



Crosslinguistic Semantic Textual Similarity of Buddhist Chinese and Classical Tibetan

RESEARCH PAPER

RAFAL FELBUR 

MARIEKE MEELLEN 

PAUL VIERTHALER 

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

In this paper we present the first-ever procedure for identifying highly similar sequences of text in Chinese and Tibetan translations of Buddhist *sūtra* literature. We initially propose this procedure as an aid to scholars engaged in the philological study of Buddhist documents. We create a cross-lingual embedding space by taking the cosine similarity of average sequence vectors in order to produce unsupervised similar cross-linguistic parallel alignments at word, sentence, and even paragraph level. Initial results show that our method lays a solid foundation for the future development of a fully-fledged Information Retrieval tool for these (and potentially other) low-resource historical languages.

CORRESPONDING AUTHOR:

Marieke Meelen

University of Cambridge, GB

mm986@cam.ac.uk

KEYWORDS:

Cross-linguistic STS;
Information Retrieval;
Buddhist Chinese; Classical
Tibetan; Translation Studies

TO CITE THIS ARTICLE:

Felbur, R., Meelen, M., &
Vierthaler, P. (2022).
Crosslinguistic Semantic
Textual Similarity of Buddhist
Chinese and Classical Tibetan.
*Journal of Open Humanities
Data*, 8(1): 23, pp. 1–14. DOI:
[https://doi.org/10.5334/
johd.86](https://doi.org/10.5334/johd.86)

1 INTRODUCTION

Buddhist *sūtra* texts, which are fundamental sources for understanding the beliefs that once dominated, and largely continue to dominate, Asian societies, present formidable challenges to the modern researcher. Like oral literature, the *sūtras* are authorless and textually fluid and their content is complex and can be rather formulaic (Silk, 2020). As a result, it is often impossible to determine the ‘original’ form of a given work. The situation is complicated further by the huge volume of these documents and the linguistic diversity of their extant versions: for most, only fragments survive in the languages of their original composition (i.e. Sanskrit or other Indic languages) and all we have are their translations, mainly into Chinese and Tibetan.

In this paper we present a novel method¹ designed to help researchers tackle these challenges more effectively than has been possible to date. This is a method for automatic detection of cross-linguistic semantic textual similarity (STS) across historical Chinese and Tibetan Buddhist textual materials. It aims to enable philologists to take any passage in a Chinese Buddhist translation text, and to quickly locate Tibetan-language parallels to it anywhere in the Tibetan Buddhist canon.

The novelty of our contribution is its cross-linguistic capability for historical, low-resource and under-researched languages. Although in both of the languages in question, Buddhist Chinese and Classical Tibetan, searching for parallel passages (i.e. monolingual alignment) is possible (Klein, Dershowitz, Wolf, Almqvist, & Wangchuk, 2014; Nehrdich, 2020, as well as, in a crude but effective way, through the user interfaces of CB Reader, in both its web-based and desktop versions, or the SAT Daizōkyō Text Database), cross-linguistic semantic textual similarity and Information Retrieval (i.e. cross-linguistic ‘alignments’) in Buddhist texts have long remained an unsolved task. For a limited number of edited texts in Sanskrit and Tibetan an attempt at automatic crosslinguistic alignment has recently been made by Nehrdich (2020)² using the YASA sentence aligner.³ However, this method depends on the availability of texts in which words and sentences have been manually pre-segmented, which is not the case for the vast majority of texts we are targeting. Furthermore, being designed for Sanskrit and Tibetan, this method is not currently applicable to our highly specific Buddhist Chinese.

In short, no advanced cross-linguistic information retrieval techniques have yet been developed for any historical languages. Both the Tibetan and Buddhist Chinese texts under investigation pose particular challenges because e.g. of their different scripts, the lack of word segmentation and sentence boundaries, as well as due to the highly specific Buddhist terms and (often deliberately) obscure double meanings etc. In this paper we build on the extant work on these languages by Vierthaler and Gelein (2019) and Vierthaler (2020) (for alignment and segmentation of Buddhist Chinese) and Meelen and Hill (2017), Faggionato and Meelen (2019) and Meelen, Roux, and Hill (2021) (for segmentation and POS tagging of Old and Classical Tibetan) to develop the first-ever Buddhist Chinese–Tibetan cross-linguistic STS pipeline, creating unsupervised cross-linguistic alignments for words, sentences, and whole paragraphs of these Buddhist texts, and potentially of contemporaneous non-Buddhist materials as well. Our proposed procedure for these highly specific Buddhist Chinese or Tibetan texts will be an important asset for anyone working with under-researched and low-resource historical languages.

2 METHOD

In recent years, large digitisation projects have provided online access to huge Buddhist Chinese and Buddhist Tibetan corpora: digitized versions of over 70,000 traditional woodblock print pages in the Tibetan case, as well as, on the Chinese side, of some 80,000 typeset print pages of the modern Taishō canon, in addition to growing quantities of other canonical and extra-canonical materials. In this section we show how we developed our procedure step-by-step. Figure 1 shows the full pipeline of our proposed procedure, starting with tokenisation of the individual Chinese and Tibetan corpora and ending with the full output ranked after clustering and optimisation of cosine similarity scores of target outputs.

¹ All code available on https://github.com/vierth/buddhist_chinese_classical_tibetan (last accessed: 8 August 2022).

² <https://github.com/sebastian-nehrdich/sanskrit-tibetan-etexts> (last accessed: 8 August 2022).

³ <http://rali.iro.umontreal.ca/rali/?q=en/yasa> (last accessed: 8 August 2022).

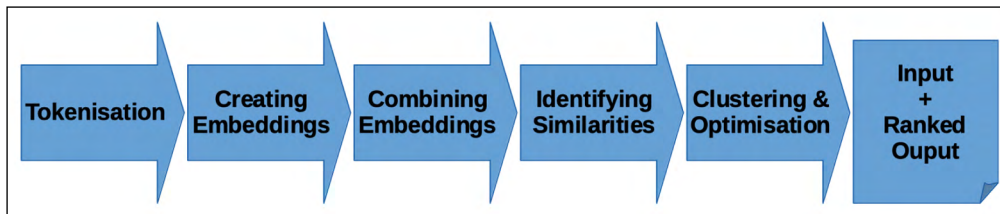


Figure 1 Pipeline for overall procedure of cross-lingual Buddhist Chinese & Classical Tibetan alignment.

2.1 TOKENISATION

While tokenisation and sentence segmentation are not usually significant hurdles when working with documents written in Western languages, in which words are delineated by white space, these are not trivial tasks for either premodern Chinese, including Buddhist Chinese, or Classical Tibetan. Neither language uses clear morphological markers or white space to indicate words, and in many cases it is not easy to even divide a text into sentences or utterances. Accordingly, before we can develop a model, we must first preprocess our corpora to include token and sentence boundaries.

Tokenisation is especially challenging on the Chinese side. For the Chinese, we use Chinese Buddhist translation texts from the Kanseki repository (Wittern, 2016).⁴ These texts are mostly provided with punctuation, which makes sentence level segmentation relatively simple. Complications arise, however, when it comes to segmentation on the word level of these materials. While much effort is currently being invested in attempts to develop tools that will segment Chinese texts into words (some of them specifically designed to segment Buddhist materials, e.g. Wang, 2020), these tools remain unusable to us, since the underlying models themselves are often not openly released, and the training data used to create them is often not available. For this reason, we had to devise our own strategy for tokenising the Chinese Buddhist translation texts. In doing so, we used three different approaches and compared their efficiency: **Word-based** tokenisation, **Character-based** tokenisation, and a **Hybrid** approach. For the first approach, we began by creating word-based embeddings on the basis of two glossaries of Buddhist terms (Inagaki, 1978; Yokoyama & Hirose, 1996). This allowed us to scan each sentence in our texts for Buddhist terms listed in these glossaries, prioritising longer sequences of characters. Once the Buddhist vocabulary was identified, the remaining sequences not found in the glossaries were parsed into words using a Classical Chinese tokeniser⁵ (see Qi, Zhang, Zhang, Bolton, & Manning, 2020). Because this word-based tokeniser introduced significant noise into our downstream tasks, we tested two other tokenisation approaches: a character-based approach that treats individual characters as tokens, and a hybrid approach that uses the word-based tokenisation described above, but which parses sequences not found in the glossaries simply as individual characters (i.e. without using the Classical Chinese tokeniser). We also enhanced the dictionaries, using more advanced glossaries by Karashima Seishi (Karashima, S., 1998, 2001, 2010) for our first test, which we will refer to as ‘Hybrid 1’, and an even further extended dictionary including the *Da zhidu lun* (Li, 2011) glossary which we will refer to as ‘Hybrid 2’.

On the Tibetan side, tokenisation was converted to a syllable-tagging and recombination task with the ACTib scripts⁶ developed by Meelen et al. (2021). As for sentence segmentation, we could use the technique developed by Meelen and Roux (2020) and optimised by Faggionato, Hill, and Meelen (2022) to create sentence boundaries in Tibetan, which is good, but not 100% accurate. Existing automatic aligners rely on sentence boundaries, so accuracy is of crucial importance. Another issue that arises in this context is the difference between the Chinese and Tibetan texts we focus on specifically, as there are often multiple Tibetan sentences corresponding to one sentence in Buddhist Chinese. For these reasons, our procedure is solely based on semantic textual similarity, thereby bypassing the need for sentence boundaries altogether.

⁴ Kanseki Repository <http://web.archive.org/web/20210418080358/http://blog.kanripo.org/> (last accessed: 8 August 2022). The texts themselves are hosted on GitHub: <https://github.com/kanripo> (last accessed: 8 August 2022) and derive from work done by the CBETA project.

⁵ As distributed through the Stanza python library. https://stanfordnlp.github.io/stanza/available_models.html (last accessed: 8 August 2022).

⁶ <https://github.com/lothelamor/actib> (last accessed: 8 August 2022).

2.2 DEVELOPING EMBEDDINGS

There are many ways to acquire useful vector representations of words, known as word embeddings, which in turn can be used to aid downstream tasks like text classification, stylometric analysis, sentiment analysis, and, crucially for us, information retrieval, and its specific application in automatic textual alignments. These ways range from the straightforward count vector models that simply track word frequency across a corpus, to more advanced algorithms like Google’s Word2Vec and Facebook’s FastText, which use neural networks to develop models that can predict words based on a set of context words (continuous bag of words, or CBOW), or that can predict context words when given an input term (skip-gram). State-of-the-art word representations can be attained using transformer-based algorithms like BERT (Devlin, Chang, Lee, & Toutanova, 2019) and ERNIE (Zhang et al., 2019), which learn word representations by predicting masked words. In our procedure, in order to balance sophistication against complexity, we have elected to use FastText to create the embeddings that will drive our approach.⁷

In addition to selecting the most adequate embedding method, it is essential to choose the most appropriate textual corpus as a basis for the embeddings. Since our goal was to create an embedding model that will be useful for the specific goal of aligning Chinese and Tibetan Buddhist translation texts, we chose a corpus that contains just the type of language that is specifically used in these texts. This is essential because the idiom and style of Buddhist texts is usually markedly different from that used in the broader language as a whole. Accordingly, for Chinese, we used Buddhist texts contained within the Kanseki repository, encompassing the Taishō edition of the Chinese Buddhist canon and a variety of supplementary materials, for a total of 4,137 documents containing 174m characters (20,775 unique). For Tibetan, we used the *sūtra* translations in the *Kangyur* (the electronic Derge version of the *eKangyur* collection), as well as electronic versions of commentarial and other texts in the entire *eTengyur* to create a corpus that is large enough to create word embeddings. The *eKangyur* consists of around 27 m tokens and the *eTengyur* consists of around 58m tokens (see Meelen & Roux, 2020); these together represent 31k unique tokens.

Because we are attempting to develop a system that is not dependent on a priori knowledge of which Chinese text ‘should’ align with which Tibetan text, we trained two separate embeddings, one on the Chinese Buddhist texts, and one on the Tibetan. That is, we took each corpus independently and fed the corpora into the FastText algorithm with the same settings, creating two independent spaces of 100 dimensions each. We then projected the resulting embeddings into the same space, creating a combined embedding space, discussed in Section 2.3.

2.3 COMBINING EMBEDDINGS

For creating the combined embedding space, we adopted the approach of Glavaš, Franco-Salvador, Ponzetto, and Rosso (2018),⁸ which is in turn an implementation of the linear translation matrix approach suggested by Mikolov, Le, and Sutskever (2013). In effect, our method takes an embedding space for each language and then relies on a bilingual glossary to create a linear projection. This projection casts the two spaces into a shared space, one which preserves internal linguistic similarity while trying to bring the glossary terms as close together as possible.⁹ Using the two embedding spaces created in the previous step, we can then apply the aforementioned Yokoyama-Hirosawa and Inagaki glossaries, which provide Chinese and Tibetan translation pairs. We then identify every pair for which we have an embedding in both Chinese and Tibetan and use all these pairs together to create a projection into a shared embedding space.

⁷ While it might be ideal to use a transformer model, there are no available models trained on Buddhist Chinese or Classical Tibetan specifically and existing models for modern Chinese or even Tibetan are not suitable for the task since the [classical] languages differ too much compared to the corresponding contemporary varieties. We therefore leave transformers for future research and use FastText rather than Word2Vec as it learns sub-word level representations of terms, which in the end creates a slightly more flexible model.

⁸ Following the method they describe in Glavaš et al. (2018), we adapted their translation matrix code (<https://bitbucket.org/gg42554/cl-sts/src/master/code/> [last accessed: 8 August 2022]) for this project.

⁹ It is possible that orthogonal constraints on the translation matrix and other normalisations could improve the resulting embedding space, as is suggested by Xing, Wang, Liu, and Lin (2015). However, this would require extensive refactoring of code and is planned for the future.

In cases where the translation glossary includes a multi-character Chinese term not found in the embedding space, but where all constituent characters are present, an embedding is derived by averaging the vectors for all the characters within the word. We can glean some insight into the quality of the new shared embedding space by looking at the cosine similarity between known translation pairs from the glossaries, as shown in Table 1.

CHINESE EMBEDDING TYPE	MOST SIMILAR	LEAST SIMILAR	MEDIAN	MEAN	STD
Character	0.9	-0.2	0.66	0.64	0.12
Hybrid1	0.9	0.19	0.66	0.65	0.11
Hybrid2	0.91	0.22	0.66	0.64	0.11
Word	0.92	0.3	0.67	0.67	0.11

The results listed in Table 1 show that the different Chinese tokenisation approaches used lead to different rates of similarity in the shared embedding space. For word-based embeddings and to a lesser extent ‘Hybrid 2’, these results also indicate that, in general, the larger the tokenisation dictionary, the higher the similarity. Although word-based tokenisation performs slightly better at this initial step, it does not work as well as the hybrid approaches for our downstream tasks, as shown in Section 3 below.

As a further sanity check, we visualised some embeddings to see whether similar words indeed exist in close proximity to each other. The resulting visualisation is presented in Figure 2, which demonstrates this for some sample vectors for animals, directions, numbers, and seasons.¹⁰ All these categories are nicely clustered together as expected. The only outlier is Tibetan *nya sha*, which was labelled as an animal, but it actually means ‘fish (as) meat’, i.e. fish that will be eaten. It is therefore not entirely surprising that it would be farther away from the rest of the animal words, which are not used as food. Figure 3 is a zoomed-in view of the “animal” cluster from Figure 2, with English translations for the vectors. This zoomed-in view shows that Tibetan and Chinese equivalents are placed relatively close together, as expected.

Table 1 Summary of cosine similarity scores of Tibetan-Chinese glossary pairs within the new embedding spaces according to Chinese tokenisation method. Shows the highest scoring pair, lowest scoring pair, and some descriptive statistics. Higher scores with lower standard deviation indicate a more accurate embedding space.

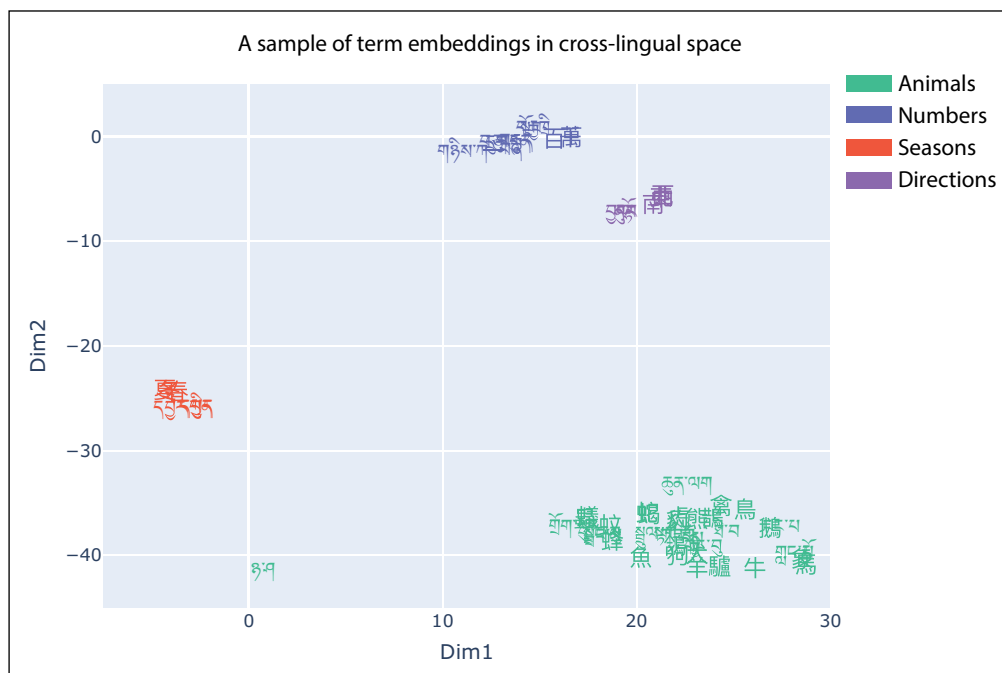


Figure 2 A sample of embeddings selected from the cross-lingual Tibetan-Chinese space. This includes a selection of animal, numerical, seasonal, and directional words.

¹⁰ The embeddings exist in 100-dimensional space, and we have used tSNE to reduce the dimensionality in order to visualize the relationships. Please note that this preserves local similarity but obscures global differences.

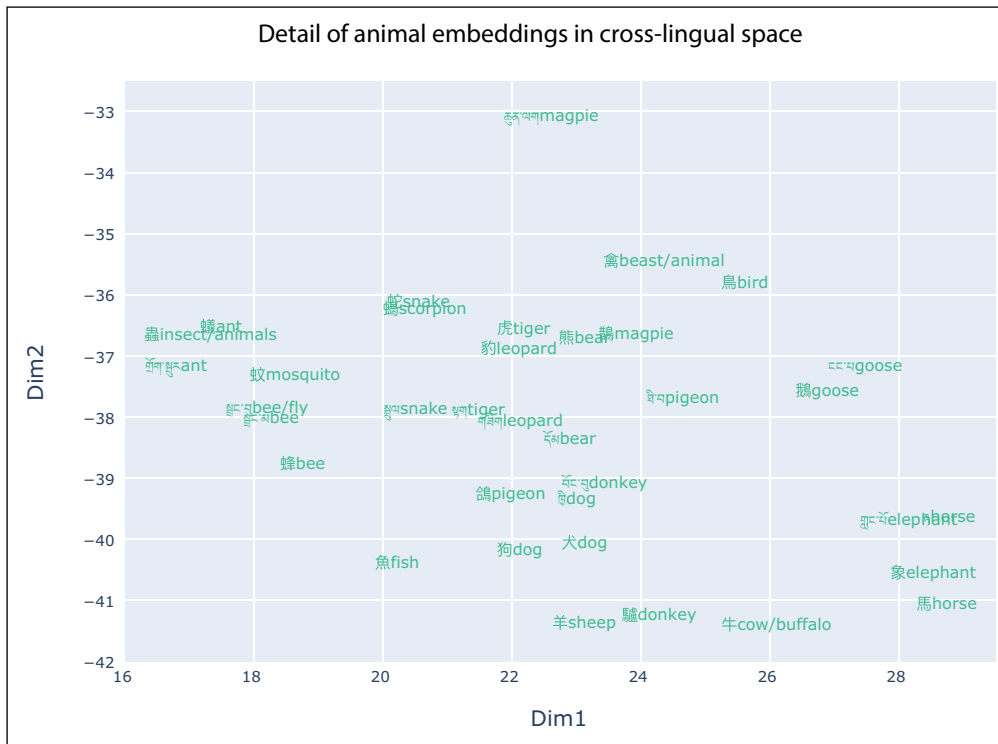


Figure 3 A zoomed in detail of some of the animal words from the cross-lingual embedding space shown in Figure 1, including English translations.

There is room for improvement in the quality of the shared embedding space, but the real test is the space’s utility for the task at hand, which is identifying textual sequences with similar semantic meaning across languages.

2.4 IDENTIFYING SIMILAR SEQUENCES

With the combined and checked word embeddings in hand, we are ready to apply our procedure to what has been the main goal all along, i.e. searching for sequences of text in both Tibetan and Chinese that carry similar meanings. In this pilot study we use as our source texts three Chinese *sūtras* from the *Maharatnakāra* (MRK) collection, which have been manually divided into sections.¹¹ We then tokenise each section into either characters, words, or Buddhist terms (as in our two Hybrid embedding approaches). Then we fetch the vector for each token in the section and average the vectors together to create a vector representation of the entire section. We then define Tibetan texts parallel to the Chinese *sūtras* as the ‘target.’ We divided this target text into sections as well: we did this by using a sliding window of text from a Tibetan candidate document, the length of which window is based on the length of the Chinese section, adjusted by some length factor. We then calculate the cosine similarity between the Chinese section in question and all Tibetan sections. Finally, we have the system rank the suggested results based on highest cosine similarity of the combined embeddings, and report the results. The highest-scoring sections are likely to have similar meaning.

2.5 PARAMETER SETTINGS, CLUSTERING & OPTIMISATION

When we looked closely at the generated results, we found that we could improve their quality by optimising the test parameter settings, specifically the length of the Tibetan search window. One reason why such optimisation proved advantageous may be the fact that the Tibetan text is always more elaborate than the Chinese, meaning that for every Chinese passage of *n* tokens, the parallel Tibetan will include roughly 50% more tokens. In order to accommodate this difference, we extended the Tibetan search window by a fixed rate (proportional rates proved inefficient, hence we rejected them), in order to ensure the results would cover the entire Chinese input. Significantly shorter Chinese input phrases required a different rate still, since they tend to be proportionally even longer in Tibetan than are longer Chinese phrases. In Section 3.3 we discuss the parameter options to optimise results for different input lengths.

¹¹ Please see section 3.1 of the Alignment Scoring Manual Handy and Meelen (2022): <https://zenodo.org/record/6782150#.Yu5UicbA5pQ> (last accessed: 8 August 2022).

2. Genghe shang youpoyi hui 恒河上優婆夷會 (T310 [31]), from the early 8th century, and the *Gang ga'i mchog gis zhus pa* (D75), roughly a century later—translations of the **Gagottara-paripṛcchā* (henceforth 'Text 2')
3. Shande tianzi hui 善德天子會 (T310 [35]), from the early 8th century, and the *Sangs rgyas kyi yul bsam gyis mi khyab pa bstan pa* (D79), roughly a century later, translations of the **Acintyabuddhaviṛaya-nirdea* (henceforth 'Text 3')

All three texts survive in their entirety only in the Chinese and Tibetan translations, with no known complete Sanskrit or other Indic language versions, they also differ in many ways. One of these ways is especially consequential for our results: Text 1 is mainly narrative, and consists of stories that illustrate moral points, while the latter two are more abstract-philosophical, and contain a narrower set of more technical metaphysical concepts. We weigh the implications of this difference in Section 3.1. For this pilot study, we use the Chinese sentence as input and let the system find Tibetan equivalents that are semantically as similar as possible, ideally capturing the exact target that the philologists identified in the gold standards.

3 RESULTS

In this section we present the results of using the different methods of creating Buddhist Chinese embeddings described above in Section 2.2. As these embeddings were not yet optimised, a comparison of the effectiveness of the different methods when applied to each of our three texts can give us further insight into which method is best suited for the task at hand. Tibetan word embeddings were already optimised (see Meelen, 2022), including the addition of specialist (Buddhist) terms. In the remainder of this section, we first present the aggregate results per text, and then zoom in on select 'interesting' results in order to discuss how they may have been affected by the different embedding methods used, as well as by the unique characteristics of the inputs qua vocabulary, style, and grammar.

3.1 RESULTS PER TEXT

Table 2 shows what percentage of outputs for each text was ranked first or in the top 5/10/15; a separate listing is given for each of the four Chinese embedding methods. Ideally, the system would automatically rank the exact Tibetan target 'first,' so that philologists can instantly find the Tibetan equivalents of the Chinese inputs they are looking for. However, since this is not likely to happen always, or even frequently, a dedicated user interface for philologists should display the top 5/10/15 (depending on preference), which the user would then go through by hand. For this reason, we list not only the percentage of target alignments that were automatically ranked first, but also those where the target was found in the top 5/10/15, as well as the average ranking of the target result and the number of cases in which the target alignment in Tibetan was not found in the top 15 (i.e. ranked 'zero').¹⁵

Table 2 shows that the results for Text 2 are always better than those for Text 1 and Text 3: the average rank is higher (ranging from 1.24 with Character embeddings to 2.48 for Word embeddings); there are no zero results with any of the embedding methods used; and it has the highest percentage of perfectly matched target results in the top ranks (with almost all targets found in the top 5 with any embedding method). In practice, this means that philologists inputting Chinese passages from Text 2 are very likely to be presented with exact Tibetan targets (i.e. semantically similar passages or target alignments as identified manually by philologists) when searching the entire text. The results for Texts 1 and 3 are not as outstanding, but are still very good, with average ranking between 3.3–4.6 (as against 1.2–2.4 for Text 2). Still, for both Text 1 and 3, we came across some problematic cases in which the system found no Tibetan equivalent in the top 15 of the ranked results, as well as ones in which the Character-embedding method yielded zero results. These problematic cases are particularly interesting to us: by looking at what went wrong we may understand how to improve our system. One example of such a problematic case with a 'zero result' is Alignment 20 in Text 1 (T1.A20), as shown in example 1. The highest-ranked match for this input based on Character-embeddings is shown in 1c.

¹⁵ Note that 'zero' could mean, for example, that the target was ranked 16th, which is not such a bad result. However, if a targeted interface for philologists only displays the top 15 results than anything ranked lower could not be considered.

3.2 THE EFFECT OF DIFFERENT CHINESE EMBEDDING METHODS

One variable parameter in our results consists of the different methods of creating Chinese embeddings, as described in Section 2.2 above. ‘Hybrid 2’ embeddings are essentially ‘Hybrid 1’ embeddings extended with additional Buddhist terms from the *Da zhidu lun* glossary. Therefore, whenever ‘Hybrid 2’ embeddings yielded better results for certain alignments than did Hybrid-1-embeddings, we expect this is because these alignments contain terminology that is only found in the *Da zhidu lun* glossary. One clear example of this is Alignment 21 in Text 1 (ranked first with Hybrid 2, but sixth with Hybrid 1). This alignment contains 如來 ‘Tathāgata’ which, among the glossaries we used, is only found in the *Da zhidu lun* glossary, and *not* in the Karashima lists upon which the ‘Hybrid 1’ embeddings were based. This example is shown in 2, along with its Tibetan target:

- (2) (a) Input: 佛告須賴言族姓子有四法具足受持若族姓子族姓女見如來者審見善見 [T1.A21]
 “The Buddha said to Sūrata: “Son of good family! There are four things that, if a son or daughter of good family should possess them and uphold them, will allow them to see the Tathāgata, to completely see him, to see him well.”
- (b) Target བོན་ལྷན་འདས་ཀྱིས་བཀའ་སྤྲུལ་པ་དེས་པ་རིགས་ཀྱི་བྱ་ཚེས་བཞི་དང་ལྷན་ན་དེ་བཞིན་གཤེགས་པ་ལེགས་པར་མཐོང་བ་ཡིན་ནོ།།
 “The Lord said: “Sūrata! A son of good family by possessing four things sees the Tathāgata.”

Figure 5 shows the results (up to top-10 ranks) from Table 1 in a chart organised by type of Chinese embedding. Though this pattern of superiority of Hybrid-2 over Hybrid-1 embeddings is expected and indeed quite common in our results, we also found one counterexample to it, namely the short Alignment 11 in Text 2 (shown in 3). In this case, Hybrid-1 performed best (target ranked 5th), while Hybrid-2-embeddings had the target ranked 11th. This is unexpected, because the input contains 攀緣 ‘in accordance with conditions,’ which is found in the Karashima lists, but not in the *Da zhidu lun* glossary. This means that this particular term was included in both Hybrid-1 and Hybrid-2 embeddings and there must be another, as of yet unidentified, reason why the Hybrid-1 embeddings yield a better result here.

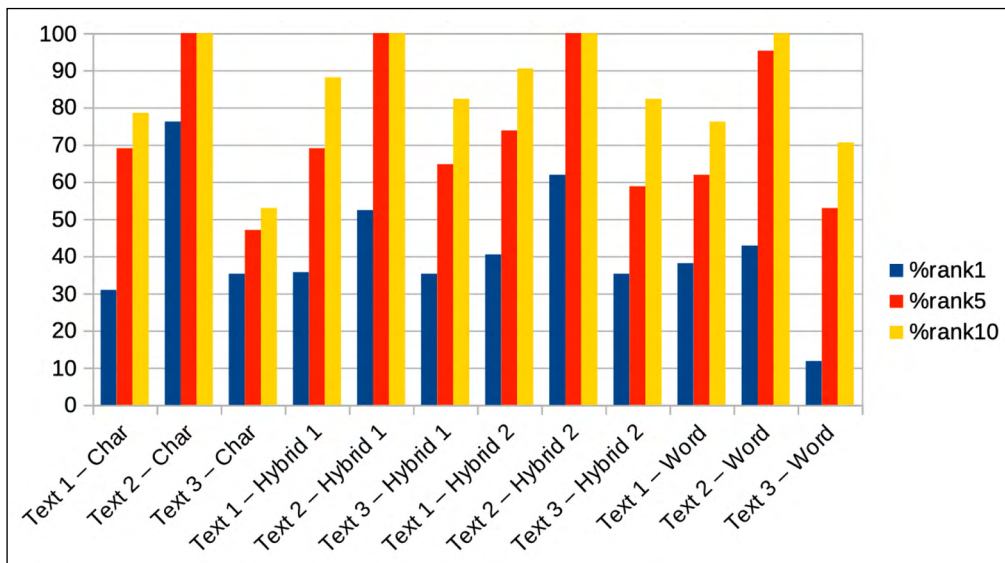


Figure 5 Top-ranked results for each Chinese embedding method by text.

- (3) (a) Input: 恒河上言: 如來豈以有所攀緣, 而致斯問 [T2.A11]
 “Gaṅgotarā said: “Surely the Tathāgata does not ask this question because he has [something] which could be referentially objectified, does he?”
- (b) གསོལ་པ་ཅི་བོའོ་ལྷན་འདས་ཀྱིས་དེ་མིགས་པ་དང་བཅས་པའི་སྤྲུལ་སྤྲུལ་ལེགས་པར།
 “She asked: ‘Does the Blessed One say this because the question is endowed with a referential object?’ ”

Another category of results consists of those in which Character embeddings performed best. In these cases we expect to be dealing with inputs that contain few multi-character proper nouns and specialist Buddhist terms, which is indeed the usual pattern. Nonetheless, we found a number of exceptions, e.g. Alignment 12 of Text 2. The input here does contain some

technical multi-character terms (世尊 ‘Bhagavan’, 能知 ‘knowable’ and 能得 ‘graspable’). This might lead one to expect that Hybrid embeddings would perform best. This, however, is not the case: Character embeddings proved superior. The reason for this is not entirely clear, although it may have something to do with the fact that all the terms listed above also make sense if they are split up into single characters (‘world-honour,’ ‘able-know,’ ‘able-grasp’ respectively). A similar explanation can be offered for Alignment 33 in Text 1 (ranked 14th with Char vs 24/35th with Hybrid-1/2 embeddings), so this phenomenon does not appear to be text-specific. Other cases of better performance of Char-embeddings include:

- Text 1: Alignments 27 (ranked 2nd with Char vs 4/7th with Hybrid-1/Word) and 32 (ranked 2nd with Char vs 7th in Hybrid-1 and Word);
- Text 3: Alignments 12 and 15 (both ranked 1st with Char vs 3rd/4th with Hybrid-1/Word), and also 7 (ranked 8th with Char vs 19/36th with Hybrid-1/Word), 13 (ranked 1st with Char vs 3rd/6th with Hybrid-1/Hybrid-2) and 14 (ranked 1st with Char vs 6th/3rd with Hybrid/Word).

Some of these cases are especially difficult to interpret. For instance, Alignments 27 and 32 of Text 1 contain multi-character proper names, like 波斯匿 ‘Prasenajit.’ These are expected to pose difficulties for Char-embeddings, for, while they can be read as individual characters, this would result in jibberish: 波-斯-匿 is ‘wave-this-conceal.’ Similarly, Alignments 12 and 15 of Text 3 contain the long phonetic transcription of a Sanskrit name, 文殊師利 ‘Mañjuśrī’, which, if read as individual characters, would make little sense (‘literature-distinct-teacher-benefit’), and which therefore can only be ‘misleading’ for alignment purposes. As for Alignments 7, 13 and 14 of Text 3, the fact that Char-embeddings performed best may be related to the fact that the inputs are extremely short, consisting only of max 7 characters (see Section 3.3). These types of unexpected examples form a minority, however, and while further analysis of such cases is a desideratum, it can only be performed at a later stage, using a larger dataset. Overall, we can conclude that in the three texts we have investigated for this pilot study, the enhanced Hybrid-2 embeddings generally perform better for alignments that contain specialist Buddhist terminology, and that in the absence of such terminology, Char embeddings perform equally well or better, which is exactly what we expected.

3.3 THE EFFECT OF INPUT LENGTH

Some texts exhibit a relatively high degree of repetition of short, generic clauses. This presents a challenge for the alignment procedure as it is unclear which passage is the target identified by philologists if multiple passages with very similar meanings are present in the text. This problem pertains especially to Texts 2 and 3, where aligned segments are relatively short. Especially in Text 3, we have short recurring inputs like ‘X said’, e.g. Alignment 7 with input 諸比丘言 ‘all the monks said’ (ranked 8th) or Alignment 11 with input 汝等應知 ‘you all should know’ (partial match ranked 12th, because the Tibetan target contains an additional vocative ‘friends!’ རྩོགས་ལོ་དག་). While short inputs pose challenges to our procedure, very long inputs usually lead to good results. One example of this is Alignment 10 in Text 3, which contains a very long tantric incantation. As input length clearly affects our results we included the option of adjusting several minor parameters in order to improve the results of variable input lengths as follows:

- The proportion by which to adjust long phrases (as they are generally longer in Tibetan than in Chinese);
- The proportion by which to adjust short phrases (as short Chinese phrases are often significantly longer in Tibetan);
- The length threshold for what constitutes a “short phrase”;
- How far apart results can be clustered together in the final analysis (results within n words of each other get reported as a single result).

Of all these minor parameters, we observed that the greatest impact on the results could be generated by adjusting the parameters for long and short phrases. This is most clearly seen in examples from Text 2. Text 2 has the longest input alignments in general (with a median length of 21 characters; Text 1 has a median of 12.5, and Text 3 a median of 10), and Alignments 4, 6 and 15 of this text demonstrate the importance of adjustments according to phrase length.

With the new settings of a 50% increased adjustment length for short phrases from Chinese to Tibetan, instead of the much longer, 130%/140%/160% options we tested before, the rankings of results improved significantly (ranking improvement of 14 → 3 for Alignment 4; 11 → 2 for Alignment 6 and 6 → 2 for Alignment 15). For some alignments, however, reducing the phrasal length settings resulted not in higher rankings, but in lower ones, although these differences were much smaller than the gains observed for the other alignments (ranking 1 → 3 for Alignment 1; 1 → 2 for Alignments 10 and 17). Our current corpora are too small to justify any generalisations here. However, based on the results of our pilot study we can conclude that it is certainly worthwhile to allow for the adjustment of additional parameters, and that the most optimal settings are a function of input length and content (i.e. how common the key terms of the input are and how often they reoccur in the text).

3.4 THE EFFECT OF MANUAL ANNOTATION

One limitation of the current pilot study lies in the manual annotation: the alignment scores of each of our texts were added by three different philologists. For Text 1, we asked the same annotator to provide scores for his alignments on two different occasions, at least 1 year apart. We observed that some alignments he had at first identified as perfect equivalents (score 5), were scored 4 in the second round of manual annotation. This shows the important issue of subjectivity in manual scoring. This issue can only be effectively addressed through rigorous and repeated large-scale inter-annotator agreement checks. However, at present such checks are almost impossible for logistical reasons: they require time- and labour-intensive participation of multiple philologists who are experts in both classical languages as well as in the highly complex Buddhist content of the texts, and such participation is extremely difficult to secure. In view of this, while in future work we hope to include at least partial inter-annotator agreement scores, in the present pilot study we had to settle for the sub-optimal single-scored method.

3.5 MEASURING THE SUCCESS OF ACTUAL SEMANTIC SIMILARITY

Alignment 20 from Text 1 illustrated in example 1 above already showed that frequently-occurring key terms could have a negative impact on ranking: whenever key terms occur repeatedly, the chances of multiple outputs with high cosine similarity scores increase, and the chances of a high ranking for just one specific output (corresponding to the target) decrease. In this section we briefly demonstrate that although lower rankings may initially indicate a bad result, this does not necessarily mean that our system is performing badly: high-ranked outputs may not be the exact target (as identified by expert philologists in our gold standard), but they could still convey the same or a very similar meaning. We can see this in particular for alignments where the average cosine similarity results are low. Consider, for example, Alignment 8 from text 1:

- (4) (a) Input: 栴檀香之塗我觀其如是。 [T1.A8]
 “The sandalwood perfumes with which you anoint yourselves, I view them like this”
 (b) Target: ལྷོང་ཡང་མེ་ཉེག་དང་འབྲུག་པ་དང་ལྷོས་དང་མེ་ཉེག་ཐོང་དང་ལྷོག་པ་དང་ལྷོའི་ཚོན་དཔྲུག་གི་ཕྱུ་མ་
 “Your bodies anointed with sandalwood paste, and your moist skin, I view as similar to [something] very filthy”
- (5) Output 1: ལྷོང་ཡང་མེ་ཉེག་དང་འབྲུག་པ་དང་ལྷོས་དང་མེ་ཉེག་ཐོང་དང་ལྷོག་པ་དང་ལྷོའི་ཚོན་དཔྲུག་གི་ཕྱུ་མ་
 “[...] thousand, flowers, incense, perfume, garlands, ointments, and gods’ sandalwood paste [...]”

The average cosine similarity of this alignment with the Hybrid-2 embeddings is only 0.80 (standard deviation of 0.02). The target is ranked 2nd with a cosine similarity of 0.85807, but the highest-ranked output shown in (5) scored 0.88005. The color coding shows that this output contains two of the key terms present in the Chinese input. Since the Chinese input is relatively short, overlap in two such highly specific terms can yield relatively high similarity and thus lead to a highly-ranked result.

4 CONCLUSION

In this paper we presented the first-ever procedure for identifying highly similar sequences of text in Chinese and Tibetan translations of Buddhist sūtra literature. Our pilot study is based on creating a cross-lingual embedding space by taking the cosine similarity of average sequence

vectors in order to produce unsupervised similar cross-linguistic parallel alignments at word, sentence, and even paragraph level. We evaluate the results of the pilot study comparing three Buddhist texts that are manually aligned by expert philologists. Initial results show that our method lays a solid foundation for the future development of a fully-fledged Information Retrieval tool for these (and potentially other) low-resource, historical languages. We will address questions of scalability and of further philological use cases in future research.

SUPPLEMENTARY FILES

Supplementary materials are deposited on Zenodo:

- Alignment Scoring Manual (Handy & Meelen, 2022): <https://doi.org/10.5281/zenodo.6782150>
- Buddhist Chinese embeddings (Vierthaler, 2022): <https://doi.org/10.5281/zenodo.6782932>
- Classical Tibetan embeddings (Meelen, 2022): <https://doi.org/10.5281/zenodo.6782247>

ACKNOWLEDGEMENTS

Thanks to the British Academy and to the European Research Council (ERC) for financial support, as well as to Gregory Forgues & Jonathan A. Silk for manual alignments.

FUNDING INFORMATION

This work was supported by the European Research Council (ERC) under the Horizon 2020 program (Advanced Grant agreement No 741884).


COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Rafal Felbur  orcid.org/0000-0002-0555-9992
Leiden University, NL

Marieke Meelen  orcid.org/0000-0003-0395-8372
University of Cambridge, GB

Paul Vierthaler  orcid.org/0000-0002-2135-9499
College of William and Mary, US

REFERENCES

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.** (2019, June). 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)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N19-1423> (last accessed: 8 August 2022) DOI: <https://doi.org/10.18653/v1/N19-1423>
- Faggionato, C., Hill, N., & Meelen, M.** (2022, June). NLP Pipeline for Annotating (Endangered) Tibetan and Newar Varieties. In *Proceedings of The Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference* (p. 1–6). Marseille, France: European Language Resources Association.
- Faggionato, C., & Meelen, M.** (2019). Developing the Old Tibetan treebank. In N. T. Angelova Mitkov (Ed.), *Proceedings of Recent Advances in Natural Language Processing* (p. 304–312). Varna: Incoma. DOI: https://doi.org/10.26615/978-954-452-056-4_035
- Glavaš, G., Franco-Salvador, M., Ponzetto, S. P., & Rosso, P.** (2018). A resource-light method for cross-lingual semantic textual similarity. *Knowledge-based systems*, 143, 1–9. DOI: <https://doi.org/10.1016/j.knosys.2017.11.041>
- Handy, C., & Meelen, M.** (2022, June). *MRK alignment scoring guidelines*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.6782150> (last accessed: 8 August 2022).
- Inagaki, H.** (1978). *Index to the Larger Sukhavatvyūha-sūtra. A Tibetan Glossary with Sanskrit and Tibetan Equivalents*. Tokyo: Nagata Bunshudo.

- Karashima, S.** (1998). *A Glossary of Dharmarakṣa's Translation of the Lotus Sutra: Zheng fahua jing ci dian*. Tokyo: The International Research Institute for Advanced Buddhology, Soka University.
- Karashima, S.** (2001). *A Glossary of Kumārajīva's Translation of the Lotus Sutra: Myōhō Rengekyō shiten*. Tokyo: The International Research Institute for Advanced Buddhology, Soka University.
- Karashima, S.** (2010). *A Glossary of Lokakṣema's Translation of the Aśvāsahasrikā Prajñāparamitā*. Tokyo: The International Research Institute for Advanced Buddhology, Soka University.
- Klein, B. E., Dershowitz, N., Wolf, L., Almogi, O., & Wangchuk, D.** (2014). Finding Inexact Quotations Within a Tibetan Buddhist Corpus. In *9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2014, Lausanne, Switzerland, 8–12 July 2014, Conference Abstracts*.
- Li, Q.** (2011). *Da zhidu lun cidian* 大智度論辭典. Electronic resource. Retrieved from <https://www.dropbox.com/s/ocsagb529k3e70v/dzdl.bgl?dl=0> (last accessed: 1 June 2021).
- Meelen, M.** (2022). Tibetan language models: from distributional semantics to facilitating Tibetan NLP. *Accepted submission to IATS 2022*.
- Meelen, M., & Hill, N.** (2017). Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2). DOI: <https://doi.org/10.5070/H916234501>
- Meelen, M., & Roux, É.** (2020). Meta-dating the parsed corpus of Tibetan (PACTib). In *Proceedings of the 19th Workshop on Treebanks and Linguistic Theories* (pp. 31–42). DOI: <https://doi.org/10.18653/v1/2020.tlt-1.3>
- Meelen, M., Roux, É., & Hill, N.** (2021). Optimisation of the largest annotated Tibetan corpus combining rule-based, memory-based, and deep-learning methods. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1), 1–11. DOI: <https://doi.org/10.1145/3409488>
- Mikolov, T., Le, Q. V., & Sutskever, I.** (2013). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168. Retrieved from <http://arxiv.org/abs/1309.4168> (last accessed: 8 August 2022).
- Nehrdich, S.** (2020). A method for the calculation of parallel passages for Buddhist Chinese sources based on million-scale nearest neighbor search. *Journal of the Japanese Association for Digital Humanities*, 5(2), 132–153. DOI: https://doi.org/10.17928/jjadh.5.2_132
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D.** (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*. DOI: <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Silk, J. A.** (2020). Tekisuto sokei no nai kōtei: Bukkyō kyōten to yudayakyō rabi bunken kenkyū ni okeru honbun hihan, soshite ‘Hirakareta bunkengaku’ dejitaru hyūmanitizu purojekuto” テキスト祖型のない校訂: 佛教經典とユダヤ教ラビ文獻研究における本文批評、そして「開かれた文獻學」デジタルヒューマニティーズプロジェクト [Editing without an Ur-text: Buddhist Sūtras, Rabbinic Text Criticism, and the Open Philology Digital Humanities Project]. *Tōyō no Shisō to Shākūyō* 東洋の思想と宗教, 37, 22–58.
- Vierthaler, P.** (2020). *A Simple Dictionary-Based Tokenizer for Classical Chinese Text*. Retrieved from https://github.com/vierth/dictionary_parser (last accessed: 8 August 2022).
- Vierthaler, P.** (2022, June). *Buddhist Chinese Word Embeddings*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.6782932> (last accessed: 8 August 2022).
- Vierthaler, P., & Gelein, M.** (2019, 3 22). A blast-based, language-agnostic text reuse algorithm with a markus implementation and sequence alignment optimized for large Chinese corpora. *Journal of Cultural Analytics*, 4(2). DOI: <https://doi.org/10.22148/16.034>
- Wang, Y.-C.** (2020). Word segmentation for Classical Chinese Buddhist literature. *Journal of the Japanese Association for Digital Humanities*, 5(2), 154–172. DOI: https://doi.org/10.17928/jjadh.5.2_154
- Wittern, C.** (2016). The Kanseki repository: A new online resource for Chinese textual studies. *Digital Scholarship in History and the Humanities*.
- Xing, C., Wang, D., Liu, C., & Lin, Y.** (2015, May–June). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies* (pp. 1006–1011). Denver, Colorado: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N15-1104> (last accessed: 8 August 2022).
- Yokoyama, K., & Hirokawa, T.** (1996). *Index to the Yogācārabhūmi, Chinese-Sanskrit-Tibetan: 漢梵藏对照瑜伽師地論總索引*. Tokyo: Sankibō Busshorin.

Felbur et al.
*Journal of Open
 Humanities Data*
 DOI: 10.5334/johd.86

TO CITE THIS ARTICLE:

Felbur, R., Meelen, M., & Vierthaler, P. (2022). Crosslinguistic Semantic Textual Similarity of Buddhist Chinese and Classical Tibetan. *Journal of Open Humanities Data*, 8(1): 23, pp. 1–14. DOI: <https://doi.org/10.5334/johd.86>

Published: 04 October 2022

COPYRIGHT:

© 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.