Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6728–6734
Marseille, 20-25 June 2022

SansTib, a Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model

Sebastian Nehrdich
Institute for Language and Information,Heinrich-Heine-Universität Düsseldorf;
Khyentse Center for Tibetan Buddhist Textual Scholarship, Universität Hamburg

nehrdich@uni-duesseldorf.de

Abstract

This paper presents the development of SansTib, a Sanskrit - Classical Tibetan parallel corpus automatically aligned on sentence-level, and a bilingual sentence embedding model. The corpus has a size of about 317,289 sentence pairs and 14,420,771 tokens and thereby is a considerable improvement over previous resources for these two languages. The data is incorporated into the BuddhaNexus database to make it accessible to a larger audience. It also presents a gold evaluation dataset and assesses the quality of the automatic alignment.

Keywords: Sanskrit, Tibetan, Sentence Alignment, Bilingual Sentence Representation, Parallel Corpus

1. Introduction

Translations of Buddhist texts have been of central importance to the formation of the Tibetan Buddhist literary tradition. The majority of these translations are from Indian Buddhist texts, of which many have been composed in Sanskrit or Middle Indic languages. The study of these Tibetan translations is not only very important for the study of the Tibetan Buddhist tradition, but also for the study of Buddhism in general, since many Indian Buddhist texts did not survive in their original language. Digitally available parallel data therefore can facilitate the research on Tibetan translations of Sanskrit texts substantially. While considerable effort has been undertaken by the community of scholars to create large bilingual dictionaries for the two languages,[1] a comprehensive parallel corpus with sentence alignment is a desideratum that this paper seeks to address.

Both Sanskrit and Classical Tibetan are low-resource languages that still require a lot of linguistic research. In recent years, comparatively large digital monolingual corpora of Sanskrit and Tibetan Buddhist literature have become available.[2] At the same time, sentence alignment algorithms supported by multilingual sentence embedding have reached increasing levels of precision for noisy texts of low resource languages (Thompson and Koehn, 2019). It is now possible to automatically align Sanskrit texts with their respective Tibetan translations and reach a sufficient level of precision to use these aligned texts either as a direct resource for philological research in the form of an online database[3] or for the training of multilingual language models and neural machine translation models (see Thompson and Koehn (2019), section 4.3). This paper makes the following contributions to this problem:

- A sentence-level aligned parallel corpus of Sanskrit Buddhist texts and their Tibetan translations with a total number of 317,289 sentence pairs, which is much larger and covers a greater variety of domains than the already available bilingual resources for these two languages (see section 4).
- Three manually aligned datasets with a combined size of 6,916 sentence pairs spanning different genres of Buddhist literature for the evaluation of sentence alignment quality.
- A bilingual sentence embedding model that can be used for information retrieval and sentence alignment.

I make the dataset available at: `https://github.com/sebastian-nehrdich/sanstib`
The bilingual sentence embedding model is available via huggingface: `https://huggingface.co/buddhist-nlp/sanstib`
In section 2 I briefly discuss the characteristics of the languages. In section 3 I discuss some prior linguistic resources and related work on sentence alignment as well as its specific challenges for Sanskrit and Tibetan. In section 4 I discuss the data that I use and what preparation steps I take. In section 5 I describe the entire pipeline for the creation of the aligned corpus. In section 6 I evaluate the quality of the sentence alignment and the bilingual sentence representation model.

2. Sanskrit and Classical Tibetan

Sanskrit is a classical language of the Indo-Aryan branch of the Indo-European languages. It was widely used as lingua franca by religious, scientific and literary communities of ancient India. Since the second

[1] See for example the Tibetan-Sanskrit dictionary by J. Negi (Negi, 1993 2005).

[2] See section 4 for examples of such monolingual resources.

[3] The results are accessible at `https://buddhanexus.net/multi/neutral`.

half of the first millennium BCE Sanskrit was used for the production of Buddhist literature. Early Buddhist works in Sanskrit show a strong influence of Middle Indian dialects, which has decreased over time (Edgerton, 1953). Sanskrit relies heavily on morphology to indicate grammatical relations and has a relatively free word order. It follows the nominative-accusative alignment, has a complex verbal system and a rich nominal declension. Nominal compounds are frequently found. Another special characteristic and challenge in the computational processing of Sanskrit is the phenomenon of Sandhi, by which the contact phonemes of neighboring word tokens are changed and merged, and which creates unseparated strings spanning multiple tokens (Hellwig and Nehrdich, 2018).

Classical Tibetan on the other hand belongs to the Sino-Tibetan language family and refers especially to the written language of the Tibetan cultural sphere until the 20th century. In this paper I focus on the Classical Tibetan language of the early canonical texts of Tibetan Buddhism that have been translated in the second half of the first millennium CE from other languages, mainly Sanskrit, that is sometimes also referred to as Old Tibetan. Classical Tibetan is classified as an ergative-absolutive language. Unlike Sanskrit, Tibetan nouns are not marked regarding grammatical gender or number. Particles are used to indicate case and are attached to whole noun phrases with the actual noun remaining unchanged. Even though Classical Tibetan started as a language of translation, it contains relatively few loanwords from Sanskrit; transliterations, especially of names of persons, places, and objects/things unknown to the Tibetan culture do occur, but are not very frequent in the majority of texts. In this paper I focus on Buddhist Sanskrit texts and those Classical Tibetan texts that are translations from Sanskrit and Buddhist Hybrid Sanskrit.

3. Related Work

While parallel corpora and treebanks are widely available for modern high-resource languages, such resources are scarce for the languages of the ancient Buddhist traditions. The most important resource for multilingual aligned ancient Buddhist texts is the Thesaurus Literaturae Buddhicae (TLB) hosted by the University of Oslo, Norwegian Institute of Palaeography and Historical Philology and directed by Jens Braarvig and Asgeir Nesøen.[4] This database features sentence-level and paragraph-level aligned texts with their respective translations featuring Sanskrit, Chinese, Tibetan, English and more depending on the text. The data of the TLB has been typed in and aligned manually. Currently (December 2021) it includes a total number of 88 texts excluding dictionaries, grammars and manuscript materials. Not all of them are aligned on the sentence level and not all of them are available completely. For example, the *Abhidharmakoṣabaṣya* features text alignment of the Sanskrit and Tibetan only for the first two and the beginning of the third chapter.

Another important digital resource is the Uma Institute for Tibetan Studies Tibetan-Sanskrit-English Dictionary (UMA dictionary) by Jeffrey Hopkins, which is widely available in digital form.[5] While the UMA dictionary does not include full parallel sentences, it is very useful to extract word and compound/short phrase pairs. Also available in digital form is the Chinese-Sanskrit–Tibetan index to the *Yogācārabhūmi* by Koitsu Yokoyama and Takayuki Hirosawa (Koitsu Yokoyama, 1996 1997).[6] Further available in digital form is the traditional Sanskrit–Tibetan dictionary *Mahāvyutpatti*.[7]

Sentence alignment has received a lot of attention in the late eighties and early nineties and is since then generally considered as a solved problem for high-resource languages. Traditional length-based aligners (Gale and Church, 1993; Brown et al., 1991) work best when the texts to be aligned are not very noisy and the languages follow similar punctuation conventions. The alignment of Sanskrit texts with their respective Tibetan translations is however a challenging task: The Tibetan translations, while for some texts of very high quality, can at times lack sentences, paragraphs or whole chapters of the Sanskrit texts that are at our disposal. The opposite case is also frequently seen: Due to difficulties in the transmission of manuscripts in India, many Sanskrit Buddhist texts are nowadays lost or transmitted only partly. A further factor is that some texts, especially of the Sūtra genre, have developed and changed over a longer period of time and the version available in Sanskrit edition might therefore differ from the Tibetan translation, at times considerably. The order of sub-clauses and whole sentences, especially when verses are translated, is sometimes inverted in Tibetan translation. For example, verse 10.340 of the *Laṅkāvatārasūtra* demonstrates this problem (corresponding text parts are colored accordingly):

Sanskrit:
rājāno rājaputrāś ca amātyāḥ śreṣṭhinas tathā |
piṇḍārthe nopadeśeta yogī yogaparāyaṇaḥ ||
Tibetan:
rnal 'byor gzhol ba'i rnal 'byor bas ||
rgyal po dang ni rgyal po'i bu ||
de bzhin blon po tshong dpon la ||
zas kyi phyir ni bsten mi bya ||

This example also demonstrates a difference in punctuation convention: In the case of Sanskrit, two pādas

[4] https://www2.hf.uio.no/polyglotta/index.php?page=library&bid=2

[5] https://glossaries.dila.edu.tw/glossaries/JHK?locale=en

[6] Available online via the Internet Archive: http://web.archive.org/web/*/http://www.buddhist-term.org/yoga-table/

[7] https://glossaries.dila.edu.tw/glossaries/MVP?locale=en

(the Sanskrit equivalent to a metrical line) are printed on one line in the edition and punctuation occurs in the form of a single daṇḍa after the second and a double daṇḍa after the fourth pāda at the end of the verse. In the Tibetan translation, each single pāda is separated with a double shad, with the same punctuation being used at the end of the verse. Since each of these languages belongs to different language families, their grammars are substantially different and their vocabulary has almost no direct overlap. These problems are also evident in the example: While the Tibetan here is clearly a translation of the Sanskrit, there is no clear orthographic or etymological resemblance between the vocabulary. The arrangement of the tokens within sentences and phrases can also vary, as the following pair of pādas demonstrates:

Sanskrit: amātyāḥ śreṣṭhinas tathā |

Tibetan: de bzhin blon po tshong dpon la ||

These differences are not the result of a faulty or incomplete translation process, but are cases of variation that are encountered regularly. The length of the Tibetan translation of a Sanskrit sentence can vary substantially and therefore an alignment based on length alone is difficult. While on average, a character of Sanskrit in roman transliteration is reproduced by 1.6 characters of Tibetan in Wylie transliteration, cases where much less or much more characters in Tibetan are used are encountered frequently. This happens either due to variations in the length of vocabulary, the use of abbreviations or the deliberate or accidental addition or omission of material on both sides.

I therefore decided to adapt YASA (Lamraoui and Langlais, 2013) for the coarse alignment process, since it can make use of both length and lexical features via bilingual dictionaries and has been shown in Lamraoui and Langlais (2013), table 2, to have strong performance on noisy parallel data. In section 6 I evaluate the performance of YASA against another widely used length- and dictionary based aligner, hunalign (Varga et al., 2005).

I also adapt the multilingual extension of SBERT, which is a sentence similarity model based on deep contextual embedding (Reimers and Gurevych, 2020). Multilingual SBERT uses an approach which they call multilingual knowledge distillation where a student model distills the knowledge of a teacher model. It requires a teacher model, a set of parallel sentences and a student model. The student model learns an embedding space that has two important properties: 1) Vector spaces are aligned across languages so that identical sentences in different languages are represented by similar vectors and 2) the vector space properties in the original source language from the teacher model are adopted and transferred to the target language(s).

With the help of SBERT, I search on corpus-level for potential further text pairs. For the final alignment I adapt vecalign (Thompson and Koehn, 2019) which currently is the sentence aligner with the strongest performance on low resource languages. I feed vecalign with the multilingual SBERT representations, which then uses cosine distance and dynamic programming to determine the optimal alignment of the sentences.

4. Data

For the Sanskrit data, I use the etexts available in the GRETIL collection.[8] This collection consists of a total number 1316 files with a combined size of 342MB in HTML format. The total number of tokens in this collection without applying Sandhi splitting and without stripping away the headers and HTML tags lies at 31,094,814. It is a diverse collection of Sanskrit texts with the oldest dating back to the 2nd millennium BCE. It includes material from various Indian religious traditions as well as epic and scholastic material. The etexts are typed versions of available modern editions of Sanskrit texts; some of these editions are not based on original manuscripts, but are reconstructions of the Sanskrit text based on their Tibetan and/or Chinese translations. The language contained in this collection is equally diverse and not limited to Classic Sanskrit; it includes Vedic as well as Buddhist Hybrid Sanskrit and other Middle Indic languages. A number of texts in this collection appear more than one time, for example when two different editions of the same text have been included or when a different sorting system of verses or a different division of chapters into files have been applied. Two sub-folders of this collection contain exclusively Buddhist material: `4_rellit/buddh/` and `6_sastra/3_phil/buddh/`. They contain 405 files with a combined size of 60MB and 5,437,267 tokens. `4_rellit/buddh/` contains 252 files with a size of 41MB and 3,722,364 tokens with the majority of the files being scriptures. `sastra/3_phil/buddh/` contains 153 files with a size of 19MB and 1,714,903 tokens. The majority of these texts are treatises. The temporal range of the Buddhist material in these two sub-folders reaches from the second half of the 1st millennium BCE to the middle of the 2nd millennium CE.

The source of the Tibetan data for the experiments are the Kangyur and Tengyur collections as they have been digitalized by the Asian Classics Input Project (ACIP).[9] The Kangyur collection consists of works that traditionally have been regarded as the Word of the Buddha in Tibetan translation while the Tengyur collection consists of treatises, commentaries and other related works in Tibetan translation.[10] The ACIP Kangyur contains 858 files with a size of 114 MB and 27,140,270 tokens. The ACIP Tengyur contains 3423 files with a size of 247MB and 58,266,067 tokens. together they contain

[8] `http://gretil.sub.uni-goettingen.de/gretil.html`

[9] `https://asianclassics.org/library/downloads/`

[10] The possible genesis of these two collections has been recently discussed by O. Almogi (Almogi, 2021).

4284 files, 361MB and 85,406,337 tokens.
To aid the YASA aligner, I extract word and phrase pairs from the UMA dictionary, the *Yogācārabhūmi* index and the *Mahāvyutpatti* and combine them into one dictionary. I extract 2,000 manually aligned sentence pairs of the second chapters of the *Abhidharmakoṣabaṣya* from the TLB as an evaluation dataset for the treatises genre (AKBh). I also take 2,770 manually aligned sentence pairs of the *Vimalakīrtinirdeśa* from the TLB as an evaluation dataset for the scripture genre (VkN). Additionally, I manually aligned the Sanskrit treatise *Pañcaskandhakavibhāṣā* with its Tibetan translation yielding 2,146 sentence pairs (PSkVBh). During the manual alignment I aimed at producing a bitext that reflects the punctuation conventions of both languages as close as possible.

4.1. Data Preperation

As the first step I removed the headers and HTML markup from the GRETIL files. The next step was to segment the Sanskrit and Tibetan files into sentences and put each of the sentences on a separate line. This task is not without problems since the punctuation conventions between Sanskrit and Tibetan are usually not the same. In the case of GRETIL, one finds a variety of punctuation conventions: Some editions (especially those printed in Devanagari) follow Indian punctuation conventions and only apply the daṇḍa-sign. Other editions follow the punctuation of Western languages and use full-stop, question mark, exclamation mark and more. In order to simplify the process, I decided to regard all punctuation marks (daṇḍa, comma, full-stop etc.) in the GRETIL files as a sentence delimiter and split the texts according to these marks. The average sentence length therefore can vary a lot between the different GRETIL texts depending on their punctuation scheme. In the case of Tibetan, I regarded the occurrence of single or double shad as decisive sentence boundary. These differences in punctuation make frequent many-to-one and one-to-many alignments necessary to arrive at optimal solutions.
The GRETIL texts have then been processed with a joint Sandhi+Compound splitting tool (Hellwig and Nehrdich, 2018). This type of tokenization makes it possible for the sentence alignment algorithms to access the individual lexical units and match them with their Tibetan counterparts. In the case of the Tibetan files I stripped away all punctuation marks and folio markings. The original form of the texts in both languages have been stored in a database to make it possible to reconstruct them into their original form at a later point. As a final step I sorted the GRETIL files according to the Kangyur and Tengyur collections and their respective sub-units ('Dul ba, 'Bum etc.). The GRETIL files have been further paired with their respective Tibetan translations for those cases where this was obviously possible based on the works' titles.

5. Alignment Process

The full pipeline is demonstrated in figure 1. The individual steps are as follows: After the data preparation of both corpora I use the YASA aligner with the support of the previously created dictionary to do a coarse sentence alignment of the manually determined text pairs. Since the average character sentence length ratio between Sanskrit and Tibetan sentences from the gold data lies at 1:1.6 I remove all pairs of sentences where the ratio lies above 1:3 or below 1:0.5, to mitigate the influence of badly aligned sections. This yields a total number of about 114,000 sentence pairs. Next I train a RoBERTa-model (Liu et al., 2019) on the Buddhist GRETIL texts and the full ACIP data. This model has 6 layers and a hidden dimensionality of 768. This RoBERTa model is then used as a teacher model in combination with the sentence pairs from the coarse YASA alignment and the word/phrase pairs from the dictionary for the training of a multilingual SBERT model. The multilingual SBERT model also has 6 layers and a hidden dimensionality of 768.
I then create a HNSW index (Johnson et al., 2021; Nehrdich, 2020) of the SBERT representations of all Buddhist Sanskrit sentences and query this index with the SBERT representations of the sentences of each of the Tibetan files contained in ACIP. In this way I can determine which Tibetan text has the most matches with which Sanskrit texts according to the multilingual SBERT model and yield more possible text pair candidates for further alignment.
I then use vecalign with the SBERT representations to do the fine alignment. I align all text pairs used for the coarse alignment and all text pair candidates from the HNSW search with vecalign. From the results I once again remove all sentence pairs with a length ratio above 1:3 or below 1:0.5 yielding a total number of about 300,000 sentence pairs. These sentence pairs are used together with the UMA dictionary data to retrain the multilingual model. This model is then used to do the final alignment of all text pair candidates. Table 1 shows the statistics of the resulting dataset. To be noted is that the alignment quality of these texts is not always consistent and depending on the desired task, further filtering might be necessary. For the training of a machine translation or sentence similarity model, I advise to filter out sentence pairs where the length ratio lies above 1:3 or below 1:0.5.

6. Evaluation

I use the previously described datasets AKBh, VkN and PSkVBh to evaluate the quality of the sentence alignment and the bilingual SBERT model. For the evaluation of the sentence alignment, I summarize the precision and recall ratios into the F-measures F_A and F_S (Lamraoui and Langlais, 2013). F_A is computed on alignment and F_S on the sentence level. Each correctly identified bisegment gets a positive score in the

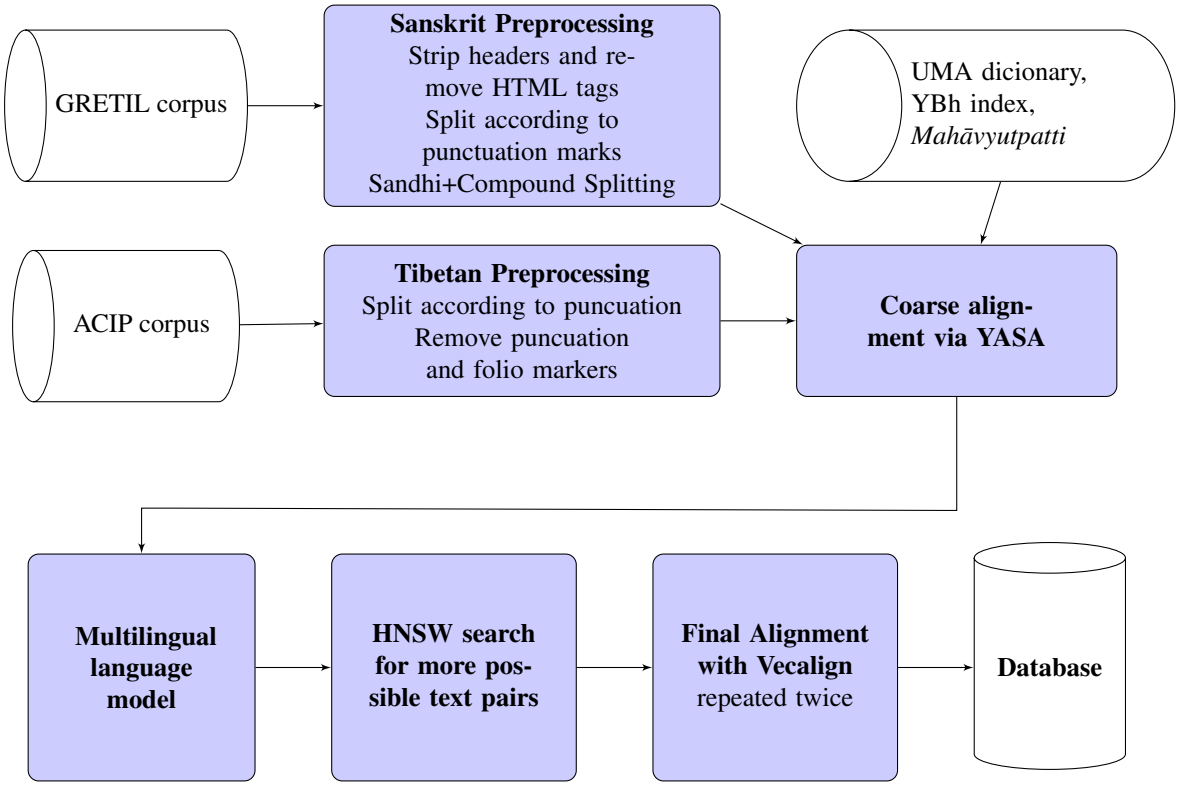

Figure 1: Diagram of the entire pipeline for the generation of a parallel Sanskrit–Tibetan corpus.

Collection	Category	Text pairs	Aligned sentence pairs
Scriptures (Kangyur)	Vinaya	4	9513
	Prajñāpāramitā	21	62775
	Ratnakūta	3	3160
	Avataṁsaka	1	6053
	Sūtra	26	55453
	Tantra	19	22651
	Total	74	159605
Treatises (Tengyur)	Tantra	6	481
	Prajñāpāramitā	4	10967
	Madhyamaka	26	32684
	Sūtra Commentaries	3	1529
	Yogācāra	18	38843
	Abhidharma	4	37440
	Jātaka	2	12295
	Lekha	5	1037
	Pramāṇa	19	22408
	Total	87	157684
Total		**161**	**317289**

Table 1: Statistics of the bilingual dataset

alignment, the accuracy and recall are then calculated accordingly (see Lamraoui and Langlais (2013), section 3). F_A is a rather unforgiving measure since it only considers those bisegments that are aligned exactly in the same way as in the gold dataset, thus ignoring partly right alignments. I therefore also show F_S which reports how many actual sentences have been aligned correctly regardless of wether the bisegments are identical to that of the gold dataset or not. I split the Tibetan sentences of the evaluation datasets according to their punctuation into separate units before the alignment process and use these together with the unaltered Sanskrit sentences as input for the aligners. It is therefore necessary for the aligners to make many-to-one decisions on the side of Tibetan to match the Sanskrit segments. In this setup, many-to-many decisions will result in a decrease in F_A score even if the aligned sentences are parallel. The F_S score is not affected by many-to-many decisions as long as the bisegments are parallel. In the extreme case, if the whole Tibetan and Sanskrit text would be regarded as a single bisegment containing all sentences in a many-to-many alignment, F_S would be 100.0, while F_A would be 0.0. It is therefore important to take both measures into account when evaluating the performance of the aligners. I compare the performance of hunalign, YASA and vecalign. Hunalign and YASA are both supported by a bilingual dictionary. I report the results in table 2. Vecalign outperforms all other algorithms by a considerable margin. YASA is on second place and hunalign shows the weakest performance with the exception of the F_S for PSkVBh, where it is stronger than YASA. All algorithms show their strongest performance on PSkVBh. This is not very surprising since PSkVBh was manually aligned with the goal in mind to produce a bitext that resembles the punctuation of the original languages closely. All aligners show their second strongest performance on AKBh. This can be explained by the facts that the AKBh is a Sanskrit text of the treatise Abhidharma category that is comparatively well represented in the training dataset and that it is known for the high quality of its Tibetan translation. The aligners show

their weakest performance on VkN, a text of the scripture Sūtra category that brings its own challenges with it, such as many-to-one and one-to-many alignments in the case of long lists that are punctuated differently in the two languages. As expected, F_A is the score that the aligners struggle the most with. The comparatively good results of PSkVBh indicate that punctuation and a meaningful pre-segmentation of the material play a big role in achieving high alignment scores.

For the evaluation of the bilingual SBERT model I report the precision score by selecting the Tibetan sentence with the lowest cosine distance for each Sanskrit sentence. All datasets have been limited to 2,000 sentences to make the results more comparable. The results are reported in table 3. The precision in this experiment is in a similar region for all three datasets, with VkN being slightly less good than AKBh and PSkVBh. Since unlike the sentence alignment experiment, punctuation does not play a role in this case, the difference between the datasets is not as pronounced.

	hunalign		YASA		vecalign	
Dataset	F_A	F_S	F_A	F_S	F_A	F_S
AKBh	44.6	75.4	56.0	78.1	**82.8**	**94.3**
VkN	30.1	63.5	49.0	73.0	**75.1**	**90.6**
PSkVBh	63.6	83.1	66.1	80.4	**92.6**	**97.3**

Table 2: Accuracy of the sentence alignment

Dataset	Precision
AKBh	85.6
VkN	83.1
PSkVBh	85.1

Table 3: Precision of bilingual SBERT

7. Conclusions

This paper presented the development of a parallel corpus of Sanskrit–Tibetan texts with sentence-level alignment. The corpus has a total size of 317,289 sentence pairs and 14,420,771 tokens. The paper also presented three different datasets for the evaluation of Sanskrit–Tibetan sentence alignment and tested three different aligners on these datasets. The highest scoring aligner, vecalign, is able to achieve an F_S score above 90% for all three datasets. I terefore used vecalign to align the data of the corpus. While I cannot assume that an F_S score of more than 90% is met for all texts in the data, the evaluation results lead me to assume that the majority of the aligned texts have a reasonable good quality. The presented data can be used as a resource on its own for philological research and has already been incorporated into the BuddhaNexus database. The development of a digital Sanskrit–Tibetan dictionary with references to the original texts is possible based on the aligned sentence pairs. It is also possible to use this data for the training of machine translation systems or multilingual language representation models.

I also presented a multilingual SBERT model that can be used for bilingual information retrieval tasks of Sanskrit and Tibetan. It is able to achieve a precision of more than 80% for all three datasets when 2,000 sentences are considered. Since the Buddhist textual tradition is characterized by a high occurence of textual reuse, multilingual information retrieval can be used to locate such instances of textual reuse and identify them systematically. For the near future, I plan to continue with this task and will present the additionally discovered data in the BuddhaNexus database as well.

Since the quality of the alignment of the presented data depends strongly on the punctuation conventions followed in both languages, I believe that standar-dized methods for recognizing sentence boundaries for Sanskrit could help to improve the performance of the alignment algorithms.

Almogi, O. (2021). The Old sNar thang Tibetan Buddhist Canon Revisited, with Special Reference to dBus pa blo gsal's bsTan 'gyur Catalogue. *Revue d'Etudes Tibétaines*, 58:165–207.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning Sentences in Parallel Corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176, Berkeley, California, USA, June. Association for Computational Linguistics.

Edgerton, F. (1953). *Buddhist Hybrid Sanskrit Grammar and Dictionary*. Motilal Banarsidass.

Gale, W. A. and Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1):75–102.

Hellwig, O. and Nehrdich, S. (2018). Sanskrit Word Segmentation Using Character-level Recurrent and Convolutional Neural Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium, October-November. Association for Computational Linguistics.

Johnson, J., Douze, M., and Jégou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Koitsu Yokoyama, T. H. (1996-1997). *Index to the Yogācārabhūmi, Chinese-Sanskrit-Tibetan*. Sankibō Busshorin.

Lamraoui, F. and Langlais, P. (2013). Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France, September 2-6.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 [cs.CL].

Negi, J. (1993-2005). *Tibetan-Sanskrit Dictionary*, volume I-VI. Dictionary Unit Central Institue of Higher Tibetan Studies.

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.

Reimers, N. and Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel Corpora for Medium Density Languages. In *Proceedings of the RANLP 2005,*, pages 590–596, 9.