Breakthroughs in Tibetan NLP & Digital Humanities*

Marieke Meelen
University of Cambridge

Sebastian Nehrdich
Heinrich Heine University Düsseldorf;
University of California, Berkeley

Kurt Keutzer
University of California, Berkeley

The field of Digital Humanities has been transformed in recent years, not just because of advances in computing software and hardware, but in particular because of breakthroughs in Natural Language Processing (NLP). In this introduction to this Special Issue related to the Tibetan NLP and Technology panel at the conference of the International Association of Tibetan Studies (IATS) in Prague in 2022, we give a brief overview of the so-called 'state-of-the-art' of NLP and Digital Humanities tools for Tibetan Studies in particular. We aim to provide accessible introductions to the contributions in this Special Issue by other panel members as well as other recent developments in the field of 'Tibetan Tech' that could benefit any scholars in the field.

1. Introduction

In the Humanities, Social Sciences, Cultural Heritage and literary communities, there is increasing interest in, and demand for, Digital Humanities and Natural Language Processing (NLP) methods to enhance our data and facilitate new lines of research. The IATS 2022 panel 'Tibetan digital humanities and natural language processing' aimed to bring together researchers from all

* Marieke Meelen, Sebastian Nehrdich & Kurt Keutzer, "Breakthroughs in Tibetan NLP & Digital Humanities", *Revue d'Etudes Tibétaines*, no. 72, Juillet 2024, pp. 5-25.

The authors would like to acknowledge support from the AHRC (Grant ref. AH/V011235/1), the ELDP (Grant ref. G114548) as well as the BDRC, Esukhia and MonlamAI. This research is furthermore partially funded by the European Union (ERC, PaganTibet, 101097364). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

these areas to discuss new technologies, NLP methods and digital humanities tools related to any aspect of Tibetan Studies (language, linguistics, culture, history, literary studies, etc.).

Recent years have seen great progress in these areas with a wide range of research projects focusing on various aspects of Tibetan NLP and Digital Humanities. The panel encouraged discussions on both the technical as well as the applied side, so that both developers and users of these new tools can collaborate and enhance their research. In 2022 already, there were major Tibetan text corpora that could be digitally mined. Nonetheless, digital techniques have not reached everyone in the field of Tibetan Studies yet. In addition, there remained a gap even between those technologically engaged Tibetologists working on resource creation (e.g. library scientists) and those working on task-based tools (e.g. computational linguists) as well as as the wider field at large. This panel provided an excellent forum for researchers working across the domains of Tibetan digital humanities, natural language processing, and library science to share results and build collaborations.

Developments in the field of NLP and Digital Humanities are quick even for languages like Tibetan that are traditionally classified as 'low-resource' and 'under-researched'. The panel and this Special Issue demonstrate, however, that rapid progress in the field has opened up a wide range of opportunities from digitising textual and audiovisual resources to get data (Section 2), to enriching data through linguistic annotation (Section 3), retrieving information from digitised and annotated resources (Section 4) as well as Machine Translation (Section 5) and language learning or textual editing (Section 6).

2. Getting Data: ASR & OCR/HTR

Advanced textual analysis tools require a corpus of digitised texts in a Unicode format, henceforth called "e-texts". Data for Tibetan languages comes from various resources. For historical data, we mainly rely on manuscripts or xylographs, whereas new data for modern Tibetan languages can be collected in both written and audiovisual format.[1]

2.1 Optical and Handwritten Text Recognition (OCR/HTR)

In recent years computational methods have been employed in order to digitise written or printed data in a more efficient way through the use of Optical and Handwritten Text Recognition that take images with a textual component

1 In more recent years, especially Modern Standard/Lhasa Tibetan data has become available as original e-texts as well (i.e. so-called 'born-digital'), which makes it easier to study this variety. In this article, however, we focus on materials that are not yet digitised.

as an input and outputs a Unicode transcription of the text that is automatically searchable. Optical Character Recognition (OCR) was originally used for typed books (depending on white space between characters), but has now become a more general term for any form of automatic recognition of any type of texts. Handwritten Text Recognition (HTR) focuses specifically on manuscripts that are handwritten and therefore present a more challenging task. When it comes to Tibetan, progress was initially made for the (printed) *uchen* script, starting in the late 1980s in a collaboration between Bell Laboratories and the University of Virginia (Baird & Lofting 1990), and a range of Chinese OCR implementations in the following decades (e.g. Wang & Ding (2003: 5296), Drup, Zhao, Ren, Sanglangjie, Liu & Bawangdui (2010)) as well as efforts to use multi-language OCR software like Tesseract and Abbyy, and, finally, custom-made software like Namsel, which was used at the Tibetan Buddhist Resource Center (TBRC) to support the production, review, and distribution of searchable Tibetan texts at a large scale (Rowinski & Keutzer 2016). Most recently, the Buddhist Digital Resource Centre (BDRC) has collaborated with the Google Cloud Vision team, and a number of scholars, to speed up the creation of e-texts from their digital image library, focusing on OCR as well as automated post-correction.

Queenie Luo's contribution to this Special Issue gives an excellent example of a post-correction pipeline. In her article, she reports on the outcomes of a joint BDRC-Harvard-Berkeley project creating a Buddhist manuscript database using Natural Language Processing (NLP) algorithms. BDRC has digitised over 8,000 volumes of Tibetan Buddhist texts in recent years, most of which were processed by Google Cloud Vision's OCR engine. Since these OCR engines do not create a perfect e-text corpus, Luo shows how using Tibetan language models (e.g BERT, Bidirectional Encoder Representations from Transformers, GRU (Gated recurrent unit) and LSTM (Long-Short Term Memory) can not only facilitate auto-correction, but also develop further tools to retrieve named entities, identify topics as well as different genres. Her workflow combines automated computational as well as human efforts to optimise the results by first automatically mapping the OCR-ed e-texts with a Tibetan dictionary, and then using a spelling-check model to auto-correct the misspelled words based on their context. Following that, human experts validated and edited all machine-corrected texts, which were then made publicly available through the BDRC online BUDA platform.

While Google OCR[2] performs reasonably well for printed books in the standard Tibetan *uchen* script, block prints usually prove more challenging. This most commonly-found form of Tibetan literature exhibits more irregu-

2 https://cloud.google.com/use-cases/ocr

larities and digital images are not as clear as modern printed publications. Early efforts on OCR of blockprint identified the key challenges with this medium, but were not very successful at addressing them (Hedayati, Chong & Keutzer 2011). The BDRC has therefore developed a model for both layout analysis as well as OCR for block prints.[3]

Meanwhile, digitising handwritten manuscripts is an even more challenging task, due to various reasons ranging from a wide variety of cursive scripts, as well as typical features of writing styles such as overlap in characters (i.e. lack of white space separating them), lack of straight lines and the use of marginalia. In her contribution to this Special Issue, **Rachael Griffiths** discusses these challenges in the context of her work for the *The Dawn of Tibetan Buddhist Scholasticism (11th-13th)* (TibSchol) project at the Austrian Academy of Sciences, which is utilising the Transkribus platform to explore possible solutions to automate the transcription of Tibetan cursive scripts, in particular Tibetan *ume*. Transkribus is a user-friendly platform for transcribing, annotating, and searching historical manuscripts, which can be run on a local machine or in an online web interface (Muehlberger, Seaward, Terras, Oliveira, Bosch, Bryan, Colutto, Déjean, Diem, Fiel et al. 2019). She presents a detailed methodology and workflow for recognising the text fields (layout analysis) as well as different models to transcribe the text after providing the system with manual training data ("Ground Truth").

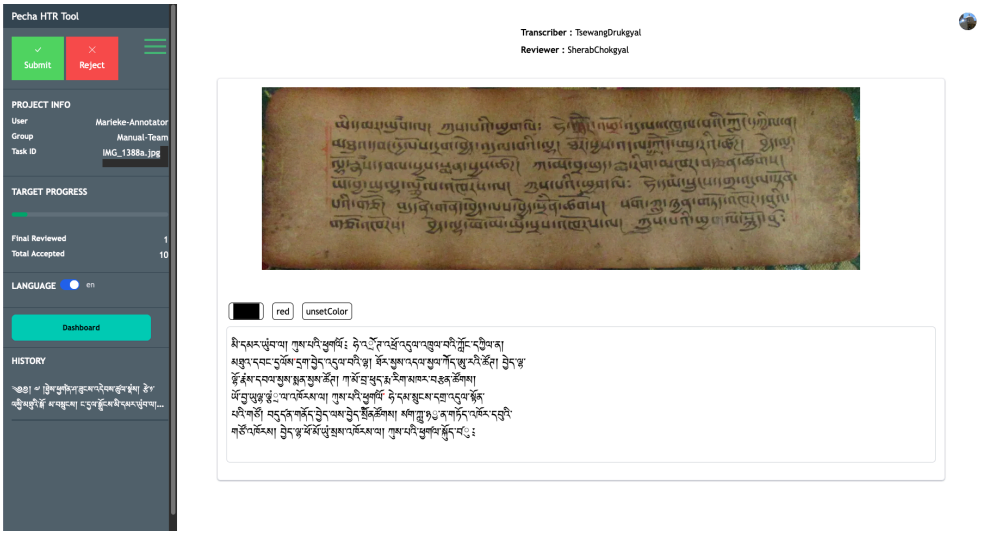

Fig. 1 – Screenshot of the OCR/HTR input view in Pecha Tools

3 https://huggingface.co/BDRC

Fig. 2 – Screenshot of the OCR/HTR correction tool with list of conventions

The results form an important step forward for anyone working with historical Tibetan documents that are handwritten, since they can form the starting point for any future HTR model that aims to transcribe any type of Tibetan cursive writing, as shown by the application of the *TibSchol* models to new materials from a very different period and region investigated in the ERC-funded PaganTibet project led by Charles Ramble. For this project, a new HTR input and correction workflow and tool was developed in collaboration with Esukhia and the BDRC to facilitate transcription. Figure 1 shows an example of an easy way to input transcriptions, while Figure 2 shows an example of the HTR tool that enables manual correction of automatically-generated transcriptions. Both tools have a built-in list of conventions for transcribing particularly difficult graphemes, ligatures and abbreviations (shown in its expanded version in Figure 2).

Finally, work on the generation of synthetic data in Tibetan cursive is undertaken by BDRC in collaboration with Esukhia & MonlamAI means future training of HTR models for these script varieties should be even more straightforward. Overall, these new methods to automatically create and correct searchable e-texts form a great leap forward in efforts to make written Tibetan materials accessible to both academics and the general public anywhere in the world through collaborations with the BDRC and its BUDA platform where new manuscript images and e-texts are (and will be) stored even after

these projects have ended.

2.2 Automatic Speech Recognition (ASR)

In addition to OCR and HTR, much progress has been made with Automatic Speech Recognition (ASR) of audio(visual) data. Developing good ASR models facilitates the creation of any Speech-to-Text (STT) or Text-to-Speech (TTS) tools. In the last decade deep-learning methods have overtaken traditional hybrid models that consisted of a lexicon with a custom phoneme set for each language, handcrafted by phoneticians. The more recent end-to-end deep-learning models, on the other hand, directly map the acoustic input onto a sequence of transcribed words without the need for force-aligned data or a language-specific lexicon.

For Modern Standard (Lhasa) Tibetan, Esukhia and MonlamAI have led efforts to collect and transcribe data that can be used to train ASR system. Their 'Tibetan Voice' data set currently contains >950 hours of Tibetan films, religious teachings, natural speech, audio books as well as children's speech. Their latest model (OpenPecha run 10) achieves results of 20.42 Character Error Rate (CER) on the benchmark.[4] Based on these, MonlamAI have also built Speech-to-Text and Text-to-Speech tools, which can transcribe recordings or produce spoken Tibetan from text online.[5]

For non-English ASR in general, but in particular for any non-standard Tibetan variety for which no large transcribed and time-aligned audio datasets exist, creating high-quality end-to-end STT systems has been challenging until recent developments in multilingual deep learning. Baevski, Zhou, Mohamed & Auli (2020) show that their Wav2Vec2 model, which learns representations from speech audio alone can outperform earlier methods when it is fine-tuned on transcribed speech in any target language. Similarly, OpenAI's Whisper trained on 680k hours of multilingual web data (about a third of which is non-English) has enabled transcription in multiple other languages. Since the proportion of English data is much larger in Whisper than in Wav2Vec2, the latter proves more successful for finetuning languages like Tibetan. Because most of the training data in these models consists of European languages, transcription in a regular, romanised script with a straightforward 1-to-1 mapping of sounds and graphemes is easier for these models. Standard (Lhasa) Tibetan audio is generally transcribed in Tibetan Unicode script, which even in its romanised (Wylie) conversion is far removed from its pronunciation. Fine-tuning of these ASR systems for languages like Tibetan is

4 This dataset and their models can be found on https://huggingface.co/openpecha.

5 Speech-to-Text: https://monlam.ai/model/stt and Text-to-Speech: https://monlam.ai/model/tts.

therefore most successful when used in combination with language-specific pre- and post-processing tools that can convert scripts and use language-specific dictionaries and spell-checkers.

The real strength of these multilingual models, however, lies in their ability to enable fine-tuning of extremely low-resource and endangered languages, like most non-standard modern Tibetan varieties. For example, O'Neill, Meelen, Coto-Solano, Phuntsog & Ramble (2023), based on earlier work on endangered languages by Coto-Solano (2021) and Coto-Solano, Nicholas, Datta, Quint, Wills, Powell & Feldman (2022), show that fine-tuning a Wav2Vec2 model for Dzardzongke (South Mustang Tibetan) can be particularly useful in language preservation, as it forms an efficient way to address the well-known transcription bottleneck in endangered language documentation (Shi, Amith, Castillo García, Sierra, Duh & Watanabe 2021). Meelen, O'Neill & Coto-Solano (2024) demonstrate that results can be further improved through transfer learning (i.e. using converted Standard Lhasa Tibetan data to enhance the dataset) as well as signal and output modifications. For example, pitch and amplitude modifications yield Character Error Rates (CER) of <10 with transcribed input of less than two hours of Dzardzongke data. Adding a post-correction dictionary (even just one built-up automatically from just three hours of input data) further improves results with a CER of 8 and a Word Error Rate (WER) of 32. These results are extremely promising for other non-standard varieties of Tibetan, especially those that are in danger of disappearing in the near future.

3. *Linguistic Annotation*

Transcription of materials is generally not sufficient for research into language variation and change, or any other form of linguistics. Especially for historical stages of the language, where native speakers are not available, it is essential to have access to well-annotated corpora to get the most out of scarce data. Linguistic annotation can also facilitate research beyond linguistics such as history and religious studies (cf. Krishna, Vidhyut, Chawla, Sambhavi & Goyal (2020) for an investigation of large, annotated religious corpora) or literature (cf. Reiter, Gius, Strötgen & Willand (2017) on performance gains in finding narratives structures when the corpus is accurately annotated). In general, good word and sentence segmentation is often essential to feed into off-the-shelf digital humanities tools for topic modelling, document classification, information retrieval, etc. (see also Section 4 below).

3.1 Normalisation, Tokenisation & Segmentation

Linguistic annotation can be done on various levels, from surface-level normalisation, tokenisation and (sentence) segmentation to mid-level morphosyntactic annotation as well as the addition of semantic and pragmatic features in deeply-annotated corpora. Especially in the absence of a reliable automatic lemmatisation tool, which groups together different inflected or conjugated forms of the same word, it can be useful to preprocess historical texts by normalising the orthography and/or standardising certain aberrant features. For Old and Classical Tibetan, for instance, this can be done using Faggionato & Garrett's 2019 Constraint Grammar Formalism. When spelling variation is the focus of research, these steps can (and should, probably) be skipped unless links to the original versions are kept, e.g. through multi-layered XML or JSON formats that preserve diplomatic transcriptions alongside normalised forms. Preprocessing and normalisation specifically, however, is very beneficial for any subsequent annotation tasks, such as segmentation, Part-of-Speech (POS) tagging or parsing.

Tokenisation (splitting into meaningful words or tokens) and sentence segmentation are non-trivial tasks in languages like Tibetan in which the script does not indicate meaningful word or sentence boundaries. Early tokenisation attempts focusing on meaningful linguistic units in particular include the syllable-tagging and recombination method developed by Meelen & Hill (2017), building on earlier work by Garrett, Hill, Kilgarriff, Vadlapudi & Zadoks (2015). Since the Tibetan script marks syllable boundaries, either by a *tsheg* ་ or by a *shad* །, these can be used to automatically split syllables. To facilitate multisyllabic meaningful units or 'words' alongside monosyllables, Meelen & Hill (2017) recast tokenisation as a syllable-tagging task with labels for beginning, middle and end syllables with a postprocessing procedure that combines these into meaningful linguistic units. More recent Tibetan tokenisers such as OpenPecha's Botok[6] use dictionaries to split a text into meaningful words or tokens. Depending on the specific Tibetan variety, the wordlist can be adjusted to get better results for different dialects. In addition to word segmentation, Botok also has the option to split sentences and/or paragraphs. Similarly, the ACTib segmenter and POS tagger[7] can do both word and sentence segmentation for Old and Classical Tibetan, using a combination of the Botok tokeniser and a syllable-tagging method with a number of post-processing rules that focus on getting meaningful linguistic segments on the word and sentence level (cf. Meelen, Roux & Hill (2021) and Faggionato, Hill & Meelen (2022)). Consistent and well-thought-through normalisation &

6 https://github.com/OpenPecha/Botok
7 https://github.com/lothelanor/actib

segmentation are often essential to improve word embeddings and large language models (like those forming the basis of tools like ChatGPT, cf. section 4 below). Sentence segmentation in particular is also crucial for the development of well-working machine translation tools.

3.2 Adding morphosyntactic and other information

For more sophisticated linguistic analyses, normalisation and segmentation of the data are not nearly enough. Adding detailed and reliable morphosyntactic information is not only useful for both synchronic, comparative and diachronic linguistic research, it can also enhance other NLP tools such as word embeddings (cf. Garcia-Silva, Denaux & Gomez-Perez (2021)).

In a series of papers, Edward Garrett, Nathan Hill and Abel Zadoks and colleagues presented one of the first attempts of adding morphosyntactic tags to each token (i.e. Part-of-Speech 'POS' tagging) of a small Classical Tibetan corpus . They used a rule-based tagger to disambiguate Tibetan verb stems Garrett, Hill & Zadoks (2013) and POS tag four Classical Tibetan texts (Garrett, Hill & Zadoks (2014) and Garrett et al. (2015)). They present a detailed set of morphosyntactic tags, going much beyond the usual set of 10-15 Universal Dependency tags[8] to facilitate more detailed linguistic research. Meelen & Hill (2017) built on this first manually-corrected, POS-tagged corpus to train a memory-based tagger, achieving 95% Global Accuracy in a ten-fold cross-validation with a tagset consisting of 79 morphosyntactic tags, which Faggionato & Meelen (2019) extend to Old Tibetan as well.[9] Meelen et al. (2021) optimise the annotation pipeline and test neural-based methods for annotation as well, while Meelen & Roux (2020a) focus on adding crucial metadata as well as constituency parses (i.e. syntactic information) to a very large diachronic corpus (Meelen & Roux 2020b). Although this corpus is not manually corrected yet, with over 166 million tokens from over 5000 texts across 11 centuries, it has opened up a wide range of research opportunities for anyone with an interest in the history of the Tibetan language. The latest version of the above-mentioned ACTib POS tagger furthermore includes additional morphological information for over 100 specific Tibetan auxiliary and light verbs to facilitate more complex diachronic linguistic research in particular.

In his contribution to this Special Issue **Christian Faggionato** demonstrates how to add syntactic information in the form of dependency parses to historical Tibetan texts. He shows how to implement a rule-based dependency parser written in the Constraint Grammar (CG-3) formalism to create the first Classical Tibetan treebank that can be part of the Universal Depen-

8 https://universaldependencies.org/u/pos/
9 https://zenodo.org/records/4727552

dencies (UD) collection.[10] Having syntactic relations encoded (either in dependency or in hierarchical phrase-structure format) facilitates more complex linguistic annotation in other domains too. Faggionato et al. (2022), for example, show how semantic and information-structural annotation can be added to Tibetan corpora by making use of an even further extended POS tag set and with syntactic annotation, Darling, Meelen & Willis (2022) show that this can be used for coreference resolution tracking noun phrases in Early Irish, which can be transferred to languages like historical Tibetan where omission of arguments is prevalent too. With categories like the animacy of noun phrases as well as the distribution of foci and topics in the sentence, more fine-grained and complex questions on the emergence of egophoric and/or switch-reference marking can be answered (cf. Meelen & Hill (2023)).

Linguistic annotation is not just beneficial to linguistics, but also forms a stronger basis for other research in a wide range of fields by enhancing opportunities for Information Retrieval, discussed in the next section.

4. Information Retrieval

Information Retrieval (IR) is the NLP task of gathering specific information from a corpus, based on any research question. It can range from simple queries like "Which people or place names can be found in which text?" to more complex tasks like Text Classification and identifying parallel content, but not necessarily identical passages in different corpora.

4.1 Named-Entity Recognition

The first question can be addressed through automatised Named-Entity Recognition (NER), i.e. adding appropriate labels to proper names, organisations, dates, institutions and other 'named entities' like 'Tashi', 'Lhasa', 'Tibetan New Year' or 'United Nations', etc. Because the Tibetan Unicode script does not use capital letters, detailed morphosyntactic annotation, e.g. distinguishing proper nouns like the personal name 'Nyima' from the regular noun 'nyima' meaning 'sun', is essential to facilitate automatic NER (cf. Suzuki, Komiya, Sasaki & Shinnou (2018)). This can be used in combination with Semantic Role Labelling (SRL), which identifies the relations with other named entities and/or verbs in the sentence, e.g. 'Tashi' is a personal name + the undergoer ('patient') of the action in the sentence 'Tashi was pushed aside by his brother.' (Zhang, Xia, Zhou, Jiang, Fu & Zhang 2022). In addition, 'his brother' can be linked to Tashi as a close family relationship. This type of information is not

10 https://universaldependencies.org/

just beneficial for linguists, but crucial for anyone doing historical research if NER and SRL annotation is provided for diachronic corpora in particular. In addition, scholars of religious studies, sociology and anthropology can benefit from this type of richly-annotated data as it enables them to extract named entities together with geographic and temporal indicators from a body of texts as well as what roles the protagonists potentially played and how they were related to each other.

For Tibetan, Liu & Wang (2018) presented one of the first NER methods, but acknowledge that the lack of good training data hindered progress. Barnett, Faggionato, Meelen, Yunshaab, Samdrup, Hill & Diemberger (2021a) and Barnett, Hill, Diemberger & Samdrup (2021b) aimed to remedy that presenting the a detailed annotation scheme of 17 Named-Entities ranging from the more conventional DATE, PLACE and PERSON to more specific TITLE, RELIGIOUS ORGANISATION and IDEOLOGY as well as a unique new set of training data consisting of almost 10k annotated terms.[11]

4.2 Text similarity & classification

A real breakthrough in the field of NLP came with development of computational methods that could 'understand', not just in frequencies, forms and structures, but also *meaning*. To get closer to letting computers gain insight into distributional semantics, i.e. deriving the meaning of words from their context, Mikolov, Yih & Zweig (2013) developed so-called 'word embeddings' using Word2Vec, a neural network-based algorithm that learns numerical representations of words. For historical languages and data sets with more orthographic variation, Meta built a character-based extension of this,[12] which Meelen (2022) used to train the first semantic model for Classical Tibetan.[13] Word embeddings are representing words for text analysis in the form of real-valued, static vectors of numbers, that can be extended to dynamic models and full-blown Language Models if enough data (i.e. billions of words with little or no orthographical variation) are available. With larger amounts of data from (Early) Modern Tibetan, for example, Engels, Erhard, Barnett & Hill (2023) developed a Language Model for SpaCy.[14]

These numerical representations of large amounts of texts are useful for more complex NLP tasks since the come closer to a semantic model of the language. Even when only smaller amounts of data are available and large language models are not an option, well-curated static word embeddings can

11 https://zenodo.org/records/4536516
12 https://fasttext.cc/
13 https://zenodo.org/records/6782247
14 https://zenodo.org/records/10148636

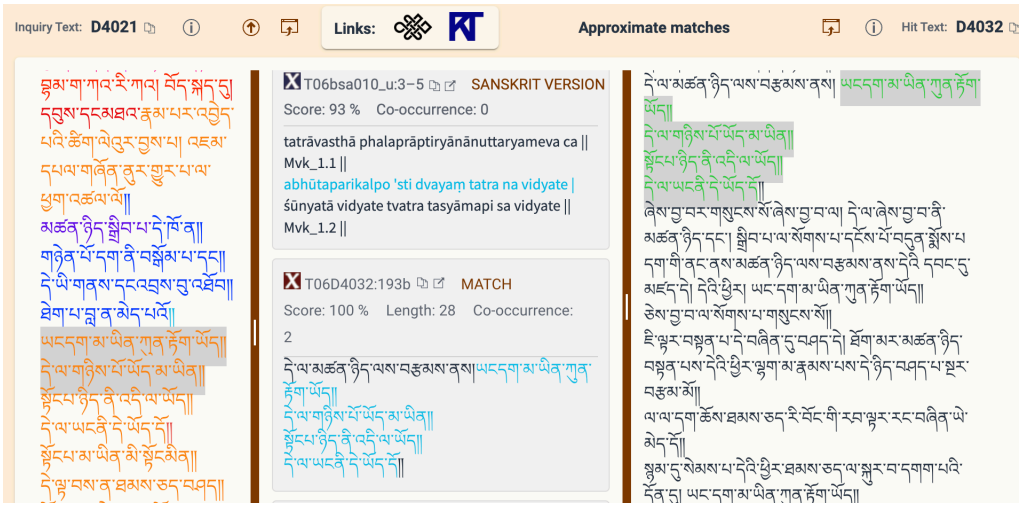

Fig. 3 – Buddhanexus text-view display with a Tibetan text on the left hand side, called inquiry text, a match in Sanskrit and a match in Tibetan in the middle column and the Tibetan text that has a match with the inquiry text on the right hand side, called hit text.

facilitate tasks like text classification, as Meelen (2022) shows in a pilot study that aims to classify texts based on different types of religious features, ranging from Buddhist to Bön and Pagan, on the other end of the scale. In addition to finding differences, both word embeddings and large language models help Information Retrieval tasks like finding textual similarity. Felbur, Meelen & Vierthaler (2022) show that if Classical Tibetan embeddings are combined with Classical Chinese models, it is possible to retrieve similar passages from Tibetan canonical texts based on Chinese input.

On a larger and user-friendly scale, the Khyentse Center for Tibetan Textual Scholarship, University of Hamburg launched the BuddhaNexus platform shown in Fig. 3.[15] This is a text-matching database that provides advanced functionality for the research of intertextual matches in Pāli, Sanskrit, Tibetan, and Chinese. BuddhaNexus provides not only a huge dataset of more than 120 million textual matches, but also advanced filter and different display options to make working with such a large dataset as comfortable as possible.

On the technical level, BuddhaNexus uses word embeddings in combina-

15 https://buddhanexus.kc-tbts.uni-hamburg.de/

tion with *k* nearest neighbor search (kNN) in order to calculate intertextual links within a large textual database efficiently (cf. Nehrdich (2020)). BuddhaNexus also features bilingual matches between Sanskrit texts and their Tibetan translations. These matches have been created using a combination of contextual word embeddings of the BERT type and vector-based sentence alignment (cf. Nehrdich (2022)). Since its launch, a number of studies have appeared that used the unique capabilities of BuddhaNexus in order to conduct intertextuality research on a scale that was not possible before (e.g. Dorjee (2021), Almogi (2022a), Almogi (2022b), Cheung (2023), and Nehrdich (2023)). In addition to researchers at western institutions, Buddhanexus is also used extensively by Tibetan khenpos.

5. *Machine Translation*

When it comes to automatic translation, in April 2023, the Dharmamitra project started at the Berkeley AI Research Lab (BAIR), UC Berkeley, under the guidance of Kurt Keutzer together with Sebastian Nehrdich. The goal of Dharmamitra is the creation of a range of Large Language Model enabled tools to help academics and translators interact with material in Tibetan and other Classical Asian languages. These include machine translation, semantic search, and Question/Answer systems. The first outcome of the Dharmamitra project, the Monlam-Mitra machine translation model, was made public through the MonlamAI website in Autumn 2023.[16]. This model is based on Meta's *No Language Left Behind* (NLLB Team 2022) Large Language Model, which was then finetuned on more than two million Tibetan-English sentence pairs collected by the Monlam AI team. Figure 4 shows the Monlam AI website providing Tibetan to English machine translation. In February 2024, the Dharmamitra team created a new version based on MADLAD-400 (Kudugunta, Caswell, Zhang, García, Choquette-Choo, Lee, Xin, Kusupati, Stella, Bapna & Firat 2023), which, in addition to English, also supports translation from Sanskrit, Chinese and Pāli into and from Tibetan. Figure 5 shows how this model provides translations from Classical Buddhist Chinese into English. The Tibetan-English model is currently accessible in interactive applications.[17]

A natural question is: What does this tool offer its users? Before its public release, feedback on Monlam-MITRA was solicited from twenty experienced translators and scholars. There is a natural positive confirmation bias in those who responded, as those who saw no particular value were unlikely to respond. First it should be made clear that no one indicated that the resulting machine translations were immediately usable. Everyone felt errors had to be

16 https://monlam.ai/

17 See monlam.ai and dharmamitra.org for the multilingual model.

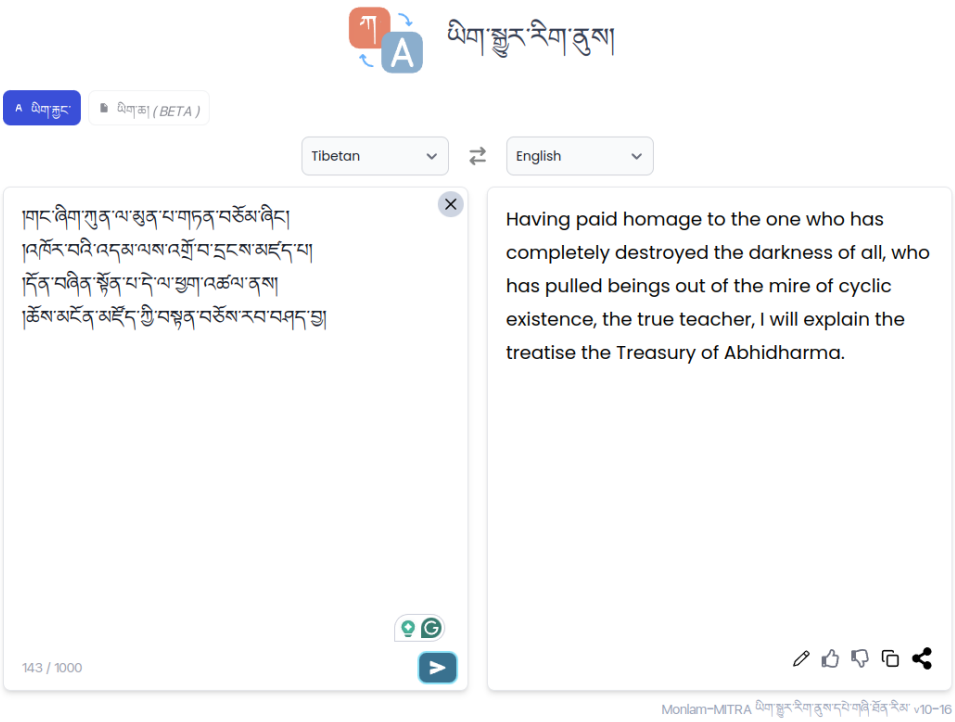

Fig. 4 – monlam.ai machine translation from Tibetan into English.

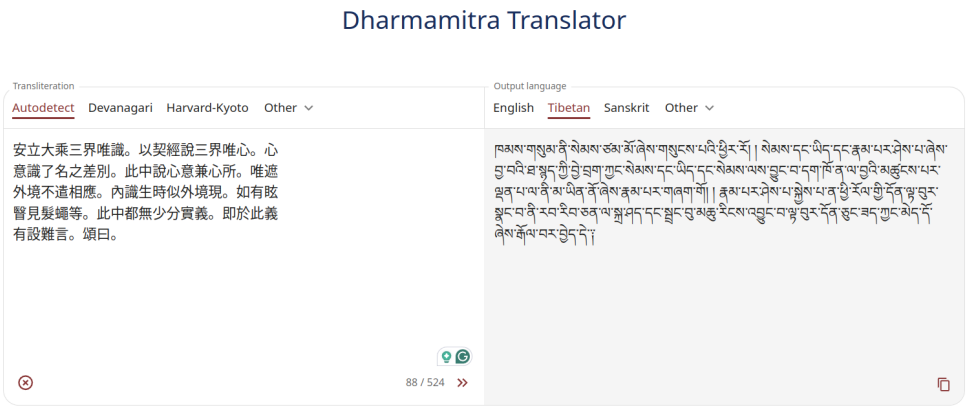

Fig. 5 – dharmamitra.org machine translation from Chinese into Tibetan.

corrected and translations reworked. Nevertheless, many translators felt that the initial translation gave a great point of departure. Surprisingly, even experienced translators found consistent productivity gains using the Monlam-MITRA translation tool. They found it very easy to quickly identify mistakes in the machine translation and move on. In addition to being faster in translation, scholars found that the interactive translation capability, with instantaneous results, made it easy to navigate unfamiliar literature. One translator, working closely with ethnic Tibetan translators, observed that Tibetans with modest English skills greatly benefited from the tool as well, making it easier for them to obtain a draft translation in the target language. Another encouraging sign of use is that there are 20,000 individual translations done per day now.

Regarding the shortcomings of the system, the most recurrent errors were incomplete translations (i.e. portions of the original entirely missing in the translation), oscillatory hallucinations in which portions of translations are repeated in a fugue-like manner (e.g. "May all sentient beings find comfort and joy. May all sentient beings find comfort and joy"). Translations of material with missing training data, such as particular types of Tibetan medicine, were non-sensical. All of these issues are fairly straightforwardly addressable, however, and improvements have been seen even over the last couple months.

6. *Learning & Editing*

In addition to the above-mentioned NLP applications, Tibetan studies is also moving forward in areas of language learning as well as text collation and editing. In his contribution to this Special Issue, **Dirk Schmidt** introduces *Dakje*,[18] a Tibetan education tool. Since pronunciation of most Modern Tibetan varieties is far removed from the orthography due to diglossia in the spoken and written language (Ferguson 1959), he explains that learning how to read and write Tibetan is a challenging task, with high rates of illiteracy from 21% to 34% (Reddy & Bhole 2023, Textor 2022). Building on earlier work (Schmidt 2020), he shows how *Dakje* aims to address these issues by unpacking the learning-to-read process in Tibetan step-by-step. Using two Esukhia speech corpora (the Nanhai Corpus (Esukhia 2020) and The Children's Speech Corpus (Esukhia 2022a)), he proposes a new data collection technique for writing beginning reading material, the creation of story-specific 'mini speech corpora', and how to use word segmentation in both *Dakje* (Esukhia 2022b) and *Botok* (OpenPecha 2023) for editing and reading level identifica-

18 https://dakje.io/

tion. This provides readers who wish to write, edit, or analyse early reading materials with the practical information, tools, and resources they need.

When it comes to reading and editing historical texts in digital environments, recent years have seen progress here too in a variety of ways. First, an ever-increasing number of digital images and eTexts are available through the Buddhist Digital Resource Center's (BDRC) BUDA platform.[19] These texts are not only provided with detailed metadata, but also linked to other platforms through IIIF and Linked Open Data (LOD) protocols. In addition, BUDA facilitates searching not just for textual strings, but also for people, places, topics, collections, works, etc. making it the largest digital library of Buddhist and other Tibetan material in the world.

When it comes to creating editions of texts, the BDRC and Esukhia have collaborated to develop *Pydurma*, a tool that can create a clean e-text version of any Tibetan work from multiple sources.[20] It can very efficiently collate different texts (in multiple formats) and has a configurable system of weights to choose a preferred reading as the one presented in the diplomatic edition (most common reading, best OCR confidence index, conformance to spelling standards, etc.). The result is a new clean version called a 'vulgate edition'. This is useful for publishers who need clean copies of texts, but also developers who need clean data to train new computational models.

7. Conclusion and further developments

Computer-aided methods in the Humanities and Social Sciences based on Natural Language Processing (NLP) techniques have led to considerable breakthroughs in recent years. The field of Tibetan Studies has already seen a number of changes as more state-of-the-art tools have become available not just to those with a technical background, but also as user-friendly online applications that facilitate research. This brief introduction and the Special Issue in general are not meant to be exhaustive, but merely aims to provide a taste of the wide range of opportunities these tools have to offer.

Bibliography

Almogi, Orna. 2022a. Editors as canon-makers: The formation of the tibetan buddhist canon in the light of its editors' predilections and agendas. In Orna Almogi (ed.), *Evolution of scriptures, formation of canons: The buddhist case*, vol. 13 Indian and Tibetan Studies Series, 351–458. Hamburg: Department of Indian and Tibetan Studies, Universität Hamburg.

19 https://library.bdrc.io/
20 https://github.com/buda-base/pydurma

Almogi, Orna (ed.). 2022b. *Evolution of scriptures, formation of canons: The buddhist case*, vol. 13 Indian and Tibetan Studies Series. Hamburg: Department of Indian and Tibetan Studies, Universität Hamburg.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.

Baird, Fossey Henry S., Henry S. & P. Lofting. 1990. The typestyle jockey: Putting the horse out front in Devanagari and Tibetan. In *Nordic Institute of Asian Studies report*, 5–30.

Barnett, Robert, Christian Faggionato, Marieke Meelen, Sargai Yunshaab, Tsering Samdrup, Nathan Hill & Hildegard Diemberger. 2021a. *Named Entity Recognition (NER) for Tibetan and Mongolian Newspapers*. Poster presented at the Cambridge Language Sciences Symposium. doi:10.33774/coe-2021-xhw9l-v2.

Barnett, Robert, Nathan Hill, Hildegard Diemberger & T Samdrup. 2021b. Named-Entity Recognition for Modern Tibetan Newspapers: Tagset, Guidelines and Training Data [Data set]. doi:https://doi.org/10.5281/zenodo.4536516.

Cheung, Daisy. 2023. "Madhyamakanising" Tantric Yogācāra: The Reuse of Ratnākaraśānti's Explanation of maṇḍala Visualisation in the Works of Śūnyasamādhivajra, Abhayākaragupta and Tsong Kha Pa. *Journal of Indian Philosophy* 51(5). 611–643.

Coto-Solano, Rolando. 2021. In *Proceedings of the first workshop on natural language processing for indigenous languages of the americas. association for computational linguistics.*, 173–184. Explicit tone transcription improves ASR performance in extremely low-resource languages: A Case Study in Bribri.

Coto-Solano, Rolando, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell & Isaac Feldman. 2022. Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori. In *Proceedings of the thirteenth language resources and evaluation conference*, 3872–3882. https://aclanthology.org/2022.lrec-1.412.

Darling, Mark, Marieke Meelen & David Willis. 2022. Towards coreference resolution for Early Irish. In *Proceedings of the CLTW 4 @ LREC2022*, 85–93. European Language Resources Association (ELRA).

Dorjee, Khenpo Tashi. 2021. *Chos bzang rigs pa'i rnam dpyod | Rong zom ma hā paṇḍita'i theg chen tshul 'jug rjod byed zhib dpyad zhu dag lung khungs ngos 'dzin dang | brjod bya gnas lugs rig par rtsad zhib tshom bu |*, vol. 3 sNga 'gyur rnying ma'i zhib 'jug. Bylakuppe, Mysore: Ngagyur Nyingma Institute, Ngagyur Nyingma Research Centre.

Drup, Ngo, Dongcai Zhao, Puts Ren, Daluo Sanglangjie, Fang Liu & Bian Bawangdui. 2010. Study on printed Tibetan character recognition. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, 280–285. IEEE.

Engels, James, Xaver Erhard, Robert Barnett & Nathan Hill. 2023. Tibetan for Spacy 1.1 [Data set]. https://doi.org/10.5281/zenodo.10148636.

Esukhia. 2020. The nanhai corpus. https://github.com/Esukhia/Corpora/tree/master/Nanhai.

Esukhia. 2022a. Children's Stories Speech Corpus. https://github.com/Esukhia/Corpora/tree/master/Childrens_Stories.

Esukhia. 2022b. Dakje. https://github.com/Esukhia/dakje-desktop.

Faggionato, Christian & Edward Garrett. 2019. Constraint Grammars for Tibetan Language Processing. In *Nealt proceedings series 33:3*, 12–16.

Faggionato, Christian, Nathan Hill & Marieke Meelen. 2022. 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*, 1–6. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.eurali-1.1.

Faggionato, Christian & Marieke Meelen. 2019. Developing the Old Tibetan treebank. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 304–312. Varna, Bulgaria: INCOMA Ltd. doi:10.26615/978-954-452-056-4_035.

Felbur, Rafal, Marieke Meelen & Paul Vierthaler. 2022. Crosslinguistic Semantic Textual Similarity of Buddhist Chinese and Classical Tibetan. *Journal of Open Humanities Data* doi:10.5334/johd.86.

Ferguson, Charles A. 1959. Diglossia. *WORD* 15(2). 325–340. doi:10.1080/00437956.1959.11659702.

Garcia-Silva, Andres, Ronald Denaux & Jose Manuel Gomez-Perez. 2021. On the impact of knowledge-based linguistic annotations in the quality of scientific embeddings. *Future Generation Computer Systems* 120. 26–35.

Garrett, Edward, Nathan W Hill, Adam Kilgarriff, Ravikiran Vadlapudi & Abel Zadoks. 2015. The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries. *Revue d'Études Tibétaines* 32. 51–86.

Garrett, Edward, Nathan W Hill & Abel Zadoks. 2013. Disambiguating Tibetan verb stems with matrix verbs in the indirect infinitive construction. *Bulletin of Tibetology* 49(2). 35–44.

Garrett, Edward, Nathan W Hill & Abel Zadoks. 2014. A rule-based part-of-speech tagger for Classical Tibetan. *Himalayan Linguistics* 13(2).

Hedayati, Fares, Jike Chong & Kurt Keutzer. 2011. Recognition of tibetan

wood block prints with generalized hidden markov and kernelized modified quadratic distance function. In *Proceedings of the 2011 joint workshop on multilingual ocr and analytics for noisy unstructured text data*, 1–14.

Krishna, Amrith, Shiv Vidhyut, Dilpreet Chawla, Sruti Sambhavi & Pawan Goyal. 2020. SHR++: An interface for morpho-syntactic annotation of Sanskrit corpora. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 7069–7076.

Kudugunta, Sneha, Isaac Caswell, Biao Zhang, Xavier García, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna & Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *ArXiv* abs/2309.04662. https://api.semanticscholar.org/CorpusID:261682406.

Liu, Fei-Fei & Zhi-Juan Wang. 2018. Active Learning for Tibetan Named Entity Recognition based on CRF. In Jinhua Du & Mihael Arcan (eds.), *Lrec 2018 workshop mlp–moment*, 18–45.

Meelen, Marieke. 2022. Tibetan word embeddings: from distributional semantics to facilitating Tibetan NLP. *International Association of Tibetan Studies - Tech Panel presentation* .

Meelen, Marieke & Nathan Hill. 2017. Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics* 16(2). 64–89.

Meelen, Marieke & Nathan Hill. 2023. From co-reference to evidentiality: how syntax, semantics and information structure interact to create a new grammatical feature. *Himalayan Languages Symposium* doi:10.13140/RG.2.2.22106.72646.

Meelen, Marieke, Alexander O'Neill & Rolando Coto-Solano. 2024. ASR for endangered languages in Nepal. In *Proceedings of the Comput-EL workshop at the EACL*, 83–93.

Meelen, Marieke & Élie Roux. 2020a. Meta-dating the PArsed Corpus of Tibetan (PACTib). In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, 31–42.

Meelen, Marieke & Elie Roux. 2020b. The Annotated Corpus of Classical Tibetan (ACTib) - Version 2.0 (Segmented & POS-tagged) [Data set]. doi:https://doi.org/10.5281/zenodo.3951503.

Meelen, Marieke, Élie Roux & Nathan Hill. 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* 20(1). 1–11. doi:10.1145/3409488.

Mikolov, Tomáš, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational lin-*

guistics: Human language technologies, 746–751.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, Hervé Déjean, Markus Diem, Stefan Fiel et al. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of documentation* 75(5). 954–976.

Nehrdich, Sebastian. 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.

Nehrdich, Sebastian. 2022. SansTib, a Sanskrit - Tibetan Parallel Corpus and Bilingual Sentence Embedding Model. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the thirteenth language resources and evaluation conference*, 6728–6734. Marseille, France: European Language Resources Association.

Nehrdich, Sebastian. 2023. Observations on the intertextuality of selected abhidharma texts preserved in chinese translation. *Religions* 14(7).

NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv* abs/2207.04672. https://api.semanticscholar.org/CorpusID:250425961.

O'Neill, Alexander, Marieke Meelen, Rolando Coto-Solano, Sonam Phuntsog & Charles Ramble. 2023. Language Preservation through ASR. doi:10.33774/coe-2023-rm6wq-v2.

OpenPecha. 2023. Botok. https://github.com/OpenPecha/Botok.

Reddy, Rahul K. & Omkar Bhole. 2023. Analysing China's Census Report .

Reiter, Nils, Evelyn Gius, Jannik Strötgen & Marcus Willand. 2017. A Shared Task for a Shared Goal: Systematic Annotation of Literary. In *Digital humanities*, .

Rowinski, Zach & Kurt Keutzer. 2016. Namsel: An optical character recognition system for Tibetan text. *Himalayan Linguistics* 15(1). 12–30.

Schmidt, Dirk. 2020. Grading Tibetan Children's Literature: A Test Case Using the NLP Readability Tool "Dakje". *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19(6). doi:10.1145/3392046. https://doi.org/10.1145/3392046.

Shi, J., J. D. Amith, R. Castillo García, E. G. Sierra, K. Duh & S. Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yolóxochitl Mixtec. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* 1134–

1145.

Suzuki, Masaya, Kanako Komiya, Minoru Sasaki & Hiroyuki Shinnou. 2018. Fine-tuning for named entity recognition using part-of-speech tagging. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, 632–640.

Textor, C. 2022. Illiteracy rate in China in 2021, by region. https://www.statista.com/statistics/278568/illiteracy-rate-in-china-by-region/.

Wang, Hua & Xiaoqing Ding. 2003. New statistical method for multifont printed Tibetan/English OCR. In *Document recognition and retrieval xi*, vol. 5296, 155–165. SPIE.

Zhang, Yu, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu & Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4212–4227.