White Paper Report

Report ID: 104542

Application Number: HD5138311

Project Director: Stefan Baums (baums@berkeley.edu)

Institution: Mangalam Research Center for Buddhist Languages

Reporting Period: 9/1/2011-2/28/2013

Report Due: 5/31/2013

Date Submitted: 5/31/2013

White Paper

Grant Number: HD-51383-11

Buddhist Translators Workbench

Principal Investigator: Stefan Baums

Grantee Institution: Mangalam Research Center for Buddhist Languages

May 31, 2013

1. Background and Goals

The earliest Buddhist texts were transmitted in oral form, and the evidence indicates that as far back as the time of the Buddha himself, his followers used a variety of Indian dialects for spreading his word. While the use of Sanskrit – the sacred language of the Brahmans – was initially avoided, around the second century CE it too began to be used for the composition of new Buddhist texts, and the old canonical texts were translated into Sanskrit. At the same time, Chinese-speaking Buddhists from Central Asia and China began preparing the first translations of Buddhist texts into a non-Indian language. This process of linguistic translation and cultural transposition continued unabated until it reached a culmination in the eighth century CE, when Tibetan translators refined and codified their sophisticated system of collaborative translation.

The Buddhist Translators Workbench (BTW) project found its initial inspiration in this Tibetan effort, and in particular in the attempt to standardize the translation of Sanskrit Buddhist terms into the Tibetan language. Two works are prominent in this regard: the *Mahāvyutpatti* and the *sGra sbyor bam po gnyis pa*. Both were composed by committees of scholars who meticulously considered the semantic range of Sanskrit terms and painstakingly sought out Tibetan expressions that could faithfully capture the most universal sense of key Buddhist terms. Unlike these ancient translators who were at liberty to create a more or less artificial Buddhist idiom in their target language, modern translators of Buddhist texts cannot aim for a one-to-one mapping of technical terms in their source and target languages. Instead of trying to establish such mechanically fixed conventions for the modern translation of Buddhist terms, the aim has to be a contextually appropriate translation of Buddhist terms in their individual contexts.

To facilitate the determination of the contextually most appropriate translation choices, the concept of a Buddhist Translators Workbench emerged in the late autumn of 2010 as an extension of a multi-lingual digital Buddhist glossary project initiated at the Mangalam Research Center for Buddhist Languages (MRC) earlier the same year. The concrete aim of BTW is the development of a set of digital tools (the 'workbench') to support academic as well as non-academic translators of ancient Buddhist text into modern language. The tools are developed with close attention to the traditional workflow of translators in this field – discussed, observed and implemented initially at a series of workshops and symposia at the Mangalam Research Center (see section 3). Conceptually, the set of tools falls into two parts: (1) a Translation Terminology tool (with Article Editor), and (2) a Multilingual Search tool (with Dictionary Lookup tool).

The Translation Terminology tool provides a database for individual translators or members of coordinated translation teams to document and justify their translation choices for Buddhist technical terms. The database is designed to be maximally open on the input side, including to members of the general public, but to provide powerful filtering mechanisms on the lookup side, so that a translator working on a specific text has full control

over consulting all and only those translation choices of his predecessors that are relevant for the task at hand. In addition to the free contribution of translation terminology, the BTW team is engaged in entering translation choices made by recognized leaders in the field that were previously only available in the form of inconsistent glossaries or embedded into the full translated texts.

The Multilingual Search tool addresses two fundamental facts about the early tradition of Buddhist texts: (1) The Middle Indo-Aryan dialects and Sanskrit that were the vehicle for Indian Buddhist literature are to a very large degree mutually intelligible, and the early Buddhist communities of India transformed their texts freely from one to the other. Consequently, when wishing to search the available early Buddhist corpus in Indian languages, careful scholars cannot restrict themselves to any one of the dialects in question. (2) Very large parts of the early Buddhist literature of India have been lost in their original language and are now only available in ancient translations into Chinese, Tibetan and several minor languages of Central Asia. For this reason, scholars wishing to settle a point of content rather than form by means of a corpus search cannot restrict themselves to Indian languages either, but must equally take into account those texts only preserved in translation. Since very few scholars are equally competent in all the relevant languages, and since the total amount of relevant textual material is vast, the Multilingual Search tool provides support for this type of balanced multilingual search. Drawing on the cross-linguistic terminological equivalences established in the Translation Terminology tool's database, the Multilingual Search tool will eventually make it possible to search for a term in any of the main classical languages of Buddhism (Sanskrit, Pali, Chinese, Tibetan) or any of a number of modern languages (English, French, German, Spanish, Japanese) and also receive results in any of the other languages covered by the database.

In addition to these two central tools, BTW also provides a number of convenience functions supporting other aspects of the translation task, such as a Dictionary Lookup tool. In its final implementation, it will also offer strong collaborative functions (already implemented in part in the Article Editor).

2. The BTW Team

The founding team of BTW consisted of Stefan Baums (then a visiting professor at the University of California, Berkeley, now at the University of Munich and the Bavarian Academy of Sciences), David Mellins (senior research fellow at the Mangalam Research Center and visiting scholar at the University of California, Berkeley), and Jack Petranker (director of the Mangalam Research Center). Baums and Mellins served as co-directors of the BTW project, and Baums as principal investigator of the NEH startup grant. This core team was soon joined by Luis Gómez (professor at the Colegio de México and academic director of MRC), Alberto Todeschini, Ligeia Lugli and Andrew McGarrity (postdoctoral fellows at MRC and visiting scholars at UC Berkeley). The technical design and programming work for the BTW system was carried out by Simon Wiles (doctoral student in

Buddhist Studies at Stanford University) and Louis-Dominique Dubeau (visiting professor at Dickinson College). Michelle Baird served as grant coordinator, and Ann Bergfors provided administrative support at MRC.

3. BTW Products I: Translation Terminology and Multilingual Search

By the end of its NEH startup-grant phase (March 2013), the BTW project had established an information website at http://btw.tilaa.nl/ as well as on the main MRC website at http://www.mangalamresearch.org/projects-resources/buddhist-translators-workbench/. The BTW tools are made available for general download from GitHub and are tested by the project team on a development server at http://46.19.36.89/btw/. Please refer to the appendix of this white paper for illustrative screenshots.

The current implementation of the Translation Terminology tool (developed by Wiles) contains lexical data in Sanskrit, Chinese and Tibetan collected in part at the Mangalam Research Center's Triṃśikā workshop in October/November 2011 (see below).

The current implementation of the Multilingual Search tool (also developed by Wiles) allows for regular-expression searches over the separate language corpora, and contains small sample corpora of Chinese and Pali Buddhist texts for illustration. The next step will be providing broad coverage of canonical texts in Chinese (provided by CBETA), Tibetan (provided by the Tibetan Buddhist Resource Center), Pali (provided by the Vipassana Research Institute) and Sanskrit (provided by the University of the West). This measure already provides the first online search facility for Buddhist texts in multiple languages, constituting a significant advance in the field.

To approach the goal of true multilingual searches as described above, following the period of the NEH startup grant, Mangalam Research Center staff will seed the Translation Terminology tool with Sanskrit-Chinese glossaries (provided by Seishi Karashima, Soka University), Sanskrit-Tibetan glossaries (provided by the Tibetan and Himalayan Library) and a Sanskrit-Pali glossary (extracted from the PTSD dictionary). This will in effect establish Sanskrit as the linguistically central member of multilingual searches, reflecting its central role in the history of Buddhist literature as it spread beyond India. While searches on any of the other languages will be possible, it is anticipated that searches taking Sanskrit as their starting point will yield the most targeted results for the foreseeable future.

In addition to these two main tools, the BTW development server already provides a dictionary search function. This currently allows the pasting of any text in Chinese or Tibetan, mouse-over definitions being provided from three freely available dictionaries (two for Chinese, one for Tibetan, with a total of 270,000 entries).

4. BTW Products II: 'Salve' and 'Wed'

In August 2012, at roughly the mid-point of the NEH startup-grant period, the BTW project evaluated the needs of the scholarly team against the state of the BTW software prototypes developed to that date. This evaluation highlighted the fact that, whereas previous software development had focused on functions primarily useful to people searching through BTW's data, what BTW's scholarly team needed to create this data was software designed to assist them in authoring and publishing their work. Therefore, the focus of software development shifted to address this immediate need, and Dubeau took over the management and maintenance of the prototypes that had been developed by Wiles. Dubeau designed new editing functions for BTW and a storage schema for encoding the articles created by the BTW team, and tested this schema against actual data.

An evaluation of existing software for online collaborative editing of structurally-rich documents showed that the contenders most likely to present the functionality needed for the BTW project were ANGLES, Ace, CodeMirror and XET. It turned out, however, that these projects did not provide the authoring assistance needed for BTW. ANGLES initially appeared to be a project that would provide an appropriate level of assistance for BTW's scholarly team, but using it for BTW would have required scholars to author their articles directly in XML without much assistance from the authoring software. Requiring our contributors to bear the burden of authoring in XML would have been unrealistic.

To overcome these limitations, Dubeau began designing and developing 'salve,' the foundational library upon which all other authoring functions of BTW depend. This library is able to validate an XML document against a RelaxNG schema. It also provides context-sensitive assistance for modifying such documents in a way that maintains the validity of the XML structure. Next Dubeau began prototyping and designing 'wed,' the component of BTW that provides the core of authoring assistance for creating articles for BTW. This component hides the XML structure of a document behind a user-friendly interface. Whenever the document is changed, wed automatically verifies (by using salve) that the document still has a valid structure. It automatically numbers those elements that require numbering and renumbers them appropriately whenever needed, automatically numbering references to such elements. It also provides context-sensitive editing assistance. Both 'salve' and 'wed' have been released under an open source license on GitHub, in the expectation that they will be found useful for other digital humanities projects besides BTW. They are available for public download from the following URLs: https://github.com/mangalam-research/salve and https://github.com/mangalam-research/wed.

A software-development agency was contracted to produce a component of BTW designed to manage references to secondary sources. This component provides a graphical user interface to help BTW's authors search through bibliographical information when they want to insert a new reference in their article or change a reference already present. The bibliographical information itself is managed by Zotero, while the component works as a conduit between the Zotero database of works and BTW. By using Zotero in this way, the BTW project avoids

having to develop in-house software to manage bibliographical information.

5. Communication and Collaborations

Project directors Baums and Mellins gave formal introductory presentations on the BTW project at the NEH ODH Project Directors' Workshop in Washington, September 27, 2011, and at the Mangalam Research Center's Triṃśikā Workshop in Berkeley, October 31 – November 3, 2011. The following additional presentations on BTW were made during the report period:

Staff presentation at Yogācāra Symposium at Mangalam Research Center, November 2011.

Presentation by Jack Petranker, "The Voice of the Dharma and the Future of Buddhism," at the Global Buddhist Congregation, New Delhi, December 2011.

Presentation by Simon Wiles, "The Buddhist Translators Workbench, An Integrated Suite of Tools for Working with Buddhist Languages," 2nd Int. Association of Buddhist Universities Conference, Maha-chulalongkornrajavidyalaya University, Thailand, May 30 – June 2, 2012.

In addition, throughout the spring of 2012, the project directors and Todeschini regularly participated in the UC Berkeley Digital Humanities Working Group to exchange insights and results with other members of the Berkeley Digital Humanities community.

The BTW project also pursued institutional collaborations with the Sanskrit Library project (headed by Peter Scharf, Université Paris Diderot) on the sandhi resolution and parsing of Sanskrit texts, and with the University of the West on the provision of digitized Sanskrit Buddhist texts.

From March 8 – 10, 2013, the Mangalam Research Center hosted the conference "Advances in Digital Humanities for Buddhist Studies," marking the conclusion of the NEH Startup Grant and in part funded by it. Eleven leading practitioners of Digital Humanities approaches to Buddhist Studies, from the United States and Europe, assembled to discuss the use of digital tools in their own projects and the potential that the Buddhist Translators Workbench holds for the field. Video for several of the presentations can be found at http://www.mangalamresearch.org/projects-resources/videos/btw/. After a welcoming address (by MRC director Petranker), an introductory presentation of the BTW project (by co-director Mellins), and an in-depth discussion of the BTW translation tools (by MRC academic director Gómez and post-doctoral fellow Lugli), presentations covered the following topics:

Online research and publication in Buddhist Studies (Birgit Kellner and Eric Decker, University of Heidelberg)

The contribution of optical character recognition to the development of large corpora for the use of BTW (Kurt Keutzer, University of California, Berkeley)

Translation tools for the Tibetan Tengyur (Paul Hackett, Columbia University)

Digital collation and digital editions of Chinese Buddhist texts (Marcus Bingenheimer, Temple University, and John Kieschnick, Stanford University)

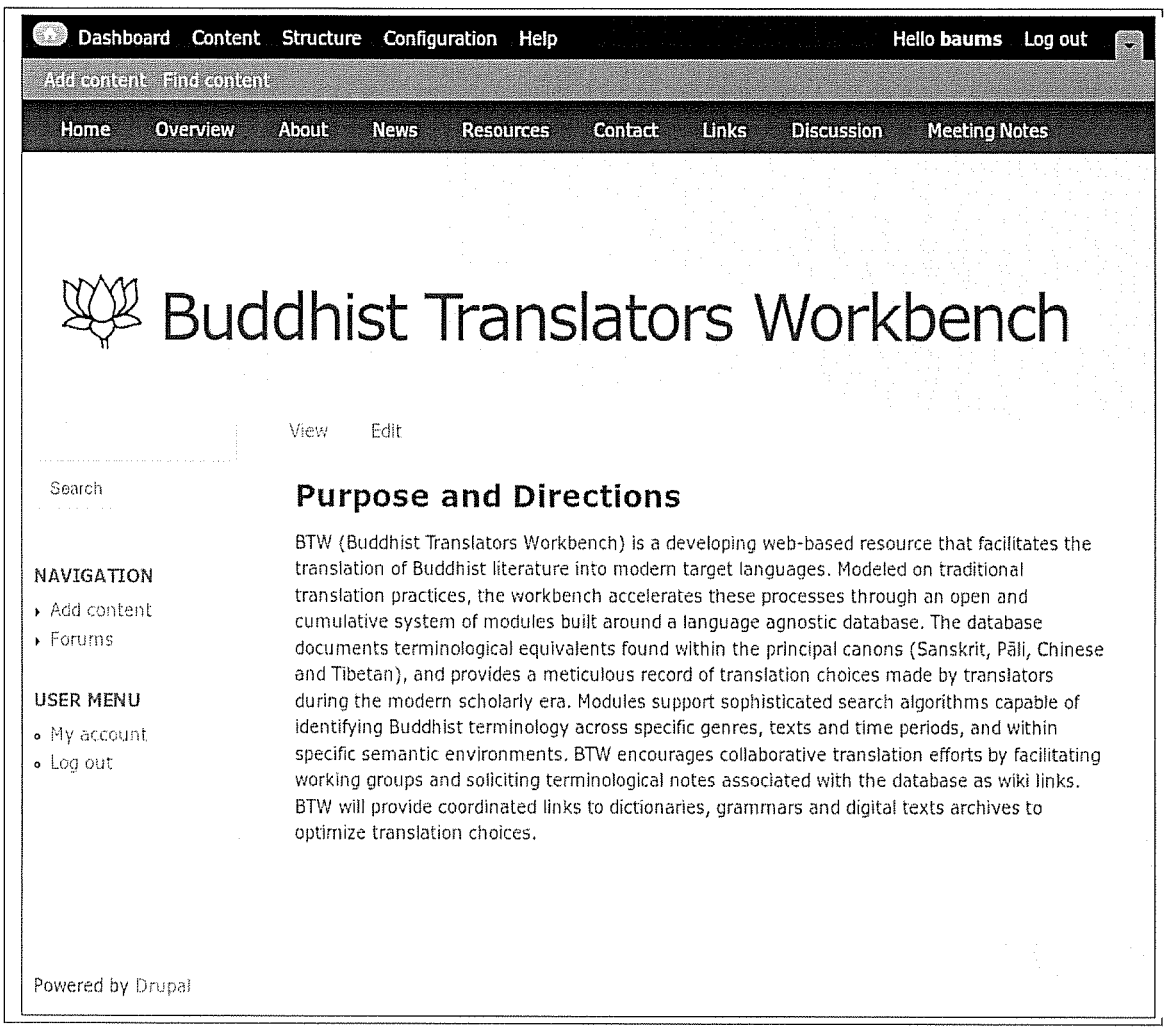
The integration of Tibetan texts with lexical and geographical resources (Steven Weinberger, University of Virginia)

Digital tools for collaborative research on Gāndhārī texts (Stefan Baums, University of Munich).

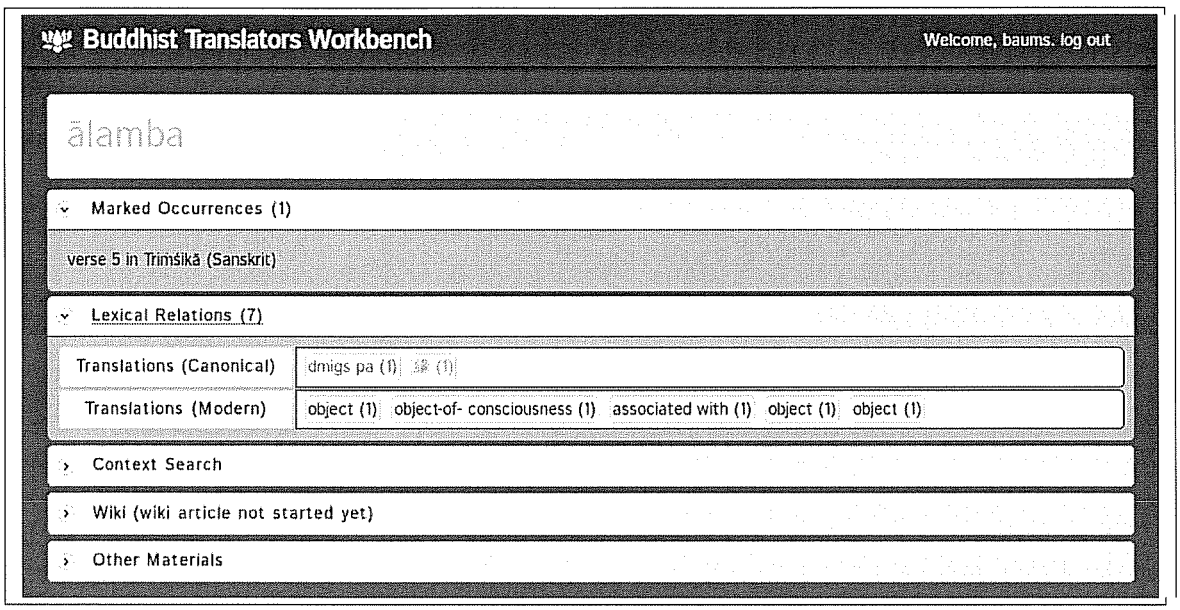
A concluding discussion highlighted the importance of open data and well-defined formats for the exchange and long-term usability of source texts and research data in the field of Buddhist studies, and discussed the role that BTW can play in developing such formats as the project unfolds beyond its startup phase.

Appendix: Screenshots of BTW Tools

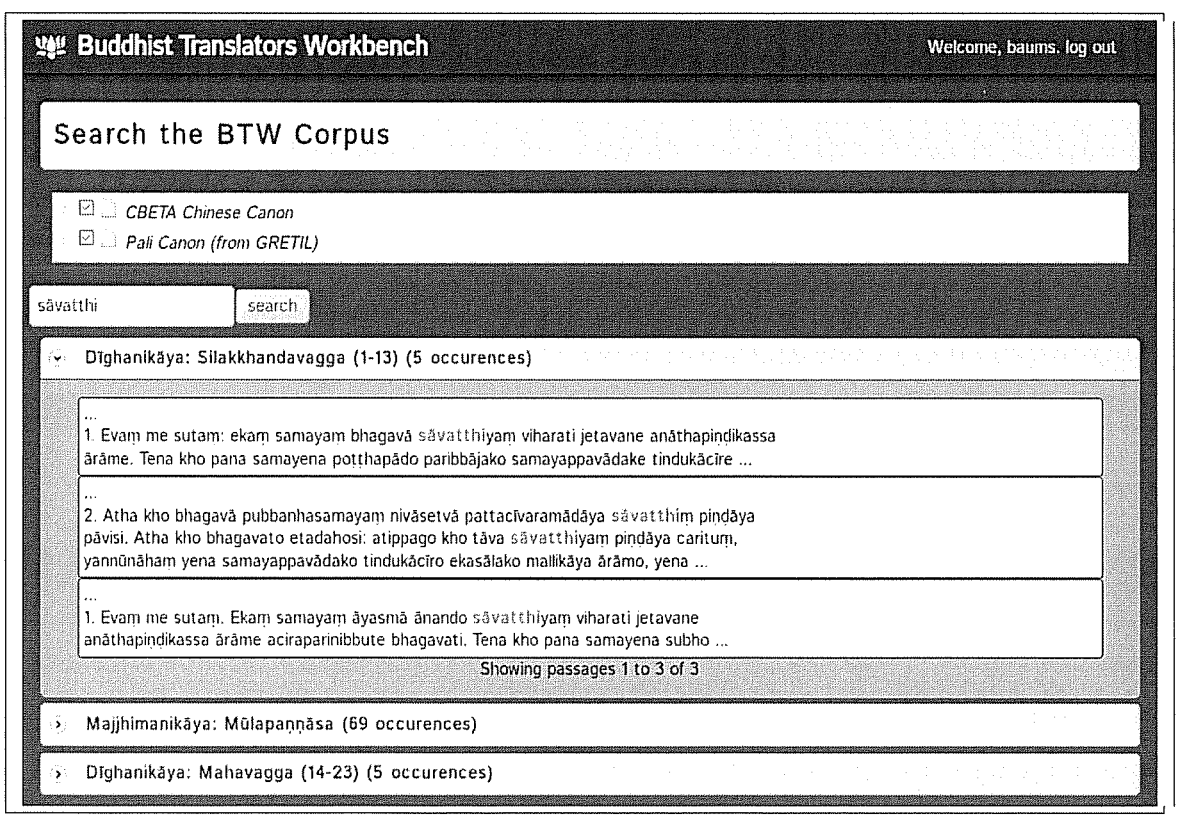
The Buddhist Translators Workbench informational website (http://btw.tilaa.nl/):



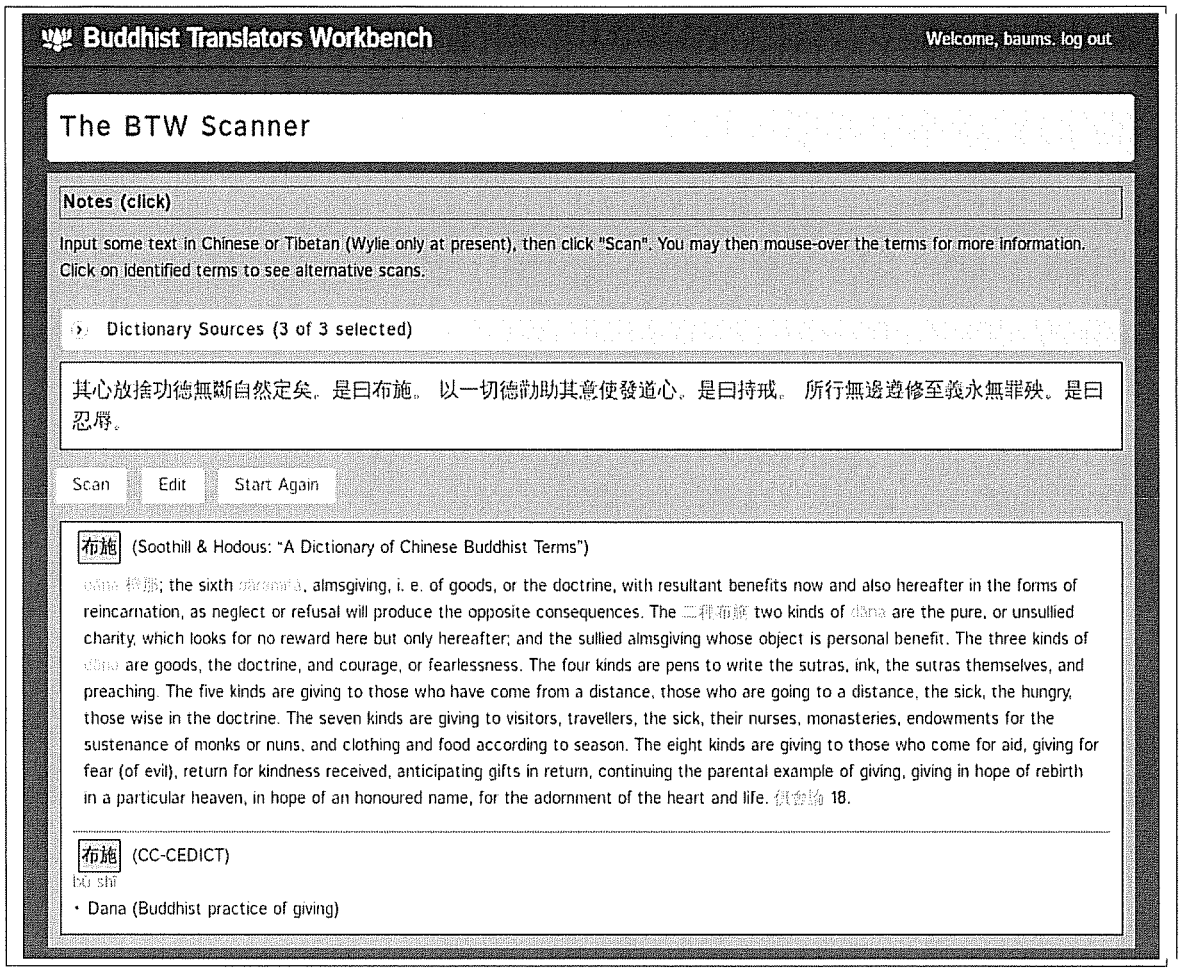
The BTW Translation Terminology Tool:



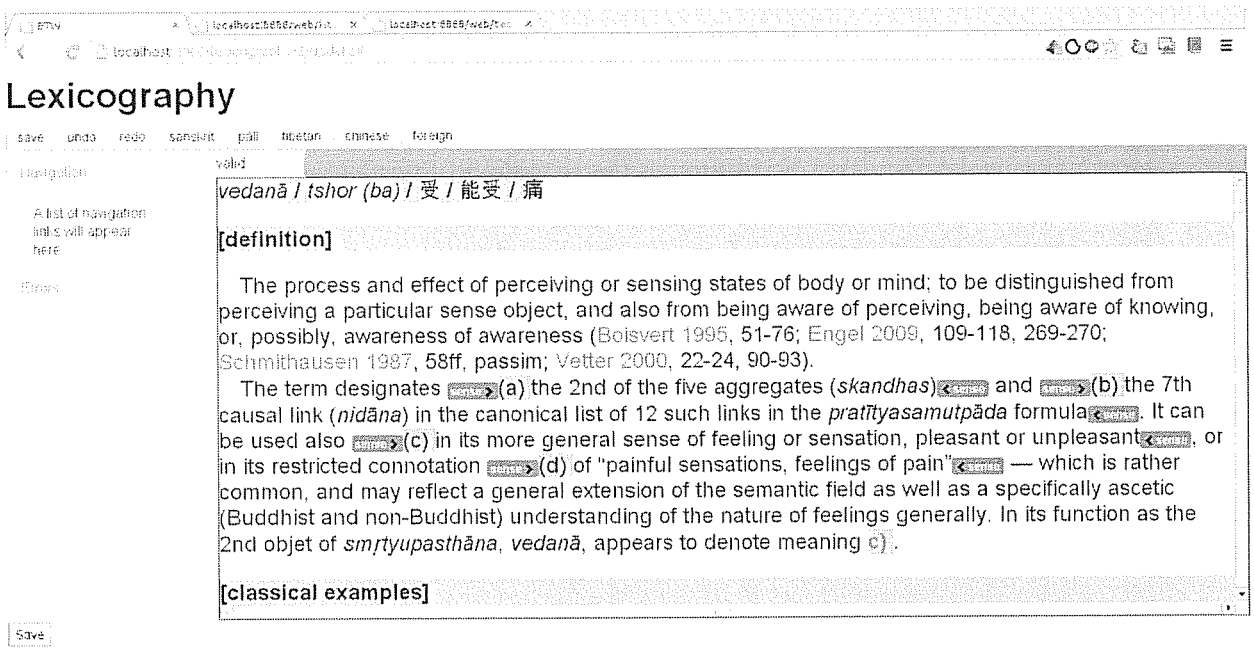
The BTW Multilingual Search Tool:



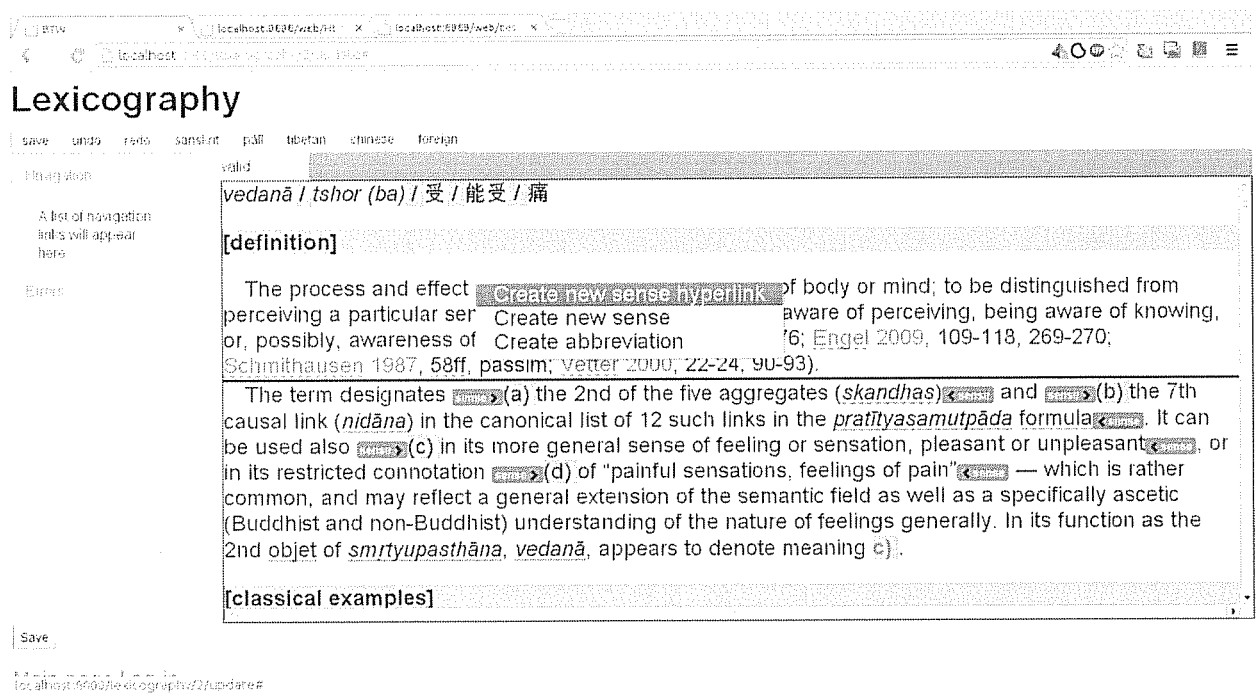
The BTW Dictionary Lookup Tool:

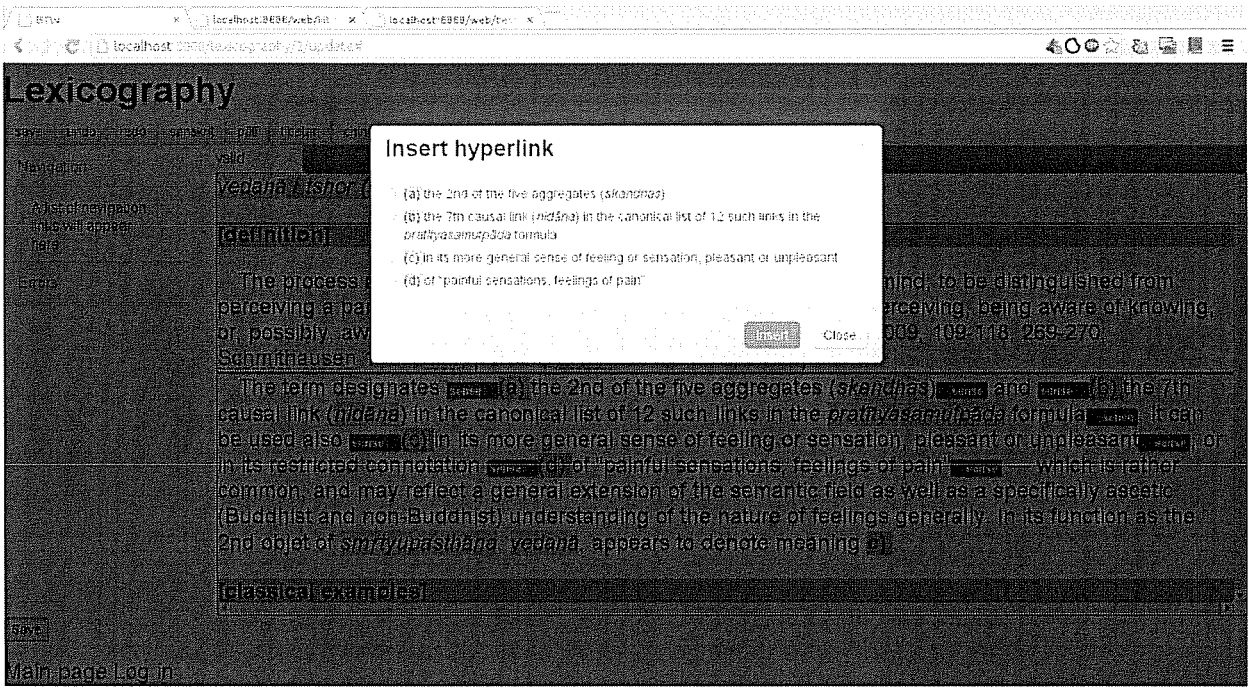


The BTW Article Editor: The green bar indicates that the document has been fully verified and is valid. Text in gray is text that cannot be directly edited by the author, either because it is immutable by design (e.g. the first section is titled "[definition]" in all articles) or because BTW (through Wed) automatically manages that text (e.g. the "sense" elements all begin with labels "(a)", "(b)", "(c)", etc.).



The user may bring up a contextual menu to know what editing actions are possible in a given context. Depending on where the user clicks, the list of possible actions changes.



This screenshot shows the dialog that BTW presents if the user selects "Create new sense hyperlink." Once the

user selects one sense for hyperlinking, BTW will update the XML structure behind the scenes to encode a hyperlink from the XML ptr element representing the link to the XML sense element representing the sense.