**HUNG — Digital Corpus**

<2 figures>

**Digital Corpus and Toolset for Performing Text Analysis on Chinese Translation of Buddhist Scriptures**

Hung, J.

The success of Buddhism in China can be partly attributed to the great number of texts translated from Indian languages during the Eastern Han dynasty to the Tang dynasty (618–907 CE). The vast amount of Buddhist texts that were translated from both Sanskrit and other Indian languages between the 2nd and the 11th centuries not only had a far-reaching cultural impact on Chinese society but also made itself felt on Chinese language in terms of syntactic and lexical patterns of change.

Over the years, scholars have leveraged traditional text-critical methods to research language change in Chinese as evinced in the corpus of Buddhist texts. Philologists have investigated doctrinal, literary, and linguistic aspects in order to ascertain the authorship, dating, authenticity, and so forth of these texts. It has been found that many of the authorship or translatorship attributions of early Chinese sutras are unreliable.

Traditional philology has its limits, however, and the advantages of qualitative, computational approaches to the analysis of Buddhist texts are increasingly explored. Digital philology as a branch of the digital humanities, with its application to Buddhist materials, stands to open new horizons for research. For European languages, text analysis has been a very successful field. As for Chinese Buddhist texts, a digital version of the Taishō edition of the Chinese Buddhist canon is freely available through the efforts of the Chinese Buddhist Electronic Texts Association (CBETA). The availability of such a corpus in digital form enables researchers to apply statistical methods or artificial intelligence algorithms to text analysis. However, this type of research is only rarely applied to the study of Chinese Buddhist texts. We believe there are two main issues behind the present under-exploitation of these resources.

First, the design of the CBETA XML markup is mainly aimed at providing a correct and comfortable display of the text on the computer screen rather than preparing the text for future possible uses for the sake of quantitative analysis. Second, the performance of quantitative analysis on digital texts still requires high-level skills in computer programming and advanced statistical knowledge, which creates a high barrier for scholars in the humanities who are now attempting to navigate these tools. In order to assist researchers we develop various resources for the computational analysis of the CBETA corpus. The main tasks include:

1. Creation of a ‘text-analysis friendly’ version of the CBETA corpus.1 This dataset has following features:

• The markup is compliant with TEI P5.

• All information that is not critical for corpus analysis has been taken out—e.g., critical apparatus, markup of menu items, etc.

• The representation of text structures has been simplified and unified; only the <div> element with ‘type’ attribute is allowed for representing the text structure.

• Each textual block is wrapped with an <ab> element with a type attribute, which is used to distinguish text block with different types (prose, verse . . .).

• Every non-Unicode character is assigned a unique code point in the Unicode private-use area (‘undisplayable’ is not a good category; you can display everything, of course).

2. As an alternative to the XML format, we created the N-gram statistic dataset.2 The dataset is generated by transforming the XML file into plain text format and removing all punctuation from the text. Then the long text strings are cut with fixed lengths to generate n-grams, and the occurrences of each gram are calculated.

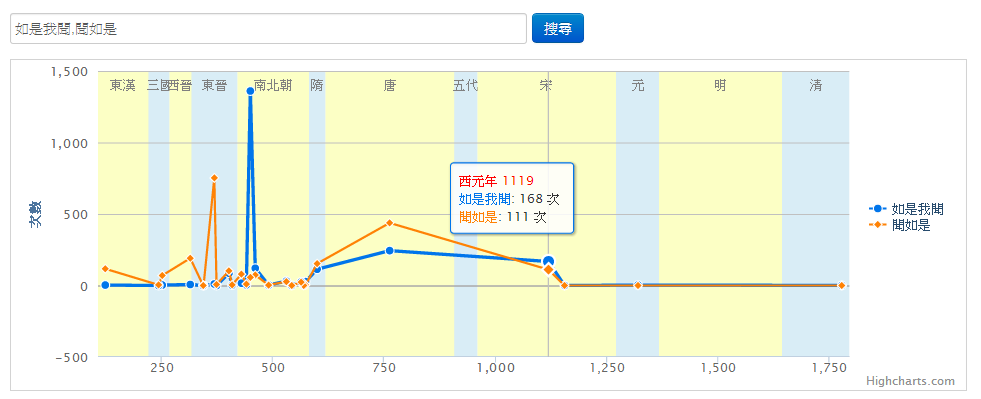
3. We developed an online tool called the ‘Buddha N-gram Viewer’.3 This tool allows users to visualize the overtime occurrences of phrases in Chinese Buddhist texts. It also provides detailed lists of the matches of phrases in the text, which enables the researcher to trace the text back to the original source and look it up there.

Figure 1. Occurrences of ‘Thus have I heard’ as ‘如是我聞’ and ‘聞如是’ in the Buddha N-gram Viewer.

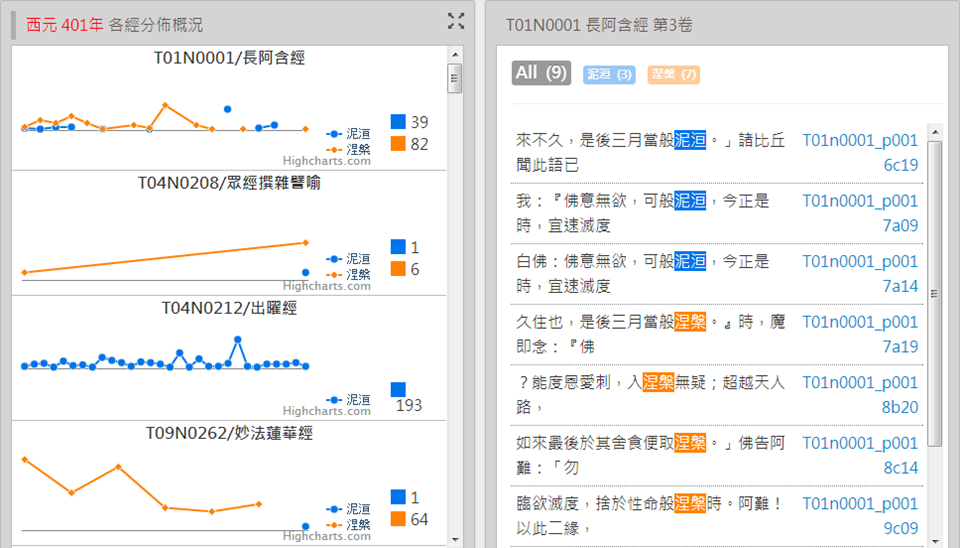


Figure 2. Nirvana as ‘泥洹’ and ‘涅槃’ in the third fascicle of the 長阿含經 (the Chinese translation of the Dīrgha-āgama; T 1).

**Notes**

1. The dataset is available at https://github.com/ddbc/CBETA\_TAFxml.

2. The dataset is available at https://github.com/ddbc/CBETA-ngram.

3. The tool is available at http://dev.ddbc.edu.tw/BuddhaNgramViewer/.