A03 Computational Methods und Tibetan Studies (Forschungsstand)

Stand: 03.09.2026

1. Einleitung

Es existiert eine verhältnismäßig große Anzahl an aktiven Initiativen/Projekten gibt, die sich mit der Digitalisierung, Bereitstellung und Analyse von tibetischen Texten und anderen tibetischen Quellen befassen. Ebenfalls existieren zahlreiche Tools und Modelle für die Verarbeitung und Analyse tibetischer Texte, was die Vorverarbeitung der Texte für computergestützte Analysen vereinfacht.

2. Tools und Forschungsstand

Die OpenPecha Initiative hat wichtige Tools für die Vorverarbeitung tibetischer Texte und deren computergestützte Analyse erstellt, die auch in der Machbarkeitsstudie genutzt werden. Dazu gehören der **Tokenizer** Botok, der **Wylie Konvertierer** pyewts sowie zahlreiche Machine Learning Modelle, die via HuggingFace verfügbar sind. Es existieren außerdem Modelle für **spaCy** (siehe folgender Link).

Im Bereich **LLMs** gibt es speziell für tibetische Sprache und Kultur trainierte Modelle, darunter **Banzhida** (Pan et al. 2025, basierend auf Qwen 2.5) und **Llama Sun-Shine** (Huang et al. 2024). Außerdem existiert mit **Monlam Think** ein vom Monlam Tibetan IT Research Center auf mehr als 72 GB Daten trainiertes generatives Modell.[1] Fraglich bei letzterem Modell ist, ob dieses frei verfügbar ist (im Sinne von Open Weights). Auf der Website direkt verfügbar ist eine Web Schnittstelle für Übersetzungen und andere Services (OCR). Auch Open Pecha bietet, wie oben bereits erwähnt, eine breite Palette an Modellen auf HuggingFace an, darunter ein Embedding-Modell, das auch in der Case-Study verwendet wurde.

Im Bereich **HTR/OCR** wird in einigen Studien Transkribus genutzt (Erhard 2025). Aber auch das Dharmamitra Projekt bietet eine eigene OCR Schnittstelle auf der Website, die auch die Umwandlung von größeren Dokumenten erlaubt. Zu dem Thema hat u.a. *Rachael Griffiths* gearbeitet (siehe Konferenzbeitrag „Towards transcribing handwritten Tibetan manuscripts with HTR“ sowie Griffiths 2024). Auch *Élie Roux* des Buddhist Digital Resource Center (BDRC, siehe unten) hat einen Vortrag zum Thema „OCR at the Buddhist Digital Resource Center: status and perspectives“ gehalten.

Im Bereich **Image Recognition** und **Visual Analytics** wurde von Li, Channa, Marco Peer and Kaifeng Yang versucht, die Autorschaft von Dunhuang-Manuskripten mit Hilfe von ML-Ansätzen zu

[1] https://phayul.com/pioneering-tibetan-it-outfit-launches-monlam-think-claims-it-outperformed-global-ai-models-in-tibetan-language/

untersuchen (Konferenzabstract „Looking for Chos grub: Proposing An AI-Assisted Method for Recognizing Scribal Hands of Dunhuang Tibetan Manuscripts“).

Im Bereich **NLP** ist u.a. Marieke Meelen aktiv und hat verschiedene Case-Studies zu tibetischen Texten und NLP durchgeführt (siehe u.a. das Konferenzabstract zu „NLP for Tibetan corpus creation: how to get more out of our data?“).[2] Das Dharmamitra Projekt bietet verschiedene Modelle und Tibetan-Sanskrit Dictionaries zum Download an (siehe u.a. https://dharmamitra.github.io/dharmamitra-guides/mitra_dictionaries/). Ein zentraler Beitrag in diesem Kontext stammt von Meelen, Nehrdich & Keutzer (2024) unter dem Titel „Breakthroughs in Tibetan NLP“. Für Tokenization, Segmentierung, POS Tagging usw. wird neben Botok (s.u.) auch ACTib genutzt (https://github.com/mariekemeelen/actib). Das Thema NER scheint noch recht offen zu sein, und der Übersichtsbeitrag von Meleen et al. (2024) hat einige Referenzen, u.a. zu Liua & Wang (2018).

Es gibt eine Pipeline namens **tibetan-intertext-lab**,[3] das aus Berkeley u.a. von Kurt Keutzer entwickelt wird. Es bietet unter anderem Satz-Segementierung, Embeddings für Sätze und Texte und direkte Embedding-basierte Vergleiche zwischen Texten.

2.1 Machine Learning

Oliver Hellwig hat Bayesian Modelle trainiert, um Fragen der Chronologie vedischer Texte zu beatnworten (Konferenzabstrakt „Dating text corpora with Bayesian models“). *Sebastian Nehrdich*, der u.a. für das Dharmamitra Projekt verantwortlich ist, hat sich u.a. mit Shared Vector Spaces befasst (siehe Konferenzabstrakt „Creating a Shared Semantic Vector Space for Buddhist Languages“), um buddhistische Texte in verschiedenen Sprachen miteinander zu vergleichen. Ein weiterer wichtiger Name im Kontext von Dharmamitra ist *Kurt Keutzer*, Prof. Informatik emeritus aus Berkeley. Es gibt ein Embeddings-Model für spaCy (https://doi.org/10.5281/zenodo.10148636). Über das Dharmamitra Projekt werden auf HuggingFace viele Modelle angeboten: https://huggingface.co/buddhist-nlp

Ein besonders spannendes Modell ist *gemma-mitra-2-e*,[4] das auch im Tibetan Intertext Lab Tool verwendet wird (Nehrdich and Keutzer 2026).

3. Initiativen/Forschungsprojekte

1. **Buddhist Digital Resource Center** (BDRC)[5] mit dem Buddhist Universal Digital Archive (BUDA)[6]. Das ACTib Corpus mit über 185 Mio. Tokens basiert auf dem BUDA Corpus (siehe

[2] Siehe auch Marieke Meelen / Nathan Hill: Segmenting and POS tagging Classical Tibetan using a memory-based tagger. In: Himalayan Linguistics 16 (2018), H. 2. 17.01.2018.

[3] https://github.com/ten-jampa/tibetan-intertext-lab.

[4] https://huggingface.co/buddhist-nlp/gemma-2-mitra-e

[5] „Current programs focus on the preservation of Tibetan literary heritage, the preservation of manuscripts in Southeast Asia, and the development of technology to support digital archiving and scholarly research.“

[6] „BDRC developed the Buddhist Digital Archives (BUDA) as a collaborative platform for Buddhist texts in order to benefit the Buddhist community and BDRC's many users in academia. With new features and an innovative design, BUDA greatly improves access to a vast collection of Tibetan Buddhist works, as well as to collections of

Meelen und Roux 2018).

2. **Resources for Kanjur & Tanjur Studies** (rKTs)[7]: Bietet auf der Website Zugriff auf ein großes Corpus an Texten mit vielfältigen Analysemöglichkeiten und Suchfunktionen.
3. Eng verbunden mit dem rKTS ist auch das **Tibetan Manuscript Project Vienna** https://tmpv.univie.ac.at/
4. **Tibetan and Himalayan Library** (THL): Eine sehr breite Initiative, die vielfältige Quellen im Zusammenhang mit Tibet archiviert und verfügbar macht.
5. **OpenPecha** (siehe oben).
6. **PaganTibet**: Ein in Frankreich basiertes Forschungsprojekt, das ebenfalls Gebrauch von digitalen Methoden macht und ein eigenes Corpus erstellt hat.
7. **Dharmamitra Projekt**: Ein Projekt geleitet u.a. von Sebastian Nehrdich, zu dessen Vorläufern das BuddhaNexus Projekt gehörte. Erreichbar unter https://dharmamitra.org/. Es wird außerdem das Corpus DharmaNexus angeboten (siehe Corpora).
8. **Pydurma** ist eine auf Python basierende Kollations-Engine, die für die **automatisierte Auswahl von Textvarianten** in textuellen Traditionen entwickelt wurde, insbesondere für Sprachen wie Tibetisch. Sie ist auf **Geschwindigkeit, Modularität und Skalierbarkeit** optimiert und eignet sich daher besonders für groß angelegte Projekte, etwa das Zusammenführen von OCR-Ausgaben oder die Erstellung automatisierter Editionen. https://github.com/OpenPecha/Pydurma Pydurma gehört mit zur OpenPecha Initiative.
9. **tibetan-intertext-lab**: Berkeley u.a. von Kurt Keutzer entwickelt wird. Es bietet unter anderem Satz-Segementierung, Embeddings für Sätze und Texte und direkte Embedding-basierte Vergleiche zwischen Texten. https://github.com/ten-jampa/tibetan-intertext-lab Für beispielhafte Ergebnisse, siehe hier: https://ten-jampa.github.io/tibetan-intertext-lab/
10. **buddhist-nlp** bei HuggingFace: Von Sebastian Nehrdich und Dharmamitra, viele Modelle https://huggingface.co/buddhist-nlp

4. Corpora

Wahrscheinlich gibt es große Überschneidungen zwischen den Corpora, die in einer detaillierteren Analyse nachvollzogen werden müssten.

1. **Old Tibetan Corpus**: https://github.com/tibetan-nlp/old-tibetan-corpus

Sanskrit, Chinese, Pali, Burmese, and Khmer materials."

[7] „The Resources for Kanjur & Tanjur Studies (rKTs) is an open-access digital platform that is currently the largest database of Tibetan canonical literature. Developed by the Tibetan Manuscript Project Vienna (TMPV), a long-term research initiative based at the University of Vienna's Department of South Asian, Tibetan and Buddhist Studies, rKTs offers a range of advanced tools to support the study and analysis of Tibetan sources." (https://digitalorientalist.com/2025/02/14/resources-for-kanjur-and-tanjur-studies-introduction-and-review/)

2. **Annotated Corpus of Classical Tibetan (ACTib)**[8]: https://doi.org/10.5281/zenodo.3951503
3. **TIB-STC Dataset** (Datensatz zum Training von Sun-Shine, siehe Huang et al. 2024): https://github.com/Vicentvankor/sun-shine
4. **rKTS Corpus**: Im e-Repository der Website kann ebenfalls auf eine große Anzahl an Texten zugegriffen werden: http://www.rkts.org/etexts/repository.php
5. **Keutzer et al. Corpus**: Das initiale Corpus, das auch in Ausschnitten für die Machbarkeitsstudie verwendet wurde. https://github.com/keutzer/bo-corpus-analytics
6. **Pagan Tibet Corpus**: Das Projekt hat ebenfalls ein eigenes Corpus: https://www.pagantibet.com/data-corpora
7. **Corpus des Text Surrounding Projects**: Ein Projekt (2019-2023), das verschiedene und vornehmlich tamilische und Sanskrit Texte aus Bibliotheken und Archiven in Deutschland und Frankreich (u.a. Hamburg) digitalisiert hat. https://tst-project.github.io/

Das **DharmaNexus Corpus** (https://dharmamitra.github.io/dharmamitra-guides/dharmanexus/) umfasst tibetische, Pali, Sanskrit und chinesische Texte und gehört mit zum Dharmamitra Projekt (siehe oben).

5. Links

Botok Tokenizer: https://github.com/OpenPecha/Botok

Buddhist Digital Resource Center: https://www.bdrc.io/

Dharmamitra Projekt: https://dharmamitra.org/translate

Open Pecha Initiative: https://openpecha.org/

Open Pecha Hugging Face: https://huggingface.co/openpecha

Pagan Tibet: https://www.pagantibet.com/

Resources for Kanjur & Tanjur Studies: http://www.rkts.org/

Tibetan and Himalayan Library: https://worldhistorycommons.org/tibetan-and-himalayan-digital-library

Wyle Konvertierer pyewts: https://github.com/OpenPecha/pyewts

[8] Über 185 Millionen Tokens; basiert auf den digitalen Textsammlungen des *Buddhist Digital Resource Center* (BDRC) BUDA. „The Annotated Corpora of Classical Tibetan (ACTib) version 2.0 is a Tibetan corpus containing 170 million words. The corpus consists of Classical Tibetan texts and was built as part of the Tibetan in Digital Communication project (2012-2015). The Annotated Corpus of Classical Tibetan is a collection of Tibetan electronic texts compiled by the Buddhist Digital Resource Center and can be downloaded from this Zenodo repository.“ (https://www.sketchengine.eu/tibetan-corpus/)

6. Literatur

Erhard, Franz Xaver. 2025. “Text and Information Extraction for Modern (Post-1950) Tibetan Newspapers and Print Publications with Transkribus.” Revue d’Etudes Tibétaines, no. 74 (February): 128–71.

Griffiths, Rachael. 2024. „Handwritten Text Recognition (HTR) for Tibetan Manuscripts in Cursive Script“. *Revue d’Etudes Tibétaines*, Nr. 72 (Juli): 43–51.

Huang, Cheng, Fan Gao, Nyima Tashi, et al. 2026. “TFD: A Comprehensive Structured Tibetan Foundation Dataset for Low-Resource Language Processing and Large-Scale Modeling.” arXiv:2503.18288. Preprint, arXiv, February 14. https://doi.org/10.48550/arXiv.2503.18288.

Huang, Cheng, Fan Gao, Nyima Tashi, Yutong Liu, et al. 2024. “Sun-Shine: A Large Language Model for Tibetan Culture.” arXiv Preprint arXiv:2407.10671.

Huang, Cheng, Nyima Tashi, Fan Gao, et al. 2025. “Tibetan Language and AI: A Comprehensive Survey of Resources, Methods and Challenges.” Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2510.19144.

Krishna, Ravi, Norman Mu, and Kurt Keutzer. 2021. “Applying Text Analytics to the Mind-Section Literature of the Tibetan Tradition of the Great Perfection.” ACM Transactions on Asian and Low-Resource Language Information Processing 20 (2): 1–32. https://doi.org/10.1145/3392047.

Kyogoku, Yuki, Franz Xaver Erhard, Robert Barnett, and Nathan Hill. 2024. “Basic Modern Tibetan SpaCy Model.” Version 0.1.2. Preprint, Zenodo, December 15. https://doi.org/10.5281/ZENODO.14494472.

Li, Yan, Xiaomin Li, Yiru Wang, Hui Lv, Fenfang Li, and La Duo. 2022. “Character-Based Joint Word Segmentation and Part-of-Speech Tagging for Tibetan Based on Deep Learning.” ACM Transactions on Asian and Low-Resource Language Information Processing 21 (5): 1–15. https://doi.org/10.1145/3511600.

Liang, Siyu, and Zhaxi Zerong. 2025. “The Tonogenesis Continuum in Tibetan: A Computational Investigation.” arXiv:2510.22485. Preprint, arXiv, October 26. https://doi.org/10.48550/arXiv.2510.22485.

Luo, Queenie, und Leonard W. J. van der Kuijp. 2024. „Norbu Ketaka: Auto-Correcting BDRC’s E-Text Corpora Using Natural Language Processing and Computer Vision Methods“. *Revue d’Etudes Tibétaines*, Nr. 72 (Juli): 26–42.

Meelen, Marieke, and Nathan Hill. 2018. “Segmenting and POS Tagging Classical Tibetan Using a Memory-Based Tagger.” Himalayan Linguistics 16 (2). https://doi.org/10.5070/H916234501.

Meelen, Marieke, and Élie Roux. 2020. “The Annotated Corpus of Classical Tibetan (ACTib) - Version 2.0 (Segmented & POS-Tagged).” Zenodo, May 4. https://doi.org/10.5281/ZENODO.3951503.

Meelen, Marieke, Élie Roux, and Nathan Hill. 2021. “Optimisation of the Largest Annotated

Tibetan Corpus Combining Rule-Based, Memory-Based, and Deep-Learning Methods.” ACM Trans. Asian Low-Resour. Lang. Inf. Process. (New York, NY, USA) 20 (1). https://doi.org/10.1145/3409488.

Meelen, Marieke, Sebastian Nehrdich, und Kurt Keutzer. 2024. „Breakthroughs in Tibetan NLP & Digital Humanities“. Revue d’Etudes Tibétaines, Nr. 72 (Juli): 5–25.

Nehrdich, Sebastian, and Kurt Keutzer. 2026. “MITRA: A Large-Scale Parallel Corpus and Multilingual Pretrained Language Model for Machine Translation and Semantic Retrieval for Pāli, Sanskrit, Buddhist Chinese, and Tibetan.” arXiv:2601.06400. Preprint, arXiv, January 10. https://doi.org/10.48550/arXiv.2601.06400.

Pan, Leiyu, Bojian Xiong, Lei Yang, et al. 2025. “Advancing Large Language Models for Tibetan with Curated Data and Continual Pre-Training.” arXiv:2507.09205. Preprint, arXiv, July 28. https://doi.org/10.48550/arXiv.2507.09205.