Thesuarus Literaturae Buddhicae (TLB): Its Scope, and a Description of Its Routines

Jens Braarvig
University of Oslo, Norway

It is an imperative to make a complete collection of all the extant canonical scriptures of Buddhism. With the new possibilities of publishing on the internet, and the enormous possibilities for access to any part of a collection by various kinds of search routines, one should make a complete "canon" of all classical Buddhist texts available to believers and scholars alike.

This, however, has as a prerequisite an electronic format of all the collections of the Buddhist canons which is easy accessible, and one that has a lexicographical component, which makes it possible to search in a way that it can be ready accessible and able to yield information for various kinds of users.

Indeed Buddhist literature, the *Dhammapitaka*, is the literature that connects the whole of Asia, Buddhism being the Asian culture which more than any has spread all over Asia, and has remained for a period of two thousand years. Thus the complete and multilingual exposition of its literature will constitute a tool in conducting research on this magnificent culture, and the way it has provided common ways of understanding among Eastern peoples.

The languages of Buddhism, at least in the concrete and linguistic sense, are quite numerous, though the most important ones, historically, may be said to be Pali, Sanskrit, Chinese and Tibetan, though of course, Buddhist Chinese has been translated into Japanese, Tibetan into Mongolian, and Sanskrit and Pali into a lot of Central Asian languages.

Thus to understand the linguistic variations on the common theme of Buddhism is important in the attempt to understand East Asian Culture on the whole. From a cultural linguistic viewpoint, one can study how the original Buddhist ideas, the original Indian conceptual schemes, take on differing forms of language. Further one can study how the languages, or rather idioms thus generated by Buddhist ideas, influence the languages of the host cultures, how Buddhist ideas are transformed into Chinese, Tibetan, etc., and how they are changed, and influence the languages where they are taken in.

Such reflections and investigations have a general cultural linguistic interest, but indeed also

in understanding the histories of the languages where such processes take place, as e.g. in the Chinese language, and the Tibetan. The Tibetan written and learned language was to a great extent constructed on the basis of Sanskrit, on the basis of Sanskrit grammarians and commentators, which undertaking also completely transformed Tibetan language, religion, culture and organization. The Tibetan culture to which Buddhism arrived, was thus much less complex than the Tibetan culture which ensued from the rather well planned introduction of Buddhism to Tibet.

When Buddhism arrived in China, though, China was already a sophisticated culture, Buddhism met in China a culture which had a rich language and literature expressing the various modes of human existence and behaviour. Thus the process of “importing” Buddhism took quite another way in China than the Tibetan import of the same. In the earliest attempts to translate Buddhist scriptures and their conceptual world into Chinese, the classical and Taoist Chinese idioms were employed, but quite soon an idiom developed which was peculiar to Buddhism, a kind of Chinese to some extent constructed to suite the new kind of thinking and the way of expressing thought peculiar to Buddhism. But this kind of Chinese, “Buddhist Hybrid Chinese” as some prefer to call it, has still had great influence on the general Chinese language, and thus its study is essential for understanding the history of Chinese language.

Thus there is all reason that the translation techniques of this grand intellectual effort, which the translation of Buddhism really was, should be studied in a detailed and comprehensive way. The great undertaking of transferring Buddhist thinking from Indian soil to the whole of Central and East Asia deserves such a study, and the format of the planned thesaurus could be a help for such a project. It could also make more sophisticated corpus linguistics in the field possible, including various kinds of syntactical and other grammatical studies. The Thesaurus, as envisaged, could be an essential tool of giving a complete description of the grammar of Buddhist Hybrid Chinese, to understand how the categories of Sanskrit language and grammar have influenced the Chinese language, and to understand how the Chinese translators related to the Sanskrit language and transferred the Indo-European grammatical structures onto a language that different.

Further, a complete and ready accessible multilingual Internet Buddhist Canon will make the rich culture, philosophy and religion of Buddhism available in an unprecedented way. Thus for the study of Buddhism, its philosophical and religious richesse, it is certainly the time to gather the manpower and means of for compiling a Thesaurus Literaturae Buddhicae.

In recent decades several similar thesauri have been made available, among them, and probably the earliest to be constructed on a larger scale, those of Thesaurus Lingua Graeca (TLG) and Thesaurus Lingua Latinae (TLL), both of which have been available for a long time on CD, and with a number of search machines to suit different kinds of users. The TLG is also being developed into a bilingual version, in the Perseus project, so as through English translation give easier access to the classical Greek literature. Other thesauri, more recently compiled, are those of the Sumerian literature, the classical Jewish literature, the Nordic literature, and several others.

Indeed much has also been done for Buddhist literature in this respect, not the least by persons and institutions represented in this meeting, and great efforts have been done to make Buddhist scriptures available on the net. It has thus been my intention to build on these efforts to try to make a thesaurus that might contribute to the study of cultural, linguistic, lexicographical and philological aspects of this literature, in addition to contributing generally to the study of the thinking and message of Buddhism as a religion and a philosophy.

With this perspective in mind it would be a great advantage make a *multilingual* thesaurus, including both Sanskrit/Pali, Chinese, Tibetan languages, and that with English translations all along, to make the wonderful texts of Buddhism accessible to a greatest possible number of readers and users.

What I propose, then, is that a complete quadrilingual corpus should be constructed, paragraph by paragraph, containing the corresponding passages of Sanskrit/Pali (where extant), with the Chinese, Tibetan translations (where extant), and then with an English translation. These paragraphs should be rather short, so it will be possible to view them in their quadrilingual context, so that one easily can see which words and expressions are equivalent in the different languages.

Now, when considering the enormous size of the Chinese, the Tibetan, the Pali and even the extant Sanskrit texts of Buddhism, this seems an almost immeasurable task to complete – however, in the spirit of Buddhism itself, that of immeasurable energy and other virtues, it could be achieved as an international collective effort, that with the full use and practice of all the six *påramitås*. A merit of the the electronic form is of course also that the process of producing such a thesaurus is a cumulative process, text by text can be made available on the internet as soon as it is ready in according with the principles of the thesaurus. And, indeed, to some extent, starting this work, is also standing on the shoulders of Giants, since much of the materials are already

input, it is already there to be given the proposed format. The good internet lexicographical work already done, as that of Digital Dictionary of Buddhism (DDB), should also be linked up to the project if possible. As a background also the lot of good indices in book form, made by Japanese scholars mainly, might in some form be integrated, if permissible by copyrights, and if available in electronic form.

There is, though, all reason to try to cooperate with the whole of the scholarly world studying Buddhism, as well as Buddhist organizations themselves. As recently done by myself in lectures on Buddhist literature, one can read parallel versions in the classes and input them along with teaching their grammar, analysis and understanding, thus bringing the students into the project. This may have a beneficial influence on the students, thus being part of an extensive international project beside learning their curricula. Reading and inputting in the way envisaged, paragraph by paragraph, and with a translation, would necessitate a rather thorough understanding of the texts, and will be a good philological training for both students of Sanskrit, Pali, Buddhist Chinese language, and the classical Tibetan language. Of course, one cannot expect every student to know all of these languages, though a few might reach a high level of ability in reading all. To participate it would be enough to know two, Chinese and Sanskrit, or Tibetan and Sanskrit, etc. – in this way the work could be shared. As such the project may have some effect in boosting the study of Buddhist philologies in our universities, where one often wish philology as a discipline would have more support. Anyway, the production of such a thesaurus will have to involve a collective effort of a certain size and extent.

With this as an introduction to the idea of a Thesaurus Literaturae Buddhicae, the next points of this paper will be concerned with the formats of the thesaurus, then at the end there are given a few examples. The following, then, is an attempt to set down a number of rules which can be used for organizing the TLB.

I) The user can search for a word, an element of word, or a sequence of words, in Skt, Chin, Tib, Eng, by entering it into a search form on the screen. The search may be for a number of non-consecutive words within a the 2-5 lines of a paragraph. The search entry can be in Skt, Chin, Tib, or Eng. The search can be limited to an indefinite number of individual texts picked out by the user, or to a group of texts, e.g. those having been translated by a certain Chinese translator, “Vinaya texts,” “Prajnaparamita texts,” “Mahayana sutras,” or similar categories. A search for a word in Skt will, if desired, search for all inflectional forms, as well as for all sandhi forms, as far as these are noted in the database, and systematized in the lexicon. A search for a Chin

character will, if desired, search for all graphic variants of the character concerned, as these are noted in the lexicon. Search request are constructed according to Boolean logic and can be cumulative. Search results can be stored for later use, along with the search criteria used to obtain them.

> Option 1: A search will lead to a list of references to the text, with indications of page and line, (e.g. Sbv 4,3 [=Samgabhedavastu p. 4, line 3]; Knjr ca 123b2 [=Kanjur]; T2014 59b8 [=Taisho]; and lex s.v.), though if the references are too many (the limit to be set by the user) only the titles of the texts concerned occur – from this list the occurrences of the search term in individual texts can be accessed.

> Option 2: A concordance can be generated in which the search term (with variants) appears as a key word in context, with the lines where it appears successively for easy study. Provisions are made for searches for one word within the context of another word, the context being specified by the user in terms of words or characters to the left and right.

> lex(icon): Every form, grammatical or the result of sandhi, will be listed systematically, and can be searched in the lexicon, as subsumed under a main Skt "headword." This headword, or main Skt entry, has the form of the basic grammatical form, as root (with prefixes), stem, etc. All the extant grammatical forms will be listed in accordance with traditional grammatical principles (cases, tenses, modes, etc.), and will thus provide a complete grammatical documentation for each word of the input Skt texts. Listed under each Skt (sub-)entry, or form, are listed also all equivalents in Chin, Tib. Eng is listed only under the headword. As a general principle concepts in Tib and Chin are defined as semantic units on the basis of their Skt counterpart.

> Equivalence relations: Each Skt word in the texts is linked with its equivalents in Chin, Tib and Eng – the point of departure is thus Skt throughout. The equivalencies will be linked by reference to the lexicon files: one equivalence will consist of one Skt lexical entry linked to one Chin and one Tib lexical entry. Grammatical particles in Tib and Chin are given as equivalents the Skt grammatical function. When a Tib and Chin particle expresses a grammatical structure in Tib or Chin which has no counterpart in Skt, it is marked with the abbreviation *part.* for *particle* (e.g. *abl[ative] part.*; *isol[ating] part.*). When it refers directly to a grammatical function in Skt, it is named after that grammatical function, (e.g. *abl[ative]*, *acc[usative]*, *pres[ent] tense*, *plur[al]*), or, usually, by a documented Skt word including its, case, inflection, etc. Equivalencies will be established when inputting parallel paragraphs, but the number of

new equivalents will decrease rapidly, as most equivalents will already be input after a rather limited number of texts. Lex will include all lexical forms that occur in the treated Skt texts, and has to be controlled all the way to ensure that all forms in the treated texts are included, when the actual form is not input.

> record: a [up to] quadrilingual unit. One can search for equivalencies involving words in any of the languages concerned, e.g. search for all the Tib and Chin equivalents of a certain Skt term, in its various forms. From lists of such equivalencies one can immediately access the texts in which the equivalencies are evidenced and study them in context.

II) Click on reference (text-page-line).

> Text appears in its quadrilingual context as a record. The quadrilingual setup has four windows, one for Pali/Skt, one for Chin, one for Tib and one for Eng and makes up one record. If there exists more than one translation, which is often the case in Chin and sometimes the case in Tib, these will appear as parallel version paragraph by paragraph in the same window. When commentaries (Pali/Skt, Chin, Tib) are extant in only one language, they appear also in the same window as the root-text. All text will be input with one quadrilingual paragraph on each record. Eng is envisaged as a guide to understand the other Buddhist languages, but will itself constitute a full Buddhist canonical scripture in Eng, if all texts are translated. One also has the possibility to input more tranlations paragraph by paragraph, even also French and German, but one Eng should be the norm. The paragraph filling each item in a record should not be longer than it can make clear the parallellity of the texts. The line-changes in a record should be those of the editions used, so as to be able to refer directly to the standard text edition employed – usually more than one line will be quoted from these, be they Skt, Chin, Tib or Eng. Each record must have a full set of references to the standard editions employed.

III) Click on any word appearing in a record:

> Takes the user to the formed clicked on as it appears in lex, in Skt as subsumed under its basic grammatical form with all possible equivalents in Chin, Tib, Eng. This helps the user to identify the equivalents in a record. Chin and Tib syllables clicked on takes the user to this syllable in lex so as to view the Chin and Tib words in which it appears and its extant possible equivalents in Skt.

IV) Click on any word in lex.

> References to this word appear in the same way as in I, Option 1, excepting the lex s.v. function.

V) Click on any abbreviation.

> When clicking on a text abbreviation one accesses documentation on the standard edition, to which text the references given in the thesaurus refer, then further bibliography. For Skt the best available edition is employed; for Pali PTS; for Chin T(aisho); for Tib Derge, but with concordances accessible. When Tib or Chin have critical ed., these should be employed, with references to side and line; for Eng, the standard translation should be employed, but with reference only to page. Copyright problems will have to be addressed. It is desirable that many translations into Eng are produced for the TLB itself. Clicking on any abbreviation will access its explanation.

By the five routines defined by these rules one may move between all levels of the thesaurus, and thus access multiple language structures.

Envisaging such routines is a rather easy discipline compared to the discipline of applying them in concreto on the given materials. Indeed a huge scholarly contribution within the field of Buddhist philologies is needed, but also a huge programming work to make the enormous databank function in accordance with the intentions, to make the proposed routines function in the machine in an effective way and with a speed acceptable to the user. In this connection also the question of standards, so that the thesaurus can work on both Mac and PC and whatever systems that may be relevant, has to be addressed.

The first texts in the process of being input are the Samghabhedavastu, the Shikshasamuccaya, the Yogacarabhumi, the Ratnagotravibhagabhashya, the Samadhiraja, the Akshayamatinirdesha, and more.

A search machine is being developed by Dr. Jens Oestergaard Pedersen, Department of Asian Studies, University of Copenhagen. In this way Dr. Oestergaard Pedersen has made a great

contribution to TLB project, being also a Sinologist. The search machine will be developed to integrate the principles listed above, and it will, in it present state of development, be demonstrated at the EBTI-meeting in Seoul, May 2001.

To sum up:

1. TLB is quadrilingual in such a way that the same passage can be compared in Pali/Sanskrit, Chinese, Tibetan. In addition there will be English translations.
2. TLB is lexicographical, listing terminological correspondences between these languages exhaustively.
3. TLB is morphologically analytic in that occurrences of all derived forms from a derived stem can be found by searching the stem as keyword.
4. TLB is chronologically specific in the sense that all text entries are dated to the extent this dating is possible.
5. TLB is text-critical in the sense that those texts which are entered should be provided with textual variants.
6. TLB is analytical in the sense that the lexical equivalencies will not only be listed but analyzed and sub-classified according to relevant criteria.
7. TLB is cooperative and international, possible collaborators from Japan, USA, Germany, Russia, Hungary, Denmark, Norway have shown interest. We hope more will join the effort.
8. TLB will start formally at the Institute for Advanced Study in Oslo in January 2002, where a research group on Buddhist studies will work for the academic year 2001/2 under the leadership of the author of this paper.
9. The project will be constructed with a relational database management system. This approach is considered to be more flexible than the ordinary approach which uses tagging for mark-up. This procedure also keeps the original text more “intact” for the user, when he wishes to import the texts into his own work.