Note on the first BETA-version publication of the Database of Medieval Chinese Texts (DMCT)

Introduction

DMCT is a collaborative project between the **Centre for Buddhist Studies at Ghent University (Christoph Anderl), Dharma Drum Institute of Liberal Arts (DILA, Joey Hung / Lin Ching-hui), and Marcus Bingenheimer / Zhang Bo-yong**. Although it is an ongoing and open-ended project, we have decided to make a BETA-version public with a focus on the manuscripts which have been digitized and marked-up so far. The DB also includes other analytical parts (on Late Medieval Chinese syntax and sentence analysis), however, these parts will be made public at a later date. Currently, ca. 750 LMC function words are registered and analyzed, illustrated by ca. 700 parsed example sentences. Another part of the DB on phonetic loan characters in vernacular Dunhuang texts will likewise be made public at a later date.

In addition to the texts which were directly digitized in the framework of DMCT (financed by research foundations of Flanders, Ghent University, DILA / Chung-hwa Institute of Buddhist Studies, and the Tianzhu Foundation; main encoder: Lin Ching-hui [JH] 林靜慧), Marcus Bingenheimer also contributed the results from his project "**Four Early Chan Texts from Dunhuang**" (funded by the Chung-hwa Institute of Buddhist Studies; 2014-2017; main encoder: Zhang Bo-yong [ZBY] 張伯雍). The latter project was also published in book form. [1] For an introduction to the project and the texts, see here.

The selection of the texts digitized in the DMCT project (2015-) has been motivated by the research interests of the participating scholars and ongoing PhD projects at GCBS. Since the initial impetus was linguistic research, many of the Dunhuang texts were selected based on their vernacular / colloquial features (syntax and semantics). As the project progressed, the focus gradually shifted to variant character forms, yitizi 異體字. As such, this Beta version will also contain a **Database of Variants** extracted from the digitized manuscripts. This dataset is continuously growing parallel to the digitization of new manuscripts. Most of the variant forms registered are directly linked to the manuscript line they occur in. Up to now, more than **32.000 variant character** - text passage links have been collected in the Variant DB. Please be patient when entering the Variant DB. For better performance the overview of all variants will be preloaded, this process might take a few seconds.

Since the variant forms collected in the dataset are based on different projects and slightly different approaches to the mark-up, they are registered and visualized in various forms. Since the focus was not on registering every specific form, but rather on the **structure** of the variants, whenever possible, the images of variant forms contained in the Taiwanese DB 異體字字典, YTZZD, were used and referred to (identifiable by their number as registered in this dictionary). In the "Four Early Chan Texts from Dunhuang" project, further **images of variant forms** were created as part of the project (for those forms not contained in YTZZD).

In the DMCT project we took these two datasets of variant forms as point of departure and used them whenever possible, however, for variants not contained in either of them, the original form as found in the respective manuscript is reproduced.

How to use the Database?

The DB is multi-functional and presents editions of a selection of important Medieval Chinese texts, based on meticulous (and very time-consuming) mark-up and a comparison with previous editions. Please note, that the texts presented are **not critical editions**, but the aim was to reproduce the features of the original texts as faithful as possible. As such, occasionally, we digitized several versions of the same text as represented in different manuscripts (sometimes, in the footnotes a critical apparatus is added, with comparative notes and annotations). The XML marked-up version of each text is visualized in a twofold way, a **diplomatic transcription**, and a **regularized version**. Whereas the diplomatic transcription tries to reflect as many original features as possible (including character variants), the regularized version tries to produce a "clean" and readable text. In the diplomatic transcription, by gliding the cursor over a character form marked in light-orange color, the variant form will appear on the right upper corner of the screen. This variant form is either a form imported from the YTZZD, or – if not registered there – based on a drawing or represented by a photograph of the variant form as it appears in the respective manuscript. The other main features of the mark-up are explained in the beginning of each marked-up text.

Some of the digitized texts already served as the basis for new editions / translations (concerning the project / publication of the 破魔變, [2] see here), as well as research articles, either already published or forthcoming. [3]

All the variant forms were collected in the **Variant Database**, designed and programmed by **Christian Bell** and **Jan Schrupp**. The aim of the DB is to collect as many variant forms as possible, based on the digitization and encoding of the manuscripts, and develop a tool to facilitate the reading of Dunhuang manuscripts which abound in rare variant characters. In addition, we hope that the DB will develop into a useful tool when researching variant forms in non-canonical Dunhuang texts. Currently, we also install a more advanced search function in the DB (already implemented in the "Bibliography" section), as well as importing the information kindly provided by William H. Baxter / Laurent Sagart concerning the historical readings of the characters.

We have also tried to implement an "interactive" element into the DB, in order to be able to register comments and corrections; this is a project in progress, and necessarily there will be mistaken readings and interpretations, alternative readings, and bugs. Currently, there is the possibility to add comments in the "comment box." In order to do so, one has to **register as a user** in the DB. Registered users will also receive information on updates, etc.

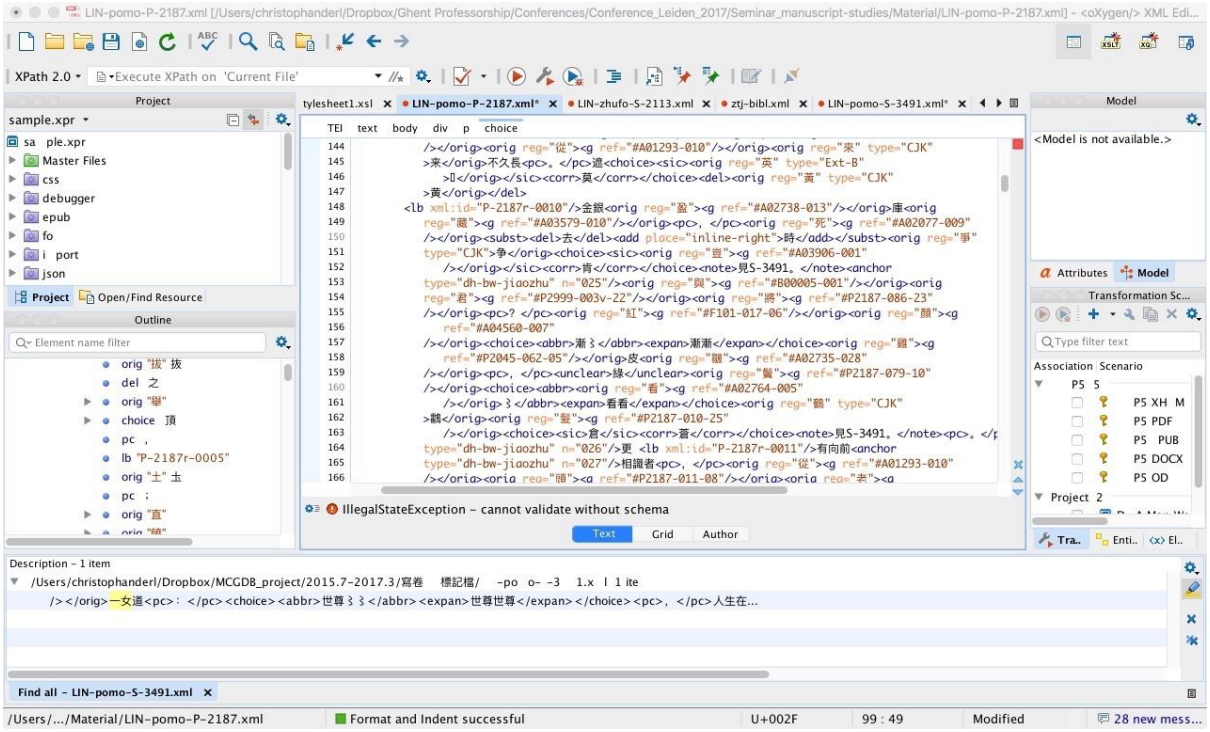
The project is designed to be "open-ended", since the amount of interesting non-canonical textual material from Dunhuang is huge, and only a small selection can be digitized based on the time-consuming nature of the mark-up work, as well as the specialized knowledge and the great patience and diligence demanded from the encoders.

In addition, the DB is adapted to the needs of ongoing research projects; currently, we are

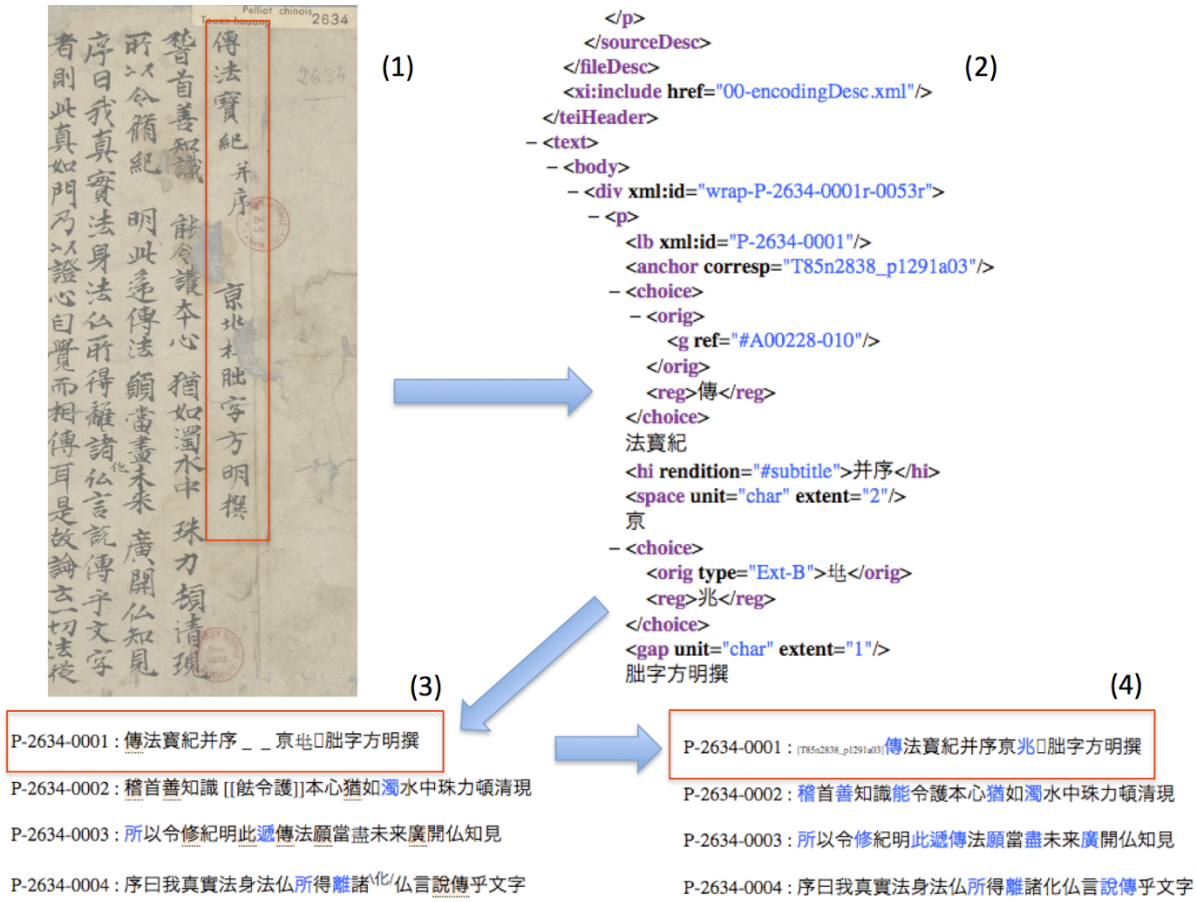
implementing a tool for registering and cross-referencing “Chan phrases and sayings”, based on a joint research project with Sichuan University.
During the last two years the DB was also used in Master courses (e.g., students jointly translated and annotated the Taizi chengdao jing 太子成道經), as well as in 3-months internships for Master students. As such, the DB has a “pedagogical” function in teaching advanced Master and PhD students (for a selection of PhD research projects in Ghent, see here).

Some notes on the technical design

The texts in the DB are marked up according to TEI standards. In the DMCT project, we basically followed the conventions formulated by Marcus Bingenheimer for the project on “Four Early Chan Texts from Dunhuang”, with certain adjustments.



The mark-up is done in oXygen (here, a short portion of the Po Mo bian 破魔變 from Pelliot 2187).



Example of the process of transcription and encoding of line 1 of ms. Pelliot 2634: (1) The first line of the manuscript facsimile is indicated by the red box; (2) shows the encoding of the first line into TEI-compatible XML in oXygen; (3) the red box shows the transformation of (2) into a html webpage, directly reflecting the features of the manuscript ("diplomatic version"); (4) shows a html transformation of (2) into "normalized" text ("regularized version"); this version is the basis for further work on the manuscripts (e.g., annotations; translations; grammatical analysis).

Currently, the technical framework is as follow:

- In earlier versions of the DB we used eXist DB for storing all data as XML files, but ca. one year ago we switched to MySQL. MySQL is a relational database, which is organized in tables;
- MySQL can use different storage engines; depending on the specific table, we use InnoDB or MyISAM;
- MyISAM is used for all tables which are designed for full-text searches. InnoDB is used for all other tables like user management tables, etc.;
- The program logic is implemented in PHP, using object-oriented programming (OOP) and other modern interfaces like PDOs;
- The view is designed with CSS;
- Further supporting languages are HTML5 and JavaScript;
- Since we deal with XML files, but the database itself does not store XML files, we implemented an XML import/export function;

- All previous features of the old database have been implemented in the new system, in addition to new features such as the module for Variant Characters, commentary functions in all modules, and several other features including sophisticated input masks. A more advanced user management, several search functions including global search;
- The new system is considerably faster, more stable, and it is much easier to add new features (including new input masks and analytical modules) and fix bugs;
- The development process is optimized: Source Code is managed via a code versioning system (Subversion) and the deployment is automatized.

Despite its current shortcomings and probably numerous bugs, we very much hope that you will enjoy using this Beta-version!

For the editors,

Christoph Anderl

Footnotes:

- [1] Bingenheimer, Marcus (馬德偉) and Chang Po-Yung 張伯雍 (eds.): Four Early Chan Texts from Dunhuang – A TEI-based Edition 早期禪宗文獻四部 —— 以 TEI 標記重訂敦煌寫卷：楞伽師資記，傳法寶紀，修心要論，觀心論. Taipei: Shin Wen Feng 新文豐. 3 Vols. Vol. 1: Facsimiles and Diplomatic Transcription 摹寫版 (ISBN: 978-957-17-2274-0), Vol. 2: Parallel, Punctuated and Annotated Edition 對照與點注版 (ISBN: 978-957-17-2275-7), Vol. 3: Calligraphy Practice 抄經版 (ISBN: 978-957-17-2276-4).
- [2] 林靜慧,Anderl, C., and 洪振洲. <破魔變>中英對照校注 [Pò Mó Biàn Critical Edition with Annotated Translations into Modern Chinese and English]. Taipei: Fagu wenhua 法鼓文化, 2017.
- [3] For example: Anderl, C. 2018. “Linking Khotan and Dūnhuáng: Buddhist Narratives in Text and Image.” Entangled Religions 5: 250-311. Anderl C. and Sørensen, H. Northern Chán and the Siddhaṁ Songs. Forthcoming in Dūnhuáng and Beyond: Texts, Manuscripts, and Contexts – In Memory of John McRae, edited by C. Anderl and C. Wittern. Numen Series. Leiden: Brill. Anderl, C. Metaphors of ‘Sickness and Remedy’ in Early Chán Texts from Dūnhuáng. In Reading Slowly: A Festschrift for Jens E. Braarvig, edited by L. Edzard, J. W. Borgland, and U. Hüsken, pp. 27-46. Wiesbaden: Harrassowitz.