

Towards the *Roof of the World*: NLP Progress in Tibetan Languages

Xi Cao[♡1,♡3,♠], **Yijie Li**[♡2,♠], **Yuan Sun**[♡2,♡3,♠,✉],
Shiwei Dong[♢,♣], **Zeguang Huang**[♢,♣], **Nuo Qun**[♢,♣]
[♡1]School of Philosophy and Religious Studies, Minzu University of China
[♡2]School of Information Engineering, Minzu University of China
[♡3]Institute of National Security, Minzu University of China
[♠]National Language Resource Monitoring & Research Center | Minority Languages Branch
[♢]School of Information Science and Technology, Xizang University
[♣]The State Key Laboratory of Tibetan Intelligence
caoxi@muc.edu.cn, sunyuan@muc.edu.cn

Abstract

Tibetan, spoken by some seven million people on and around the Qinghai-Xizang Plateau, remains underrepresented in natural language processing (NLP). We present, to our knowledge, the first peer-reviewed systematic bibliometric survey of Tibetan NLP: 140 publications from 2004 to 2026 in the ACL Anthology and DBLP, annotated for venue type, research area, first-affiliation country, and resource release. Output has grown rapidly, with 84% of the publications appearing since 2018, and has expanded bottom-up along the NLP pipeline, from corpora and word segmentation through script, speech, and applications to large language models (LLMs). Researchers based in China first-authored 87% of the publications, which span all 25 represented research areas, whereas work from other countries concentrates on Classical Tibetan. We argue that this full-pipeline paradigm, realized under the compounded constraints of a complex script, dialectal diversity, and geographic remoteness, offers a reference model for other low-resource languages of the Global South. We also identify gaps in dialect coverage, responsible NLP, native and comparable evaluation, cross-community visibility, and resource release (only 29% of the publications release data, code, or models), and recommend open science, shared standards, and community participation for the LLM era.

1 Introduction

The benefits of NLP remain unevenly distributed: most of the world's languages have little or no labelled data (Joshi et al., 2020), and LLMs, whose capabilities depend on each language's digital text, threaten to widen this gap. The risk is compounded in remote highlands, where geographic isolation, dialectal fragmentation, and limited digital infrastructure add to data scarcity; the Qinghai-Xizang Plateau, the world's highest, known as the *Roof of the World* and the *Third Pole*, is an extreme instance.

Tibetan, the principal language of the plateau, is spoken by some seven million people in the Xizang Autonomous Region and the provinces of Qinghai, Sichuan, Gansu, and Yunnan in China, and by communities in Pakistan, India, Nepal, and Bhutan. Its script does not mark word boundaries, its spoken varieties differ substantially, and its digital text is scarce (§2.1); in the taxonomy of Joshi et al. (2020), it belongs to Class 1 ("The Scraping-Bys"), with some unlabelled but almost no labelled data, a class that, the authors argue, needs "a solid, organized movement" to collect labelled data. Tibetan thus tests whether such a movement can bring a low-resource language into the LLM era.

Although Tibetan NLP has been studied for more than two decades, its literature is fragmented across communities and venues: 103 of the 140 publications we identify (74%) lie outside the ACL Anthology, the sole source of a recent survey of Sino-Tibetan languages (Liu and Best, 2025), and the closest Tibetan-specific overview is a brief editorial (Long and Hill, 2021). We address this gap with a bibliometric survey of these publications (§3), guided by four research questions:

RQ1. How has Tibetan NLP output evolved over time and across venues? (§4.1)

RQ2. Which research areas have received the most and least attention, and how have priorities shifted? (§4.2)

RQ3. Which countries have contributed, and what role has China played? (§4.3)

✉ Corresponding Author

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1088–1119
November 6-10, 2026 ©2026 Association for Computational Linguistics

RQ4. What are the key open challenges, and what could accelerate progress? (§§5–6)

Our central finding is that Tibetan NLP exemplifies what we call a *full-pipeline paradigm*: starting from annotated corpora, lexicography, and word segmentation, the field has successively added script and speech processing, applications, and, since 2022, pre-trained and large language models, while continuing to build resources. We argue that this paradigm offers a reference model for other low-resource languages of the Global South, identify the gaps to be closed in the LLM era, and release our annotated bibliography with an audit of publicly released resources.

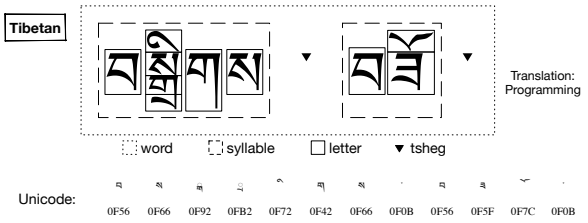


Figure 1: The Tibetan word བཀྲམ་བདེ་ (*programming*), illustrating the three-level structure of Tibetan script: letters (30 consonants and 4 vowel signs, each encoded by a Unicode code point, shown at the bottom) combine into syllables according to strict orthographic rules, and syllables, each followed by the inter-syllable mark *tsheg* (་), form words whose boundaries are not marked.

2 Background

2.1 The Tibetan Language

Tibetan belongs to the Tibeto-Burman branch of the Sino-Tibetan language family. Written continuously since the seventh century, its literary language, rooted in Classical Tibetan, serves as a shared written standard across three main dialect groups, Ü-Tsang, Kham, and Amdo, whose spoken forms differ so much in phonology that mutual intelligibility is limited (Schmidt, 2020; Ren et al., 2026); even lexical tone varies, from tonal Ü-Tsang through Kham to largely atonal Amdo (Liang and Zerong, 2026). We use *Tibetan* collectively for the classical and modern written language and its spoken varieties, often classified as distinct Tibetic languages (hence the plural in our title). This diglossia challenges NLP: models trained on written text may transfer poorly to speech, and variety-specific systems generalize poorly.

As Figure 1 shows, Tibetan script is an abugida in which 30 consonant letters and 4 vowel signs (Liu et al., 2014) combine under strict orthographic rules to form syllables. Syllables are delimited by the *tsheg*, but word boundaries are not marked, so word segmentation has long come first in text processing (Liu et al., 2011), although neural models can operate directly on syllables (Lai et al., 2018). Even syllable boundaries can be implicit: several case particles fuse with the preceding syllable and resemble ordinary final letters, a major source of errors in segmentation, syllable analysis, and subword machine translation (MT) (Li and Yu, 2013; Zhu et al., 2019; Gyatso et al., 2025). Encoding raises further problems: unlike Unicode, Chinese national extension standards encode whole stacks, 7,198 in all, as single characters (Liu et al., 2014; Wang et al., 2019); stacks may be stored in variant code-point orders (Faggionato and Meelen, 2019); and replacing the *tsheg* with spaces degrades fine-tuned models (Cao et al., 2023). Tibetan is verb-final with ergative marking (An and Long, 2021), case-particle forms depend on the preceding syllable's final letter (Khysru et al., 2021), and verbs have up to four stems (Khysru et al., 2019); its conservative orthography retains prefixed, superscribed, and subscribed letters and leaves Lhasa tones unwritten (Yang et al., 2015), both of which grapheme-to-phoneme conversion must model. Together with scarce digital text, each property has spawned a sustained line of research (§4.2).

2.2 Related Surveys

Surveys of geographically defined, under-resourced language groups, such as African languages (Adebara and Abdul-Mageed, 2022; Alabi et al., 2025), the indigenous languages of Latin America (Tonja et al., 2024), and the languages of India (Kumar et al., 2026), inform our design. Our survey differs from the two closest overviews of Tibetan NLP in scope, coverage of the LLM era, sources, and annotation (Table 1).

3 Methodology

Scope and inclusion criteria. We include peer-reviewed publications in which Tibetan, modern or classical, is the primary language of study or a principal experimental object, and whose contribution concerns computational processing of Tibetan text, speech, or script; multilingual studies qualify when Tibetan is among their principal experimental languages. Coverage runs from 2004, the year of the earliest Tibetan NLP paper in our sources, through 00:00 UTC, September 25, 2026.

Feature	Our Survey	NAACL Survey (Liu and Best, 2025)	TALLIP Editorial (Long and Hill, 2021)
Scope	**Tibetan only**	Sino-Tibetan family (49 low-resource languages)	Tibetan only
Coverage Period	**2004–2026**: from the earliest identified paper through 00:00 UTC, September 25, 2026	1952–2022: the ACL OCL corpus (Rohatgi et al., 2023); venue trends for 2000–2021	2016–2021: editorial for a special issue of *ACM TALLIP*
Corpus	**140 publications**, annotated for venue type, research area, country, and resource release	253 papers (274 language–paper matches for 23 of the 49 languages), annotated for topic	No fixed count: narrative overview of key developments
Sources	**Hybrid**: ACL Anthology BibTeX and DBLP Search API	ACL Anthology only, via the ACL OCL corpus	Selected by the special-issue editors

Table 1: Comparison of our survey with the two most closely related overviews of Tibetan NLP.

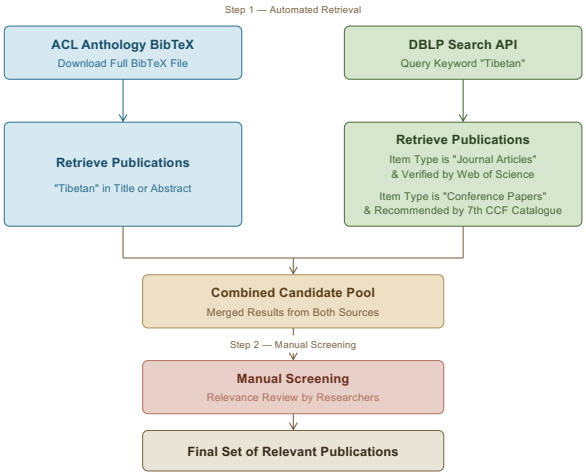


Figure 2: Two-step construction of the survey corpus. **Step 1 (automated retrieval):** entries with *Tibetan* in the title or abstract are retrieved from the full ACL Anthology BibTeX file, and publications matching the keyword *Tibetan* are retrieved through the DBLP Search API, retaining journal articles verified by Web of Science and conference papers from venues recommended in the 7th edition of the CCF catalogue. **Step 2 (manual screening):** the merged and deduplicated candidate pool is reviewed for relevance, yielding the final set of 140 publications.

Retrieval and screening. We proceeded in two stages (Figure 2). First, we retrieved all entries with the keyword *Tibetan* in the title or abstract from the ACL Anthology's full BibTeX export[1]. To cover research outside it, we queried the DBLP Search API[2] with the same keyword and, to restrict results to established peer-reviewed venues, retained only Web of Science[3] indexed journal articles and papers from international conferences recommended in the 7th edition of the China Computer Federation (CCF) catalogue[4]. Second, four researchers independently screened the deduplicated candidates against the inclusion criteria, considering only English-language papers and resolving disagreements by discussion. The final corpus has 140 publications: 37 from the ACL Anthology (25 conference and 12 workshop papers; Appendix A) and 103 from DBLP (29 conference papers, 74 journal articles; Appendix B).

Annotation. Each publication was annotated along three dimensions: (1) *venue type* (conference, workshop, or journal); (2) *country* of the first author's primary affiliation (Appendix C); and (3) *research area*: up to two of the 29 areas of the ACL Rolling Review (ARR) taxonomy[5] (Appendix D). In practice, 115 publications received two areas and 25 received one, yielding 255 area assignments, so area counts sum to more than 140. We also recorded whether the authors release code, data, or models at a public URL (Appendix F).

4 Results

4.1 Publication Trends by Year and Venue

Annual output (Figure 3) was sporadic for nearly two decades, with three papers in 2004–2006, none in 2007–2009, and 19 in 2010–2017, before rising sharply: 118 publications (84.3%) date from 2018 onward, and 61 (43.6%) from 2024–2026 alone, although 2026 is only partially covered. This growth coincides with the wider field's adoption of deep learning and, later, of pre-trained Transformers and LLMs. Several annual peaks, however, reflect clusters in single venues: five of the nine papers of 2018 appeared at PRCV 2018, seven of the 13 of 2019 in *IEEE Access*, eight of the ten of 2021 in TALLIP, six of them in the special issue introduced by Long

[1] https://aclanthology.org/faq/bib/
[2] https://dblp.org/faq/13501473.html
[3] https://www.webofscience.com
[4] https://www.ccf.org.cn/Academic_Evaluation/By_category/
[5] https://aclrollingreview.org/areas

and Hill (2021), and five of the 21 of 2026 in one issue of *Data Intelligence*. Dedicated venues have thus shaped growth, and yearly fluctuations should not be over-interpreted.

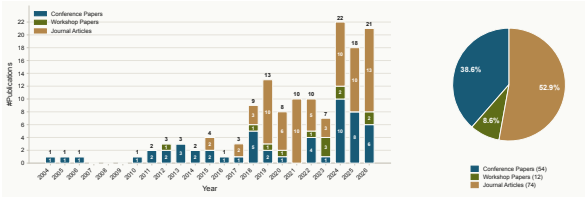


Figure 3: Annual number of Tibetan NLP publications by venue type (left; stacked bars) and overall share of each venue type (right; counts in the legend), 2004–2026. The year 2026 is only partially covered. The full-scale version is Figure 7 in Appendix E.

Journal articles form the largest share of the corpus (74; 52.9%), followed by conference papers (54; 38.6%) and workshop papers (12; 8.6%), dispersed across 10 workshops, 22 conferences, and 25 journals, led by TALLIP (19 articles) and *IEEE Access* (14 articles) (Figure 4). Only 37 publications (26.4%) appear in the ACL Anthology, but 17 of these date from 2024–2026, including papers at ACL, EMNLP, AACL, COLING, and LREC: Tibetan NLP is entering the core NLP venues, although most of its literature circulates elsewhere.

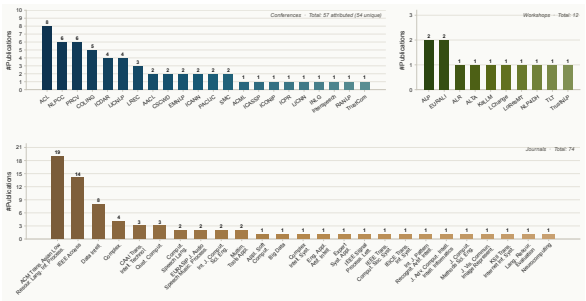


Figure 4: Number of publications per conference (top left), workshop (top right), and journal (bottom). Papers at joint conferences (ACL-IJCNLP 2015, IJCNLP-AACL 2025) are counted under each constituent venue, so the conference counts sum to 57 for 54 papers. The full-scale version is Figure 8 in Appendix E.

4.2 Research Areas

The 255 area assignments (Figures 5, left, and 6) cover 25 of the 29 ARR areas but are highly skewed: Multilingualism and Cross-Lingual NLP (41 publications) and Resources and Evaluation (40) account for 31.8% of all assignments and the five largest areas for 56.1%, whereas 11 areas have between one and four publications and four (Dialogue and Interactive Systems; Ethics, Bias, and Fairness; LLM Agents; and NLP and Code Models) have none. We discuss them in four groups corresponding broadly to layers of the NLP pipeline.

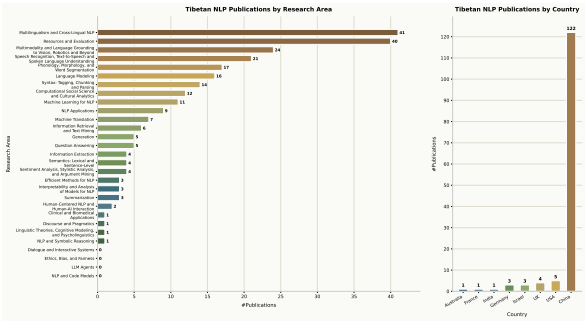


Figure 5: Number of publications per ARR research area (left; descending order) and per country of the first author's affiliation (right; ascending order). As a publication can belong to more than one area, area counts sum to 255. The full-scale version is Figure 9 in Appendix E.

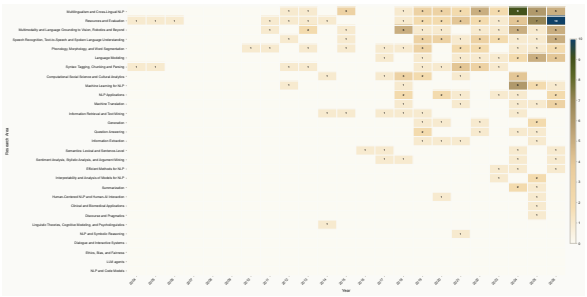


Figure 6: Number of publications per ARR research area and year, 2004–2026; darker cells indicate more publications. Areas are ordered by total count (descending). The full-scale version is Figure 10 in Appendix E.

Foundations: resources, segmentation, and syntax. Resources and Evaluation is the most persistent area, active in 16 of the 20 years with any Tibetan NLP publication, from a syntactically annotated corpus and lexicographic resources (Wagner and Zeisler, 2004; Jiang et al., 2005; Jiang, 2006) to the training and evaluation data of the LLM era (Zhang et al., 2024, 2025a; Gao et al., 2025; Zhuang et al., 2025a; Zhuang and Sun, 2025; Li et al., 2026). Corpus scale has grown from 35M syllables of news text (Liu et al., 2012) to the 166M-token diachronic PACTib (Meelen and Roux, 2020) and 72 GB for LLM pre-training (Yang et al., 2026), and 17 of its 40 publications appeared in 2025–2026. Phonology, Morphology, and Word Segmentation (17) centres on the absence of word boundaries. For lack of a public segmented corpus, early work

trained a conditional random field segmenter on machine-segmented text (Liu et al., 2011), and Chinese national standards for Tibetan word segmentation and part-of-speech tags followed only in 2018 (Zhu et al., 2019; Li et al., 2022b). The area has since moved from syllable tagging (Liu et al., 2011; Li and Yu, 2013) and unknown-word identification (Nuo et al., 2015) to morphology-aware language models (Khysru et al., 2019), subword methods for MT (Li et al., 2021; Gyatso et al., 2025), joint segmentation and part-of-speech tagging (Li et al., 2022b), neural segmentation (Brown, 2024), and segmentation evaluation (Zhuoma and Zhijie, 2026). Syntax: Tagging, Chunking and Parsing (14) follows the opposite trajectory: anchored in treebanks of Old and Classical Tibetan (Faggionato and Meelen, 2019; Meelen and Roux, 2020) and in dependency parsing (An and Long, 2021; An, 2023b), it has no publications after 2023, mirroring its estimated decline across the ACL Anthology from 20% to 2% of papers in 2000–2021 (Rohatgi et al., 2023).

Script and speech. Nearly a third of the corpus (45 publications; 32.1%) processes visual or acoustic input. Multimodality and Language Grounding (24) is dominated by script recognition: online and offline handwriting (Ma et al., 2011; Ma and Wu, 2015; Cai and Wang, 2018; Wang et al., 2019; Li et al., 2024; Tashi et al., 2024; Zhou et al., 2026), historical documents in woodblock and manuscript styles (Wang et al., 2018b; Zhao et al., 2018; Wang et al., 2018a; Li et al., 2022c; Zhao et al., 2024), scene text (Ke et al., 2025; Lhundrub et al., 2026), and, recently, vision–language and multimodal optical character recognition (OCR) resources (Xu et al., 2026; Jie et al., 2026). Speech Recognition, Text-to-Speech and Spoken Language Understanding (21) has progressed from isolated-word recognition with at most 20 speakers (Zhao et al., 2012) to end-to-end recognition for the Lhasa and Amdo varieties and in multi-dialect settings (Zhao et al., 2019b; Qin et al., 2022b; Liu et al., 2023; Zhu and Huang, 2020; Liu et al., 2024), alongside speech synthesis (Yang et al., 2015; Zhao et al., 2019a; Xu et al., 2021; Zhang et al., 2026b), speech translation (Fan et al., 2026; Ren et al., 2026), and word-level evaluation (Moore et al., 2026). A recurring strategy exploits the compositional structure of the script: decomposing stacks into components reduces 562 handwritten character classes to 120 components (Ma and Wu, 2013, 2015) and aids few-shot printed-text recognition (Han et al., 2018), and sub-syllable radicals serve as units for language modelling and speech recognition (Shen et al., 2017; Qin et al., 2022b,a). Both areas' prominence plausibly reflects a large manuscript and print heritage and spoken varieties poorly served by written resources.

Cross-lingual transfer and language modelling. Multilingualism and Cross-Lingual NLP (41) is the largest area and has gained the most publications since 2022 (27). Early work uses the Tibetan–Chinese pair to project base noun phrases without a Tibetan tagger or treebank (Nuo et al., 2012), to gloss keywords in OCR output in Chinese in lieu of unreliable MT (Su et al., 2013), and for bilingual clustering and cross-lingual speech synthesis (Ye et al., 2015; Yang et al., 2015), while Sanskrit–Tibetan resources serve the classical tradition (Nehrdich, 2022). Since 2024, the area has been shaped by multilingual corpora and benchmarks for China's minority languages (Zhang et al., 2024, 2025a; Li et al., 2026) and by adapting multilingual LLMs (Gurgurov et al., 2024; Yang et al., 2026). Language Modeling (16; 13 since 2022) recapitulates the wider field's arc, from recurrent models exploiting radicals and morphology (Shen et al., 2017; Khysru et al., 2019), through BERT-style models (Liu et al., 2022; Liang et al., 2024; Zhang et al., 2025c), to LLaMA-based LLMs (Lv et al., 2025; Sha et al., 2025), alignment datasets (Zhuang et al., 2025a; Chen et al., 2026), and studies of vocabulary size and continual pre-training (Han et al., 2026; Yang et al., 2026). Machine Learning for NLP (11; nine since 2024) and Efficient Methods for NLP (3; all since 2023) reflect a turn to few-shot, prompt-based, and parameter-efficient adaptation (Bao et al., 2024; Tang et al., 2025; Zhou et al., 2024b).

Applications. Computational Social Science and Cultural Analytics (12) and Information Retrieval and Text Mining (6) emerged with mining Tibetan web and microblog text in 2014–2019 (Ye et al., 2014; Qiu et al., 2017; Xu et al., 2018), and the former has returned with offensive-language and rumour detection (Ren et al., 2024; Yu et al., 2024). Eight of the nine publications in NLP Applications concern historical or classical materials, such as interlinearized ancient texts (Long and An, 2023) and OCR spelling correction (Luo and Chuang, 2024). Machine Translation (7), has moved from syllable- and subword-based neural MT (Lai et al., 2018; Li et al., 2021) to filtering noisy parallel data with a

cross-lingual model (Zhou et al., 2024a), speech translation from Amdo or all three dialect groups (Ren et al., 2026; Fan et al., 2026), and comparisons of dedicated MT engines with a prompted LLM (Wang et al., 2026). Question Answering (5), Generation (5), Information Extraction (4), and Summarization (3) remain small (e.g., Dan and Sun, 2024; Xu et al., 2025; Zhang et al., 2025b); all summarization studies use news titles, leads, or LLM-generated summaries as references (Ouyang et al., 2024; Hao et al., 2024b; Wang et al., 2025).

Persistent gaps. Besides the four empty areas, topics central to the responsible deployment of language technology are nearly absent. Interpretability and Analysis of Models (3) is represented only by adversarial-robustness studies from a single research team (Cao et al., 2023, 2025b,a), Human-Centered NLP (2) by a readability tool and human-in-the-loop adversarial-text generation (Schmidt, 2020; Cao et al., 2025b), and Clinical and Biomedical Applications (1) by medical named-entity recognition, where F1 falls to 71% from 94% on news (Zhang et al., 2025b); Discourse and Pragmatics, Linguistic Theories, and Symbolic Reasoning have one each.

A layered trajectory. Figure 6 and Table 2 show that the field has expanded broadly bottom-up along the NLP pipeline, adding new layers rather than replacing old ones: resources and syntax appeared in 2004; segmentation, script, speech, and cross-lingual work in 2010–2012; web mining in 2014; language modelling, machine translation, question answering, generation, and information extraction in 2017–2019; and efficient methods, interpretability, summarization, clinical applications, and discourse only in 2023–2025. Earlier layers persisted: resources remain the second most frequent area in the LLM era, and segmentation research continues. Only syntax has declined, as across the ACL Anthology, where resources have also waned but dialogue, interpretability, and ethics have grown (Rohatgi et al., 2023), the very areas in which Tibetan NLP lags.

4.3 Country-Level Analysis

Research is highly concentrated (Figure 5, right): 122 publications (87.1%) are first-authored in China (hereafter, China-affiliated publications), and the remaining 18 in the USA (5), the UK (4), Germany (3), Israel (3), and Australia, France, and India (one each). China's share has exceeded 80% in every phase (Table 2), consistent with its lead in research teams and public funding for Tibetan information processing (Long and Hill, 2021).

Phase	Publications	Areas	China
2004–2017	22	13	95.5%
Top areas: Resources (7), Phon./Morph./Seg. (5), Multilingual (5), Syntax (4), Multimodality (4)			
2018–2021	40	18	82.5%
Top areas: Resources (9), Multilingual (9), Multimodality (7), Speech (7)			
2022–2026	78	23	87.2%
Top areas: Multilingual (27), Resources (24), Language Modeling (13), Multimodality (13)			

Table 2: Three phases of Tibetan NLP. **Areas**: ARR areas represented; **China**: share first-authored in China; **Top areas**: four most frequent areas, with ties. Boundaries: the 2018 output rise and the first BERT-style model, TiBERT (Liu et al., 2022), in 2022.

China's contribution is broad as well as large: China-affiliated publications cover all 25 represented areas, 15 of them exclusively, including Language Modeling, Multimodality, Machine Translation, Question Answering, and Generation. The 18 publications from other countries fall into ten areas, mainly Resources and Evaluation (9), Syntax (5), NLP Applications (5), and Multilingualism (5), and most concern Old or Classical Tibetan and Buddhist texts (cf. Long and Hill, 2021): UK-based researchers built treebanks and annotated corpora (Faggionato and Meelen, 2019; Meelen and Roux, 2020; Meelen et al., 2021); Israel-based researchers, with German Tibetologists, developed evaluation resources and methods for Buddhist texts (Hashiloni et al., 2025; Cohen et al., 2026; Bilitski et al., 2026); and researchers in Germany, France, and the USA contributed a Sanskrit–Tibetan parallel corpus, e-text tools, text analytics, and OCR spelling correction (Nehrdich, 2022; Trinley et al., 2021; Krishna et al., 2021; Luo and Chuang, 2024). US-based work also covers modern Tibetan segmentation, speech recognition evaluation, and tonogenesis (Brown, 2024; Moore et al., 2026; Liang and Zerong, 2026).

The two groups also publish in different venues: two-thirds of the publications from outside China (12 of 18) appear in the ACL Anthology, seven of them at workshops, whereas 97 of the 122 China-affiliated publications (79.5%) appear in journals or in conferences outside it. Tibetan NLP thus comprises two partly separate communities that differ in research focus as well as in venue, although

exchange is productive where it occurs, as in the 2021 TALLIP special issue with papers from China, France, and the USA (Long and Hill, 2021).

5 Discussion

5.1 A Full-Pipeline Paradigm and Its Limits

Our results characterize Tibetan NLP by *concentration* (87.1% first-authored in China), *breadth* (China-affiliated work covers all 25 represented areas), and *layering* (new layers added, earlier ones kept). We call this pattern a full-pipeline paradigm: research spans every layer from script processing to LLMs, though not every task in each layer. The paradigm also has limits. First, it is predominantly engineering-oriented, with little fundamental research (cf. Long and Hill, 2021). Second, it is only partly visible: most China-affiliated research lies outside the ACL Anthology, and hence outside surveys restricted to it (Liu and Best, 2025). Third, its outputs are only partly reusable: 41 of the 140 publications (29.3%) release data, code, or models at a public URL (Appendix F). Release has become more common, from none of the 31 publications of 2004–2018 to 27 of the 61 of 2024–2026 (44.3%), but remains less frequent.

5.2 Lessons for the Global South

Many low-resource languages of the Global South face obstacles usually studied in isolation: dispersed speakers, dialectal diversity, scripts that complicate encoding and segmentation, and limited infrastructure and investment. All co-occur on the Qinghai-Xizang Plateau, yet Tibetan research has advanced across the entire pipeline, suggesting four transferable lessons: (1) invest first in the base layers of script processing, segmentation, and annotated corpora, on which later work depends; (2) treat resource and benchmark construction as a continuing frontier, not a one-off step (§4.2); (3) exploit a high-resource contact language, here Mandarin Chinese, for cross-lingual transfer, while guarding against translated data and benchmarks, which can lose cultural content and misstate difficulty (Zhuang and Sun, 2025; Li et al., 2026); and (4) use dedicated venues to consolidate a research community (§4.1). Transfer is conditional: Tibetan NLP has had sustained public funding (§4.3), whereas African NLP has grown through community initiatives and targeted funding (Alabi et al., 2025); communities with neither will rely more on international collaboration and open resources.

5.3 The Transition to LLMs

Since 2022, the field has pivoted toward pre-trained models and LLMs, from three separate LLaMA-2 extensions with Tibetan vocabularies and 6–26 GB of text (Lv et al., 2025; Sha et al., 2025; Zhuang et al., 2025a) to dense and mixture-of-experts models trained on a curated 72 GB corpus (Yang et al., 2026), vocabulary sizing (Han et al., 2026), reinforcement learning that limits forgetting of Chinese (Su et al., 2026), and vision–language models (Xu et al., 2026). Two obstacles recur. Tokenization is unresolved: a syllable vocabulary of 35K types still leaves 15% of tokens unknown (Liu et al., 2022), whereas multilingual tokenizers fragment Tibetan, so that a 27-token English sentence takes 100 or more in Tibetan (Zhang et al., 2025a), and expanding byte-level byte-pair-encoding vocabularies helps only beyond about 3,000 new tokens per language (Han et al., 2026). Data are noisy and partly closed: 42% of sampled Tibetan CulturaX pages retain headers or footers (Zhang et al., 2024), CINO's pre-training corpus is not public (Zhuang and Sun, 2025), and 41 GB of the 72 GB corpus are private (Yang et al., 2026).

Evaluation reveals a large gap. GPT-3.5 reaches 3.6% F1 on Tibetan WikiAnn named-entity recognition, the lowest of ten low-resource Sino-Tibetan languages (Liu and Best, 2025); on TLUE, most LLMs perform below a random baseline, and Qwen2.5-72B drops from 84.7% on the Chinese CMMLU to 16.5% on its Tibetan counterpart (Gao et al., 2025); and general multimodal LLMs recognize 49–66% of characters on TibOCR-Bench, against 92% for a 7B model tuned on 4.5K Tibetan images (Jie et al., 2026). Evaluation is itself a bottleneck. Encoders are compared chiefly on one news-classification corpus, TNCC (Liu et al., 2022; Zhou et al., 2024b; Hao et al., 2024a; Cao et al., 2025a), and task data are increasingly obtained by pivoting through other languages, as in offensive-language and sentiment datasets translated from English or Chinese (Ren et al., 2024; Zhang et al., 2026a) and summaries bootstrapped through Chinese with GPT-3.5 (Hao et al., 2024b). Alignment data, too, are wholly or partly translated, from GPT-4 answers to English prompts (Zhuang et al., 2025a) or Chinese preference data (Chen et al., 2026). Many LLM benchmarks are likewise translated (Gao et al., 2025; Xu et al., 2026; Yang et al., 2026), yet machine-translated test items depress scores, most of all for Tibetan (Zhang et al., 2025a), and

tasks built from native minority-language knowledge are 11.4–13.9 points harder than translated ones (Li et al., 2026). The move to LLMs thus risks outpacing its foundations, with dialects underserved and evaluation standards only emerging (§6).

5.4 Classical and Modern Tibetan

The corpus contains two research traditions, centred on modern Tibetan and its spoken varieties and on Old and Classical Tibetan, which differ in who conducts them and where they appear (§4.3). China-affiliated work on historical materials has focused on document-image analysis (Wang et al., 2018a; Ma et al., 2020; Li et al., 2022c), the frontier of character recognition once OCR of modern print was in effect solved (Long and Hill, 2021), and on interlinearization (Long and An, 2023). Shared archives already link the traditions: Buddhist Digital Resource Center scans underlie both a China-built character dataset (Li et al., 2022c) and US-based OCR post-correction (Luo and Chuang, 2024). Transfer, however, works more readily within writing than for speech: Old Tibetan has been normalized to Classical orthography to reuse its tools (Faggionato and Meelen, 2019), and a single pipeline annotates texts from the 11th to the 21st century because literary Tibetan has changed little (Meelen and Roux, 2020), whereas pre-training on Classical text did not help segment spoken Standard Tibetan (Brown, 2024), and a segmenter with a classical and literary dictionary marks 87% of syllable boundaries in transcribed Lhasa speech as word boundaries, against 41% for an annotator (Moore et al., 2026). Domain also matters within writing: general-domain Tibetan–Chinese MT drops from about 50 BLEU to at most 5 on Buddhist text (Gyatso et al., 2025). Classical treebanks and parallel corpora could support modern written Tibetan, and modern OCR and parsing the digitization of classical texts, while spoken varieties need their own resources.

6 Recommendations

Open science and cross-community exchange. As fewer than a third of surveyed publications release resources (§5.1), we encourage public release, not only on request, of datasets, models, and code under documented licences, and of pre-training corpora where rights permit (as in Trinley et al., 2021; Zhang et al., 2024), following the practices in Appendix F. To make work findable, authors should name the varieties studied in titles and abstracts, and venues could record them as metadata (Liu and Best, 2025); joint workshops, special issues, and shared tasks could connect China-affiliated research with research elsewhere, and the classical with the modern tradition (§5.4).

Shared standards and native evaluation. To make results comparable across teams, the field needs shared conventions for encoding normalization, segmentation, and unit terminology; public test sets for core tasks; and reporting of splits, metrics, and the varieties and scripts covered (Kumar et al., 2026). Shared segmentation tasks offer a basis, but their single gold segmentation covers only 22% of valid granularities (Zhuoma and Zhijie, 2026). LLM benchmarks should increasingly be built from native material with expert involvement, as in CMiLBench's minority-knowledge tasks and DharmaBench (Li et al., 2026; Hashiloni et al., 2025), rather than by translation alone.

Community participation. No surveyed publication assesses Tibetan speakers' needs, yet needs assessments show that priorities vary across speaker groups, even across dialects of one language (Al Kautsar et al., 2025); German dialect speakers, for instance, favour tools that take spoken dialect input over those producing dialect output (Blaschke et al., 2024), a preference worth testing in diglossic Tibetan. Resource creation and evaluation should involve speakers of all three major dialect groups, including diaspora communities, throughout research (Tonja et al., 2024), as handwriting datasets with up to 235 writers from five regions (Tashi et al., 2024; Zhou et al., 2026) and a readability tool built on children's speech (Schmidt, 2020) begin to do.

Responsible NLP. Beyond the gaps noted in §4.2, safety is evaluated only within recent benchmarks, which show that it does not carry over to Tibetan: most LLMs fall below chance on TLUE's safety set (Gao et al., 2025), and safety alignment transfers unreliably across minority languages (Li et al., 2026). Bias is unmeasured, although zero-shot LLMs tend to label Tibetan news as politics or religion (Tang et al., 2025). Applications that mine user-level social-media data for public-opinion monitoring or to locate user groups (Ye et al., 2014; Qiu et al., 2018b; Xu and Gao, 2024), microblog corpora stored with user names (Qiu et al., 2017, 2018a), speaker verification motivated partly by criminal investigation (Gan et al., 2022), and an OCR bench-

mark with identity cards and bank receipts but no stated anonymization (Jie et al., 2026) raise privacy and dual-use questions that the absence of ethics research leaves unexamined and that call for the consent and privacy safeguards urged for African languages (Adebara and Abdul-Mageed, 2022). Bias, safety, privacy, and human-centred evaluation need study before Tibetan LLMs are deployed in education, healthcare, or public services.

Dialects and related languages. Only eight publications name Lhasa, Amdo, or a multi-dialect setting in their titles, all on speech recognition, synthesis, or translation. Open speech data for all three groups total about 80–100 hours, against over 10,000 for Mandarin or English (Liu et al., 2024), and Kham is their smallest, hardest part: the main multi-dialect corpus, TIBMD@MUC, was released with 8.7 hours of Kham speech against 30 of Lhasa speech (Zhao et al., 2020), and even the best recognizer trained on it errs on 23% of Kham characters, against 3–5% for the other groups (Liu et al., 2024). Kham and Amdo resources should be prioritized, publicly released, and evaluated per variety. Dzongkha and Ladakhi use Tibetan script, so OCR and segmentation tools may transfer directly, and Burmese could share low-level resources (Fang et al., 2024).

7 Conclusion

We have surveyed 140 Tibetan NLP publications from 2004 to 2026. Since 2018, the field has grown rapidly and expanded bottom-up along the NLP pipeline while continuing to build resources, with 87% of the work first-authored in China. This full-pipeline paradigm can serve as a reference model for other low-resource languages of the Global South; its limits in dialect coverage, responsible NLP, native and comparable evaluation, resource release, and cross-community visibility set the agenda for keeping plateau-region languages from being left behind in the LLM era.

Limitations

This survey has several limitations. First, its coverage is bounded by its sources: we include only publications indexed in the ACL Anthology or DBLP, so work in Chinese-language venues, in venues indexed by neither source, and in preprints is not covered, and topics published mainly in such venues may be underrepresented. Second, keyword-based retrieval may miss multilingual studies that include Tibetan without naming it in the searched metadata fields; this risk is greater for DBLP, whose records do not include abstracts. Third, assigning research areas required interpretive judgement, especially for work spanning several topics. Fourth, the country-level analysis relies on the first author's primary affiliation, which may not reflect all contributors or the national context of the research. Fifth, publication counts measure research output rather than impact or quality, and we did not perform a citation analysis; our audit of released resources (Appendix F) records whether resources are publicly available at a URL given by their authors, not whether they are complete, documented, or maintained, and may miss resources released elsewhere or whose links have lapsed. Sixth, our qualitative synthesis, including the discussion of evaluation practice and of the LLM era, relies on results reported by the surveyed publications, which use different models, data, and metrics and which we did not reproduce. Finally, the survey reflects the literature available up to 00:00 UTC, September 25, 2026, so 2026 is only partially covered. We release the full bibliography, in raw and refined versions, to facilitate community-driven correction and extension, and we will continue to track new work: `https://github.com/metaphors/TibetanNLPSurvey`.

Ethical Considerations

This survey analyses publicly available bibliographic metadata and peer-reviewed publications; it involves no human subjects and no personal data beyond author names and affiliations as they appear in the public record. Country-level statistics describe institutional affiliations, not the nationality or ethnicity of researchers, and differences in resource release rates may reflect venue norms and legitimate constraints, such as consent and privacy requirements for speech data or missing permission to share collected data, rather than the practices of individual teams. Where we discuss the data practices or stated applications of individual publications, we do so to motivate shared norms, not to single out particular teams. We do not reproduce or redistribute copyrighted material beyond standard academic citation practices. We have aimed to describe research trends accurately without advancing normative political positions, and we recognize that decisions about language technology for Tibetan should involve the communities who speak it.

Acknowledgments

This work is supported by the Science and Technology Strategic Consulting Project of the Chinese Academy of Engineering (2025-XZ-16-06), the Project of the China Tibetology Literature Resources Data Center (2025SJ003), the National Social Science Foundation of China (22&ZD035), the Key Project of the Natural Science Foundation of Xizang Autonomous Region (XZ202401ZR0040), and the Major Project of the Science and Technology Program of Xizang Autonomous Region (XZ202601ZD0010).

References

Praveen Acharya, Rupak Raj Ghimire, Prakash Poudyal, Balaram Prasain, and Bal Krishna Bal. 2026. English–Nepali–Tamang: A trilingual parallel corpus and benchmark for low-resource machine translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 23–24, Tilburg, The Netherlands. European Association for Machine Translation.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Dehan Al Kautsar, Lucky Susanto, Derry Tanti Wijaya, and Fajri Koto. 2025. What do Indonesians really need from language technology? A nationwide survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7220–7245, Suzhou, China. Association for Computational Linguistics.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Bo An. 2023a. Prompt-based for Low-Resource Tibetan Text Classification. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(8):207:1–207:13.

Bo An. 2023b. Tibetan dependency parsing with graph convolutional neural networks. In *Proceedings of the Ancient Language Processing Workshop*, pages 213–221, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bo An and Congjun Long. 2021. Neural Dependency Parser for Tibetan Sentences. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):30:1–30:16.

P. V. Arun. 2014. A semantic parsing approach for Bhutanese language of Dzongkha. *Int. J. Speech Technol.*, 17(2):161–165.

Mana Ashida, Seunghun Lee, and Kunzang Namgyal. 2020. Building a part-of-speech tagged corpus for drenjongke (bhutia). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 57–63, Suzhou, China. Association for Computational Linguistics.

Xun Bao, Weilan Wang, Xiaojuan Wang, Guanzhong Zhao, Huarui Li, and Meiling Liu. 2024. Toward Better Generalization of Cross-Domain Few-Shot Classification in Tibetan Character With Contrastive Learning and Meta Fine-Tuning. *IEEE Access*, 12:134439–134452.

Guy Bilitski, Lev Shechter, Sonam Jamtsho, Nir Marciano, Nicola Bajetta, Rebecca Sunden, Omri Drori, Kai Golan Hashiloni, Orr Zwebner, Asaf Shina, Orna Almogi, Dorji Wangchuk, and Kfir Bar. 2026. Automatic segmentation of Classical Tibetan texts into autochthonous and allochthonous regions. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 1017–1030, Palma de Mallorca, Spain. ELRA Language Resource Association.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Collin J. Brown. 2024. Improved neural word segmentation for standard Tibetan. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 12–17, Torino, Italia. ELRA and ICCL.

Zhengqi Cai and Weilan Wang. 2018. Online Handwriting Tibetan Character Recognition Based on Two-Dimensional Discriminant Locality Alignment. In *Pattern Recognition and Computer Vision - First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III*, volume 11258 of *Lecture Notes in Computer Science*, pages 88–98. Springer.

Xi Cao, Dolma Dawa, Nuo Qun, and Trashi Nyima. 2023. Pay attention to the robustness of Chinese minority language models! Syllable-level textual adversarial attack on Tibetan script. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 35–46, Toronto, Canada. Association for Computational Linguistics.

Xi Cao, Quzong Gesang, Yuan Sun, Nuo Qun, and Tashi Nyima. 2025a. TSCheater: Generating High-Quality

Tibetan Adversarial Texts via Visual Similarity. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Xi Cao, Yuan Sun, Jiajun Li, Quzong Gesang, Nuo Qun, and Nyima Tashi. 2025b. Human-in-the-loop generation of adversarial texts: A case study on Tibetan script. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 9–16, Mumbai, India. Association for Computational Linguistics.

Binaya Kumar Chaudhary, Bal Krishna Bal, and Rasil Baidar. 2020. Efforts towards developing a Tamang Nepali machine translation system. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 281–286, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Long Chen, Gengga Duojie, and Yuan Sun. 2026. Dataset Used for Reinforcement Learning Tasks of Tibetan Large Models. *Data Intell.*, 8(2):20250044.

Chungku Chungku, Jurmey Rabgay, and Gertrud Faaß. 2010. Building NLP resources for Dzongkha: A tagset and a tagged corpus. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 103–110, Beijing, China. Coling 2010 Organizing Committee.

Shay Cohen, Jingyi Yang, Gal Rabinovitz, Sonam Choden, Ofir Shtrosberg, Nicola Bajetta, Goody Ben Horin, Rebecca Sundén, Omri Drori, Sonam Jamtsho, Dorji Wangchuk, Kfir Bar, Orna Almogi, and Shai Fine. 2026. Scaling sentence similarity for Classical Tibetan with automatic annotations. In *Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities*, pages 150–166, San Diego, USA. Association for Computational Linguistics.

Zhengcuo Dan and Yuan Sun. 2024. TibetanQA2.0: Dataset with Unanswerable Questions for Tibetan Machine Reading Comprehension. *Data Intell.*, 6(4):1158–1167.

Tirthankar Dasgupta, Manjira Sinha, and Anupam Basu. 2012. Forward transliteration of Dzongkha text to Braille. In *Proceedings of the Second Workshop on Advances in Text Input Methods*, pages 97–106, Mumbai, India. The COLING 2012 Organizing Committee.

Parshu Ram Dhungyel and Janis Grundspenkis. 2017. Analysing the Methods of Dzongkha Word Segmentation. *Appl. Comput. Syst.*, 21(1):61–65.

Qianji Di, Ekaterina Vylomova, and Tim Baldwin. 2019. Modelling Tibetan verbal morphology. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 35–40, Sydney, Australia. Australasian Language Technology Association.

Christian Faggionato, Nathan Hill, and 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*, pages 1–6, Marseille, France. European Language Resources Association.

Christian Faggionato and Marieke Meelen. 2019. Developing the Old Tibetan treebank. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 304–312, Varna, Bulgaria. INCOMA Ltd.

Hetong Fan, Yue Zhao, Jing Yu, Zairan Gong, Xiaona Xu, and Haizhou Li. 2026. Tibetan-Chinese Speech-to-Speech Translation Based on Large Speech Models. *Data Intell.*, 8(2):20250137.

Jiajun Fang, Wentao Huang, Aimin Yang, Dong Zhou, and Nankai Lin. 2024. MLCL: A Framework for Reducing Language Imbalance in Sino-Tibetan Languages through Adapter Structures. In *Asian Conference on Machine Learning, 5-8 December 2024, Hanoi, Vietnam*, volume 260 of *Proceedings of Machine Learning Research*, pages 431–446. PMLR.

Zhenye Gan, Yue Yu, and Min Luo. 2022. A Tibetan-dependent speaker recognition method based on deep learning. *Multim. Tools Appl.*, 81(21):30821–30840.

Fan Gao, Cheng Huang, Yutong Liu, Nyima Tashi, Xiangxiang Wang, Thupten Tsering, Ban Ma-bao, Renzeng Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Xiao Feng Cd, Yongbin Yu, and Hao Wang. 2025. TLUE: A Tibetan language understanding evaluation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35071–35097, Suzhou, China. Association for Computational Linguistics.

Rupak Raj Ghimire, Bipesh Subedi, Balaram Prasain, Prakash Poudyal, Praveen Acharya, Nischal Karki, Rupak Tiwari, Rishikesh Kumar Sharma, Jenny Poudel, and Bal Krishna Bal. 2026. NepTam: A Nepali-Tamang parallel corpus and baseline machine translation experiments. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 8850–8861, Palma de Mallorca, Spain. ELRA Language Resource Association.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics.

Kalzang Gyatso, Sonam Tshering, Tashi Norbu, Nyima Tashi, Tong Xiao, Jingbo Zhu, Garma Tashi, and Gadeng Luosang. 2025. An Improved Byte Pair Encoding Method for Tibetan. *J. Adv. Comput. Intell. Intell. Informatics*, 29(6):1273–1282.

Ao Han, Andong Chen, Yuan Sun, and Xiaobing Zhao. 2026. Scaling laws or threshold effects: Exploring the optimal vocabulary size for balancing performance and efficiency in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 31741–31758, San Diego, California, United States. Association for Computational Linguistics.

Yuehui Han, Weilan Wang, Yiqun Wang, and Xiaojuan Wang. 2018. Research on the Method of Tibetan Recognition Based on Component Location Information. In *Pattern Recognition and Computer Vision - First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III*, volume 11258 of *Lecture Notes in Computer Science*, pages 63–73. Springer.

Minghui Hao, Xiaodong Yan, and Xinpeng Ouyang. 2024a. A Tibetan Text Classification Method Based on Hybrid Model and Channel Attention Mechanism. In *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*, pages 1522–1527. IEEE.

Yanrong Hao, Bo Chen, and Xiaobing Zhao. 2024b. TiLTS:Tibetan Long Text Summarization Dataset. In *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part IV*, volume 15362 of *Lecture Notes in Computer Science*, pages 265–276. Springer.

Kai Golan Hashiloni, Shay Cohen, Asaf Shina, Jingyi Yang, Orr Meir Zwebner, Nicola Bajetta, Guy Bilitski, Rebecca Sundén, Guy Maduel, Ryan Conlon, Ari Barzilai, Daniel Mass, Shanshan Jia, Aviv Naaman, Sonam Choden, Sonam Jamtsho, Yadi Qu, Harunaga Isaacson, Dorji Wangchuk, and 3 others. 2025. DharmaBench: Evaluating language models on Buddhist texts in Sanskrit and Tibetan. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2088–2110, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Di Jiang. 2006. The current status of sorting order of Tibetan dictionaries and standardization. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 231–236, Huazhong Normal University, Wuhan, China. Tsinghua University Press.

Di Jiang, Congjun Long, and Jichuan Zhang. 2005. The verbal entries and their description in a grammatical information-dictionary of contemporary Tibetan. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Lama Jie, Manla Cairang, Kuntharrgyal Khysru, Hua Guo Cai Rang, Jiao Jiahui, Yue Yingkai, and Tan Qian. 2026. TibOCR-Bench: A Comprehensive Benchmark and Training Pipeline for Tibetan Multimodal OCR. *Data Intell.*, 8(2):20250140.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Wenjun Ke, Yutian Liu, Xiaokang Yang, Jianguo Wei, and Qingzhi Hou. 2025. Align, enhance and read: Scene Tibetan text recognition with cross-sequence reasoning. *Appl. Soft Comput.*, 169:112548.

Kuntharrgyal Khysru, Di Jin, and Jianwu Dang. 2019. Morphological Verb-Aware Tibetan Language Model. *IEEE Access*, 7:72896–72904.

Kuntharrgyal Khysru, Di Jin, Yuxiao Huang, Hui Feng, and Jianwu Dang. 2021. A Tibetan Language Model That Considers the Relationship Between Suffixes and Functional Words. *IEEE Signal Process. Lett.*, 28:459–463.

Kuntharrgyal Khysru, Yangzom, Wenjie Tang, and Jianguo Wei. 2025. Research on speech synthesis technology based on Tibetan rhythmic features. *Expert Syst. Appl.*, 277:127181.

Chunwei Kong, Xueqiang Lv, Le Zhang, Haixing Zhao, Zangtai Cai, and Yuzhong Chen. 2024. RPEPL: Tibetan Sentiment Analysis Based on Relative Position Encoding and Prompt Learning. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(12):167:1–167:18.

Ravi Krishna, Norman Mu, and Kurt Keutzer. 2021. Applying Text Analytics to the Mind-section Literature of the Tibetan Tradition of the Great Perfection. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):21:1–21:32.

Zewang Kuanzhuo, Li Lin, and Zhao Weina. 2020. SimpleNLG-TI: Adapting SimpleNLG to Tibetan. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 86–90, Dublin, Ireland. Association for Computational Linguistics.

Raghvendra Kumar, Devankar Raj, and Sriparna Saha. 2026. BhashaSutra: A task-centric unified survey of Indian NLP datasets, corpora, and resources. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11998–12064, San Diego, California, United States. Association for Computational Linguistics.

Wen Lai, Xiaobing Zhao, and Wei Bao. 2018. Tibetan-Chinese neural machine translation based on syllable segmentation. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 21–29, Boston, MA. Association for Machine Translation in the Americas.

Gyaltsan Lhundrub, Yong Zhou, Mohan Long, Wuping Ke, Geng Chen, Jianshu Wang, and Desheng Zheng. 2026. TibSTR-Former: Multi-scale feature integration and attention mechanisms for Tibetan scene text recognition. *J. Vis. Commun. Image Represent.*, 120:104907.

Binglin Li, Jie Zhu, and Dongcai Zhao. 2024. Tibetan Handwriting Recognition Method Based on Structural Re-Parameterization ViT and Vertical Attention. In *Pattern Recognition and Computer Vision - 7th Chinese Conference, PRCV 2024, Urumqi, China, October 18-20, 2024, Proceedings, Part VII*, volume 15037 of *Lecture Notes in Computer Science*, pages 156–169. Springer.

Fenfang Li, Hui Lv, La Duo, Binbin Yong, and Qingguo Zhou. 2022a. Sentence Boundary Disambiguation for Tibetan Based on Attention Mechanism at the Syllable Level. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 21(6):1–18.

Yachao Li, Jing Jiang, Jia Yangji, and Ning Ma. 2021. Finding Better Subwords for Tibetan Neural Machine Translation. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):24:1–24:11.

Yachao Li and Hongzhi Yu. 2013. Study on Tibetan Word Segmentation as Syllable Tagging. In *Natural Language Processing and Chinese Computing - Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013, Proceedings*, volume 400 of *Communications in Computer and Information Science*, pages 363–369. Springer.

Yan Li, Xiaomin Li, Yiru Wang, Hui Lv, Fenfang Li, and La Duo. 2022b. Character-based Joint Word Segmentation and Part-of-Speech Tagging for Tibetan Based on Deep Learning. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 21(5):95:1–95:15.

Yijie Li, Xi Cao, Yuan Sun, Quulgan Minggad, Abdulla Ablikim, and Jia Qing Cai Wang. 2026. Diversity in unity, theory in practice: Hierarchical multitask benchmarks for Chinese minority languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36353–36373, San Diego, California, United States. Association for Computational Linguistics.

Zhenjiang Li, Weilan Wang, Yiqun Wang, and Qianxue Zhang. 2022c. Character recognition of Tibetan Historical document in Uchen font: Dataset and bench mark. *J. Comput. Methods Sci. Eng.*, 22(5):1779–1794.

Siyu Liang and Zhaxi Zerong. 2026. The tonogenesis continuum in Tibetan: A computational investigation. In *The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26)*, pages 27–34, Rabat, Morocco. Association for Computational Linguistics.

Yatao Liang, Hui Lv, Yan Li, La Duo, Chuanyi Liu, and Qingguo Zhou. 2024. Tibetan-BERT-wwm: A Tibetan Pretrained Model With Whole Word Masking for Text Classification. *IEEE Trans. Comput. Soc. Syst.*, 11(5):6268–6277.

Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He. 2011. Tibetan word segmentation as syllable tagging using conditional random field. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 168–177, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Huidan Liu, Minghua Nuo, and Jian Wu. 2014. Zipf's law and statistical data on Modern Tibetan. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 322–333, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Huidan Liu, Minghua Nuo, Jian Wu, and Yeping He. 2012. Building large scale text corpus for Tibetan natural language processing by extracting text from web pages. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 11–20, Mumbai, India. The COLING 2012 Organizing Committee.

Huidan Liu, Weina Zhao, Minghua Nuo, Li Jiang, Jian Wu, and Yeping He. 2010. Tibetan number identification based on classification of number components in Tibetan word segmentation. In *Coling 2010: Posters*, pages 719–724, Beijing, China. Coling 2010 Organizing Committee.

Shuheng Liu and Michael Best. 2025. A survey of NLP progress in Sino-Tibetan low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7804–7825, Albuquerque, New Mexico. Association for Computational Linguistics.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. TiBERT: Tibetan Pre-trained Language Model. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2022, Prague, Czech Republic, October 9-12, 2022*, pages 2956–2961. IEEE.

Yigang Liu, Yue Zhao, Xiaona Xu, Liang Xu, and Xubei Zhang. 2023. Multi-task Pre-training for Lhasa-Tibetan Speech Recognition. In *Artificial Neural Networks and Machine Learning - ICANN 2023 - 32nd International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, September 26-29, 2023, Proceedings, Part IX*, volume 14262 of *Lecture Notes in Computer Science*, pages 78–90. Springer.

Yigang Liu, Yue Zhao, Xiaona Xu, Liang Xu, Xubei Zhang, and Qiang Ji. 2024. Exploring task-diverse meta-learning on Tibetan multi-dialect speech recognition. *EURASIP J. Audio Speech Music. Process.*, 2024(1):37.

Congjun Long and Bo An. 2023. On the development of interlinearized ancient literature of ethnic minorities: A case study of the interlinearization of ancient written Tibetan literature. In *Proceedings of the Ancient Language Processing Workshop*, pages 222–231, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Congjun Long and Nathan W. Hill. 2021. Recent Developments in Tibetan NLP. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):19:1–19:3.

Congjun Long, Xuewen Zhou, and Maoke Zhou. 2021. Recognition of Tibetan Maximal-length Noun Phrases Based on Syntax Tree. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):25:1–25:13.

Queenie Luo and Yung-Sung Chuang. 2024. Cleansing Jewel: A Neural Spelling Correction Model Built On Google OCR-ed Tibetan Manuscripts. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(5):73.

Hui Lv, Hao Lv, Liu Yang, Jun Shen, La Duo, Yan Li, Qingguo Zhou, and Binbin Yong. 2024. Improved Tibetan Word Vectors Models Based on Position Information Fusion. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(11):153:1–153:21.

Hui Lv, Chi Pu, La Duo, Yan Li, Qingguo Zhou, and Jun Shen. 2025. T-LLaMA: a Tibetan large language model based on LLaMA2. *Complex Intell. Syst.*, 11(1).

Long-Long Ma, Huidan Liu, and Jian Wu. 2011. MRG-OHTC Database for Online Handwritten Tibetan Character Recognition. In *2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18-21, 2011*, pages 207–211. IEEE Computer Society.

Long-Long Ma, Congjun Long, Lijuan Duan, Xiqun Zhang, Yanxing Li, and Quanchao Zhao. 2020. Segmentation and Recognition for Historical Tibetan Document Images. *IEEE Access*, 8:52641–52651.

Long-Long Ma and Jian Wu. 2013. Semi-automatic Tibetan Component Annotation from Online Handwritten Tibetan Character Database by Optimizing Segmentation Hypotheses. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*, pages 1340–1344. IEEE Computer Society.

Long-Long Ma and Jian Wu. 2015. Online handwritten Tibetan syllable recognition based on component segmentation method. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 46–50. IEEE Computer Society.

Marieke Meelen and Élie Roux. 2020. Meta-dating the PArsed corpus of Tibetan (PACTib). In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 31–42, Düsseldorf, Germany. Association for Computational Linguistics.

Marieke Meelen, 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.*, 20(1):7:1–7:11.

Jacob Moore, Sheng Li, and Paula Lauren. 2026. Evaluating Tibetan ASR With Segmented Word Error Rate: Beyond Character-Level Metrics. *IEEE Access*, 14:101790–101805.

Sebastian Nehrdich. 2022. SansTib, a Sanskrit - Tibetan parallel corpus and bilingual sentence embedding model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6728–6734, Marseille, France. European Language Resources Association.

Sithar Norbu, Pema Choejey, Tenzin Dendup, Sarmad Hussain, and Ahmed Muaz. 2010. Dzongkha word segmentation. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 95–102, Beijing, China. Coling 2010 Organizing Committee.

Ming Hua Nuo, Hui Dan Liu, Wei Na Zhao, Long Long Ma, Jian Wu, and Zhi Ming Ding. 2012. Tibetan base noun phrase identification framework based on Chinese-Tibetan sentence aligned corpus. In *Proceedings of COLING 2012*, pages 2141–2158, Mumbai, India. The COLING 2012 Organizing Committee.

Minghua Nuo, Huidan Liu, Congjun Long, and Jian Wu. 2015. Tibetan unknown word identification from news corpora for supporting lexicon-based Tibetan word segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 451–457, Beijing, China. Association for Computational Linguistics.

Minghua Nuo, Congjun Lun, and Huidan Liu. 2016. Tibetan Multi-word Expressions Identification Framework Based on News Corpora. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, volume 10102 of *Lecture Notes in Computer Science*, pages 16–26. Springer.

Xinpeng Ouyang, Xiaodong Yan, and Minghui Hao. 2024. A Pre-Trained Language Model Based on LED for Tibetan Long Text Summarization. In *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*, pages 992–997. IEEE.

Siqing Qin, Longbiao Wang, Sheng Li, Jianwu Dang, and Lixin Pan. 2022a. Improving low-resource Tibetan end-to-end ASR by multilingual and multilevel unit modeling. *EURASIP J. Audio Speech Music. Process.*, 2022(1):2.

Siqing Qin, Longbiao Wang, Sheng Li, Yuqin Lin, and Jianwu Dang. 2022b. Finer-grained Modeling units-based Meta-Learning for Low-resource Tibetan Speech Recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2133–2137. ISCA.

Lirong Qiu, Qiao Lei, and Zhen Zhang. 2018a. Advanced Sentiment Classification of Tibetan Microblogs on Smart Campuses Based on Multi-Feature Fusion. *IEEE Access*, 6:17896–17904.

Lirong Qiu, Jia Yu, Jie Li, and Haoran Jia. 2018b. Tibetan Weibo User Group Division Based on Semantic Information in the Era of Big Data. *Complex.*, 2018:5809043:1–5809043:11.

Lirong Qiu, Huili Zhang, Zhen Zhang, and Qiumei Pu. 2017. Tibetan Microblog Emotional Analysis Based on Sequential Model in Online Social Platforms. *Complex.*, 2017:5342601:1–5342601:8.

Hang Ren, Jin Yang, Shengqiao Ni, Qin Yang, Jing Zhang, and Nuo Qun. 2024. A Method for Tibetan Offensive Language Detection Based on Prompt Learning and Information Theory. In *International Joint Conference on Neural Networks, IJCNN 2024, Yokohama, Japan, June 30 - July 5, 2024*, pages 1–8. IEEE.

Zengzhuoma Ren, Liping Zhu, Xiuying Yang, and Xiaobing Zhao. 2026. Research on the Quality Evaluation of Amdo Tibetan-Chinese Speech Translation Dataset. *Data Intell.*, 8(2):20250172.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Dirk Schmidt. 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):75:1–75:19.

Jiu Sha, Mengxiao Zhu, Chong Feng, and Jizhuoma Ci. 2025. Tibetan-LLaMA 2: Large Language Model for Tibetan. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(12):141:1–141:19.

Muhammad Sharif, Bin Liu, Sardar Shan Ali Naqvi, Zeeshan Abbas, and Cheng-Lin Liu. 2025. Balti-Tamko: A Spoken Words Dataset Development for Automatic Recognition of the Endangered Balti Language. In *Neural Information Processing - 32nd International Conference, ICONIP 2025, Okinawa, Japan, November 20-24, 2025, Proceedings, Part V*, volume 16313 of *Lecture Notes in Computer Science*, pages 140–150. Springer.

Tongtong Shen, Longbiao Wang, Xie Chen, Kuntharrgyal Khysru, and Jianwu Dang. 2017. Exploiting the Tibetan Radicals in Recurrent Neural Network for Low-Resource Language Models. In *Neural Information Processing - 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II*, volume 10635 of *Lecture Notes in Computer Science*, pages 266–275. Springer.

Shumin Shi, Dan Luo, Xing Wu, Congjun Long, and Heyan Huang. 2021. Multi-level Chunk-based Constituent-to-Dependency Treebank Transformation for Tibetan Dependency Parsing. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):26:1–26:12.

Bing Su, Xiaoqing Ding, Liangrui Peng, and Changsong Liu. 2013. Cross-Language Sensitive Words Distribution Map: A Novel Recognition-Based Document Understanding Method for Uighur and Tibetan. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*, pages 255–259. IEEE Computer Society.

Zeli Su, Ziyin Zhang, Zhou Liu, Xuexian Song, Zhankai Xu, Longfei Zheng, Xiaolu Zhang, Rong Fu, Guixian Xu, and Wentao Zhang. 2026. Reinforcement learning with semantic rewards enables low-resource language expansion without alignment tax. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17772–17786, San Diego, California, United States. Association for Computational Linguistics.

Yuan Sun, Andong Chen, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2021. A Joint Model for Representation Learning of Tibetan Knowledge Graph Based on Encyclopedia. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):27:1–27:17.

Yuan Sun, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2019a. QuGAN: Quasi Generative Adversarial Network for Tibetan Question Answering Corpus Generation. *IEEE Access*, 7:116247–116255.

Yuan Sun, Sisi Liu, Zhengcuo Dan, and Xiaobing Zhao. 2022. Question generation based on grammar knowledge and fine-grained classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6457–6467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuan Sun, Like Wang, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2019b. Improved Distant Supervised Model in Tibetan Relation Extraction Using ELMo and Attention. *IEEE Access*, 7:173054–173062.

Yuan Sun and Tianci Xia. 2019. A Hybrid Network Model for Tibetan Question Answering. *IEEE Access*, 7:52769–52777.

Chao Tang, Zelin Tan, and Xiaobing Zhao. 2025. Tibetan Text Classification based on Prompt Learning and Ensemble Learning. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(2):19:1–19:17.

Dorje Tashi, Bingtian Chen, Tianying Sheng, Yongbin Yu, Xiangxiang Wang, Jin Zhang, Lobsang Yeshi, Rinchen Dongrub, Thupten Tsering, and Nyima Tashi. 2026. Tibetan Data Augmentation via GAN-Based Handwritten Text Generation. *CAAI Trans. Intell. Technol.*, 11(1):55–65.

Dorje Tashi, Tianying Sheng, Bingtian Chen, Renzeng Duojie, Rinchen Dongrub, Yongbin Yu, Xiangxiang Wang, and Nyima Tashi. 2024. Construction of a Tibetan Handwriting Khyug-yig Dataset. *Data Intell.*, 6(3):870–887.

Irshad Ahmad Thukroo, Rumaan Bashir, and Kaiser J. Giri. 2023. A Hybrid Convolutional Bi-Directional Gated Recurrent Unit System for Spoken Languages of JK and Ladakhi. *J. Inf. Knowl. Manag.*, 22(4):2350028:1–2350028:23.

Atnafu Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP progress in indigenous Latin American languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987, Mexico City, Mexico. Association for Computational Linguistics.

Ngawang Trinley, Tenzin, Dirk Schmidt, Helios Hildt, and Tenzin Kaldan. 2021. Taming the Wild Etext: Managing, Annotating, and Sharing Tibetan Corpora in Open Spaces. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 20(2):22:1–22:23.

Andreas Wagner and Bettina Zeisler. 2004. A syntactically annotated corpus of Tibetan. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Hui Wang, Fei Gao, Yue Zhao, Li Yang, Jianjian Yue, and Huilin Ma. 2020a. Multitask Learning with Local Attention for Tibetan Speech Recognition. *Complex.*, 2020:8894566:1–8894566:10.

Qiufen Wang, Ying Wang, Mansour Amini, Ming Xian, and Ziqiang Fu. 2026. Quality assessment of Tibetan-Chinese poetry translation: Integrating automated metrics and qualitative insights through a cross-system comparison of dedicated NMT engines and a prompted LLM. *Lang. Resour. Evaluation*, 60(2):42.

Weilan Wang, Zhengjiang Li, Zhengqi Cai, Xiaobao Lv, Caike Zhaxi, and Yuehui Han. 2019. Online Tibetan Handwriting Recognition for Large Character Set on New Databases. *Int. J. Pattern Recognit. Artif. Intell.*, 33(10):1953003:1–1953003:21.

Xiaojuan Wang, Weilan Wang, Zhenjiang Li, Yiqun Wang, Yuehui Han, and Zhanjun Hao. 2018a. A Recognition Method of the Similarity Character for Uchen Script Tibetan Historical Document Based on DNN. In *Pattern Recognition and Computer Vision - First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III*, volume 11258 of *Lecture Notes in Computer Science*, pages 52–62. Springer.

Yiqun Wang, Weilan Wang, Zhenjiang Li, Yuehui Han, and Xiaojuan Wang. 2018b. Research on Text Line Segmentation of Historical Tibetan Documents Based on the Connected Component Analysis. In *Pattern Recognition and Computer Vision - First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part III*, volume 11258 of *Lecture Notes in Computer Science*, pages 74–87. Springer.

Yiwen Wang, Yanrong Hao, Bo Chen, Yang Xu, and Xiaobing Zhao. 2025. DiSG: A Discourse Structure-Aware Multi-stage Approach for Long Tibetan Text Summarization. In *Natural Language Processing and Chinese Computing - 14th National CCF Conference, NLPCC 2025, Urumqi, China, August 7-9, 2025, Proceedings, Part IV*, volume 16105 of *Lecture Notes in Computer Science*, pages 159–171. Springer.

Zhijuan Wang, Wenguang Fang, Xiaobing Zhao, Wei Song, Yinghui Feng, and Yining Chang. 2020b. Transliteration recognition of Tibetan person name based on Tibetan cultural knowledge. *Int. J. Comput. Sci. Eng.*, 22(2/3):305–312.

Karma Wangchuk, Panomkhawn Riyamongkol, and Rattapoom Waranusast. 2022. Next syllables prediction system in Dzongkha using long short-term memory. *J. King Saud Univ. Comput. Inf. Sci.*, 34(6 Part B):3800–3806.

Shiqi Wu, Yue Zhao, Jing Yu, Xiaona Xu, and Haizhou Li. 2026. Efficient and Natural Tibetan Speech Synthesis via Gaussian Noise-Improved Monotonic Alignment Search and Lightweight iSTFT Multiband Decoder. *Data Intell.*, 8(2):20250128.

Guixian Xu and Wenhui Gao. 2024. Tibetan microblogging user data analysis and topic identification. In *23rd IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2024, Sanya, China, December 17-21, 2024*, pages 1077–1082. IEEE.

Guixian Xu, Yide Liang, Zeli Su, Xuexian Song, Ziyin Zhang, Yushuang Dong, Ting Zhang, and Xu Han. 2026. FTibSuite: A comprehensive resource suite for Tibetan vision–language modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 18143–18159, San Diego, California, United States. Association for Computational Linguistics.

Guixian Xu, Zeli Su, Ziyin Zhang, Jianing Liu, Xu Han, Ting Zhang, and Yushuang Dong. 2025. CMHG: A dataset and benchmark for headline generation of minority languages in China. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12339–12346, Suzhou, China. Association for Computational Linguistics.

Guixian Xu, Chang-Zhi Wang, Li-Hui Wang, Yu-Hong Zhou, Wei-Kang Li, Hao Xu, and Qing Huang. 2017.

Semantic classification method for network Tibetan corpus. *Clust. Comput.*, 20(1):155–165.

Guixian Xu, Changzhi Wang, Haishen Yao, and Qi Qi. 2019a. Research on Tibetan hot words, sensitive words tracking and public opinion classification. *Clust. Comput.*, 22(Supplement):9977–9990.

Guixian Xu, Haishen Yao, Dongming Wu, Yuan Li, Deguang Ouyang, and Gaofeng Chen. 2019b. Public opinion classification and text alignment based on Chinese and Tibetan corpus. *Clust. Comput.*, 22(Supplement):10263–10274.

Guixian Xu, Ziheng Yu, and Qi Qi. 2018. Efficient Sensitive Information Classification and Topic Tracking Based on Tibetan Web Pages. *IEEE Access*, 6:55643–55652.

Xiaona Xu, Li Yang, Yue Zhao, and Hui Wang. 2021. End-to-End Speech Synthesis for Tibetan Multidialect. *Complex.*, 2021:6682871:1–6682871:8.

Hongwu Yang, Keiichiro Oura, Haiyan Wang, Zhenye Gan, and Keiichi Tokuda. 2015. Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis. *Multim. Tools Appl.*, 74(22):9927–9942.

Lei Yang, Leiyu Pan, Bojian Xiong, Renren Jin, Shaowei Zhang, Yue Chen, Ling Shi, Jiang Zhou, Junru Wu, Zhen Wang, Jianxiang Peng, Juesi Xiao, Tianyu Dong, Zhuowen Han, Zhuo Chen, Yuqi Ren, and Deyi Xiong. 2026. From curated data to scalable models: Continual pre-training of dense and MoE large language models for Tibetan. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40167–40182, San Diego, California, United States. Association for Computational Linguistics.

Chengxu Ye, Wushao Wen, and Chang-Dong Wang. 2015. Chinese-Tibetan bilingual clustering based on random walk. *Neurocomputing*, 158:32–41.

Chengxu Ye, Wushao Wen, and Ping Yang. 2014. TM-ToT: An Effective Model for Topic Mining from the Tibetan Messages. In *Natural Language Processing and Chinese Computing - Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings*, volume 496 of *Communications in Computer and Information Science*, pages 438–448. Springer.

Lisu Yu, Fei Li, Lixin Yu, Wei Li, Zhicheng Dong, Donghong Cai, and Zhen Wang. 2024. A novel Chinese-Tibetan mixed-language rumor detector with multi-extractor representations. *Comput. Speech Lang.*, 87:101625.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2024. MC^2: Towards transparent and culturally-aware NLP for minority languages in China. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8832–8850, Bangkok, Thailand. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025a. MiLiC-eval: Benchmarking multilingual LLMs for China's minority languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11086–11102, Vienna, Austria. Association for Computational Linguistics.

Guowei Zhang, Weilan Wang, Ce Zhang, Penghai Zhao, and Mingkai Zhang. 2023. HUTNet: An Efficient Convolutional Neural Network for Handwritten Uchen Tibetan Character Recognition. *Big Data*, 11(5):387–398.

Jin Zhang, Ziyue Zhang, Lobsang Yeshi, Dorje Tashi, Xiangshi Wang, Yuqing Cai, Yongbin Yu, Xiangxiang Wang, Nyima Tashi, and Gadeng Luosang. 2025b. Tibetan Medical Named Entity Recognition Based on Syllable-Word-Sentence Embedding Transformer. *CAAI Trans. Intell. Technol.*, 10(4):1148–1158.

Rui Zhang, Hongkai Liu, Gege Li, Zhaoyang Wang, Zhonglin Ye, Zhen Liu, and Cairang Jia. 2026a. Tibetan Sentiment Analysis Based on Cross-Language Sentence Vector Contrastive Distillation Learning. *KSII Trans. Internet Inf. Syst.*, 20(6):2851–2873.

Weizhao Zhang, Mengjuan Wang, Junzhi Li, and Hongwu Yang. 2026b. Cro-MTVITS: An end-to-end cross-lingual speech synthesis model for Mandarin and multi-dialect Tibetan based on VITS. *Comput. Speech Lang.*, 100:101956.

Weizhao Zhang and Hongwu Yang. 2023. Improving Sequence-to-sequence Tibetan Speech Synthesis with Prosodic Information. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(9):225:1–225:13.

Weizhao Zhang, Hongwu Yang, Xiaolong Bu, and Lili Wang. 2019. Deep Learning for Mandarin-Tibetan Cross-Lingual Speech Synthesis. *IEEE Access*, 7:167884–167894.

Ziyue Zhang, Yongbin Yu, Xiangxiang Wang, Xiao Feng, Yuze Li, Jiarun Shen, Dorje Tashi, Jin Zhang, Lobsang Yeshi, Lei Li, Nyima Tashi, and Jingye Cai. 2025c. Tibetan Few-Shot Learning Model With Deep Contextualised Two-Level Word Embeddings. *CAAI Trans. Intell. Technol.*, 10(5):1394–1410.

Guanzhong Zhao, Weilan Wang, Xiaojuan Wang, Xun Bao, Huarui Li, and Meiling Liu. 2024. Incremental Recognition of Multi-Style Tibetan Character Based on Transfer Learning. *IEEE Access*, 12:44190–44206.

Quanchao Zhao, Long-Long Ma, and Lijuan Duan. 2018. A Touching Character Database from Tibetan Historical Documents to Evaluate the Segmentation Algorithm. In *Pattern Recognition and Computer Vision - First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part IV*, volume 11259 of *Lecture Notes in Computer Science*, pages 309–321. Springer.

Yue Zhao, Panhua Hu, Xiaona Xu, Licheng Wu, and Xiali Li. 2019a. Lhasa-Tibetan Speech Synthesis Using End-to-End Model. *IEEE Access*, 7:140305–140311.

Yue Zhao, Xiaona Xu, and Guosheng Yang. 2012. Unsupervised Tibetan speech features Learning based on Dynamic Bayesian Networks. In *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, pages 2319–2322. IEEE Computer Society.

Yue Zhao, Xiaona Xu, Jianjian Yue, Wei Song, Xiali Li, Licheng Wu, and Qiang Ji. 2020. An open speech resource for Tibetan multi-dialect and multitask recognition. *Int. J. Comput. Sci. Eng.*, 22(2/3):297–304.

Yue Zhao, Jianjian Yue, Xiaona Xu, Licheng Wu, and Xiali Li. 2019b. End-to-End-Based Tibetan Multitask Speech Recognition. *IEEE Access*, 7:162519–162529.

Maoke Zhou, Xiaoke Qi, Wei Bao, and Xiaobing Zhao. 2026. TibHCR: A Large-Scale Dataset for Offline Handwritten Tibetan Character Recognition. *IEICE Trans. Inf. Syst.*, 109(1):193–205.

Mingjun Zhou, Quzong Gesang, Nuo Qun, Tashi Nyima, and Rinchen Dongrub. 2024a. Tibetan-Chinese Machine Translation Enhanced on Cross-Lingual Pre-Trained Model. In *IEEE International Conference on Systems, Man, and Cybernetics, SMC 2024, Kuching, Malaysia, October 6-10, 2024*, pages 1618–1623. IEEE.

Mingjun Zhou, Daiqing Zhuoma, Nuo Qun, and Tashi Nyima. 2024b. Efficient Fine-Tuning for Low-Resource Tibetan Pre-trained Language Models. In *Artificial Neural Networks and Machine Learning - ICANN 2024 - 33rd International Conference on Artificial Neural Networks, Lugano, Switzerland, September 17-20, 2024, Proceedings, Part VII*, volume 15022 of *Lecture Notes in Computer Science*, pages 410–422. Springer.

Jie Zhu, Shugen Wang, Yanru Wu, Meijing Guan, and Yang Feng. 2019. A New Algorithm for Component Decomposition and Type Recognition of Tibetan Syllable. In *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, volume 11839 of *Lecture Notes in Computer Science*, pages 381–392. Springer.

Xiaojun Zhu and Heming Huang. 2020. End-to-End Amdo-Tibetan Speech Recognition Based on Knowledge Transfer. *IEEE Access*, 8:170991–171000.

Wenhao Zhuang, Dawa Cairen, and Yuan Sun. 2025a. TIFD: Tibetan Instruction-Following Dataset for Large Language Models Supervised Fine-Tuning. *Data Intell.*, 7(3):776–785.

Wenhao Zhuang and Yuan Sun. 2025. CUTE: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10037–10046, Abu Dhabi, UAE. Association for Computational Linguistics.

Yan Zhuang, Yuan Sun, Yijie Li, Sisi Liu, and Xiaobing Zhao. 2025b. Tibetan Question Generation Based on Key Sentence and Knowledge Graph. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(5):42:1–42:16.

Cairang Zhuoma and Cai Zhijie. 2026. A multi-granularity word segment evaluation set construction method for Tibetan language. *Eng. Appl. Artif. Intell.*, 181:115301.

A Tibetan NLP Conference/Workshop Papers Found from ACL Anthology

A.1 Conference Papers

Li et al. (2026); Yang et al. (2026); Su et al. (2026); Xu et al. (2026); Han et al. (2026); Bilitski et al. (2026); Hashiloni et al. (2025); Cao et al. (2025b); Xu et al. (2025); Gao et al. (2025); Zhang et al. (2025a); Zhuang and Sun (2025); Zhang et al. (2024); Sun et al. (2022); Nehrdich (2022); Kuanzhuo et al. (2020); Faggionato and Meelen (2019); Nuo et al. (2015); Liu et al. (2014); Nuo et al. (2012); Liu et al. (2011, 2010); Jiang (2006); Jiang et al. (2005); Wagner and Zeisler (2004).

A.2 Workshop Papers

Cohen et al. (2026); Liang and Zerong (2026); Gurgurov et al. (2024); Brown (2024); An (2023b); Long and An (2023); Cao et al. (2023); Faggionato et al. (2022); Meelen and Roux (2020); Di et al. (2019); Lai et al. (2018); Liu et al. (2012).

A.3 Conference List

· AACL: Asia-Pacific Chapter of the Association for Computational Linguistics
· ACL: Annual Meeting of the Association for Computational Linguistics
· COLING: International Conference on Computational Linguistics
· EMNLP: Conference on Empirical Methods in Natural Language Processing
· IJCNLP: International Joint Conference on Natural Language Processing
· INLG: International Natural Language Generation Conference
· LREC: International Conference on Language Resources and Evaluation
· PACLIC: Pacific Asia Conference on Language, Information and Computation
· RANLP: International Conference on Recent Advances in Natural Language Processing

A.4 Workshop List

· ALP: Workshop on Ancient Language Processing
· ALR: Workshop on Asian Language Resources
· ALTA: Australasian Language Technology Association Workshop
· EURALI: Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia
· KaLLM: Workshop on Knowledge Graphs and Large Language Models
· LChange: International Workshop on Computational Approaches to Historical Language Change
· LoResMT: Workshop on Technologies for MT of Low Resource Languages
· NLP4DH: International Workshop on Natural Language Processing for Digital Humanities
· TLT: International Workshop on Treebanks and Linguistic Theories
· TrustNLP: Workshop on Trustworthy Natural Language Processing

B Tibetan NLP Conference Papers / Journal Articles Found from DBLP

B.1 Conference Papers

Cao et al. (2025a); Wang et al. (2025); Xu and Gao (2024); Hao et al. (2024a); Ouyang et al. (2024); Zhou et al. (2024b); Fang et al. (2024); Zhou et al. (2024a); Ren et al. (2024); Hao et al. (2024b); Li et al. (2024); Liu et al. (2023); Qin et al. (2022b); Liu et al. (2022); Zhu et al. (2019); Cai and Wang (2018); Han et al. (2018); Wang et al. (2018a,b); Zhao et al. (2018); Shen et al. (2017); Nuo et al. (2016); Ma and Wu (2015); Ye et al. (2014); Ma and Wu (2013); Su et al. (2013); Li and Yu (2013); Zhao et al. (2012); Ma et al. (2011).

B.2 Journal Articles

Tashi et al. (2026); Zhang et al. (2026b); Chen et al. (2026); Fan et al. (2026); Jie et al. (2026); Ren et al. (2026); Wu et al. (2026); Zhuoma and Zhijie (2026); Moore et al. (2026); Zhou et al. (2026); Lhundrub et al. (2026); Zhang et al. (2026a); Wang et al. (2026); Sha et al. (2025); Tang et al. (2025); Zhuang et al. (2025b); Ke et al. (2025); Zhang et al. (2025b,c); Lv et al. (2025); Zhuang et al. (2025a); Khysru et al. (2025); Gyatso et al. (2025); Kong et al. (2024); Luo and Chuang (2024); Lv et al. (2024); Yu et al. (2024); Dan and Sun (2024); Tashi et al. (2024); Liu et al. (2024); Bao et al. (2024); Zhao et al. (2024); Liang et al. (2024); An (2023a); Zhang and Yang (2023); Zhang et al. (2023); Li et al. (2022a,b); Qin et al. (2022a); Li et al. (2022c); Gan et al. (2022); An and Long (2021); Krishna et al. (2021); Li et al. (2021); Long et al. (2021); Meelen et al. (2021); Shi et al. (2021); Sun et al. (2021); Trinley et al. (2021); Xu et al. (2021); Khysru et al. (2021); Schmidt (2020); Wang et al. (2020a); Ma et al. (2020); Zhu and Huang (2020); Wang et al. (2020b); Zhao et al. (2020); Xu et al. (2019a,b); Khysru et al. (2019); Sun et al. (2019a,b); Sun and Xia (2019); Zhang et al. (2019); Zhao et al. (2019a,b); Wang et al. (2019); Qiu et al. (2018b,a); Xu et al. (2018, 2017); Qiu et al. (2017); Yang et al. (2015); Ye et al. (2015).

B.3 Conference List

· ACML: Asian Conference on Machine Learning
· CSCWD: International Conference on Computer Supported Cooperative Work in Design
· ICANN: International Conference on Artificial Neural Networks
· ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing
· ICDAR: International Conference on Document Analysis and Recognition
· ICONIP: International Conference on Neural Information Processing
· ICPR: International Conference on Pattern Recognition
· IJCNN: International Joint Conference on Neural Networks
· Interspeech: Annual Conference of the International Speech Communication Association
· NLPCC: CCF International Conference on Natural Language Processing and Chinese Computing
· PRCV: Chinese Conference on Pattern Recognition and Computer Vision
· SMC: IEEE International Conference on Systems, Man, and Cybernetics
· TrustCom: IEEE International Conference on Trust, Security and Privacy in Computing and Communications

B.4 Journal List

· ACM Transactions on Asian and Low-Resource Language Information Processing
· Applied Soft Computing
· Big Data
· CAAI Transactions on Intelligence Technology
· Cluster Computing
· Complex & Intelligent Systems

· Complexity
· Computer Speech & Language
· Data Intelligence
· Engineering Applications of Artificial Intelligence
· EURASIP Journal on Audio, Speech, and Music Processing
· Expert Systems with Applications
· IEEE Access
· IEEE Signal Processing Letters
· IEEE Transactions on Computational Social Systems
· IEICE Transactions on Information and Systems
· International Journal of Computational Science and Engineering
· International Journal of Pattern Recognition and Artificial Intelligence
· Journal of Advanced Computational Intelligence and Intelligent Informatics
· Journal of Computational Methods in Sciences and Engineering
· Journal of Visual Communication and Image Representation
· KSII Transactions on Internet and Information Systems
· Language Resources and Evaluation
· Multimedia Tools and Applications
· Neurocomputing

C Publications Classified by Country of First Author's Affiliation

C.1 Australia

Di et al. (2019).

C.2 China

Li et al. (2026); Yang et al. (2026); Su et al. (2026); Xu et al. (2026); Han et al. (2026); Cao et al. (2025b); Xu et al. (2025); Gao et al. (2025); Zhang et al. (2025a); Zhuang and Sun (2025); Zhang et al. (2024); An (2023b); Long and An (2023); Cao et al. (2023); Sun et al. (2022); Kuanzhuo et al. (2020); Lai et al. (2018); Nuo et al. (2015); Liu et al. (2014); Nuo et al. (2012); Liu et al. (2012, 2011, 2010); Jiang (2006); Jiang et al. (2005); Wang et al. (2025); Cao et al. (2025a); Li et al. (2024); Hao et al. (2024b); Ren et al. (2024); Zhou et al. (2024a); Fang et al. (2024); Zhou et al. (2024b); Ouyang et al. (2024); Hao et al. (2024a); Xu and Gao (2024); Liu et al. (2023, 2022); Qin et al. (2022b); Zhu et al. (2019); Zhao et al. (2018); Wang et al. (2018b,a); Han et al. (2018); Cai and Wang (2018); Shen et al. (2017); Nuo et al. (2016); Ma and Wu (2015); Ye et al. (2014); Li and Yu (2013); Su et al. (2013); Ma and Wu (2013); Zhao et al. (2012); Ma et al. (2011); Lhundrub et al. (2026); Zhang et al. (2026a); Zhuoma and Zhijie (2026); Wu et al. (2026); Ren et al. (2026); Jie et al. (2026); Fan et al. (2026); Chen et al. (2026); Wang et al. (2026); Tashi et al. (2026); Zhang et al. (2026b); Zhou et al. (2026); Sha et al. (2025); Tang et al. (2025); Zhuang et al. (2025b); Ke et al. (2025); Zhang et al. (2025b,c); Lv et al. (2025); Zhuang et al. (2025a); Khysru et al. (2025); Gyatso et al. (2025); Kong et al. (2024); Lv et al. (2024); Yu et al. (2024); Dan and Sun (2024); Tashi et al. (2024); Liu et al. (2024); Bao et al. (2024); Zhao et al. (2024); Liang et al. (2024); An (2023a); Zhang and Yang (2023); Zhang et al. (2023); Li et al. (2022a,b); Qin et al. (2022a); Li et al. (2022c); Gan et al. (2022); An and Long (2021); Li et al. (2021); Long et al. (2021); Shi et al. (2021); Sun et al. (2021); Xu et al. (2021); Khysru et al. (2021); Zhao et al. (2020); Wang et al. (2020b,a); Ma et al. (2020); Zhu and Huang (2020); Xu et al. (2019a,b); Khysru et al. (2019); Sun et al. (2019a,b); Sun and Xia (2019); Zhang et al. (2019); Zhao et al. (2019a,b); Wang et al. (2019); Qiu et al. (2018b,a); Xu et al. (2018, 2017); Qiu et al. (2017); Yang et al. (2015); Ye et al. (2015).

C.3 France

Trinley et al. (2021).

C.4 Germany

Gurgurov et al. (2024); Nehrdich (2022); Wagner and Zeisler (2004).

C.5 India

Schmidt (2020).

C.6 Israel

Cohen et al. (2026); Bilitski et al. (2026); Hashiloni et al. (2025).

C.7 UK

Faggionato et al. (2022); Meelen and Roux (2020); Faggionato and Meelen (2019); Meelen et al. (2021).

C.8 USA

Liang and Zerong (2026); Brown (2024); Moore et al. (2026); Luo and Chuang (2024); Krishna et al. (2021).

D Publications Classified by Research Area in ARR

D.1 Clinical and Biomedical Applications

Zhang et al. (2025b).

D.2 Computational Social Science and Cultural Analytics

Xu and Gao (2024); Ren et al. (2024); Ye et al. (2014); Yu et al. (2024); Liang et al. (2024); Krishna et al. (2021); Xu et al. (2019a,b); Qiu et al. (2018b,a); Xu et al. (2018); Qiu et al. (2017).

D.3 Dialogue and Interactive Systems

None.

D.4 Discourse and Pragmatics

Wang et al. (2025).

D.5 Efficient Methods for NLP

Han et al. (2026); Zhou et al. (2024b); Zhang et al. (2023).

D.6 Ethics, Bias, and Fairness

None.

D.7 Generation

Xu et al. (2025); Sun et al. (2022); Kuanzhuo et al. (2020); Zhuang et al. (2025b); Sun et al. (2019a).

D.8 Human-Centered NLP and Human-AI Interaction

Cao et al. (2025b); Schmidt (2020).

D.9 Information Extraction

Zhang et al. (2025b); Sun et al. (2021); Wang et al. (2020b); Sun et al. (2019b).

D.10 Information Retrieval and Text Mining

Xu and Gao (2024); Ye et al. (2014); Xu et al. (2019a, 2018, 2017); Ye et al. (2015).

D.11 Interpretability and Analysis of Models for NLP

Cao et al. (2025b, 2023, 2025a).

D.12 Language Modeling

Yang et al. (2026); Su et al. (2026); Han et al. (2026); Ouyang et al. (2024); Liu et al. (2022); Shen et al. (2017); Chen et al. (2026); Sha et al. (2025); Tang et al. (2025); Zhang et al. (2025c); Lv et al. (2025); Zhuang et al. (2025a); Liang et al. (2024); An (2023a); Khysru et al. (2021, 2019).

D.13 LLM agents

None.

D.14 Linguistic Theories, Cognitive Modeling, and Psycholinguistics

Liu et al. (2014).

D.15 Machine Learning for NLP

Gurgurov et al. (2024); Hao et al. (2024a); Ren et al. (2024); Cai and Wang (2018); Zhao et al. (2012); Tashi et al. (2026); Tang et al. (2025); Zhang et al. (2025c); Lv et al. (2024); Bao et al. (2024); Zhao et al. (2024).

D.16 Machine Translation

Lai et al. (2018); Zhou et al. (2024a); Fan et al. (2026); Ren et al. (2026); Wang et al. (2026); Gyatso et al. (2025); Li et al. (2021).

D.17 Multilingualism and Cross-Lingual NLP

Li et al. (2026); Yang et al. (2026); Su et al. (2026); Hashiloni et al. (2025); Gao et al. (2025); Zhang et al. (2025a); Zhuang and Sun (2025); Zhang et al. (2024); Gurgurov et al. (2024); Cao et al. (2023); Nehrdich (2022); Faggionato et al. (2022); Nuo et al. (2012); Hao et al. (2024a); Zhou et al. (2024b); Fang et al. (2024); Zhou et al. (2024a); Li et al. (2024); Liu et al. (2023); Qin et al. (2022b); Liu et al. (2022); Han et al. (2018); Ma and Wu (2015); Su et al. (2013); Zhang et al. (2026b,a); Sha et al. (2025); Lv et al. (2025); Yu et al. (2024); Liu et al. (2024); Qin et al. (2022a); Trinley et al. (2021); Xu et al. (2021); Wang et al. (2020a); Zhu and Huang (2020); Wang et al. (2020b); Xu et al. (2019b); Zhang et al. (2019); Zhao et al. (2019b); Yang et al. (2015); Ye et al. (2015).

D.18 Multimodality and Language Grounding to Vision, Robotics and Beyond

Xu et al. (2026); Fang et al. (2024); Li et al. (2024); Cai and Wang (2018); Han et al. (2018); Wang et al. (2018a,b); Zhao et al. (2018); Ma and Wu (2015, 2013); Su et al. (2013); Ma et al. (2011); Tashi et al. (2026); Jie et al. (2026); Zhou et al. (2026); Lhundrub et al. (2026); Ke et al. (2025); Tashi et al. (2024); Bao et al. (2024); Zhao et al. (2024); Zhang et al. (2023); Li et al. (2022c); Ma et al. (2020); Wang et al. (2019).

D.19 NLP and Code Models

None.

D.20 NLP and Symbolic Reasoning

Sun et al. (2021).

D.21 NLP Applications

Cohen et al. (2026); Bilitski et al. (2026); Long and An (2023); Wang et al. (2018a,b); Luo and Chuang (2024); Krishna et al. (2021); Schmidt (2020); Ma et al. (2020).

D.22 Phonology, Morphology, and Word Segmentation

Liang and Zerong (2026); Brown (2024); Di et al. (2019); Lai et al. (2018); Nuo et al. (2015); Liu et al. (2011, 2010); Zhu et al. (2019); Shen et al. (2017); Li and Yu (2013); Zhuoma and Zhijie (2026); Gyatso et al. (2025); Li et al. (2022a,b, 2021); Khysru et al. (2021, 2019).

D.23 Question Answering

Sun et al. (2022); Zhuang et al. (2025b); Dan and Sun (2024); Sun et al. (2019a); Sun and Xia (2019).

D.24 Resources and Evaluation

Li et al. (2026); Xu et al. (2026); Bilitski et al. (2026); Hashiloni et al. (2025); Xu et al. (2025); Gao et al. (2025); Zhang et al. (2025a); Zhuang and Sun (2025); Zhang et al. (2024); Long and An (2023); Nehrdich (2022); Meelen and Roux (2020); Faggionato and Meelen (2019); Liu et al. (2014, 2012); Jiang (2006); Jiang et al. (2005); Wagner and Zeisler (2004); Cao et al. (2025a); Hao et al. (2024b); Zhao et al. (2018); Ma and Wu (2013); Ma et al. (2011); Chen et al. (2026); Jie et al. (2026); Ren et al. (2026); Zhuoma and Zhijie (2026); Moore et al. (2026); Zhou et al. (2026); Wang et al. (2026); Zhuang et al. (2025a); Dan and Sun (2024); Tashi et al. (2024); Li et al. (2022c); An and Long (2021); Meelen et al. (2021); Shi et al. (2021); Trinley et al. (2021); Zhao et al. (2020); Wang et al. (2019).

D.25 Semantics: Lexical and Sentence-Level

Cohen et al. (2026); Nuo et al. (2016); Lv et al. (2024); Xu et al. (2017).

D.26 Sentiment Analysis, Stylistic Analysis, and Argument Mining

Zhang et al. (2026a); Kong et al. (2024); Qiu et al. (2018a, 2017).

D.27 Speech Recognition, Text-to-Speech and Spoken Language Understanding

Liang and Zerong (2026); Liu et al. (2023); Qin et al. (2022b); Zhao et al. (2012); Zhang et al. (2026b); Fan et al. (2026); Wu et al. (2026); Moore et al. (2026); Khysru et al. (2025); Liu et al. (2024); Zhang and Yang (2023); Qin et al. (2022a); Gan et al. (2022); Xu et al. (2021); Wang et al. (2020a); Zhu and Huang (2020); Zhao et al. (2020); Zhang et al. (2019); Zhao et al. (2019a,b); Yang et al. (2015).

D.28 Summarization

Wang et al. (2025); Ouyang et al. (2024); Hao et al. (2024b).

D.29 Syntax: Tagging, Chunking and Parsing

An (2023b); Faggionato et al. (2022); Meelen and Roux (2020); Faggionato and Meelen (2019); Nuo et al. (2012); Jiang et al. (2005); Wagner and Zeisler (2004); Li and Yu (2013); Li et al. (2022a,b); An and Long (2021); Long et al. (2021); Meelen et al. (2021); Shi et al. (2021).

E Full-Scale Versions of Figure 3, 4, 5, and 6

Figure 7, 8, 9, and 10 are the full-scale versions of Figure 3, 4, 5, and 6, respectively.

F Open-Source Resources Released by the Surveyed Publications

Bibliographic metadata cannot reveal whether the datasets, models, and code described in a publication are publicly released. To complement the bibliometric analysis, we manually examined each of the 140 publications for resources released by its authors; Table 3 lists the results.

Compilation. A resource is included when the authors of a publication release code, data, or models at a publicly accessible URL; resources available only on request and third-party re-uploads are excluded. For each entry, the table gives the name of the resource, the publications that introduce it, its URLs, and a one-sentence summary of its content and, where reported, its scale and licence. A resource described in several publications is listed once, under the earliest of them: this applies to the TIBMD@MUC speech corpus (Zhao et al., 2019b, 2020) and to the PACTib/ACTib corpora and annotation pipelines (Meelen and Roux, 2020; Meelen

et al., 2021; Faggionato et al., 2022). Entries are ordered by year of publication and then alphabetically by first author.

Coverage and trend. The table contains 38 entries from 41 of the 140 publications (29.3%). Public release is a recent practice in Tibetan NLP. None of the 31 publications from 2004–2018 releases a resource that meets our criteria, compared with 14 of the 48 publications from 2019–2023 (29.2%) and 27 of the 61 from 2024–2026 (44.3%), including 12 of the 18 publications of 2025. This rise coincides with the transition to pre-trained and large language models, a period in which releasing corpora, models, and benchmarks has become common practice across NLP.

What is released. The released resources span every layer of the pipeline described in Section 4.2. At the foundational layer, they include a lexicon of verbal paradigms (Di et al., 2019), the PACTib/ACTib treebank and annotation pipelines (Meelen and Roux, 2020; Meelen et al., 2021; Faggionato et al., 2022), a knowledge graph (Sun et al., 2021), a word segmenter (Gyatso et al., 2025), and expert word-boundary annotations (Moore et al., 2026). Script processing is represented by character datasets of historical documents, handwriting, and cursive script (Li et al., 2022c; Tashi et al., 2024; Zhou et al., 2026), code for few-shot character recognition (Bao et al., 2024), a multimodal OCR benchmark (Jie et al., 2026), and a vision–language model (Xu et al., 2026). Speech is represented by the TIBMD@MUC multi-dialect corpus (Zhao et al., 2019b, 2020) and a package for measuring recognition error at the word level (Moore et al., 2026). Since 2022, releases have increasingly consisted of PLMs and LLMs (Liu et al., 2022; Liang et al., 2024; Lv et al., 2025; Sha et al., 2025; Zhuang and Sun, 2025), the data used to train and align them (Zhang et al., 2024; Zhuang et al., 2025a), and benchmarks for evaluating them; six benchmarks date from 2025–2026 alone (Gao et al., 2025; Xu et al., 2025; Hashiloni et al., 2025; Zhang et al., 2025a; Li et al., 2026; Jie et al., 2026). Three further releases, all from one research team, support the evaluation of adversarial robustness (Cao et al., 2023, 2025a,b). Finally, Classical Tibetan and Buddhist textual heritage accounts for a distinct cluster of releases: a Sanskrit–Tibetan parallel corpus (Nehrdich, 2022), tools and text analytics for Tibetan Buddhist literature (Trinley et al., 2021; Krishna et al., 2021), and recent resources for benchmarking, distinguishing native from translated passages, and measuring sentence similarity (Hashiloni et al., 2025; Bilitski et al., 2026; Cohen et al., 2026).

Release rates by research area. The share of publications that release resources varies widely across ARR areas (Appendix D). It is highest in Resources and Evaluation (22 of 40; 55.0%), Language Modeling (7 of 16; 43.8%), and Multilingualism and Cross-Lingual NLP (16 of 41; 39.0%). Among the larger areas, it is lowest in Speech Recognition, Text-to-Speech and Spoken Language Understanding (3 of 21; 14.3%) and in Phonology, Morphology, and Word Segmentation (2 of 17; 11.8%), and none of the publications in Information Retrieval and Text Mining (6), Sentiment Analysis, Stylistic Analysis, and Argument Mining (4), or Summarization (3) releases a resource. Speech and word segmentation are the areas most directly shaped by two defining properties of Tibetan, dialectal diversity and unmarked word boundaries, yet they have the fewest shared resources. Consent and privacy considerations can legitimately restrict the release of speech data, but the result is that TIBMD@MUC is the only entry covering the three main dialect groups, and no text resource in the table targets a specific dialect through genres such as proverbs and folk song lyrics.

Release rates by country and venue. Release rates also differ by the country of the first author's affiliation (Appendix C). Of the 18 publications from outside China, 13 (72.2%) release resources, compared with 28 of the 122 China-affiliated publications (23.0%). Publication venue accounts for part of this difference. Publications in the ACL Anthology release resources more often than those retrieved from DBLP (18 of 37, 48.6%, vs. 23 of 103, 22.3%), and among China-affiliated publications, the rate is 40.0% (10 of 25) in the ACL Anthology but 18.6% (18 of 97) elsewhere. This pattern is consistent with the artifact-release expectations of ACL venues, such as the Responsible NLP Research Checklist. In absolute terms, China-affiliated teams remain the largest contributors, accounting for 28 of the 41 releasing publications, all of the released LLMs, and five of the six recent benchmarks. All of these rates are lower bounds, because resources shared on request or through channels that the publications do not cite are not captured.

Accessibility and sustainability. GitHub[1] is the most common host (28 entries), followed by Hugging Face[2] (13 entries), and 11 entries distribute their resources across more than one platform, for example Hugging Face with GitHub (Zhou et al., 2026) or ModelScope[3] with GitHub (Han et al., 2026). Only four entries are deposited in research-data repositories that assign persistent identifiers (ScienceDB[4] and Zenodo[5]); the others depend on platforms on which repositories can be renamed, moved, or deleted. Two datasets, TIBMD@MUC and THCU, are distributed only through Baidu Netdisk[6], which can be difficult to use from outside China, while some international platforms are, conversely, not reliably accessible from mainland China. Building on the open-science recommendations in Section 6, we encourage authors of Tibetan NLP resources to (1) deposit released data and models in archives that issue persistent identifiers, (2) state an explicit licence, (3) document each resource with a datasheet or model card, and (4) mirror resources on platforms accessible to researchers both inside and outside China.

Caveats. Inclusion in Table 3 indicates that a resource has been publicly released, not that it is complete, documented, or maintained; we did not evaluate the released resources themselves. As with the rest of the survey, 2026 is only partially covered. We ensure that all links in Table 3 are accessible at 00:00 UTC, September 25, 2026, but we cannot guarantee their accessibility in the future.

G NLP Publications on Other Bodish Languages

Complementing the recommendation on related languages in Section 6, Table 4 extends the retrieval of Section 3 from Tibetan to all 82 languages of Bodish, the subgroup of Tibeto-Burman to which Ethnologue[7] assigns Tibetan. For each language, we repeated the retrieval and screening with the keyword in the third column in place of *Tibetan*. Entries that differ only in a qualifier, such as the three Tibetan or the three Tamang entries, share a keyword, and Sikkimese was also searched under its alternative name *Bhutia*.

The keyword *Tibetan* retrieves the survey corpus itself (Appendices A and B). For the other 79 languages, the search yielded only 12 publications, against 140 for Tibetan: six on Dzongkha, three on Tamang, and one each on Balti, Ladakhi, and Sikkimese. They address text processing for Dzongkha and Sikkimese (e.g., word segmentation and tagged corpora), machine translation for Tamang, and speech processing for Balti and Ladakhi. The Tamang publications, whose titles and abstracts name no variety, are listed under Eastern Tamang. The keywords of the remaining 72 languages, including all those outside Central and West Bodish, yielded none. These 12 publications are outside the survey corpus and the analyses of Section 4. As retrieval covers only the ACL Anthology and DBLP and depends on the keywords, an empty cell means that no publication was retrieved, not that no research on the language exists.

[1] https://github.com/
[2] https://huggingface.co/
[3] https://www.modelscope.cn/
[4] https://www.scidb.cn/
[5] https://zenodo.org/
[6] https://pan.baidu.com/
[7] https://www.ethnologue.com/subgroup/4085/

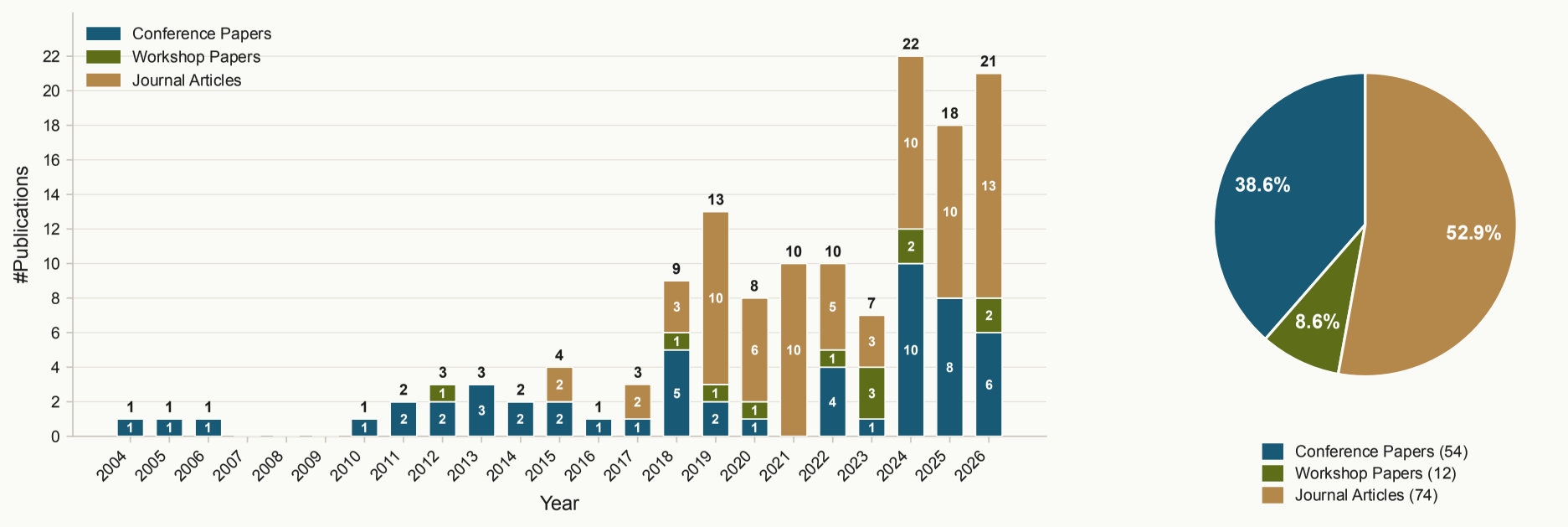


Figure 7: Full-scale version of Figure 3: distribution of the 140 Tibetan NLP publications by year and venue type, 2004–2026. **Left:** stacked bars give the annual number of conference papers (blue), workshop papers (green), and journal articles (tan); numbers inside the segments are per-type counts, and numbers above the bars are annual totals. No publications were identified for 2007–2009. Output rises from three papers in 2017 to nine in 2018 and peaks at 22 in 2024; 2026 is covered only up to 00:00 UTC, September 25, 2026. **Right:** overall share of each venue type: journal articles 52.9% (74), conference papers 38.6% (54), and workshop papers 8.6% (12).

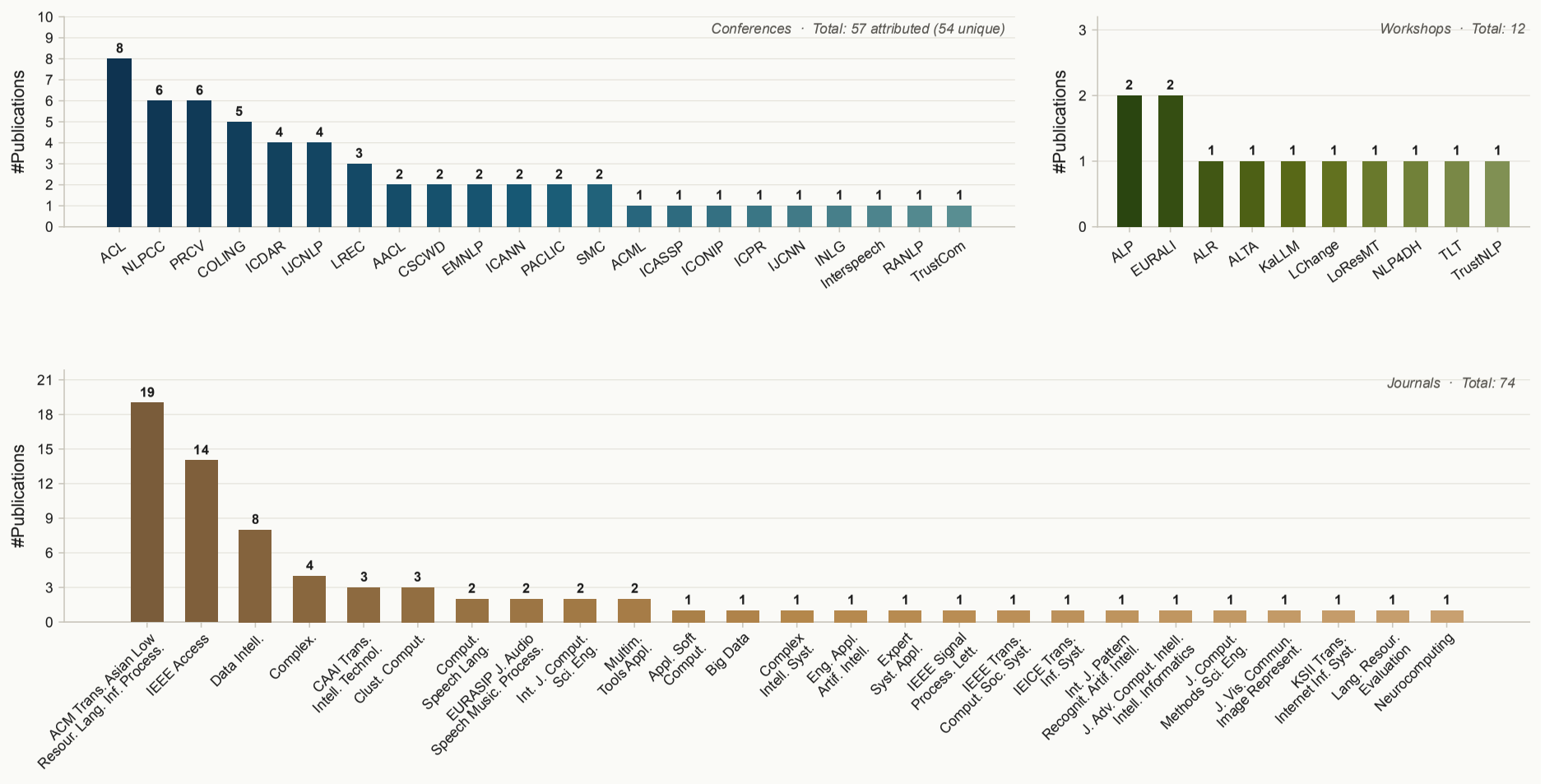


Figure 8: Full-scale version of Figure 4: number of Tibetan NLP publications per venue; within each panel, venues are sorted by count in descending order, with ties listed alphabetically. **Top left:** 22 conferences, led by ACL (8, including Findings of ACL), NLPCC (6), PRCV (6), and COLING (5). Papers at joint conferences (ACL-IJCNLP 2015 and IJCNLP-AACL 2025) are counted under each constituent venue, so the bars sum to 57 attributions for 54 conference papers. **Top right:** 10 workshops with 12 papers, led by ALP and EURALI (2 each). **Bottom:** 25 journals with 74 articles, led by *ACM Transactions on Asian and Low-Resource Language Information Processing* (19), *IEEE Access* (14), and *Data Intelligence* (8); journal names follow DBLP abbreviations. Full venue names are listed in Appendices A and B.

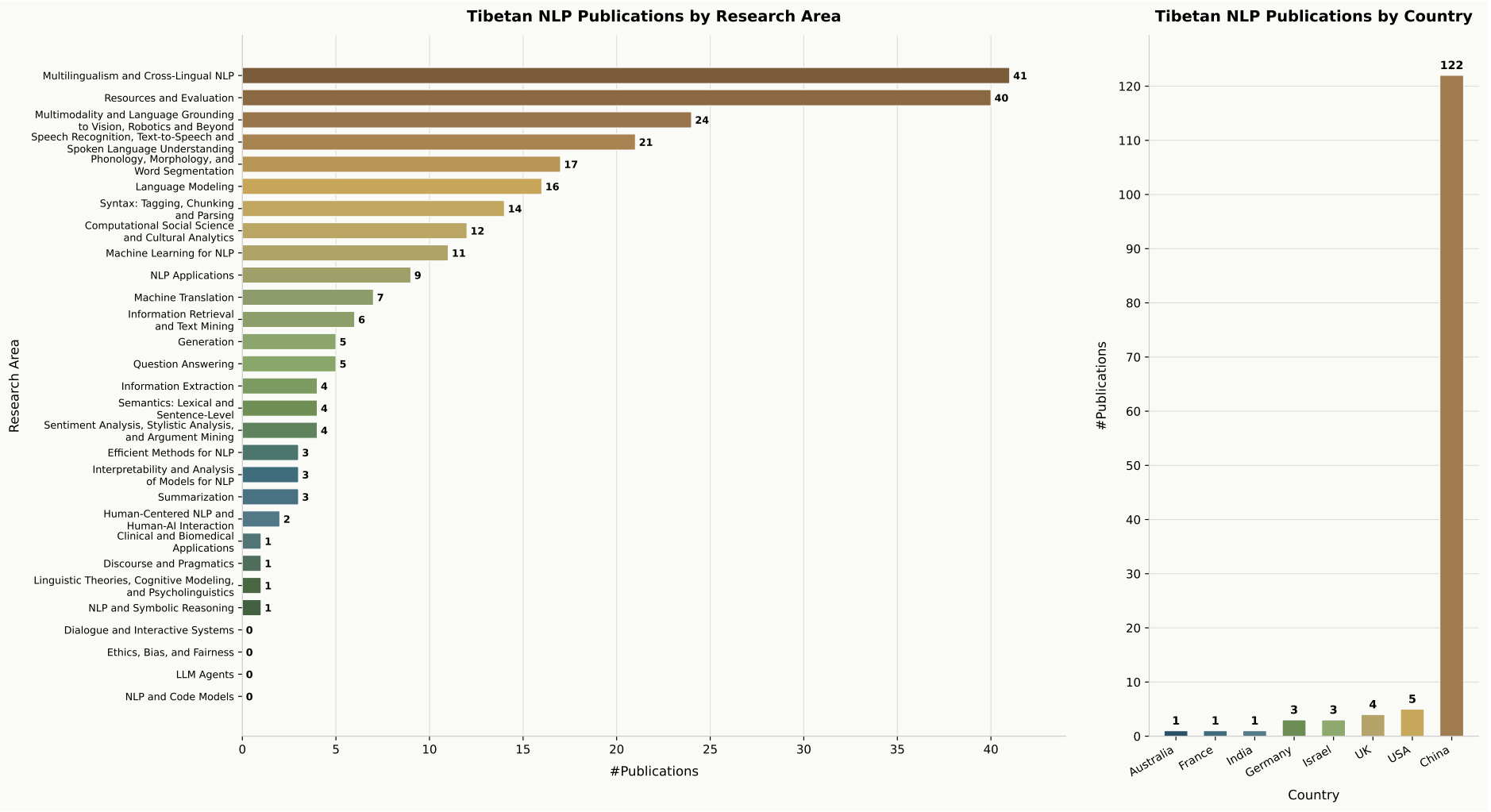


Figure 9: Full-scale version of Figure 5: number of Tibetan NLP publications by ACL Rolling Review (ARR) research area and by country of the first author's primary affiliation. **Left:** the 29 ARR research areas, sorted in descending order. Each publication was assigned up to three areas (in practice, 115 received two and 25 received one), so the counts sum to 255 rather than 140. Multilingualism and Cross-Lingual NLP (41) and Resources and Evaluation (40) lead, whereas Dialogue and Interactive Systems; Ethics, Bias, and Fairness; LLM Agents; and NLP and Code Models have no publications. **Right:** countries, sorted in ascending order: China accounts for 122 (87.1%), followed by the USA (5), the UK (4), Germany (3), Israel (3), and Australia, France, and India (one each). Per-country and per-area publication lists are given in Appendices C and D, respectively.

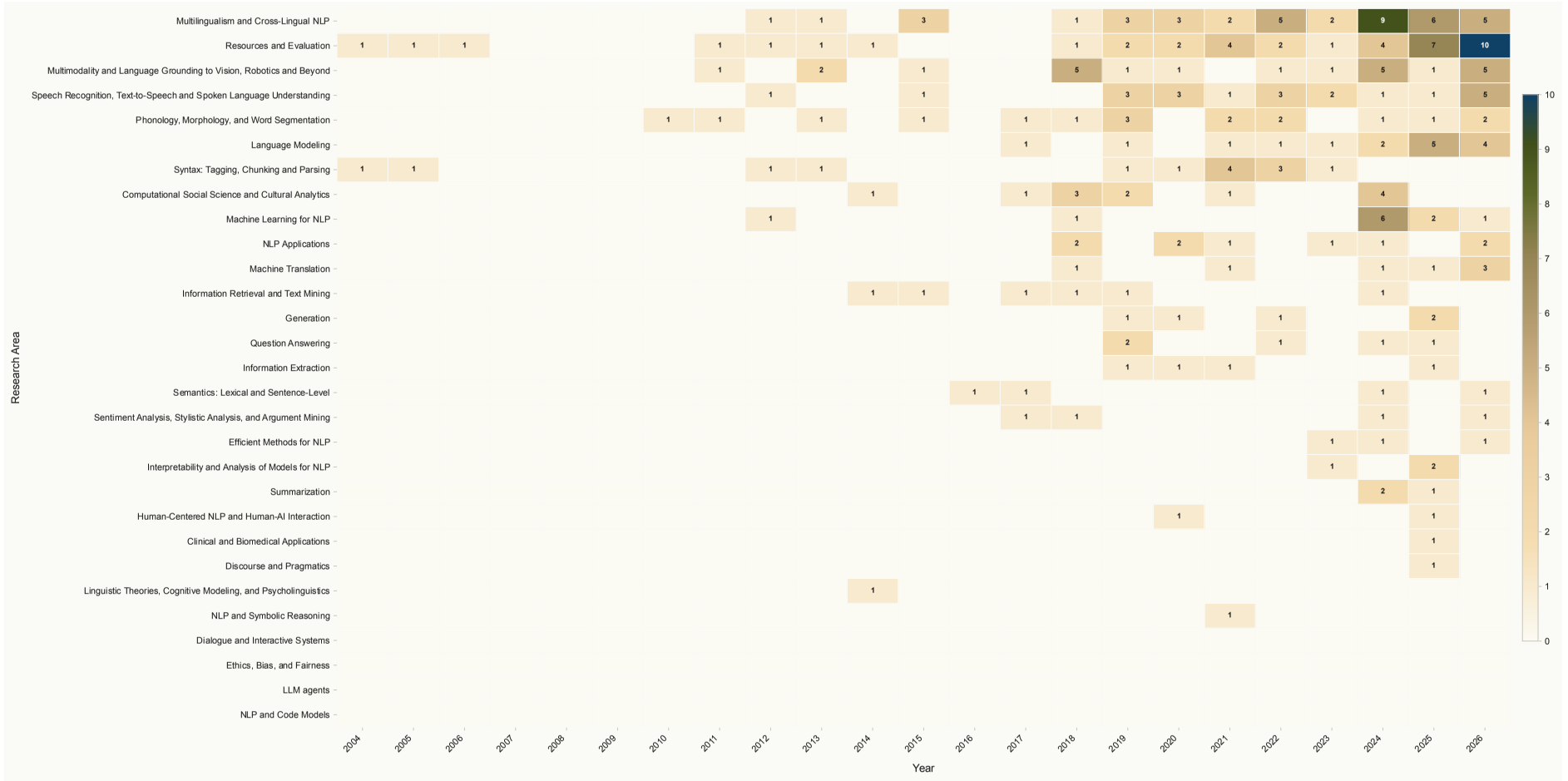


Figure 10: Full-scale version of Figure 6: number of Tibetan NLP publications in each of the 29 ARR research areas (rows) and each year (columns), 2004–2026. Rows follow the order of Figure 9 (descending total count). Cell shading darkens with the number of publications (scale 0–10); each non-empty cell shows its count, and blank cells indicate no publications. Because a publication can belong to more than one area, the cells sum to 255. The largest values are Resources and Evaluation in 2026 (10) and Multilingualism and Cross-Lingual NLP in 2024 (9); the columns for 2007–2009 are empty, and 2026 is covered only up to 00:00 UTC, September 25, 2026.

Open-source Resource	Paper	Repository URL	Description
Tibetan-Verb-Lexicon	Di et al. (2019)	https://github.com/victoriadqj/Tibetan-Verb-Lexicon	A UniMorph-style dataset of Tibetan verbal paradigms (1,433 lemmas with inflected forms) for morphological reinflection.
TIBMD@MUC	Zhao et al. (2019b, 2020)	https://pan.baidu.com/s/14CihgqjA4AFFH1QpSTjzZw	A free multi-dialect Tibetan speech corpus (Ü-Tsang, Kham and Amdo) with a phoneme set, a pronunciation dictionary and baseline recognition code, used for joint speech content recognition, dialect identification and speaker recognition.
PACTib / ACTib and annotation pipelines	Meelen and Roux (2020); Meelen et al. (2021); Faggionato et al. (2022)	https://github.com/lothelanor/actib https://zenodo.org/record/3951503	A family of Classical Tibetan corpora and tools comprising a 166M-token diachronic treebank (11th–21st c.) with segmentation, POS tags, phrase-structure parses and date metadata; ACTib v2.0, segmented and POS-tagged in small (UD) and large tag sets by a pipeline that combines rule-based, memory-based and neural methods; and an off-the-shelf pipeline from normalisation to parsing and information-structure annotation that is adaptable to Newar varieties.
Children's Stories corpus	Schmidt (2020)	https://github.com/Esukhia/children-stories	A corpus of transcribed SayMore sessions of diaspora Tibetan children's speech whose word-frequency lists the Dakje readability tool uses to grade a children's picture book.
bo-corpus-analytics	Krishna et al. (2021)	https://github.com/keutzer/bo-corpus-analytics	Unicode e-texts of the Bon and Nyingma Mind-section corpora, with notebooks for the TF-IDF, cosine-similarity, LDA and k-means analyses of Great Perfection literature.
TD50K	Sun et al. (2021)	https://github.com/andongBlue/TiData50K	A partial release of a Tibetan knowledge-graph dataset (53,797 triples, 12,573 entities and 3,285 relations) built from a Tibetan online encyclopedia.
OpenPecha Toolkit	Trinley et al. (2021)	https://github.com/OpenPecha/openpecha-toolkit	A command-line toolkit for downloading OpenPecha texts, updating and transferring annotations, and exporting views as Markdown, DOCX, PDF or EPUB.
THCU	Li et al. (2022c)	https://pan.baidu.com/s/1i1-qMcLqR0vvFJonuj5LeA (extraction code: `h5ga`)	An offline character dataset of Uchen-font Tibetan historical documents that pairs 121,214 manually annotated samples over 238 classes (THCU-M) with component-combination synthetic subsets of up to 7,238 classes (THCU-S).
TiBERT	Liu et al. (2022)	https://huggingface.co/CMLI-NLP/TiBERT	A monolingual BERT-base Tibetan pre-trained language model trained on 3.56 GB of Tibetan web text with a 30,005-token SentencePiece unigram vocabulary.
SansTib	Nehrdich (2022)	https://github.com/sebastian-nehrdich/sanstib https://huggingface.co/buddhist-nlp/sanstib	A Sanskrit–Tibetan parallel corpus of 317,289 aligned sentence pairs with 6,916 gold pairs, released with a bilingual sentence-embedding model.
TSAttacker	Cao et al. (2023)	https://github.com/UTibetNLP/TSAttacker https://github.com/metaphors/TibetanPLMsFineTuning https://github.com/metaphors/TibetanAdversarialAttack	A syllable-level black-box adversarial attack on Tibetan script for testing the robustness of Tibetan language models.
fs-TCR	Bao et al. (2024)	https://github.com/coder-bossin/fs-TCR	Code for cross-domain few-shot Tibetan character recognition that uses contrastive pre-training and meta fine-tuning with episodic synthetic query sets.
TibetanQA2.0	Dan and Sun (2024)	https://doi.org/10.57760/sciencedb.j00104.00120	A crowd-sourced Tibetan machine reading comprehension dataset of 505 passages and 1,347 unanswerable question–answer pairs over six topics, released under CC BY 4.0.
ConceptNet language adapters	Gurgurov et al. (2024)	https://github.com/d-gurgurov/Injecting-Commonsense-Knowledge-into-LLMs https://github.com/d-gurgurov/Conceptnet-Embeddings	Code, ConceptNet-extracted data and language adapters that inject knowledge-graph knowledge into multilingual LLMs for eight low-resource languages, including Tibetan.
Tibetan-PLM	Liang et al. (2024)	https://github.com/Dslab-NLP/Tibetan-PLM https://huggingface.co/Pagewood/TBERT-pytorch https://huggingface.co/Pagewood/TBERT-tensorflow https://huggingface.co/Pagewood/Tibetan-BERT-wwm	The Tibetan-BERT and Tibetan-BERT-wwm pre-trained models, the latter using whole word masking to learn word-level representations for Tibetan text classification.
Tibetan Handwriting Khyug-yig Dataset	Tashi et al. (2024)	https://github.com/13209413223/Tibetan-Handwriting-Dataset	A dataset of 9,874 labelled word images in the cursive Khyug-yig style written by 63 writers, with more data to be released incrementally.
MC^2	Zhang et al. (2024)	https://github.com/luciusssss/mc2_corpus https://huggingface.co/datasets/pkupie/mc2_corpus https://huggingface.co/pkupie/mc2-xlmr-large https://huggingface.co/pkupie/mc2-llama-13b	An open pre-training corpus for Tibetan, Uyghur, Kazakh (Arabic script) and Mongolian (traditional script).
Ti-ChTrans	Zhou et al. (2024a)	https://github.com/UTibetNLP/Ti-ChTrans	Transformer Tibetan–Chinese translation models trained on CINO-filtered parallel data, released with an expert-reviewed benchmark test set of 1,000 sentence pairs (483 common phrases and 517 news sentences).
TSCheater, TSVSDB & AdvTS	Cao et al. (2025a)	https://github.com/metaphors/TSAttack https://huggingface.co/collections/UTibetNLP/tibetan-victim-language-models-669f614ecea872c7211c121c	The TSCheater visual-similarity-based Tibetan adversarial text generator, released with the TSVSDB syllable visual-similarity database (18,833 syllables), AdvTS (the first human-proofread Tibetan adversarial robustness benchmark) and eight victim classifiers (Tibetan-BERT and CINO-small/base/large, each fine-tuned on TNCC-title and TU_SA).
HITL-GAT	Cao et al. (2025b)	https://github.com/metaphors/HITL-GAT	A human-in-the-loop system for generating adversarial texts, released under GPL-3.0.

(continued on next page)

(continued from previous page)

Open-source Resource	Paper	Repository URL	Description
TLUE	Gao et al. (2025)	https://github.com/Vicentvankor/TLUE	The first large-scale Tibetan LLM benchmark, pairing a multi-task understanding set (67 subdomains) with a safety set.
Tibetan-tokenizer	Gyatso et al. (2025)	https://github.com/gyatso736/-Tibetan-tokenizer-	A Bi-LSTM+CRF Tibetan word segmenter used to pre-segment text before the paper's agglutination-aware BPE for Tibetan–Chinese machine translation.
DharmaBench	Hashiloni et al. (2025)	https://github.com/Intellexus-DSI/DharmaBench	A 13-task benchmark on Buddhist texts in Sanskrit and Classical Tibetan, with datasets (CC-BY-4.0), code (Apache-2.0) and prompts.
T-LLaMA	Lv et al. (2025)	https://huggingface.co/Pagewood/T-LLaMA	A LLaMA 2-based Tibetan large language model with a SentencePiece-expanded Tibetan vocabulary, trained on a 2.2-billion-character Tibetan corpus.
Tibetan-LLaMA 2	Sha et al. (2025)	https://github.com/Shajiu/Tibetan-LLaMA-2 https://huggingface.co/shajiu/Yak_Llama2_7B https://huggingface.co/shajiu/Yak_Llama2_13B https://huggingface.co/datasets/shajiu/TibetanGeneral_corpus https://huggingface.co/datasets/shajiu/TibetanSft_corpus	A LLaMA 2 model extended with a Tibetan vocabulary, continually pre-trained on a 15 GB Tibetan corpus and instruction-tuned on 1M instructions, released with inference resources.
Tibetan prompt-learning text classifier	Tang et al. (2025)	https://github.com/Greenhorntc/Text-Classification-in-Tibetan-Language-based-on-Prompt-Learning	Code that recasts Tibetan text classification as prompt-based entailment and ensembles several prompt-tuned language models, evaluated on TNCC and WCM.
CMHG	Xu et al. (2025)	https://huggingface.co/datasets/KEVVVV/CMHG	A headline-generation dataset (100K Tibetan, 50K Uyghur and 50K Mongolian) with a native-speaker-annotated test set.
MiLiC-Eval	Zhang et al. (2025a)	https://github.com/luciusssss/MiLiC-Eval https://huggingface.co/datasets/pkupie/milic-eval	A 24K-instance, 9-task benchmark that evaluates LLMs on Tibetan, Uyghur, Kazakh and Mongolian in parallel.
CUTE	Zhuang and Sun (2025)	https://github.com/CMLI-NLP/CUTE https://huggingface.co/datasets/CMLI-NLP/CUTE-Datasets https://huggingface.co/CMLI-NLP/CUTE-Llama	Parallel and non-parallel 25 GB Chinese–Uyghur–Tibetan–English corpora, released with CUTE-Llama models trained on each.
TIFD	Zhuang et al. (2025a)	https://doi.org/10.57760/sciencedb.j00104.00146	A Tibetan instruction-following dataset of 11,535 instruction–input–output records, translated from GPT-4 responses and manually proofread, for supervised fine-tuning of LLMs.
ALLO/AUTO segmentation	Bilitski et al. (2026)	https://github.com/Intellexus-DSI/alloauto	The first annotated corpus for separating native Tibetan from translated passages in Classical Tibetan texts, released with trained models and scripts.
Tib-Bi-Dharma / SenSim	Cohen et al. (2026)	https://github.com/Intellexus-DSI/sensim	Tib-Bi-Dharma, a Classical Tibetan sentence-similarity model, released with an expert gold STS set, two automatically scored STS sets and the pipeline code.
Vocabulary scaling	Han et al. (2026)	https://github.com/White2360/vocab-scaling-low-resource https://www.modelscope.cn/collections/lingfeng2360/ACL-2026-Findings-Vocab-Scaling	The code repository for a study of joint Mongolian–Tibetan–Uyghur vocabulary size under BPE and byte-level BPE.
TibOCR-Bench / TibOCR-Train	Jie et al. (2026)	https://doi.org/10.57760/sciencedb.28968	A multimodal Tibetan OCR benchmark covering 30 real-world scenarios plus a preference-aligned training set, which the authors state they release together with the data pipeline, checkpoints, and training and inference code.
CMiLBench	Li et al. (2026)	https://github.com/CMLI-NLP/CMiLBench	A hierarchical benchmark (24,663 instances, 17 tasks and 5 levels) testing ability, culture and safety in Tibetan, Mongolian and Uyghur.
tibetan-wer / tibetan-word-segmentation-ds	Moore et al. (2026)	https://pypi.org/project/tibetan-wer/ https://huggingface.co/datasets/billingsmoore/tibetan-word-segmentation-ds	A Python package for character, syllable and segmented word error rate (SWER) that scores Tibetan ASR output at the word level, released with expert word-boundary annotations of 100 NICT-Tib1 utterances (2,634 inter-syllable positions) for measuring agreement between automatic segmenters and a human annotator.
FTibSuite	Xu et al. (2026)	https://huggingface.co/onedday/FTib-VLM https://huggingface.co/datasets/onedday/FTib-data	FTibVLM, a Tibetan vision–language model built on Qwen3-VL-8B-Instruct, released with FTibData for pre-training, alignment and instruction tuning.
TibHCR	Zhou et al. (2026)	https://huggingface.co/datasets/qixiaoke/TibHCR https://github.com/qixiaoke007/TibHCR	A large-scale offline handwritten Tibetan character dataset of 141,698 samples from 235 writers across 47 classes, benchmarked with CRNN and a PP-OCRv4 cross-lingual fine-tuning baseline.

Table 3: Open-source resources released by the surveyed publications, comprising corpora, datasets and benchmarks, pre-trained and large language models, and code and toolkits. The 38 entries come from 41 of the 140 publications (29.3%); only resources released by the papers' own authors at a publicly accessible URL are included. Entries are ordered by year of publication and then alphabetically by first author. A resource described in several papers (TIBMD@MUC and PACTib/ACTib) is listed once, under its earliest paper, with all related papers cited. Descriptions summarize each resource's content and, where reported, its scale and licence.

Language	ISO 639-3	Search Keyword	Publications
Bodish, no further subgroup (5)			
Gongduk	goe	Gongduk	
Lhokpu	lhp	Lhokpu	
Monpa, Kalaktang	kkf	Monpa	
Olekha	ole	Olekha	
Tshangla	tsj	Tshangla	
Central Bodish (38)			
Balti	bft	Balti	Sharif et al. (2025).
Brokkat	bro	Brokkat	
Brokpake	sgt	Brokpake	
Changthang	cna	Changthang	
Chocangacakha	cgk	Chocangacakha	
Choni	cda	Choni	
Dolpo	dre	Dolpo	
Dzongkha	dzo	Dzongkha	Wangchuk et al. (2022); Dhungyel and Grundspenkis (2017); Arun (2014); Dasgupta et al. (2012); Norbu et al. (2010); Chungku et al. (2010).
Groma	gro	Groma	
Gyalsumdo	gyo	Gyalsumdo	
Humla	hut	Humla	
Hyolmo	scp	Hyolmo	
Jad	jda	Jad	
Jirel	jul	Jirel	
Kyirong	kgy	Kyirong	
Ladakhi	lbj	Ladakhi	Thukroo et al. (2023).
Lakha	lkh	Lakha	
Layakha	lya	Layakha	
Lhomi	lhm	Lhomi	
Lhowa	loy	Lhowa	
Lunanakha	luk	Lunanakha	
Mugom-Karmarong	muk	Mugom-Karmarong	
Naaba	nao	Naaba	
Nubri	kte	Nubri	
Purig	prx	Purig	
Sherpa	xsr	Sherpa	
Sikkimese	sip	Sikkimese, Bhutia	Ashida et al. (2020).
Spiti Bhoti	spt	Spiti Bhoti	
Stod Bhoti	sbu	Stod Bhoti	
Syuba	syw	Syuba	
Tibetan, Amdo	adx	Tibetan	
Tibetan, Central	bod	Tibetan	
Tibetan, Khams	khg	Tibetan	
Tichurong	tcn	Tichurong	
Tseku	tsk	Tseku	
Tsum	ttz	Tsum	
Walungge	ola	Walungge	
Zangskari	zau	Zangskari	
East Bodish (9)			
Bumthangkha	kjz	Bumthangkha	
Chalikha	tgf	Chalikha	
Dakpakha	dka	Dakpakha	
Dzalakha	dzl	Dzalakha	
Khengkha	xkf	Khengkha	
Kurtokha	xkz	Kurtokha	
Monpa, Tawang	twm	Monpa	
Nupbikha	npb	Nupbikha	
Nyenkha	neh	Nyenkha	
West Bodish (14)			
Chantyal	chx	Chantyal	
Dura	drq	Dura	
Ghale, Northern	ghh	Ghale	
Ghale, Southern	ghe	Ghale	
Gurung	gvr	Gurung	
Kaike	kzq	Kaike	
Kuke	ght	Kuke	
Nar Phu	npa	Nar Phu	
Nyeshangte	nmm	Nyeshangte	
Seke	skj	Seke	
Tamang, Eastern	taj	Tamang	Ghimire et al. (2026); Acharya et al. (2026); Chaudhary et al. (2020).
Tamang, Eastern Gorkha	tge	Tamang	
Tamang, Western	tdg	Tamang	
Thakali	ths	Thakali	

(continued on next page)

(continued from previous page)

Language	ISO 639-3	Search Keyword	Publications
West Himalayish (16)			
Byangsi	bee	Byangsi	
Chaudangsi	cdn	Chaudangsi	
Darmiya	drd	Darmiya	
Gahri	bfu	Gahri	
Jangshung	jna	Jangshung	
Kanashi	xns	Kanashi	
Kinnauri	kfk	Kinnauri	
Kinnauri, Bhoti	nes	Kinnauri	
Kinnauri, Chhoyul	tpq	Kinnauri	
Kinnauri, Chitkuli	cik	Kinnauri	
Pattani	lae	Pattani	
Rangkas	rgk	Rangkas	
Rongpo	rnp	Rongpo	
Shumcho	scu	Shumcho	
Sunam	ssk	Sunam	
Tinani	lbf	Tinani	

Table 4: The 82 Bodish languages and the NLP publications retrieved for each. Language names and ISO 639-3 codes are taken from Ethnologue. The retrieval method follows Section 3, with the search keyword in place of *Tibetan*. Entries that differ only in a qualifier share a keyword and hence their publications: those on Tibetan form the survey corpus (Appendices A and B) and are not repeated here, and those on Tamang are listed under Eastern Tamang. Elsewhere, an empty cell means that no publication was retrieved.