BuddhaNexus

Neural network implementation to facilitate comparative research of Buddhist texts

Sebastian Nehrdich

Venerable Vimala Bhikkhunī

November 5, 2019

Contents

1 Summary

The research into parallels between Buddhist texts of all schools has given us a wealth of information about the history of these texts and the teachings of the Buddha. It has played an important role in changing our perspectives on various topics, like for instance the legality of Theravada Bhikkhunī ordination and the authenticity of Early Buddhist texts.

Traditionally, parallels were found by scholars manually, by studying the texts and noting down similarities among various Buddhist schools. With the advent of computer technology, coupled with the ongoing digitization of Buddhist texts, faster and more accurate methods could be developed.

The BuddhaNexus project aims to create a powerful tool to facilitate the study of parallels using neural networks to compare texts.

As a first step, we have concentrated on what we call "Monolingual Matching" i.e. finding parallels within the corpus of texts of the same language, be it Pāli, Sanskrit, Tibetan or Chinese.

The next step is then to train the neural network to do "Multilingual Matching" i.e. finding parallels between languages and with that between the various Buddhist schools. The neural network can be trained with existing translations between the various languages as well as with known parallels and dictionaries. This step would need far greater computer power than we currently have so for this additional financial support needs to be sought.

This document outlines our methodology and our vision for the BuddhaNexus project.

1.1 Practical information

Website: `http://buddhanexus.net/`
Repositories: `https://github.com/BuddhaNexus`
Neural network output data: `http://buddhist-db.de/suttas/` (this will be changed in the future)

Management is done using an Agile Management approach using ZenHub.

Contact details:
Email: buddhanexus.info@gmail.com

2 Introduction

2.1 Parallels research

Since the 19th century it has been recognized by scholars that many Buddhist texts have counterparts (parallels) in other collections of the various schools, often preserved in different languages. The first documentation of these parallels was published by Nanjio Bunyiu (Nanjō Bun'yū, 南條文雄) in his *A Catalogue of the Chinese Translation of the Buddhist Tripiṭaka* of 1883. Nanjio listed 24 *Dīgha Nikāya* discourses as parallels to the *Dīrghāgama* in Chinese. A year later, Samuel Beal [1884] published translations of the Pāli *pātimokkha* and the Chinese Dharmaguptaka *prātimokṣa* and concluded that these are virtually identical.

This marked the start of the work on comparative studies by many great scholars. A work that has been ongoing until the present day. This work not only had a great influence on how we understand the Buddhist scriptures in the academic field, but also on the practical applications of the Buddha's teachings and thus on the daily lives of many people.

Parallels research can contribute to a.o. dating of scriptures, author-identification of texts, finding the origins and tracing the development of certain key concepts and many mor tasks of the vast field of textual Buddhist Studies.

2.2 Practical applications

In his Comparative Study of the *Majjhima Nikāya*, Bhikkhu Anālayo [2011] has shown that all significant aspects of early Buddhist doctrine are the same across all extant textual transmission of the Suttas of the Pāli *Majjhima Nikāya*.

His work has had a great impact on the Buddhist world. Through it it became possible to distinguish the Early Buddhist Texts from later Buddhist literature, and therewith deepen our understanding of Buddhist practice (see *The Authenticity of Early Buddhist Texts* by Sujato and Brahmali [2014]).

It also had a great impact on the position of women within Buddhist communities. As in many parts of the Buddhist world, the full ordination of women and LGBTIQ was not possible before. His work on comparative research has made an immense contribution to facilitate the Bhikkhunī ordination in the Theravada tradition. (see Anālayo [2018a],Anālayo [2018b],Anālayo [2013])

Also the influence of Jain and Vedic sources on the Buddhist scriptures is a very young field of study within the field of Buddhist Studies, and proper research on parallels between texts preserved in these different languges can have far-reaching influences on how Buddhism is practiced even today (Maes [2015], Vimala [2019]).

2.3 Digital Humanities

For the research of Buddhist textual material, citations and similar passages are of major importance. Finding these parallels is however not trivial. Comparative studies have sofar relied solely on manual comparison of Buddhist texts for the identification of parallels. For this it was necessary for scholars in the past not only to be familiar with the various languages in which these texts are written, but also to have a precise knowledge of the content and vocabular of a larg number of texts. Considering the huge amount of Buddhist textual material of the various schools in existence, it is close to impossible for a human being to detect all possible parallel passages. Digital approaches however by nature seem to be the appropiate way to tackle the task of finding similar and closely related passages reliably within a large number of digitalized texts.

Important parallels are not always literally identical chunks of texts, but smaller things such as word order, used vocabulary, spelling or grammatical structures can vary. Recently, neural networks have become the tool of choice for the detection of such approximate parallels, since they can be trained to extract the semantic features of tokens and do not need to rely solely on chunks of identical strings as was he case with earlier algorithms for parallel detection. It therefore rather recently has become possible to compare large volumes of data with a much higher degree of accuracy than any human would be able to do by manually comparing texts.

Donald Sturgeon [2017] points out in his article *'Unsupervised identification of text reuse in early Chinese literature'*:

> A further observation emerging from this study is that a key advantage of digital systems over traditional printed forms of research lies in the possibility of allowing scholars to work with all data of a particular kind, rather than a useful and important subset selected by experts. With the sheer volume of parallels spread throughout the early Chinese corpus, even the most diligent of conventional studies risks omitting information that may prove relevant to the particular research questions that someone else may be interested in.

Relevant attempts to calculate parallel passages for Buddhist and related material started in 2010 (Prasad and Rao [2010]) and are ongoing since.

In 2016, Dr. Orna Almogi (University of Hamburg) a.o. (Orna Almogi and Wolf [2016]) organized a hackathon involving 17 scholars, scientists and students to develop and compare algorithms for finding parallels within the Tibetan classical corpus of texts.

Various algorithms and methods have been tried and tested subsequently (Klein et al. [2014]). The BuddhaNexus is currently using algorithms that are based on the experience gained in these previous studies while also consantly testing out and developing further new approaches.

2.4 The Buddhanexus Project

The Buddhanexus project was started by Sebastian Nehrdich and Dr. Orna Almogi to create a tool to calculate parallel passages using a neural network, especially with an focus on extracting inexact matches. A detailed description of the neural network and Sebastian's work is given in the section on Methodology.

The output of this neural network is a vast amount of data that needs to be structured and filtered in order to be used in a meaningful way. Ven. Vimala Bhikkhunī, former frontend programmer at SuttaCentral.net, became involved in the project and started to develop a web interface to display the neural network's output data, initially by just displaying the data in convenient tables.

This was further augmented by the work of Prof. Kiyonori Nagasaki, who created a sankey graph with the Tibetan data and eventually a full text display of each text including a 'heat-map' to display the sections with the highest to the lowest potential parallels was added.

The site was further developed with the help and feedback of various professionals in the field of parallels study such as Prof. Michael Radich (University of Heidelberg).

In order to manage the large amounts of data, a former programmer from SuttaCentral.net and indologist was hired to build a backend database for use with the site and speed up the loading times. A more detailed description of the website is given in the section on Website UI.

With this, Buddhanexus is now turning into a very powerful tool for the study of Buddhist texts and for advanced parallels-research which has never been seen before. However, at present Buddhanexus is only able to find potential parallels within texts of the same root language. In the future, we want to train more powerful neural models which are able to deal with the tasks of finding potential parallels between Pāli, Sanskrit and Tibetan and later on also between these and Chinese. These can be further supported by using existing dictionaries and known, verified, parallels and translations as training material. This is further described in the section on Future Vision as well as a roadmap for the development in the next year.

Cross-language comparisons will involve far larger amounts of data and much higher requirements for the GPU hardware during training and inteference. Also, a more stable platform for running the website will become necessary, for which financial backup will need to be sought. All this is described in the section on Financial Planning.

3 Biography

3.1 Sebastian Nehrdich

Sebastian Nehrdich is a master student at the University of Hamburg specializing on ancient Indian and Chinese Buddhism. His main research interest is early Yogācāra Buddhism. Currently he is preparing a new edition of the fourth chapter of the Madhyāntavibhāgaṭīkā by the Indian author Sthiramati as his thesis. Apart from philological research based on primary sources, he has a vivid interest in the application of modern computational linguistic tools on ancient Buddhist material. In 2018, he co-authored together with Oliver Hellwig the EMNLP-contribution "Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks" Since 2019, he is working for the follow-up of the project "Scholars and Scribes" by the German-Israeli Foundation for Scientific Research and Development (GIF). His main focus in this project lies on the creation of multilingual Buddhist quotation networks based on machine learning technology.
He has a constant interest in Buddhism since 2008 and joined meditation retreats of the school of S.N. Goenka regularily over the past ten years. He also spent one year at the Dharma Drum Institute for Liberal Arts in Taiwan as an exchange researcher, which is an institution that facilitates research in a monastic Buddhist environment. He started his first work on forerunners of the BuddhaNexus project in spring 2018 and cooperates with Ven. Vimala Bhikkhunī and Orna Almogi for the Buddhanexus since 2019.

3.2 Ven. Vimala Bhikkhunī

Ven. Vimala Bhikkhunī studied Geophysics (MSc 1990) at the University of Utrecht, Netherlands, and was involved in developing software for the analysis of seismic data using Pascal and Fortran. They also received an MBA from the Erasmus University in Rotterdam in 1994. In 1996 they came in contact with Buddhist practice and after some 10 years working and participating in retreats at Dhamma Pajjota, the meditation center from S.N. Goenka in Belgium, they ordained as Sayalay in Burma in 2008. Eventually they ordained as Samaneri in Perth at Dhammasara monastery and was fully ordained as Bhikkhunī in Los Angeles in 2016. Ven. Vimala started Tilorien monastery in Belgium in 2018 and is currently residing there. Since 2013, they were involved with SuttaCentral.net, initially to find and code new texts and to manage volunteers and eventually with the frontend development of the new website. Ven. Vimala has a keen interest in parallels research, without which Bhikkhuni ordination would have been far more difficult and at SuttaCentral, they were also in charge of adding new parallels. In 2019 they came in contact with Sebastian Nehrdich and is currently involved in developing the website UI for the Buddhanexus project.

3.3 Supported by University of Hamburg:

3.3.1 Dr. Orna Almogi

Orna Almogi studied Tibetology (major) and Religious Studies and Psychology (minors) at the University of Hamburg (MA 1998). She received her PhD in Tibetology from the same University in 2006 (doctoral thesis: "Rong-zom-pa' s Discourses on Traditional Buddhology: A Study on the Development of the Concept of Buddhahood with Special Reference to the Controversy Surrounding the Existence of Gnosis (ye shes: jñāna) at the Stage of a Buddha"). From 1999 until 2004 she had been working for the Nepal-German Manuscript Preservation Project (NGMPP) and the Nepalese-German Manuscript Cataloguing Project (NGMCP), where she had been responsible for the Tibetan materials. From 2008 to 2011 she has been a member of the Researcher Group "Manuscript Cultures in Asia and Africa" with the subproject "The Manuscript Collections of the Ancient Tantras (rNying ma rgyud 'bum): An Examination of Variance." Currently she is working at the "Centre for the Study of Manuscript Cultures" as a leader of the subproject "C02: Doxographical Organisational Schemes in Manuscripts and Xylographs of the Collection of the Ancient Tantras."

Her research interests extend to a number of areas connected with the Tibetan religio-philosophical traditions and Tibetan Buddhist literature, particularly that of the rNying-ma school. The primary focus of her research the past years has been the concept of Buddhahood in traditional Buddhist sources, early subclassifications of Madhyamaka, and the history of transmission of the rNying ma rgyud ' bum. Another interest of her is the culture of the book in Tibet in all its variety, specifically in connection with the compilation and transmission of collections of manuscripts and xylographs containing Buddhist works.

3.3.2 Prof. Dr. Dorji Wangchuk

Dorji Wangchuk was born in 1967 in East Bhutan. After the completion of his ten-year training (1987 - 1997) in the Tibetan monastic seminary of Ngagyur Nyingma Institute at Bylakuppe, Mysore, South India, he studied classical Indology and Tibetology, with a focus on Buddhism, at the University of Hamburg, where he received his MA (2002) and PhD (2005) degrees. He worked as an academic employee at the Department of Indian and Tibetan Studies, Asia-Africa Institute, University of Hamburg, and was teaching Tibetan language as well as being engaged in research activities. Since 2009 he is Professor for Tibetan Studies at the same Department. His special field of interest lies in the intellectual history of Tibetan Buddhism and in the Tibetan Buddhist literature.

4 Methodology

4.1 Monolingual Matching

The currently applied strategy to calculate parallel passages within monolingual corpora is based on fasttext word embedding, a pooling strategy for the representation of phrases with a fixed length and Approximate Nearest Neighbor Search (ANN) to efficiently retrieve possible parallel sequences for the size of the entire corpora. Fasttext word embedding [Bojanowski et al., 2016] is calculated via a lightweight neural network that does not require GPU processing.

The data for our research is obtained from different digitalization projects of Buddhist and related material such as GRETIL, ACIP, CBETA, SuttaCentral and the Vipassana Research Institute (Mahāsaṅgīti Tipiṭaka Buddhavasse 2500).

Permission and agreement has been sought with all these projects in advance. The data of all these projects is available in permissive licensing models (either Creative Commons or similar concepts).

Currently, calculating matches for a monolingual corpus based on fasttext and ANN takes a few hours depending on the size of the canon (about 10 hours for Chinese, which is the largest corpus at the moment).

4.2 Multilingual Matching

Multilingual matching between Tibetan and Sanskrit is currently done by the use of the deep contextual embedding model XLM [Lample and Conneau, 2019], which is a crosslingual extension of the BERT model [Devlin et al., 2019]. Contrary to fasttext, this model required GPU-hardware to be trained properly. This model was trained with supervision because in the case of Sanskrit and Tibetan, enough parallel sentences are available to apply supervised training.

For finding parallels between other languages such as Pāli and Chinese, Pāli and Tibetan or Sanskrit and Chinese, not enough parallel data is available and it is therefore necessary to train XLM without supervision. Unsupervised training is a very computationally expensive step and was thus so far not yet attempted. Given enough computational resources, the quality of unsupervised training is expected to match that of supervised training. Yu-Chun Wang from the Dharma Drum Institute of Liberal Arts (DDILA) in Taiwan has also noted the fact that for ideal results on this task, a deep contextual model seems to be the most promising approach. However the DDILA was also not yet able to train such a model because of the lack of resources.

Training a XLM-model for all of the used languages is expected to take roughly one week on 10x V100 GPUs.

5 Website UI

The generated data is of vast size and the amount of recorded parallels can be overwhelming for certain texts and passages. For this reason, a user interface with flexible display options and filters to limit the amount of presented data was designed using LIT-element webcomponents based on vanilla javascript. A python backend and ArangoDB database were added to facilitate retrieval and display of data.

5.1 Text View Mode

5.1.1 Left window partition

In the text view mode, which is the default mode of the web interface, the main text is displayed on the left hand side, with each part that has a possible parallel being highlighted and clickable. The possible parallels get displayed in a column next to the main text after clicking on a highlighted passage in the main text. Very frequently different parallels are layered on top of each other, sometimes forming clusters with hundreds of parallels in close vicinity. In order to indicate how many parallels are to be encountered at a certain position of a text, a color system is used. Light blue indicates the presence of one parallel, deep blue the presence of two parallels; as the number increases, the coloring is progressing over the gradient towards strong red, which indicates the presence of five or more quotes. The intuition here is that five or more quotes indicate a certain level of popularity of a parallel and a further progressive highlighting might not be of much practical use.

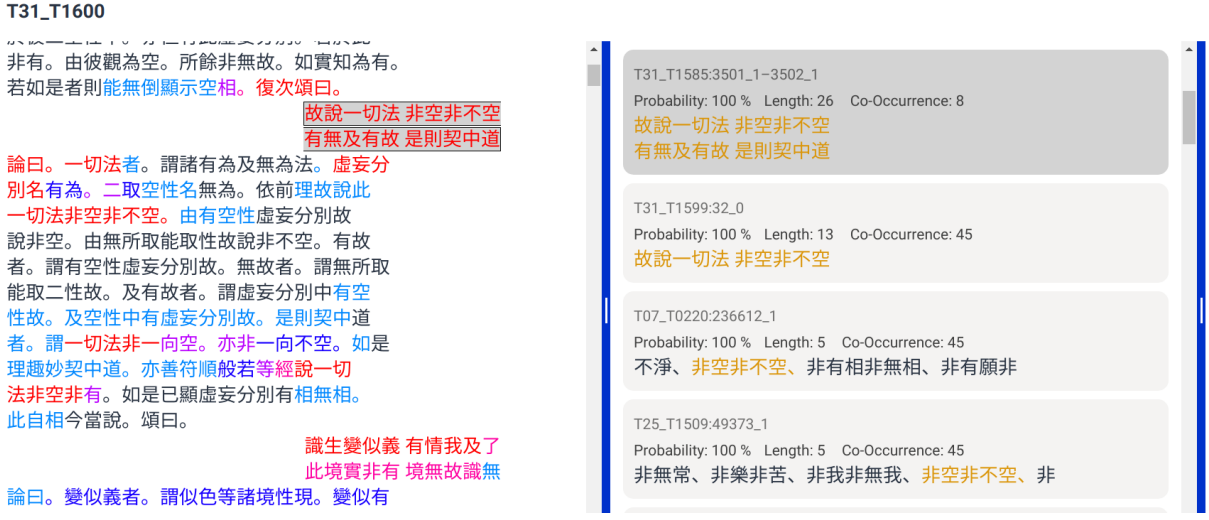


Figure 1: The website displaying T1600 (辯中邊論) as the main text in text view mode.

Figure 1 Shows how the text view view mode looks like in practice. The main text is displayed on the left side, in the column on the right side the possible parallels are displayed in descending order according to their Jaro-Winkler similarity. A full quotation

of the verse present in T1585 (成唯識論) is selected, which causes the corresponding section of the main text T1600 to be highlighted with a gray background color. This strategy of coloring the different characters according to the number of occurring parallels results in a heatmap which can give a quick impression about the location and quantity of the parallels contained within one text.[1]

5.1.2 Right window partition

After clicking on any of the parallels in the middle column in the text view mode, a third column on the right side opens up which displays the text in which the parallel that was selected in the middle column is contained as shown in figure 2.

Figure 2: Screenshot displaying T1600 with it's matching parallel in T1585 in the right column

In the third column, the currently selected parallel as well as other possible parallels that have been detected between the text in the left and the text in the right column are highlighted. When clicking on a highlighted passage on the right hand side, the corresponding parallel is displayed in the middle column and the main text on the left side is automatically scrolled to the corresponding position. In this view mode it is also possible to disable the middle column so more screen space can be devoted to the display of the two texts. This mode is providing a synoptic overview of the parallels between two texts, enabling the user to compare two texts without the need of switching windows constantly.

[1] this solution, rather by coincidence, is similar to the solution presented in (Sturgeon [2017]), where different shades of red were used to indicate layers of quotations. However, unlike (Sturgeon [2017]) in this study the differences in wordings between two passages are not visually highlighted. This is due to the fact that the data also contains examples of less verbatim quo- tations, where such a mechanical highlighting strategy becomes difficult to apply.

5.2 Table View Mode

In the table view mode it is possible to display only the parallels without the surrounding text in a table format, which then can be filtered and sorted by various criteria. This view mode is demonstrated in figure 3. Currently, it is possible to sort the parallels according to their position in the main text, by string similarity, grouped by the text in which they appear and by their length.

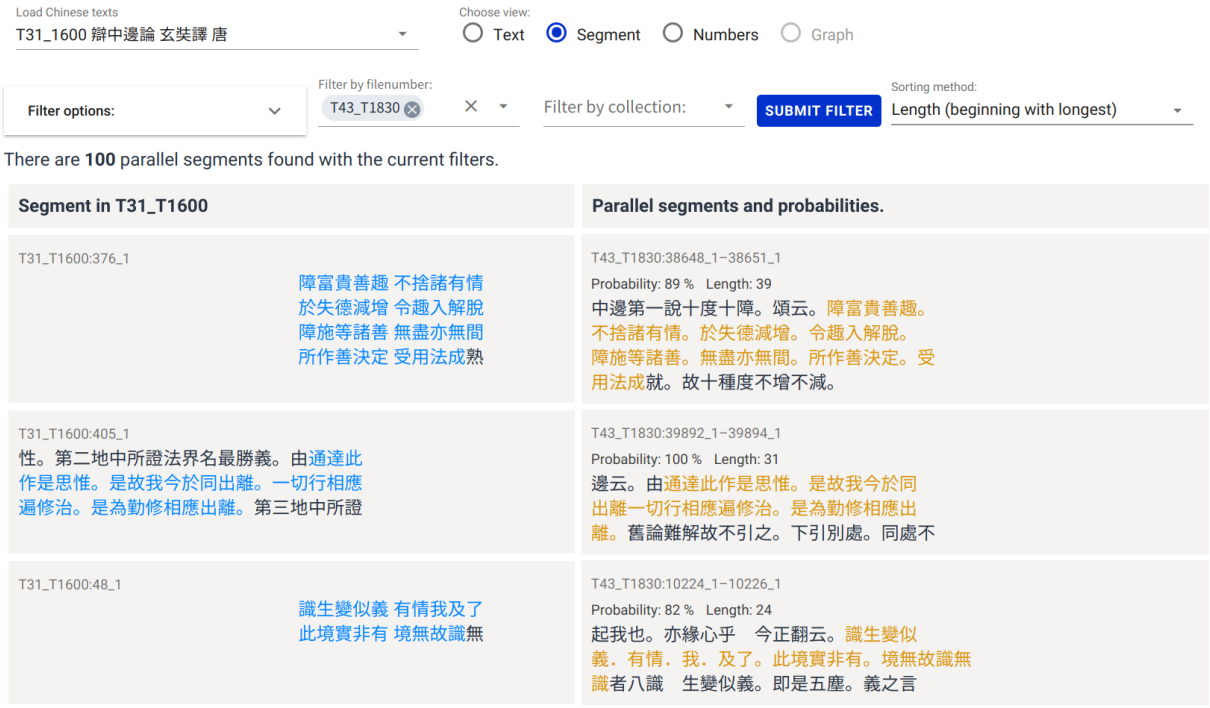
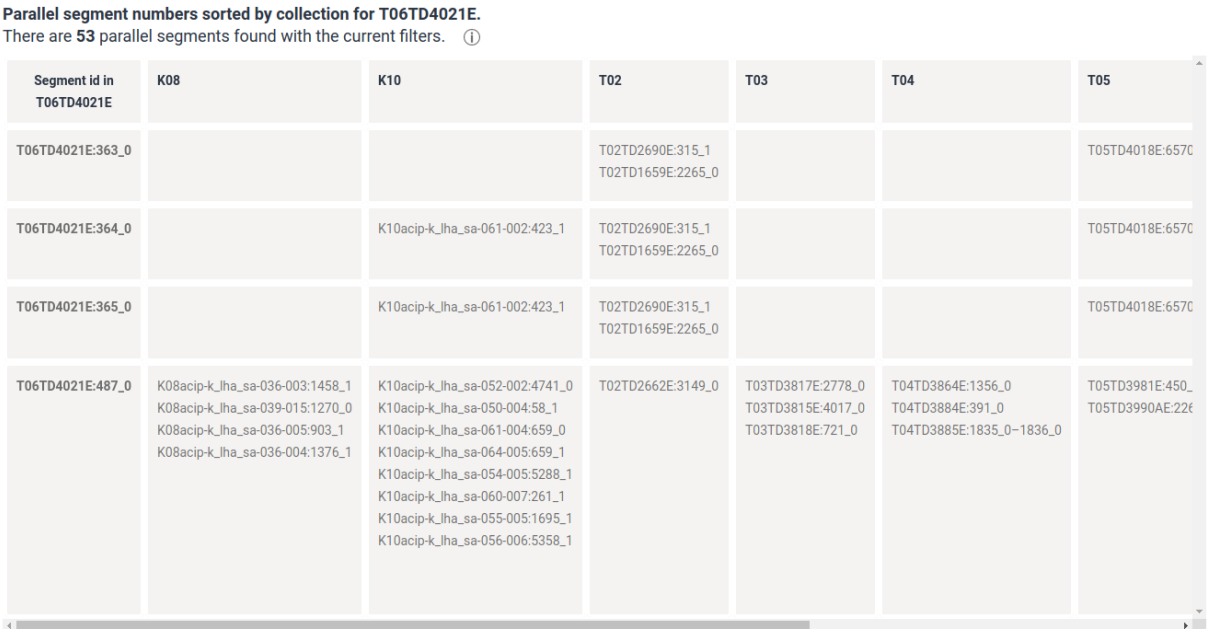


Figure 3: Screenshot of the website displaying only the parallels between T1600 and T1830 sorted by their length, beginning with the longest.

5.3 Number View Mode

The number view mode shows only the numbers of the corresponding potential parallels in a scrollable table format for a quick reference. In this mode, it is possible to get a quick overview of which texts and collections are connected with the target text. Just as in the other view modes, it is possible to narrow down or expand the results using filters. This view mode is demonstrated in figure 4, using the Tibetan text D4021.



Parallel segment numbers sorted by collection for T06TD4021E.
There are **53** parallel segments found with the current filters.

Segment id in T06TD4021E	K08	K10	T02	T03	T04	T05
T06TD4021E:363_0			T02TD2690E:315_1 T02TD1659E:2265_0			T05TD4018E:657(
T06TD4021E:364_0		K10acip-k_lha_sa-061-002:423_1	T02TD2690E:315_1 T02TD1659E:2265_0			T05TD4018E:657(
T06TD4021E:365_0		K10acip-k_lha_sa-061-002:423_1	T02TD2690E:315_1 T02TD1659E:2265_0			T05TD4018E:657(
T06TD4021E:487_0	K08acip-k_lha_sa-036-003:1458_1 K08acip-k_lha_sa-039-015:1270_0 K08acip-k_lha_sa-036-005:903_1 K08acip-k_lha_sa-036-004:1376_1	K10acip-k_lha_sa-052-002:4741_0 K10acip-k_lha_sa-050-004:58_1 K10acip-k_lha_sa-061-004:659_0 K10acip-k_lha_sa-064-005:659_1 K10acip-k_lha_sa-054-005:5288_1 K10acip-k_lha_sa-060-007:261_1 K10acip-k_lha_sa-055-005:1695_1 K10acip-k_lha_sa-056-006:5358_1	T02TD2662E:3149_0	T03TD3817E:2778_0 T03TD3815E:4017_0 T03TD3818E:721_0	T04TD3864E:1356_0 T04TD3884E:391_0 T04TD3885E:1835_0–1836_0	T05TD3981E:450_ T05TD3990AE:22(

Figure 4: Screenshot displaying number view mode for the Tibetan D4021

5.4 Filter options

In order to limit the amount of displayed parallels, various filter options are offered:

1. Similarity score
2. Length of parallel
3. Number of co-occurrences
4. Text- or collection number

(1) uses Jaro-Winkler string edit distance as a base for calculation. (2) Filters by minimum length based on the length of the parallel excluding punctuation. (3) Filters by the number of co-occurrences [1] that a parallel has. (4) Filters by either including or excluding results from selected texts or categories.

Increasing or decreasing the minimum required string similarity has the effect that more or less literal parallels are displayed. Changing the minimum required length removes or adds short parallels which might not necessarily show a strong relation to the analyzed text. Changing the number of allowed co-occurrences per parallel has a great effect when it comes to showing or hiding stock-phrases such as ' 如是我聞: 一時, 佛住王 [...]' or 'evaṃ me sutaṃ' which might be of rather little interest to the user. Filtering by text

[1] This score represents how many times a parallel is contained within other parallels. For a parallel A, it' s value is increased by one for each other parallel B m ,B m+x that contains the parallel. As a criterion, B m needs to be at least of the same length (with a tolerance of two tokens) or longer than A to affect the scoring of A.

names and categories is useful when one wants to compare a specific text against another text, set of texts or category. Being able to 'blacklist' certain texts and categories is very helpful to filter out quotes from certain sources that are not of interest to the user.

5.5 Visual charts

This section of the site presents the results in a sankey chart format. Figure 5 displays the relationships between the Tibetan Kangyur and Tengyur collections, showing clearly the level of connection between the various collections within the canon.

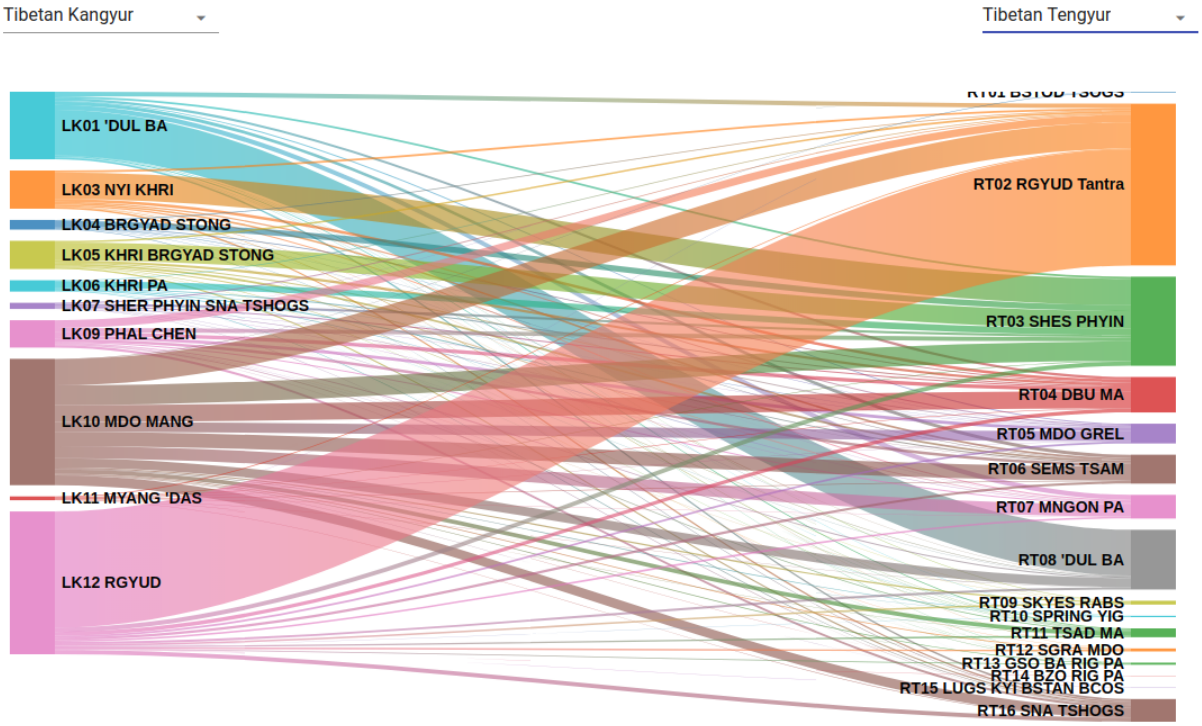


Figure 5: Screenshot displaying the sankey map comparing Tibetan Kangyur and Tengyur collections

After clicking on one of the collections, the graph changes to show a deeper level of relationships up to the level of the individual texts. When clicking on these, the text opens up in a separate tab in text view mode as outlined above.

6 Future development

In the near future, we will be able to display all Buddhist texts in Sanskrit, Chinese and Tibetan as well as the early Buddhist Pāli texts as well as their parallels in three display formats with various filters and sorting possibilities.

Further development is envisioned in the following areas:

1. Adding more texts
2. Adding cross-language parallels
3. Further development of the website UI

In the first category, later Buddhist Pāli texts can be added as well as relevant non-Buddhist texts for use in the comparison between Vedic and Jain texts and the influence on the Buddhist scriptures.

The second category is the most interesting for comparative research. Sofar we have trained neural models to just look at parallels between texts of the same language, be it Pāli, Sanskrit, Tibetan or Chinese. It is obvious that Pāli and Sanskrit are closely related and by training a deep contextual model such as XLM or BERT on both languages it will be possible to find parallels within these languages. Moreover, we have a large corpus of Sanskrit texts with Tibetan translations, so further comparison can be done between these texts and thus by inference also between Pāli and Tibetan. There also exists a large body of translations from Sanskrit and Tibetan into the Chinese language. This is a field that is little researched since the Chinese language is notoriously difficult and few Buddhist Studies scholars possess the ability to read both South Indian and Chinese languages well. For this reason, the unsupervised detection of parallels for these texts could help the Buddhist community a lot.

The main problem is that the training of large deep contextual models such as BERT or XLM requires a large increase in computer power and time. Especially the use of GPU hardware becomes necessary.

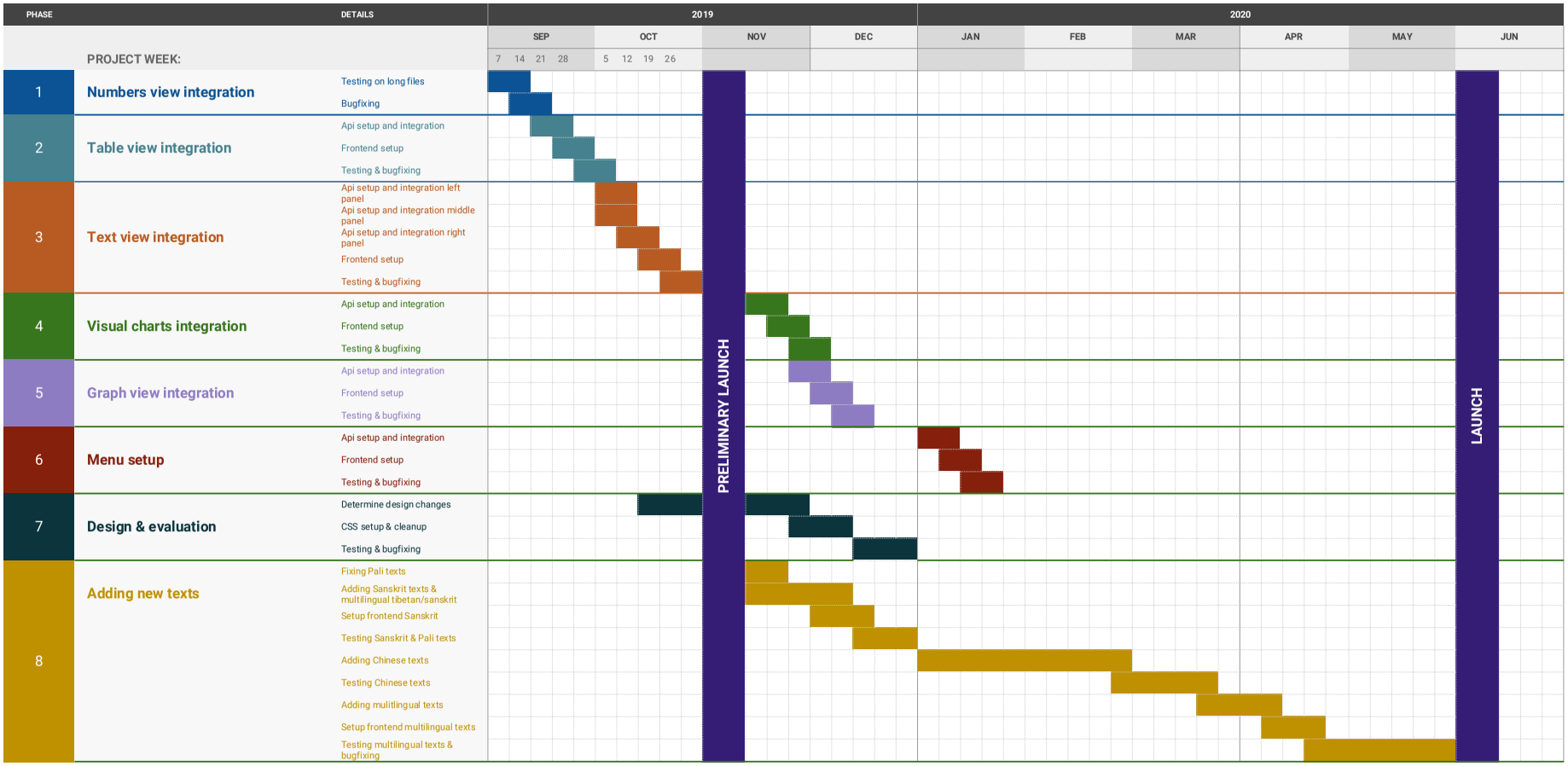
Thirdly, the ongoing development and improvement of the UI is needed to ensure a better user experience and to facilitate the ability to find parallels for comparative studies. With this we can think of adding additional graphical displays like pie-charts listing various categories of parallels for each text, but also improving on the existing filters and sorting methods. Other improvements can be in the field of transliteration and added possibilities for searching.

A roadmap for the development is shown on the next page.

BuddhaNexus Project Roadmap

PROJECT TITLE BuddhaNexus

DATE 8/9/19



7 Financial planning

In order to develop the accuracy and capabilities of the neural network and the development of the website, additional financing is needed. Up until now, we have been running on a very low budget with just one part-time paid developer, who works at lower than market rates, a basic setup of the neural network and shared hosting plan for the website.

The training of deep contextual embedding models is very necessary to develop the BuddhaNexus towards being able to cope with multilingual comparision. Currently, a single Nvidia 2080ti RTX card is used in a setup that can be extended to up to four GPUs. It is therefore desired to add additional GPUs to this setup in order to increase the capacity of the system. The addition of three more Nvidia 2080ti RTX cards will make the training of deep contextual embedding models on the current setup feasible. Since the local GPUs are not only needed for the training, but also for the inference of vector representations of a large amount of data, this is a much more cost-efficient solution than relying only on cloud-services.

Additionally to the local GPUs and optionally, some budget for renting out TPU units via the google cloud for the training of specific large models could be very helpful. The training of one BERT model for our entire dataset on multiple google cloud TPUS can be estimated to cost about 250 Dollar for each run. Another 1500 Dollar for this training would give us enough room to re-run the training with different parameters and tune the datasets to achieve optimal results.

Moreover, with the added complexity, a more robust hosting plan for the website needs to be sought out.

7.1 Annual budget estimate

The below figures represent our estimated costs for the next year.

Description	**Monthly costs (USD)**	**Annual costs (USD)**
Part-time developer	3.000	36.000
VPS Linode 4G plan	20	240
500 GB storage	50	600
3x Nvidia RTX 2080 TI GPU	-	3600
TPU training Google cloud (opt)	-	1500
Total costs		**41.940**

References

Bhikkhu Anālayo. *A Comparative Study of the Majjhima-nikāya*. Dharma Drum Academic Publisher, 2011.

Bhikkhu Anālayo. The Legality of Bhikkhunī Ordination. *Journal of Buddhist Ethics, Special 20th Anniversary Issue*, 2013.

Bhikkhu Anālayo. Bhikkhunī Ordination from Ancient India to Contemporary Sri Lanka. *Āgama Research Group*, 2018a.

Bhikkhu Anālayo. The Case for Reviving the Bhikkhunī Order by Single Ordination. *Journal of Buddhist Ethics*, 2018b.

Samuel Beal. Buddhism in China. *Society for Promoting Christian Knowledge*, 1884. URL `http://archive.org/details/buddhisminchina00commgoog`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *CoRR*, abs/1607.04606, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

Benjamin Eliot Klein, Nachum Dershowitz, Lior Wolf, Orna Almogi, and Dorji Wangchuk. Finding Inexact Quotations Within a Tibetan Buddhist Corpus. In *Digital Humanities*, pages 576–581, 2014.

Guillaume Lample and Alexis Conneau. Cross-lingual Language Model Pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Claire Maes. Dialogues With(in) the Pāli Vinaya. A Research into the Dynamicsand Dialectics of the Pāli Vinaya' s Ascetic other, with a Special Focus on the Jain Ascetic other. *Ghent University*, 2015.

Nachum Dershowitz Orna Almogi, Lena dershowitz and Lior Wolf. A Hackathon for Classical Tibetan. *Journal of Data Mining and Digital Humanities*, 2016.

A.S. Prasad and S. Rao. Citation matching in Sanskrit Corpora using Local Alignment. In *Sanskrit Computational Linguistics. Lecture Notes in Computer Science*, volume 6465, pages 124–136. Springer Berlin, 2010.

Donald Sturgeon. Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, 33(3):670–684, 11 2017.

Bhikkhu Sujato and Bhikkhu Brahmali. *The Authenticity of the Early Buddhist Texts*. Buddhist Publication Society, 2014. URL `http://ocbs.org/wp-content/uploads/2015/09/authenticity.pdf`.

Bhikkhunī Vimala. The meaning of paṇḍaka in light of the Vedic and Jain scriptures. 2019. URL `https://samita.be/en/2019/05/29/the-meaning-of-pa%e1%b9%87%e1%b8%8daka-in-light-of-the-vedic-and-jain-scriptures/`.