Proceedings of the EURALI Workshop @LREC2020, pages 1–6
Marseille, 20 June 2022

NLP Pipeline for Annotating (Endangered) Tibetan and Newar Varieties

[1]**Christian Faggionato,** [2]**Nathan Hill,** [1]**Marieke Meelen**
[1]University of Cambridge & [2]SOAS University of London and Trinity College Dublin
{cf566,mm986}@cam.ac.uk, nathan.hill@tcd.ie

Abstract

In this paper we present our work-in-progress on a fully-implemented pipeline to create deeply-annotated corpora of a number of historical and contemporary Tibetan and Newar varieties. Our off-the-shelf tools allow researchers to create corpora with five different layers of annotation, ranging from morphosyntactic to information-structural annotation. We build on and optimise existing tools (in line with FAIR principles), as well as develop new ones, and show how they can be adapted to other Tibetan and Newar languages, most notably modern endangered languages that are both extremely low-resourced and under-researched.

Keywords: Tibetan, Newar, Corpora, Segmentation, POS tagging, parsing, Information Structure

1. Introduction

There are numerous varieties of Tibetan and Newar languages of the Bodish and Himalayish branches of the Sino-Tibetan language family respectively. These varieties share common innovations, but are often not mutually intelligible. In this paper we present a comprehensive NLP pipeline to create annotated corpora of historical Tibetan texts from the earliest Old Tibetan period (8-11th c.) onwards. We aim to present off-the-shelf tools that researchers can use to create exactly the type of linguistic corpus they need, i.e. standardised & normalised text (re)converted to Tibetan Unicode script (1), text with (word and sentence) segmentation (2), with morphosyntactic annotation (3), with parsed phrase structure (4), or deeply annotated corpora including all of the preceding, but further enriched with information-structural annotation, such as animacy for noun phrases, as well as topic and focus phrases (5).[1] In Section 2, we present our tools and the three phases of our annotation pipeline, with concrete examples from the most challenging part of our historical Tibetan corpora: the Old Tibetan *Rāmāyana*. In Section 3, we first show how this pipeline can be adapted to work for related historical Tibetan varieties like South Mustang Tibetan, but also more distantly-related languages, like Classical Newar. Finally, we demonstrate how these tools can be adapted to endangered modern varieties like Sherpa and Lhomi Tibetan and Kathmandu, Dolakha and Lalitpur Newar. Our pipeline and tools are important, because they can deal with extremely low-resource and under-researched languages that are highly endangered. Off-the-shelf tools like these with instructions on how to adapt them will give researchers the opportunity to use this as a blueprint for any (Asian) language for which no resources are available.

2. Annotation Pipeline

We develop our entire three-phased pipeline (Fig. 1) in accordance with CLARIN standards and FAIR Data Principles, making our resources and tools **F**indable and **A**ccessible, whilst ensuring **I**nteroperability, and **R**eusability (Wilkinson et al., 2016). This means that wherever possible, existing tools are adapted and optimised, rather than reinvented. In addition, our pipeline is deliberately semi-supervised, with two optional stages of manual correction if perfect gold standards are required.

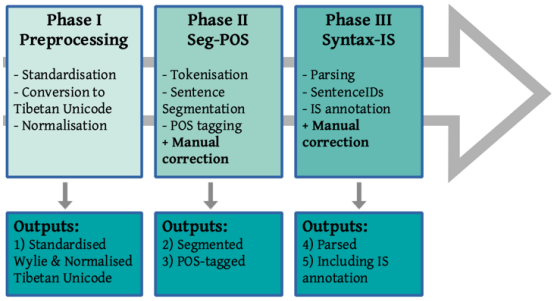

Figure 1: Historical Tibetan Pipeline and Outputs

In this section, we use example input from the Old Tibetan *Rāmāyana* (Fig. 2) to illustrate each stage of the annotation process for historical Tibetan and all output formats underneath.

2.1. Preprocessing

In the preprocessing phase of our annotation pipeline we use as input an adapted version of the Wylie transliteration system from the Old Tibetan Documents Online (OTDO) website (Fig. 2a). We standardise the OTDO Wylie to normal Wylie using a set of replacement rules, and we clean the text from the OTDO editorial conventions using a set of regular expressions (Fig. 2b). In the end we convert the standardised Old Tibetan Wylie into Old Tibetan Unicode script (Fig. 2c) through the THL's Online Tibetan Transliteration Converter, which can also be integrated into our overall pipeline using the more optimised Python implementation developed by Esukhia. The second step of the Preprocessing Phase consists of the normalisation of the Old Tibetan Unicode script. Old Tibetan presents differences in orthography compared to Classical Tibetan. Through a

[1]Code and links to corpora can be found at http://github.com/lothelanor/actib.

a)	**OTDO input**	nga	nI	tsangs pha	'I	long spyod	la	ma	chagste
b)	**Standardised**	nga	n-i	tsangs pha	'-i	long spyod	la	ma	chagste
c)	**Converted**	ང་ནི་ཙངས་ཕ་འི་ལོང་སྤྱོད་ལ་མ་ཆགས།							
d)	**Normalised**	ང་ནི་ཚངས་པ་འི་ལོང་སྤྱོད་ལ་མ་ཆགས་ཏེ							
e)	**Segmented**	ང་	ནི་	ཚངས་པ་	འི་	ལོང་སྤྱོད་	ལ་	མ་	ཆགས་	ཏེ
f)	**POS tags**	p.pers	cl.top	n.prop	case.gen	n.count	case.all	neg	v.invar	cv.sem
g)	**UD tags**	PRON	PART	PROPN	ADP	NOUN	ADP	PART	VERB	PART
h)	**Animacy**	+human		+human		+inanimate				
i)	**Parsed**	(CP-MAT-SPE (NP-TOP (NP-PRO (FS+human)(p.pers ང་))(cl.top ནི་)) (PP (NP (NP-NPR (FS+human)(n.prop ཚངས་པ་)(case.gen འི་)) (NP (FS+inanimate)(n.count ལོང་སྤྱོད་)))(case.all ལ་)) (NEGP (neg མ་)) (VP (v.invar ཆགས་)) (cv.sem ཏེ་))								
j)	**Translation**	'As for me, I'm not attached to the enjoyments of Brahma.'							

Figure 2: Example from the Old Tibetan *Rāmāyana*

set of rules written in the Constraint Grammar formalism (Cg3) and python, we deal with these differences (with $>$ 99% accuracy in 'normalisation' into Classical Tibetan) (Faggionato and Garrett, 2019) so that we can employ existing NLP tools for Classical Tibetan for further annotation. Classical Tibetan texts that are often available as eTexts in Tibetan Unicode can skip this Prepocessing Phase and go directly to Phase II (described in Section 2.2).

2.2. Segmentation & POS tagging

Since the Tibetan script does not indicate meaningful word or sentence boundaries (only syllables are marked in Fig. 2d), we first need to segment our standardised text. For word segmentation (tokenisation), we optimise a syllable-based tokeniser (Meelen and Hill, 2017), inserting missing *a chung* (transcribed as ') in cases where they were cut due to regular sandhi-like mergers in the Tibetan script. Reinsertion of these characters is effectively a form of lemmatisation of all nominal categories ending in *a chung*, e.g. *mkha'i* > *mkha' 'i* 'of the sky'. Similarly, we optimise a sentence segmentation script (Faggionato and Meelen, 2019), extending it with more detailed rules, e.g. rules which automatically capture direct speech based on common Tibetan direct speech markers like *na re, zhes*, etc. Sentence boundaries at this stage are marked by `<utt>`. To make automatic parsing and manual correction more feasible (i.e. avoiding extraordinarily long sentences that are impossible to correct on a screen), we also split consistently after semifinal particles (`cv.sem` in Fig. 2f), even though syntactically they can often function as subordinate clauses. For POS tagging, we extended an existing Tibetan tagger (Meelen et al., 2021) to facilitate downstream tasks related to the identification of information-structural (IS) features, e.g. by adding a specific tag for the topic marker *ni* (`cl.top` in Fig. 2f). In addition, we provide the option of converting the detailed tag set developed for historical Tibetan (Garrett et al., 2015) to the Universal Dependencies (UD) tag set (Fig. 2g). Since the Global Accuracy of the overall segmentation and POS tagging is >95% (especially with these improvements), the output can be fed directly to the next Phase. However, if Gold Standards or simply better downstream results are required, we recommend a round of manual correction with Pyrrha (Clérice et al., 2022). This online user-friendly annotation tool facilitates efficient manual correction by providing fixed tag lists as well as useful lists of occurrences throughout the corpus with bulk-correction options (Fig. 3).

781	ང་	p.pers	སངས་ཞེས་བྱས་པ་དང་ ། དང་སྲོང་གིས་ བཀུར་ ཐུགས་པ ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt>	11
782	ནི་	cl.focus	ཞེས་ བྱས་པ་ དང་ ། དང་སྲོང་ གིས་ བཀུར་ ཐུགས་པ ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་	75
783	ཚངས་པ་	n.prop	བྱས་པ་ དང་ ། དང་སྲོང་ གིས་ བཀུར་ ཐུགས་པ ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་ སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་	4
784	དེ་	case.gen	དང་ ། དང་སྲོང་ གིས་ བཀུར་ ཐུགས་པ ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་	290
785	ལོང་སྤྱོད་	n.count	། ། དང་སྲོང་ གིས་ བཀུར་ ཐུགས་པ ། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན་	3
786	ལ་	case.all	དང་སྲོང་ གིས་ བཀུར་ ཐུགས་པ ། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན །	205
787	མ་	neg	གིས་ བཀུར་ ཐུགས་པ ། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན ། ཐུ	120
788	ཆགས་	v.invar	བཀུར་ ཐུགས་པ ། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན ། ཐུ་ དེ་	6
789	ཏེ་	cv.sem	ཐུགས་པ ། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན ། ཐུ་ དེ་ དང་སྲོང་	43
790	।	punc	། ། ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । <utt> དབེན་པ་ དེ་ གནས་ ན ། ཐུ་ དེ་ དང་སྲོང་ སཝདཀ་	795
791	<utt>	<utt>	ང་ ནི་ ཚངས་པ་ དེ་ ལོང་སྤྱོད་ ལ་ མ་ ཆགས་ ཏེ་ । **<utt>** དབེན་པ་ དེ་ གནས་ ན ། ཐུ་ དེ་ དང་སྲོང་ སཝདཀ་ ལ་	143

Figure 3: Pyrrha - Manual Correction of POS tags, Word and Sentence Segmentation

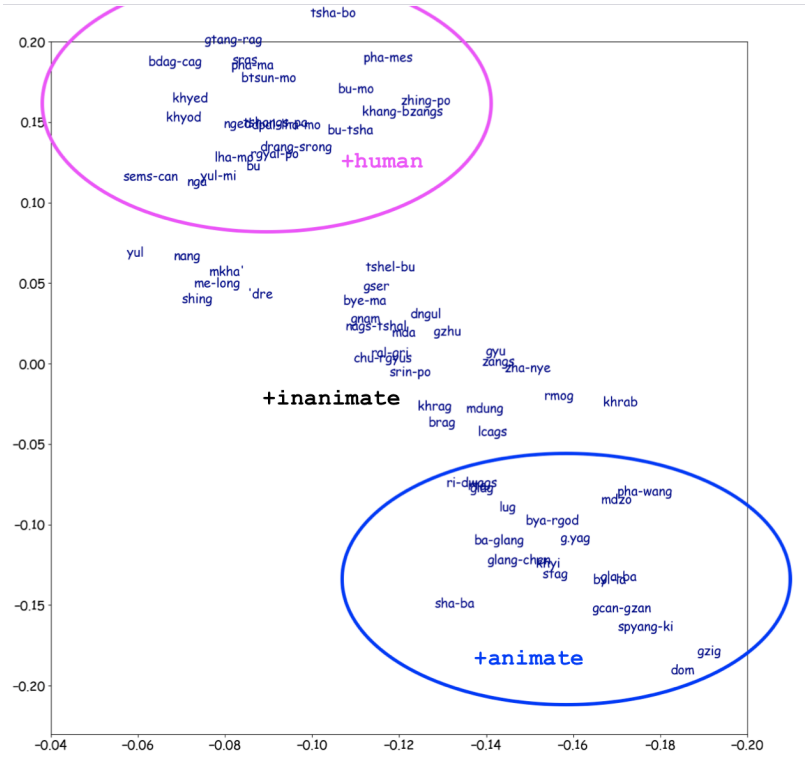

Figure 4: SVD of 100D vector representations of nouns showing Animacy clusters.

2.3. Syntax & Information Structure

We focus on constituency-based phrase structure as it is a better fit for our research questions regarding egophoricity, but dependency-based parsing is also possible (Faggionato and Garrett, 2019), (Faggionato, 2021). Conversion to/from either format is still an option at any time. Although constituency-based parsers are available for historical Tibetan (Meelen and Roux, 2020), these only provide rudimentary phrase structure. We extend these existing parsers to be able to capture complex Noun Phrases (`NP`) embedded within Postpositional Phrases (`PP`) as well as focus and topic phrases (`NP-FOC` and `NP-TOP` in Fig. 2i). The detailed tag set extensions thus help to identify important aspects of Information Structure already, i.e. **topics** and **foci**. The topic marker *ni* is classified with the POS label `cl.top`, for example (cf. Fig.2f). This is an Aboutness Topic (Frascarelli and Hinterhölzl, 2007), which can be translated into English as 'as for NP'. Similarly, Old and Classical Tibetan have focus markers, for example those marking narrow focus through particles like *kyang* 'even', which are labelled as `cl.foc`. Again, these detailed POS tags can help us derive syntactic phrase labels like `NP-FOC` automatically.

Finally, we annotate the **Animacy** of all Noun Phrases, providing them with a `+human`, `+animate` or `+inanimate` label that is integrated into the parsed bracketing format (Fig.2h and i). Animacy labels are assigned through a combination of feature-based rules and a dedicated Semantic Textual Similarity (STS) cluster-based classifier, which assigns Animacy labels based on KDTree distance measures to an average vector of tokens that are manually labelled as `+human`, `+animate` or `+inanimate`.[2] In addition, labels for certain tokens can be derived from POS tags. Tibetan personal pronouns, for example, can only refer to humans (demonstratives are used to refer to animals). Since our detailed POS tag set makes a distinction, we can automatically derive `+human` Animacy labels for NPs containing personal pronouns. Similarly, a combination of detailed POS tags and syntactic annotation allows us to automatically distinguish `+human` proper nouns, i.e. personal names, from place names (`+inanimate`), because humans typically have agentive case markers, whereas place names often occur with locatives. These rules are refined with dedicated verb classes and known argument structure information (Solmsdorf et al., 2021), (Lugli et al., 2021) and the Interactive Tibetan Valency Dictionary). Finally, we manually compiled a list of frequently-occurring animals, which allowed us to compare the semantic vector representations[3] of all new noun phrases with the labeled clusters of pronouns and personal names, animals and place names. Unseen NPs are categorised according to their highest cosine similarity to any of the clusters shown in a preliminary SVD plot in Fig. 4.

[2]For a full discussion and detailed evaluation, see (Meelen, 2022) and (Hill, 2022).

[3]These are based on FastText embeddings trained on the $185m$-token ACTib corpus (Meelen and Roux, 2020).

After automatic parsing and IS annotation, both can be manually corrected with the dedicate user-friendly tool Cesax (Komen, 2013). In addition to facilitating quick and easy correction of syntax and information structure, Cesax provides the option of semi-automatic coreference resolution based on predefined features, enhancing our IS annotation further with **topic chains**. Fig. 5 shows a screenshot of the 'tree view' option of the Cesax interface where both syntax and information structure can be corrected manually. Cesax allows for automatic conversion of parsed (`.psd`) files to TEI-compatible XML files (`.psdx`), but other outputs, e.g. FOLIA XML (compatible with ANNIS), plain text files with bracket structure shown in Fig. 2i, or UD-style CoNNL-U formats. In addition, it can export query results to R or other statistics tools. Altogether, this means the annotated corpora can be queried and analysed in many different ways (e.g. using customised XQuery or CorpusSearch (Randall et al., 2005)) catering to any kind of linguistic research.

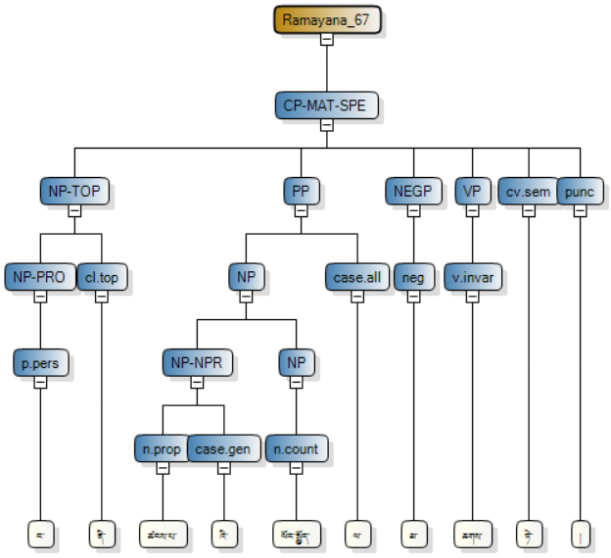

Figure 5: Syntax and IS correction in Cesax

3. Extension to Other Languages

Our pipeline with its accompanying tools can be easily adapted to other historical as well as modern, endangered Tibetan and Newar languages.

3.1. Other Historical Varieties

For **historical South Mustang Tibetan**, we can use transcriptions on the Tibetan social history project website. These transcriptions are done in Wylie, so we enter them into our pipeline at the Preprocessing Phase converting the Wylie transliterations back to Tibetan Unicode script for which we have optimised downstream NLP tools. Since historical South Mustang Tibetan is very similar to other historical Tibetan varieties, we can use the exact same tools and pipeline afterwards. For **Classical Newar**, we deal with two different sources. The first source consists of manuscript images that need to be transcribed into roman script with additional diacritics commonly used in the field. This is done using the Handwritten Text Recognition (HTR) tool Transkribus (Colutto et al., 2019), trained using Ground Truth data available for Sanskrit manuscripts written in a similar Pracalit script (Otter, nd), (Shakya and Bajracharya, 2001). The second source consists of PDF scans of romanised Newari texts. In order to properly OCR these PDFs and render all the diacritics in the texts we use Tesseract (Patel et al., 2012) with a model trained on the International Alphabet of Sanskrit Transliteration (IAST), a transliteration scheme for Indic scripts. These transcriptions are already segmented, but before running the POS tagger and applying the rest of the pipeline, we need to cut off the case suffixes from (pro)nouns to produce an accurate tokenisation similar to that of our historical Tibetan corpus.

3.2. Modern Endangered Varieties

When working with endangered or vulnerable languages there are many challenges for standard NLP pipelines. First of all, the lack of writing systems poses an intricate challenge in terms of language documentation, which creates a bottleneck at the transcription phase due to the lack of standardised conventions. Second, the limited amount of data means off-the-shelf NLP tools usually cannot be applied (Anastasopoulos et al., 2020). For Modern Tibetan and Newar varieties, all source material comes from fieldwork on language documentation projects. For this paper, we test our historical NLP pipeline for both *vulnerable* modern languages (i.e. Hile Sherpa - on the road to extinction - and Kathmandu Newar) and *endangered* ones (i.e. South Mustang Tibetan, Dolakha Newar, Lalitpur Newar and Lhomi). There are at least three different varieties of Modern Newar. Dolakha Newar spoken in a more remote region east of Kathmandu is not mutually intelligible with the varieties spoken in the Kathmandu Valley (Genetti, 2009). For **Kathmandu Newar**, we start with the fieldwork stories kindly provided by Austin Hale (Hale, nd) since the texts are in FLEx format, i.e. transcribed into IPA, segmented and glossed. This means that in our Preprocessing Phase we only need to extract the line with segmented morphemes and glosses. This gives us $10k$ tokens we can use to start training a Part-of-Speech (POS) tagger. $10k$ tokens is not nearly enough for any off-the-shelf neural-network-based taggers, but it is enough to start incrementally training a Memory-Based Tagger like the TiMBL MBT (Daelemans et al., 2003). Even though this is not a recently-developed tool, it is one of the most effective methods for developing a POS tagger from scratch since it can learn from specific features like initial and final characters as well as the context, yielding high accuracies even for extremely small data sets (Meelen et al., 2021). Once more fieldwork data (also for closely-related **Lalitpur**

Newar) has been collected, we can use this preliminary POS tagger to annotate more texts, which we will then correct with Pyrrha to create larger Gold Standards that will improve the Global Accuracy. The result can then be fed into the remaining Syntax and IS Phase of our pipeline. For **Dolakha Newar**, we can follow the same route, but only after digitising the stories published in Genetti (2009). **South Mustang Tibetan** with 1800 speakers is a severely endangered language spoken in a number of villages in Mustang, Nepal, with fieldwork data from the 1990s (Kretschmar, 1995). As these are also not available in digital format, we will collect new data in Mustang and archive it alongside romanised and IPA transcriptions after which it can enter the POS-tagging stage of our pipeline, following the same incremental annotation procedure sketched for modern Newar above. For Modern Newar varieties (as well as Sherpa and Lhomi below) the sentence segmentation is straightforward, since they are indicated with a *daṇḍa* in the case of Classical Newar and Sherpa, and a full stop in the case of Lhomi, Modern Newar and South Mustang Tibetan. **Sherpa** is a vulnerable language mainly spoken in Solukhumbu, northeast Nepal (Graves, 2007). The only Sherpa text at our disposal was a New Testament translation in Devanagari script, which is used since most Sherpa speakers read Nepali in Devanagari, but is very unsuitable for Sherpa phonotactics, which is why we convert it to romanised script (like our historical Newar). The preprocessing is then straightforward and in line with what we did for our Tibetan texts. After the script conversion, we clean it from unwanted non-textual materials (headers, footnotes, page numbers and cross-references) with a set of regular expressions. Again, similar to Classical Newar, we improve existing tokenisation by cutting off case markers from (pro)nouns, which means we can use similar downstream tagging tools. The last low-resourced Tibetan variety that we tested is **Lhomi**. Lhomi is another extremely endangered language mainly spoken in the Sankhuwa Sabha district in East Nepal. The estimated total number of speakers is in between 4000 and 7000, but this number has declined rapidly in the last 8 years (Vesalainen, 2016). The only available text is again a translation of the New Testament (NT) this time written in IPA (like the Modern Newar stories). Just like for Sherpa, Lhomi Preprocessing involves only cleaning the text with regular expressions, with the added stage of cutting off case markers. Having the same NT text available for Sherpa and Lhomi helps us in retrieving sections and verses, which are missing from the Sherpa data.

4. Conclusion

In this paper, we provided a fully-fledged annotation pipeline for historical Tibetan. The strength of our method not only lies in the fact that we build on and optimise existing tools (in line with FAIR principles), as well as develop new ones (for IS annotation in particular), but also that we can adapt these tools to other Tibetan and Newar languages in any input format (from manuscript to fieldwork data), most notably modern endangered languages that are both extremely low-resourced and under-researched. This easily adaptable pipeline will greatly help researchers working on any language for which no resources are available yet.

5. Acknowledgements

This research is AHRC-funded (AH/V011235/1).

6. Bibliographical References

Anastasopoulos, A., Cox, C., Neubig, G., and Cruz, H. (2020). Endangered languages meet Modern NLP. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 39–45, Barcelona, Spain (Online), December. International Committee for Computational Linguistics.

Clérice, T., Jolivet, V., and Pilla, J. (2022). Building infrastructure for annotating medieval, classical and pre-orthographic languages: the pyrrha ecosystem. In *Digital Humanities 2022 (DH2022)*.

Colutto, S., Kahle, P., Guenter, H., and Muehlberger, G. (2019). Transkribus. a platform for automated text recognition and searching of historical documents. In *2019 15th International Conference on eScience (eScience)*, pages 463–466. IEEE.

Daelemans, W., Zavrel, J., van den Bosch, A., and Van der Sloot, K. (2003). Mbt: Memory-based tagger. *Reference Guide: ILK Technical Report-ILK*, pages 03–13.

Faggionato, C. and Garrett, E. (2019). Constraint grammars for tibetan language processing. *Proceedings of the 22nd Nordic Conference on Computational Linguistics: 12-16.*

Faggionato, C. and Meelen, M. (2019). Developing the Old Tibetan treebank. In Nikolova Temnikova Angelova, Mitkov, editor, *Proceedings of Recent Advances in Natural Language Processing*, pages 304–312. Varna: Incoma.

Frascarelli, M. and Hinterhölzl, R. (2007). Types of topics in German and Italian. *On information structure, meaning and form*, pages 87–116.

Garrett, E., Hill, N. W., Kilgarriff, A., Vadlapudi, R., and Zadoks, A. (2015). The contribution of corpus linguistics to lexicography and the future of tibetan dictionaries. *Revue d'Etudes Tibétaines*, 32:51–86.

Genetti, C. (2009). *A grammar of Dolakha Newar*, volume 40. Walter de Gruyter.

Graves, T. E. (2007). *A Grammar of Hile Sherpa*. PhD thesis submitted to the Faculty of the Graduate School of State University of New York at Buffalo.

Hill, N. (2022). Does Tibetan have a passive voice? *International Association of Tibetan Studies - Tech Panel presentation: Prague, Czech Republic*.

Komen, E. R. (2013). Corpus databases with feature pre-calculation. In *Proceedings of the twel h workshop on treebanks and linguistic theories (TLT12). Sandra Kübler, Petya Osenova & Martin Volk (eds)*, pages 85–96.

Kretschmar, M. (1995). *Erzählungen und Dialekt aus Südmustang*, volume 1. VGH-Wissenschaftsverlag.

Meelen, M. and Hill, N. (2017). Segmenting and POS tagging Classical Tibetan using a memory-based tagger. *Himalayan Linguistics*, 16(2).

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

Meelen, M., Roux, E., and Hill, N. (2021). Optimisation of the largest annotated Tibetan corpus combining rule-based, memory-based & deep-learning methods. *Transactions on Asian and Low-Resource Language Information Processing*.

Meelen, M. (2022). Tibetan word embeddings: from distributional semantics to facilitating Tibetan NLP. *International Association of Tibetan Studies - Tech Panel presentation: Prague, Czech Republic*.

Patel, C., Patel, A., and Patel, D. (2012). Optical character recognition by open source OCR tool tesseract: a case study. *International Journal of Computer Applications*, 55(10):50–56.

Randall, B., Taylor, A., and Kroch, A. (2005). Corpussearch 2. *Philadelphia: University of Pennsylvania*.

Solmsdorf, N., Trautmann, D., and Schütze, H. (2021). Active learning for argument mining: A practical approach. *arXiv preprint arXiv:2109.13611*.

Vesalainen, O. (2016). *A Grammar Sketch of Lhomi*. SIL International.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

7. Language Resource References

Faggionato, C. (2021). Constraint grammars for Tibetan dependency parsing - DOI 10.5281/zenodo.4727200, April. Funded by the UK's Arts and Humanities Research Council (grant code: AH/P004644/1).

Hale, A. (n.d.). Collection of stories in Kathmandu Newar. *unpublished*.

Lugli, L., Garrett, E., Faggionato, C., Rode, S., Solmsdorf, N., and Pagel, U. (2021). Visual Dictionary of Tibetan Verb Valency: Data - DOI 10.5281/zenodo.5596064, October.

Meelen, M. and Roux, E. (2020). The Annotated Corpus of Classical Tibetan (ACTib) - Version 2.0 (Segmented & POS-tagged) - DOI 10.5281/zenodo.3951503. May.

Otter, F. (n.d.). Transcription of RAS Hodgson MS 23: Madhyamasvayaṃbhūpurāṇa. *unpublished manuscript*.

Shakya, M. B. and Bajracharya, S. H. (2001). Svayambhū Purāṇa. *Lalitpur: Nagarjuna Institute of Exact Methods*.