query` to the database discovers separate information about the presence and count of a given n-gram in each witness, and raw TACL results then have a separate row of data for each witness. The vast majority of the time, however (wherever there is not a variant reading), this information will be completely redundant, creating, for example, four copies of "the same" information (for, say, the Korean, Song, Yuan and Ming witnesses to the same text). It multiplies the burden on a human user to wade through these redundancies.
- Where a long string is shared by two (or more) texts in different labels, `tacl intersect` will output results including that string, but it will also include among the results every shorter string included within that string, down to the lower limit of the n-gram length included in the database, because those shorter strings are also (necessarily, logically) shared by the same texts. For example, as has been known at least since work by Lévi and Chavannes a century ago, T453 is largely verbatim identical to **Ekottarikāgama* 48.3.[8] And indeed, the 12-gram 聽我所說彌勒出現國土豐樂 is shared by these two texts, and only them, in the entire CBETA corpus, and occurs exactly once in both texts. But this same distribution is true of the two 11-grams 我所說彌勒出現國土豐樂 and 聽我所說彌勒出現國土豐, both of which are included within that 12-gram; and also of the three 10-grams 聽我所說彌勒出現國土, 我所說彌勒出現國土豐, and 所說彌勒出現國土豐樂, all included in the same 12-gram; and so on for 9-grams, 8-grams etc. Again, this phenomenon greatly increases the redundancy of raw TACL results, from the perspective of the limited aims of the human attempting to interpret the results.

In order to handle problems like these, TACL includes a range of subcommands within `tacl results`, which allow filtering and manipulation of raw results files in a number of ways. Generally, in order to address the problems outlined above, I would recommend (and usually execute myself) the following operations on a set of raw `tacl intersect` results:

- `extend` (`-e` within `tacl results`): This operation takes n-grams matching between two (or more) labels, and checks whether the larger n-grams containing those n-grams also match. Users can envisage this by imagining that the programme begins, for example, with the substring -彌勒出現國土豐- (a 7-gram) from the T453/EĀ 48.3 example above

[8] Lévi and Chavannes (1916): 191, 263, discussed in Anālayo (2010): p. 7 n. 45.

(where it is also a unique match between these two texts), and "looks either side" of the match, to see that the 8-gram 彌勒出現國土豐樂 is also a match, as is the 9-gram 說彌勒出現國土豐樂, and the 10-gram 所說彌勒出現國土豐樂, and the 11-gram 我所說彌勒出現國土豐樂, and so on; eventually arriving at the 35-gram 聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就, which is also a match. This enables the user to see the 35-gram match as a single item of information, rather than have it spread over, say, 26 10-grams (the largest unit that a database containing 2-10-grams will allows us to discover).

— `reduce` (within `tacl results`): Even if we have thus applied `extend` to find the largest contiguous matching string in this locus in our two texts, however, our results file will still also contain all its shorter constituent sub-strings. To eliminate this redundancy, the `reduce` operation takes a long string (n-gram), and checks all the shorter strings it contains (n-1, n-2, n-3…). Where the counts for the shorter string match the count of the longer string, it discards the shorter string from the results. In application to the 35-gram just "discovered" by `extend`, this function would eliminate from the results all the constituent sub-strings that also have a count of 1.

Applied in combination to `tacl intersect` results, `extend` and `reduce` thus achieve the following transformation. Results initially look like this (note that counts—in the second-to-rightmost column—are all identical):

聽我所說彌	5	T0125	base	1	EA
聽我所說彌勒	6	T0125	base	1	EA
聽我所說彌勒出	7	T0125	base	1	EA
聽我所說彌勒出現	8	T0125	base	1	EA
聽我所說彌勒出現國	9	T0125	base	1	EA
聽我所說彌勒出現國土	10	T0125	base	1	EA
我所說彌	4	T0125	base	1	EA
我所說彌勒	5	T0125	base	1	EA
我所說彌勒出	6	T0125	base	1	EA
我所說彌勒出現	7	T0125	base	1	EA
我所說彌勒出現國	8	T0125	base	1	EA
我所說彌勒出現國土	9	T0125	base	1	EA
我所說彌勒出現國土豐	10	T0125	base	1	EA
所說彌	3	T0125	base	1	EA
所說彌勒	4	T0125	base	1	EA
所說彌勒出	5	T0125	base	1	EA
所說彌勒出現	6	T0125	base	1	EA
所說彌勒出現國	7	T0125	base	1	EA
所說彌勒出現國土	8	T0125	base	1	EA
所說彌勒出現國土豐	9	T0125	base	1	EA
所說彌勒出現國土豐樂	10	T0125	base	1	EA

(...etc., for all 26 10-grams comprised within 聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就, and their constituent parts)

Extend and reduce yield a single data item like this:

聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	base	1	EA

— `zero fill` (`-z` within `tacl results`). For obvious reasons, n-grams with zero count will not be otherwise included in results for a given witness in the initial `tacl intersect` test—such a test only finds n-grams that *are* present in a text. Where an n-gram found in other witnesses for a text does not appear at all in a given witness, `zero fill` adds a row of data to results showing a zero count for that witness, which allows comparison between witnesses in which a reading does appear, and those in which it does not (philologically minded readers might like to think of this step as changing from implicit to explicit indication of silence on the reading in question). Among other things, `zero fill` is useful in preparation for the next step, `collapse-witnesses`.

— `--collapse-witnesses`: For any cases in which the counts are identical for a given n-gram across multiple witnesses, this operation collapses the results into a single row of data, and lists in one place the sigla for all witnesses with that count for that n-gram. This achieves a transformation like the following. Results initially contain a separate row for the above 35-gram for every separate witness to T125 (sigla for witnesses appear in the third-to-rightmost column), but there is in fact no variation between all witnesses (the 35-gram in question appears exactly once in each text):

聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	base	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	元	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	大	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	宋	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	明	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	明異	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教	35	T0125	磧砂	1	EA

即還就					
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	聖	1	EA
聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	麗	1	EA

`--collapse-witnesses` presents all this same information in the following more concise form:

聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還就	35	T0125	base 元 大 宋 明 明異 磧砂 聖 麗	1	EA

Here, all the information for this 35-gram is collapsed into a single row, and the sigla of the witnesses in which that n-gram appears with that count are all concentrated into a single set of information (again, third-to-rightmost cell).[9]

Thus, for this 35-gram, we now have all the information the human user would want, concentrated into a single row—the 35-gram (like all its subordinate parts!) appears once only in this text (and in T453, which would be represented in a separate row, because these are `intersect` results), and the situation is the same in all eight or nine witnesses documented in the Taishō apparatus.

A next set of problems arises from the fact that, even with the results thus cleaned up to eliminate redundancy, it is quite possible, if not likely, that without further sorting and filtering, the raw results file will be so copious in its contents, and include so much extrinsic "noise" (from the perspective of the researcher's ultimate goal), that it will be unusable for a human researcher interested only in some particular problem. To continue with the example of EĀ 48.3 and T453, our results include 3,644 instances of 比丘 (*bhikṣu*) in EĀ, but obviously, the fact that this string occurs in both EĀ and T453 does not indicate any special relationship between EĀ and T453, but rather, occurs because this word is extremely common in these texts (as many others).

We can handle this problem by using subcommands within `tacl results` to filter the results by such criteria as the length of a string, the count for the string within each text, and so on. In this instance, we might presume, for example, that shared strings of less than four characters in length are more probably recurring items of vocabulary, instead of matches in specific wording and content; and we might also estimate that the most telling evidence of such textual debts will occur when, as in our example above, a match is found precisely once in each text (and never anywhere

[9] Users should note, however, that collapse-witnesses should in fact always be performed last in any series of operations, because the results of a collapse-witnesses operation cannot be further manipulated by other TACL operations. Here, I have presented these steps out of order because it makes it easier to understand conceptually what the various operations achieve, and why. Correct sequences of operations are summarised below.

else). Thus, we might use `tacl results` with the `parameters` and `arguments --max-count 2`, `--min-size 4`, `--min-works 2`, and `--max-works 2`, in order to produce a subset of results that occur only once in each text, and are always at least four characters long.

I suggested earlier that users `pipe` the results of TACL operations, and that it is useful to save them in `csv` format. However, I have not yet explained how I ordinarily view the resulting `csv` files, for instance, to produce tables like those shown above.

The actual content of the `csv` file for the last table row shown above looks like this (as the name of the file type, `comma separated values`, indicates, the file just contains items of information separated by commas):

> 聽我所說彌勒出現國土豐樂弟子多少善思念之執在心懷是時阿難從佛受教即還
> 就,35,T0125ex(50.4),base 元 大 宋 明 明異 磧砂 聖 麗,1,EA

I, at least, find this sort of thing a bit of a clutter to look at. In order to view such results in a format I can more easily understand, I ordinarily use Microsoft Excel (though there are certainly other options for viewing and manipulating `csv` files.). Using Excel has the added advantage that it allows further sorting of the results, and this can make it possible to zero in much more quickly on items of evidential interest.

Before I give examples of sorting results in this manner, I need to mention two technical hitches that one occasionally encounters in importing results into Excel. One imports TACL results into Excel using the "From Text" button on the "Data" tab in the menus at the top of the screen, and then navigating to the TACL results one wants to display. At the next step, a dialogue window opens, and one of the options is to specify the encoding of the source file. Excel will try to automatically recognise the encoding of the source file, which should make it unnecessary for the user to specify the encoding, but often Excel often fails to identify the encoding correctly, which results in the Chinese being displayed as gibberish. Users should therefore ensure that the encoding is correctly identified at this point. The correct option is "Unicode (UTF-8)".

At the next screen in the dialogue box, it is necessary to specify what marker is used to delimit (separate) items in the data. The correct option here is "Comma", but again, Excel often automatically selects the wrong option, which results in a mess (with multiple data items crowded into one table cell), which is impossible to sort. It is therefore also important to ensure that "Comma" is selected as the delimiter.

Once the data has thus been imported, sorting with Excel (using the "Sort" button on the "Data" tab) allows the user to focus even more precisely on portions of the results that are most likely to be of interest to the research question under investigation—to find needles in haystacks.

Returning one last time to the example of EĀ 48.3 and T453: Let us assume that we are investigating a hunch that T453 might have other Chinese sources, and to test that hunch, have run a `tacl intersect` test between T453 and the remainder of the translation portion of the Chinese canon. The results listed and process above, for the 35-gram shared by T453 and EĀ 48.3, would have been among the results of such a test. But even after eliminating redundancy with

`extend`, `reduce`, and `collapse-witnesses`, and filtering to restrict results only to unique matches over a certain length, as above, we might still have an overwhelming quantity of data, which it would take a very long time to work through. (Indeed, even after performance of all the above operations, the results file for such a test still has 1524 rows; in some tests, results can have rows numbering in the hundreds of thousands.)

In the present case, then, on the assumption that the longer the string, the more likely it is to be genuinely unique and strong proof of a specific intertextual relation, rather than the result of some other phenomenon (including chance), we can maximise the chance that we will find evidence supporting our hunch by sorting to show the longest stings first. And indeed, when we do so, the topmost item among the results so sorted is a 224-gram (a string of 224 characters) which matches verbatim between EĀ 48.3 and T453:[10]

> ...訓之所致也亦由四事因緣惠施仁愛利人等利爾時阿難彌勒如來當取迦葉僧伽梨著之是時迦葉身體奄然星散是時彌勒復取種種華香供養迦葉所以然者諸佛世尊有敬心於正法故彌勒亦由我所受正法化得成無上正真之道阿難當知彌勒佛第二會時有九十四億人皆是阿羅漢亦復是我遺教弟子行四事供養之所致也又彌勒第三之會九十二億人皆是阿羅漢亦復是我遺教弟子爾時比丘姓號皆名慈氏弟子如我今日諸聲聞皆稱釋迦弟子爾時彌勒與諸弟子說法汝等比丘當思惟無常之想樂有苦想計我無我想實有空想色變之想青瘀之想...

If Lévi and Chavannes had not already told us so a century ago, then, this single piece of evidence on its own would suffice to tell us that either EĀ borrowed from T453, or *vice versa.* But it is also followed, in our sorted results file, by another 163-gram matching verbatim between the same two texts; and then by a 156-gram; and so on. In the scheme of things, then, the 35-gram we began with above is small potatoes.

To summarise the above discussion, the most useful sequence of steps for an intersect test designed to find sources of a text (e.g. to check a hypothesis that a text is of Chinese composition) is usually this:

```
tacl intersect
extend
reduce
zero fill
--max-count 2 --min-size 4 --min-works 2 --max-works 2
collapse-witnesses
```

import to Excel

sort to display longest strings first

Obviously (with perhaps some adjustment of maxima and minima), it is equally possible to use basically the same method to discover later uses, citations, or impact of a text, so long, once more, as one is sure of one's ground in determining the relative chronology of texts, and, where the same material is shared by multiple texts, which is (are) the ultimate or proximate source(s). For a brief

[10] Note, however, that this string will not be found in EĀ by a Taishō search, because it is hidden in the base text of the Taishō by one or more variant readings, and is only present in the Song, Yuan and Ming witnesses.

example of an application of TACL-like methods to such questions (the work was done before the development of TACL proper), see Radich (2012): 59 n. 64, 65; 60 (on citations of T1589 in Jizang, Zhiyi and Jingying Huiyuan); 65 n. 87; 66 n. 88, 89, 91; 67 n. 92, 95; 68 n. 96, 97, 98.

Authorship or translatorship

We have now talked through many of the aspects of the actual operation of TACL that might be unfamiliar to beginning users. In discussing application of TACL to other typical research problems, therefore, we can now concentrate more directly on methods, meaning a combination of regimens of TACL operations (including filtering by counts, etc.), sorting in Excel, and Buddhological considerations.

Another principle application of TACL is to discover text-internal evidence bearing upon questions of authorship or translatorship.[11] 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. I have undertaken studies based upon such conditions for the **Ekottarikāgama* (with Venerable Anālayo; Radich and Anālayo 2017, Radich 2017a), and of 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 for these examples). 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.

[11] 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. Space means that 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.

These are only examples. 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; logically, this scenario must exclude the possibility of anonymous translator/authorship, in which case possible translators/authors are potentially numberless), 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)]?

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.

In TACL terms, we can find a set of possible evidence answering this question by writing two catalogue files with two labels each:

> 1) TARGET TEXT and BENCHMARK CORPUS FOR CANDIDATE, and
> 2) TARGET TEXT and BENCHMARK CORPUS FOR OTHER CANDIDATE(S)

We then run a `tacl intersect` using the first catalogue, to find strings found in both the TARGET TEXT and the CANDIDATE. We will call the results of this operation "Results 1".

Next, we run a `tacl diff` using the second catalogue file, to find strings found only in the TARGET TEXT, but not in the OTHER CANDIDATE(S). It is best to make this an *asymmetric* difference test, which is achieved by passing the optional `parameter` `-a` to `tacl diff`, and an `argument` for that parameter specifying the label assigned to the TARGET TEXT in the catalogue file. This restricts the results to strings found in the TARGET TEXT only. (It is also possible to use `--remove` in `tacl results` to reduce a results file from a bilateral/symmetric difference test to only one side of the comparison, but that would be an unnecessary step if we already know that we only want results on one side, and can just confine the test to that side of the comparison from the outset.) We will call the results of this operation "Results 2".

Next, we use a TACL function called `tacl sintersect` (for "supplied intersect", i.e. an intersect to which we supply as input results from prior TACL tests), which finds the intersection between two TACL results files, to find the strings found in both Results 1 and Results 2. The new results of this `tacl sintersect` test ("Results 3") match our test conditions listed above.

Ordinarily, it would then be useful then to follow these steps to make the final results ("Results 3") as easy as possible for a human researcher to work through:

```
extend
reduce
--min-count 5, --min-works 3
collapse-witnesses
```
import to Excel

sort by:

count, diminishing order

size, ascending order

n-gram

The rationale for some of these measures will already be clear from earlier examples. Other rationales are as follows.

— Counts are predicated on the understanding that we are looking for recurring features of a regular style, and are therefore unlikely to be interested in items that occur in less than three texts (the TARGET TEXT, and at least two others due to the CANDIDATE), or in items that occur only one time in a given text. (We could also say that if our hypothesis is valid, we will hopefully find sufficient evidence above this threshold to demonstrate it persuasively.)

— The Excel sort protocol is predicated on these assumptions: We sort by diminishing counts because we presume that items that occur the largest number of times in a given corpus are the strongest evidence of a distinctive style (they are not just demonstrable habits, but habits to which a translator/author had frequent recourse). For example, the rare translation 溥首 for Mañjuśrī appears 269 times in 11 texts by Dharmarakṣa 曇無讖, but never in any other translation texts, and is thus in itself very strong evidence that all the texts in which it appears are genuine Dharmarakṣa translations. Next, we sort by ascending order of size because we assume that other things being equal, shorter n-grams are more likely to be recurring stylistic features (like words). Finally, sorting by n-gram simply keeps together all the evidence for a single n-gram.

It has to be remembered that this test will only be even-handed, and therefore rigorous, if this test is applied equally in all directions, i.e. to all candidates.

2. When we doubt a traditional ascription, but do not immediately know where else we might look for a more likely translator/author, 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,[12] Dao'an only ascribes three works still extant today to Faju (sometimes in cooperation with the even more shadowy Fali 法立): T23, T211, and T683.[13] Thus, we might take one of the 24 other

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

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

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.[14]

To perform such a test using TACL, we would first create a catalogue placing T23, T211 and T683 in one group (label), and T119 alone in a second group. We then perform the following operations:

`tacl diff` (using `-a` to restrict results to strings in T119 only)
```
reduce
zero fill
--min-count 2
collapse-witnesses
```

The results that this test yields are already possibly 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 Buddhabhadra, and the Sarvāstivāda Vinaya T1435 ascribed to *Puṇyayaśas and Kumārajīva. (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.) The same phrase occurs 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 **Sarvāstivādavinayamātṛkā* 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 among the works in which our phrase occurs most frequently.

However, the above operation 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. `tacl search` takes a list of any number of n-grams, checks whether each occurs in every text in the entire CBETA corpus, and then outputs a list of text identifiers, with a full list of all the n-grams, among those searched for, that appear in each text. Using Excel, it is then possible to order these results so that the text containing the greatest number of n-grams from the list appears first, and

[14] 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.

then other texts are listed in descending order of the number of n-grams they contain. We must allow for confounding factors like text length (a very large text like the **Mahāprajñāpāramitopadeśa* T1509, in virtue of its sheer size, will contain a great many more distinct markers than a short text like T119), or genre (for example, the *Jing lü yi xiang* 經律異相 T2121, which is a compendium of excerpts from very many texts, for this reason contains a great variety of language, and often appears high among results of `tacl search` operations).

When we perform 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,[15] 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[16] appear in T212; 28 appear in Yijing's T1442;[17] 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 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 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.

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

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

[17] 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.

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*).

In another study (Radich 2018), 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 (Radich forthcoming on T7, English and Chinese), 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 that 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.

Readers interested in further examples of the application of TACL might also consult the Appendix of Radich (forthcoming on T7), where I list multiple TACL tests used in preparing that study.

Further considerations of method, potential pitfalls, and caveats

It is important that users be aware of several other considerations.

1. As has been implicit several times already in the discussion above, TACL results by themselves provide no answers to our research questions. Rather, TACL tests, even if they have been rigorously and intelligently matched to the nature of the research question and the texts or corpora they address, at best merely provide sets of data which are likely to contain n-grams that can be used as evidence in construction 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 seeing 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.

2. When choosing texts or defining corpora as benchmarks or points of comparison, it is vital that we scrutinise our assumptions thoroughly, know as much as possible about the nature and content of the texts, and be as conservative as possible.

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;[18] or in the case of Zhu Fonian, again, perhaps as many as half of the authentic texts. Obviously, however, if we are attempting to define a style for a translator (or group) like "Zhi Qian", and 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. This point 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 liberally define a baseline that turns out to contain junk. It is therefore vital that benchmarks be defined with extreme conservatism.

Of course, it is also necessary that we apply equal conservatism on the "other side" of any comparison—that is to say, if we were trying to establish a benchmark by comparing "X" with "not X" ("Zhi Qian" with "everything else"—the remainder of the Taishō translation corpus, for example), we would also need to be equally careful not to include accidentally in "not X" ("everything else") something that in fact turns out to be "X". This means that very often, it is important that in setting up two-way comparisons, we think rigorously and systematically about a *grey zone* between the options at issue, where we place 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 corpus treated as problematic by Nattier (2008) (our most informed assessment of the external evidence).[19]

3. Likewise, it is 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

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

[19] 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.

commentary; and subsequent chapters.[20] For purposes of stylistic analysis, these three different layers of material must be treated separately. Further examples of this problem are legion.[21]

Failure to appreciate these various points could potentially lead to abuses and misapplication of the tools, and egregious error. Two examples 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 work on works ascribed to *Dharmakṣema, that markers typical of *Dharmakṣema but atypical of Paramārtha seemed repeatedly to occur 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.

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[22]—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.[23] 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.[24]

---

[20] Nattier (2008[2010]).

[21] It is possible, for the purpose of TACL analysis, to split units treated as single "text" in CBETA. However, for technical reasons, this step must be taken between implementation of `tacl prepare` and `tacl strip`, i.e. before the database is built (and the database must be rebuilt each time a new split is introduced by rerunning `tacl ngrams`). This means that users must split XML files, and the two or more XML files that result from the split must themselves be `well-formed`. This requirement complicates the use of texts that split the units defined by CBETA, if only moderately, and we will not discuss it further here.

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

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

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

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.[25] That is to say, it was true that these items of terminology of 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 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.
— 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, of 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.[26]

Final remarks, and an appeal

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 to hear from them at michael.radich@hcts.uni-heidelberg.de.

We recognise that TACL is somewhat time-consuming to use, and will entail a fairly significant learning curve for most users. (Believe me, as someone relatively computer-un-savvy, and the primary guinea-pig user to date, I know.) Some of this burden may be an inevitable result of the quantity and complexity of the data we handle in applying TACL to the Chinese canon, and perhaps,

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

[26] 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 Radich (in preparation).

therefore, must be borne as a necessary cost of addressing such problems in this manner. At the same time, it is also true that some of the complexity is due to the present set-up of the software, and could probably be alleviated to some degree, for average users, by further improvements. For example, we currently are considering the creation of a `GUI` (`graphical user interface`, i.e. a point-and-click), which, for relatively standardised and regularly recurring applications of TACL, would allow average users to avoid the current requirement to work from the `command line`; and we are also working on bundling and regularising series of operations (in recurring `algorithms`) into higher-order "meta-"operations that might simplify for average users the task of, for example, obtaining a standard set of intersect data to examine questions like "possible sources of A in B" or "possible stylistic differences distinguishing A from B". We always welcome other suggestions about how to make things simpler and more accessible.

Meanwhile, for users who are unsure whether it will be worth the investment of their time to undertake the learning necessary to install and run TACL themselves, I am always happy to consider running a trial set of tests myself and providing them (in Excel spreadsheet format), and walking the researcher through the task of analysing the results. Interested colleagues should feel free to contact me for this purpose as well.

Publications using TACL

Anālayo, Bhikkhu. "The Influence of Commentarial Exegesis on the Transmission of Āgama Literature." In *Translating Buddhist Chinese: Problems and Prospects*, edited by Konrad Meisig, 1–20. Wiesbaden: Harrassowitz, 2010.

Funayama Tōru 船山徹 [Chuanshan Che]. 2016 . "*Da fangbian Fo bao'en jing* bianzuan suoyinyong di Hanyi jingdian"《大方便佛報恩經》編纂所引用的漢譯經典 [Translated Chinese Scriptures Cited in the Composition of the *Da fangbian Fo bao'en jing*]. Translated by Wang Zhaoguo 王招國. *Fojiao wenxian yanjiu* 佛教文獻研究 2: 175-202.

Funayama Tōru 船山徹. "Ryō no Hōshō *Hikuni den* no teikei hyōgen: sensha mondai kaiketsu no tame ni 梁の宝唱『比丘尼伝』の定型表現　撰者問題解決のた." *Tōhōgaku* 東方学 135 (2018): 36-53.

Lévi, Sylvain and Édouard Chavannes. "Les seize Arhat protecteurs de la Loi (second article)." *Journal Asiatique*, ser. II, 8 (1916): 189-306.

Mizuno Kōgen 水野弘元. "Kan'yaku *Chū agon kyō* to *Zōichi agon kyō* 漢訳『中阿含経』と『増一阿含経』." *Bukkyō kenkyū* 仏教研究 18 (1989): 1-42[L]. Chinese translation: "Hanyi de *Zhong ahan jing* yu *Zengyi ahan jing* 漢譯《中阿含經》與《增一阿含經》," in Shuiye Hongyuan [=Mizuno Kōgen], *Fojiao wenxian yanjiu: Shuiye Hongyuan zhuzuo xuanji (1)* 佛教文獻研究・水野弘元著作選集(一), translated by Xu Yangzhu 許洋主, 509-579. Taipei: Fagu wenhua, 2003.

Nattier, Jan. *A Guide to the Earliest Chinese Buddhist Translations: Texts from the Eastern Han* 東漢 *and Three Kingdoms* 三國 *Periods*. Bibliotheca Philologica et Philosophica Buddhica X. Tokyo: The International Research Institute for Advanced Buddhology, Soka University, 2008.

Nattier, Jan. "Who Produced the *Da mingdu jing* 大明度經 (T225)? A Reassessment of the Evidence." JIABS 31, no. 1-2 (2008[2010]): 295-337.

Radich, Michael. "External Evidence Relating to Works Ascribed to Paramārtha, with a Focus on Traditional Chinese Catalogues." In *Shintai sanzō kenkyū ronshū* 真諦三藏研究論集 [Studies of the Works and Influence of Paramartha], edited by Funayama Tōru 船山徹, 39-102[L]. Kyoto: Kyōto daigaku jinbun kagaku kenkyūjo/Institute for Research in Humanities, Kyoto University, 2012.

Radich, Michael. "On the Sources, Style and Authorship of Chapters of the Synoptic *Suvarṇaprabhāsottama-sūtra* T664 Ascribed to Paramārtha (Part 1)." *Annual Report of The International Research Institute for Advanced Buddhology* 17 (2014): 207-244.

Radich, Michael. "Tibetan Evidence for the Sources of Chapters of the Synoptic *Suvarṇaprabhāsottama-sūtra* T664 Ascribed to Paramārtha". *Buddhist Studies Review* 32, no. 2 (2015): 245-270.

Radich, Michael. "On the *Ekottarikāgama* 增壹阿含經 T 125 as a Work of Zhu Fonian 竺佛念." *Journal of Chinese Buddhist Studies* 30 (2017a): 1-31.

Radich, Michael. "Problems of Attribution, Style, and Dating Relating to the 'Great Cloud *Sūtras*' in the Chinese Buddhist Canon (*T* 387, *T* 388/S.6916)." In *Buddhist Transformations and Interactions: Papers in Honor of Antonino Forte,* edited by Victor H. Mair, 235-289. Amherst, NY: Cambria Press, 2017b.

Radich, Michael. "A Triad of Texts from Fifth-Century Southern China: The **Mahāmāyā-sūtra*, the *Guoqu xianzai yinguo jing*, and a *Mahāparinirvāṇa-sūtra* ascribed to Faxian." *Journal of Chinese Religions*, 46, no. 1 (2018): 1-41.

Radich, Michael = He Shuqun 何書群. "Zhu Fahu shifou xiuding guo T474? 竺法護是否修訂過 T474?" *Foguang xuebao* 佛光學報, New Series, 5, no. 2 (2019): 15-38.

Radich, Michael. "Kumārajīva's 'Voice'?" Forthcoming in *China and the World – the World and China: A Transcultural Perspective*, edited by Barbara Mittler, Joachim & Natascha Gentz and Catherine Vance Yeh, 131-145. Deutsche Ostasienstudien 37. Gossenberg: Ostasien Verlag, 2019.

Radich, Michael. "Reading the Writing on the Wall: 'Sengchou's' Cave at Xiaonanhai, Early Chinese Buddhist Meditation, and Unique Portions of Dharmakṣema's **Mahāparinirvāṇa-mahāsūtra*." Forthcoming, *Journal of the International Association of Buddhist Studies.*

Radich, Michael. "Was the *Mahāparinirvāṇa-sūtra* 大般涅槃經 T7 translated by 'Faxian'? An Exercise in the Computer-Assisted Assessment of Attributions in the Chinese Buddhist Canon." Forthcoming.

Radich, Michael = He Shuqun 何書群. "*Da banniepan jing* (*Mahāparinirvāṇa-sūtra*, T no. 7) shifou you 'Faxian' suoyi? Jisuanji xiezhu yiren kanding 《大般涅槃經》（Mahāparinirvāṇa-sūtra, T no.

7）是否由「法顯」所譯？計算機協助譯人勘定.” In *Xiantangshan yu Faxian wenhua: Hanseng Faxian (337-422) qi shengping yu yichan guoji yantaohui lunwenji* 僊堂山與法顯文化： 漢僧法顯(337-422)其生平與遺產國際研討會論文集, edited by Miaojiang 妙江, Chen Jinhua 陳金華 , and Ji Yun 紀贇. Taibei: Xinwenfeng Chuban Gongsi 新文豐出版公司, 2019.

Radich, Michael and Anālayo Bhikkhu. “Were the *Ekottarika-āgama* 增壹阿含經 T 125 and the *Madhyama-āgama* 中阿含經 T 26 Translated by the Same Person? An Assessment on the Basis of Translation Style.” In *Research on the* Madhyama-āgama, edited by Dhammadinnā, 209-237. Dharma Drum Institute of Liberal Arts Research Series 5. Taipei: Dharma Drum Publishing Corporation, 2017.

Zürcher, Erik. *The Buddhist Conquest of China: The Spread and Adaptation of Buddhism in Early Medieval China*. Third Edition. Leiden: Brill, 1959 (2007 reprint).