Building a Part-of-Speech Tagged Corpus for Drenjongke (Bhutia)

Mana Ashida
Tokyo Metropolitan University
ashida-mana@ed.tmu.ac.jp

Seunghun J. Lee
International Christian University
University of Venda
seunghun@icu.ac.jp

Kunzang Namgyal
Nar Bahadur Bhandari Degree College
kunzang49348@gmail.com

Abstract

This research paper reports on the generation of the first Drenjongke corpus based on texts taken from a phrase book for beginners, written in the Tibetan script. A corpus of sentences was created after correcting errors in the text scanned through optical character reading (OCR). A total of 34 Part-of-Speech (PoS) tags were defined based on manual annotation performed by the three authors, one of whom is a native speaker of Drenjongke. The first corpus of the Drenjongke language comprises 275 sentences and 1379 tokens, which we plan to expand with other materials to promote further studies of this language.

1 Introduction

According to Joshi et al. (2020), most of the world's languages cannot benefit from the state-of-the-art methods of Natural Language Processing (NLP) due to a lack of resources. Given that the need for language technologies is equally distributed among speakers of each language, the investigation of the construction of language resources is necessary. This paper reports how the first Drenjongke machine-readable labeled data were developed.

Drenjongke (Bhutia[1]) is a Tibeto-Burman language with the status of one of the official languages in the state of Sikkim, India. Yliniemi (2019) estimates that the number of Drenjongke speakers is 25,000-30,000 and that Drenjongke falls into one of the categories of "vulnerable," "definitely endangered," and "severely endangered" within UNESCO's Language Vitality and Endangerment framework, with a rapid loss of native speakers due to the lack of economic value in speaking Drenjongke. The low economic value is due to the widespread use of Nepali in Sikkim. Most official and private business matters are conducted in Nepali or English. As a result of community-driven revitalization efforts[2], written materials in Drenjongke are being produced, including but not limited to dictionaries, folk songs, and novels. An archive was created resulting from a recent project with Drenjongke community members: "Phonetics, Phonology and New Orthographies in Roman and Indigenous Script: Helping Native Language Communities in the Himalayas (2017-2020)."[3] The archive contains spoken materials, including conversations and nursery rhymes.[4] However, the focus of the archive was speech recordings. Thus, machine-readable texts were not included.

The goal of our project is to build the first digitized resource of Drenjongke so as to serve as a basis for expanding NLP resources on Drenjongke by suggesting a set of PoS tags that can be used for processing other Drenjongke texts in the future. Moreover, we provide annotated text materials that would promote future computational research.

2 Related Work in Other Tibeto-Burman Languages

Tibeto-Burman languages share many characteristics. From NLP perspectives, some similarities include 1) orthographic systems that are not based on the Roman alphabet, and 2) agglutinative morphology that creates difficulty in determining the word boundaries often used in NLP systems targeting Indo-European languages. Even so, annotated text corpora in Burmese (Myanmar), Dzongkha

[1] Drenjongke is known as Bhutia, Sikkimese, and also Lhoke. Community members prefer the terminology Drenjongke over others. Bhutia is the official name accepted in India (Yliniemi, 2019).

[2] A native speaker shared with us the observation that the latest number of speakers may reach as high as 40,000 to 50,000 with revitalization efforts.

[3] https://phophono.aa-ken.jp/

[4] The nursery rhymes are also accessible at the following YouTube channel https://www.youtube.com/channel/UC90NszbLUgoOw7ZLEUHJ3WA/ along with their lyrics in both Drenjongke and English.

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
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

(Bhutan), and Classical and Modern Tibetan exist.

The largest existing Burmese corpus with annotations is the "Burmese (Myanmar) Treebank of Asian Language Treebank Project" (Ding et al., 2019). This corpus contains 20,000 sentences that are morphologically annotated with the annotation scheme "nova," in which *n* stands for nouns, *v* for verbs, *a* for adjectives, and *o* for others. This categorization was created for low-resource, highly analytic languages. Subsequently, Ding et al. (2020) propose the conversion procedure of the "nova" tag to the Universal Part-of-Speech (PoS) tagset.[5]

Corpora as well as PoS taggers in Classical Tibetan[6] have been developed as part of the project "Tibetan in Digital Communication (2012-2015)" hosted at SOAS.[7] The project has annotated four documents from Tibetan classical literature.[8] As for Modern Tibetan, Liu and Congjun (2018) report a collection of Tibetan text corpora (CTTC) containing 52,041 PoS-tagged sentences, which have been used in Machine Translation and Tibetan syntactic parsing.

Dzongkha also has several NLP resources, including a corpus and tokenizers. Norbu et al. (2010) propose a word segmentation method applied to Dzongkha using the dictionary-based maximal matching algorithm. Chungku et al. (2010) describe the corpus building process for Dzongkha, in which they prepared 66 PoS tags for annotation.

We found that these NLP resources on Tibeto-Burman languages often lack uniformed annotation guidelines as pointed out in Hill and List (2017). As such, when building corpora in a different low-resource Tibeto-Burman language, namely Drenjongke, information about the process of building of existing corpora was not directly transferable.

3 Corpus Construction

Our corpus contains 275 sentences and 1379 tokens taken from Bhutia Phrase Book (Drenjongpo, 2017). The corpus has two parts: the sentence level, containing (a) sentence ID, (b) the original Drenjongke sentence in Tibetan script, and (c) an English translation; and the token level, consisting

```
#sentence id = 112
#text = ཁྱན་ཀུས་ དབྱིན་ཇི་ ཀུསས་པོ་ གནང་ག་ ཡག།
#trans = Do you speak English?
1   ཁྱན་ཀུས་    PRON        lhan gye    you (honorific)
2   དབྱིན་ཇི་    NOUN        y'in j'i    english
3   ཀུསས་པོ་    VERB        kyap po     doing
4   གནང་ག་     AUXHON+Q    nang ga     ?
5   ཡག།         HON         lâ
```

Figure 1: Excerpt from the Drenjongke corpus

of (a) token, (b) PoS tag, (c) romanization of the token in phonological Drenjongke (Section 4.2), and (d) an English gloss of each token. An excerpt from the corpus is shown in Figure 1. The entire corpus is publicly available.[9]

We built the corpus using the following steps: (a) correcting errors in the text obtained by running Optical Character Reading (OCR) of scanned pages, (b) tokenizing sentences using spacing, and (c) manually annotating PoS tags of each token.

3.1 OCR and correcting OCR errors

We used Google OCR to recognize the Tibetan texts from scanned pages. The input method of the Tibetan script is complex because typing a syllable requires typing a root character that has additional superscript, subscript, and vowel diacritics. Since the size of our corpus is not large, we could have typed all the data, but we opted for using the OCR method instead. Testing the OCR method was beneficial because we found that the OCR-ed texts contained errors due to the "tsha-lag" ◌ྫ marker, which is used to mark the pronunciation of [bʲ] in Drenjongke. The use of this marker is unique to Drenjongke because Tibetan does not have the sound [bʲ]. Any errors with ◌ྫ were manually corrected. The knowledge of this shortfall of the Tibetan OCR for our corpus will help us when we add more text to the current corpus.

3.2 Tokenization

For tokenization, space was set as a delimiter. Drenjongke script is marked by a syllable marker called "tsheg" ་, and has a space between potential morpheme or word boundaries. The use of space in the orthography is specific to Drenjongke as other Tibetan languages do not utilize spacing in a sentence. We assume that either a morpheme boundary or a word boundary is demarcated by a space. The tokenization process revealed some inconsistencies in the use of spacing. Thus, it is not im-

[5] https://universaldependencies.org/u/pos/

[6] https://github.com/tibetan-nlp

[7] https://www.soas.ac.uk/cia/tibetanstudies/tibetan-in-digital-communications/

[8] https://zenodo.org/record/574878#.XvBraGozZZ1

[9] https://github.com/ICULingLab/drenjongke

mediately obvious whether spacing directly represents morpheme or word boundaries.

In addition to the spacing, defining words, affixes, and clitics in Drenjongke is not an easy task because monosyllabic morphemes can readily concatenate to the head of a word. For example, one verb + one monosyllabic affix can be counted as a word, while they can be two different words or a word and a clitic depending on the definition of the word. As such, spacing was used in the tokenization of the sentences from the operational point of view. In our annotation scheme, tokens with multiple PoS tags are joined with the "+" symbol: for example, VERB+INF (an infinitive marker).

3.3 Annotation

The annotation of the current corpus was done by the authors; the first two authors have worked on the language for a number of years, and the last author is a native speaker. Due to the lack of solid annotation guideline, we failed to locate other native speakers who can annotate the corpus. As such, no independent annotation verification could be performed.

The PoS tag design is based on the gloss used in Yliniemi (2019). Though applying the Universal PoS tagset can be one solution, using gloss-based tags makes the corpus more informative, especially because Drenjongke has multiple functional particles which would be reduced to PART (particles) or ADP (adpositions) in the Universal PoS tags. Though the use of the Universal features would be suitable, not all the features are covered in the list.[10] Thus, we are planning to develop a language-specific documentation of those two categories when applying the Universal Dependencies framework in the near future.

In this section, we list the PoS tags proposed for the Drenjongke corpus in 3.3.1. Section 3.3.2 lists meta information on the tokens and PoS tags. Section 3.3.3 presents three examples of multi-functional morphemes and investigates disambiguation criteria for each morpheme. We also discuss some difficulties in the PoS annotation of the Drenjongke language in 3.3.4 and in 3.3.5.

3.3.1 Part-of-Speech tag set

This section lists the 34 PoS tags proposed and used in the Drenjongke Corpus. Both 8 freestanding closed class PoS tags, 18 modifying closed class PoS tags and 8 open class PoS tags were used in the tagging process. The closed class tags mark relatively fixed vocabulary, while the open class unit can be labelled on newly created words. The modifiers PoS tags are always attached to other segments such as VERB or NOUN and can not appear on their own. Freestanding PoS tags do not have such restriction.

Closed Class (freestanding)

AUX Copulas.

AUXHON AUXHON stands for honorific auxiliary. Drenjongke has a complex honorific system, and the verb *give* is frequently used to express the honorific meaning. In this case, *give* is a secondary verb labelled as AUXHON, which will be explained in detail in 3.3.3.

AUXQ Copulas used in the interrogatives.

CCONJ Coordinating conjunction, དང་ 'd'ang' (and). Moreover, Drenjongke employs a different form as a verbal coordination conjunction སྟེ་ 'te'.

HON The honorific particle ལགས་ 'hnang/nang/hnâ' is often found at the end of a sentence or phrase, as well as attached to a person's name, indicating that the sentence is honorific.

NEGAUX Drenjongke has negative auxiliaries that are not easily decomposed into simply a negation part and a copular part.

NUM Numerals written with both digits and with characters. No ordinal numbers appear in the corpus.

PRON Pronouns, which can be followed by case markers, and in that case, they are marked PRON+POSS, PRON+DAT, PRON+LOC.

Closed Class (modifiers)

DAT Dative case marker.

DET Definite and indefinite articles: འདི་ 'di' (this) and ཁྱེ་ལས་ 'khye le' (all).

IMP A marker of imperative as a verbal suffix.

INF A marker of infinitive as a verbal suffix.

IPFV A marker of imperfective as a verbal suffix.

LOC Locative case marker.

MOD Modality, ཆུ་གས་ 'chu' (can) and དགོས་ 'gö' (need to). In most cases, these are attached to the verb; however, they can also follow a noun. In that case, the modality behaves as a normal verb "want, need" instead of "want to, need to." For example, ཆུ་ དགོས་པོ་ 'chu gö bo' (water need-INF).

[10] https://universaldependencies.org/u/feat/index.html

NEG Affixes that attach verbs and adjectives to render their meaning negative.

NMLZ A nominalizer that attaches to an auxiliary verb or a verb. English equivalent *-er*.

NPST A marker of non-past tense as a verbal suffix.

POSS Possessive case marker.

PREP Prepositions.

Q Question particles.

REFL Reflexives such as *own* preceded by a pronoun.

REL A relativizer meaning *where it does V*.

SCONJ Subordinate conjunction, ན་ 'na' (if), which follows verbs. ན་ itself is used in many different cases, and then needs to be distinguished when annotating (Section 3.3.3).

SIM A marker of simultaneity. It can be an adverbial element by itself or a post-posed element of the tensed verb form.

SUG Suggestive particle as a verbal suffix. It is equivalent to *let's* in English.

Open Class

ADJ Adjectives. Some adjectives share the morpheme -རྒས་.

ADV Adverbs.

INTJ Interjections, such as *Yes*, *No*, and *Hello*.

MWE Multiword expressions. One example of MWE in the corpus is བཀྲ་ཤིས་ བདེ་ལེགས། 'tra shi de lek' (Congratulations). བཀྲ་ཤིས་ means auspicious and བདེ་ལེགས་ means goodness, happiness.

NOUN Nouns. As with pronouns, nouns are also followed by case markers. The corresponding annotation is NOUN+POSS, NOUN+DAT, NOUN+LOC.

PROPN Proper nouns, place, and person names. Possibly followed by case markers as nouns.

UNK Unknowns. Annotation is made by consulting the English translation of each token, as well as looking up examples from the descriptive Drenjongke grammar (Yliniemi, 2019) glossed in similar categories with our PoSs. Nevertheless, it is sometimes not clear which PoS corresponds to a token or a morpheme since the word never appears in that grammar book. In those cases, we annotate them UNK.

VERB Verbs in Drenjongke are often followed by morphemes that indicate TAME (tense, aspect, mood, evidentiality).

3.3.2 Meta Information on PoS tag Structure

Our corpus contains 275 sentences and 1379 tokens. The small number of tokens in each sentence is due to 1) the type of genre of the Drenjongke text, and 2) the agglutinative morphology of Drenjongke. The original text comes from a book for a Drenjongke person so that they can learn basic phrases in Drenjongke. Sentences are conversational and thus are short in general. In particular, a large number of sentences are presented as pairs of questions and responses. The responses such as "Yes, we can" are shorter than full sentences.

Another factor that makes fewer tokens per sentence is the synthetic characteristics of Drenjongke morphology. It uses various verbal and nominal affixes to express tense, aspect, mood, evidentiality, etc. In the PoS tags employed in our corpus, these functional elements are analyzed as a unit, which further lessened the number of unique tokens.

Figure 2 shows the top-10 PoS-tagged sequence patterns observed in a token separated by spaces. A token is marked with more than one PoS tag. The high appearance frequency of the HON tag is one of the characteristics of Drenjongke. The HON tag often appears as one token or a part of a token preceded by an auxiliary verb.

3.3.3 Disambiguation of functional morphemes

Drenjongke has a variety of monosyllabic morphemes that have more than one function. In the annotation scheme, these morphemes need to be disambiguated, as in the three examples below.

ན་ 'na' (LOC vs. SCONJ)

ན་ is a morpheme that attaches to nouns, verb,s or adjective wh-words (such as *which*). When attached to a noun, ན་ functions as a locative morpheme (annotated as LOC). After a verb, it functions as a subordinating conjunction (annotated as SCONJ).

- ང་ སྒང་རྟོག་ན་ སྡོད་ཀྱི་ ཨིན་ལགས།ངོ་ གནང་ལགས།
  nga gangtok na dö to ing la
  "I live in Gangtok. (LOC)"

- ཁཱག་ངལ་ ཡོད་ན་ ངའི་རྩང་ ཇོན།
  ka ngal yod na ngê tsä j'on
  "If you have any difficulty then come to me. (SCONJ)"

ཨོ་ 'she' (NPST vs. NOUN)

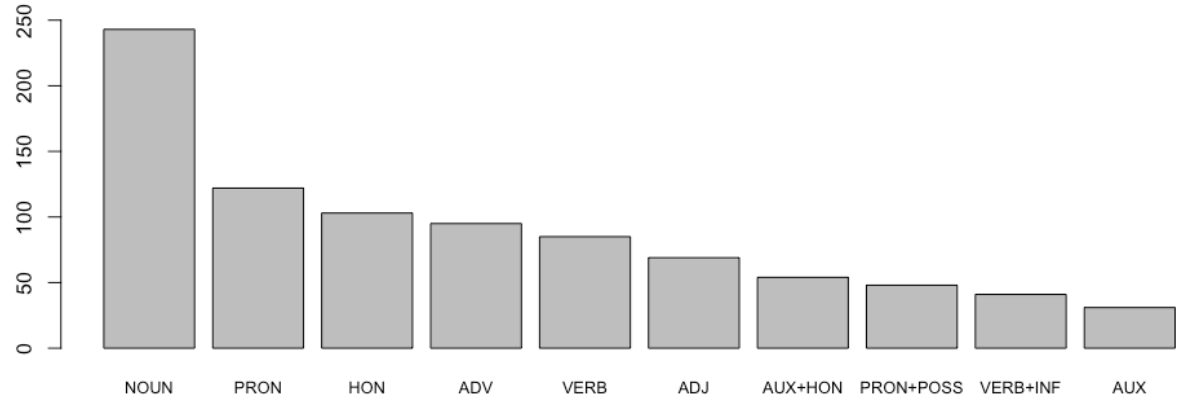

Figure 2: Top-10 PoS-tagged Sequence Patterns

ཤད་ only attaches to a verb. As a clause suffix, it indicates non-past tense (annotated as NPST). ཤད་ as a derivational suffix marks a gerund (annotated as NOUN reflecting the result of becoming a gerund).

- གཟིམ་པ་ བྱོན་ཤད་ ལ་གས།
  zim pa j'on she la
  "Going to sleep? (NPST)"
- འདི་ཁར་ གནས་ རྩ་ཆེན་ མཇལ་ཤད་ གན་ ཡོད་ལ་གས།
  di khâ hne tsä ch'e j'ä s'he gan yô la
  "Are there any important holy places around here? (NOUN)"

གནང་ 'hnang' (VERB vs. AUXHON)

གནང་ is employed as a main verb as well as a secondary verb. When used as a main verb, གནང་ is annotated as VERB. The usage as a secondary verb is annotated as AUXHON because it expresses an honorific register.

- ལྷན་རྒྱས་ ད་ལྟ་ གན་གནང་ བདེ་ལ་གས།
  lhan gye däng ga na j'on bö lâ
  "What are you doing right now? (VERB)"
- དེ་རིང་རང་ གཡོག་འདི་ རྫོང་ཡོ་ གནང་།།
  d'a ring ra yö di nyô po nang
  "Finish this work by today. (AUXHON)"

3.3.4 Annotation challenges

In our corpus, spacing in the written text was set as a delimiter. When a token has more than two morphemes, the PoS tags are concatenated by "+." This "+" method is often used in the Dzongkha corpus (Chungku et al., 2010) and Korean corpora[11], two languages with morphological structures akin to Drenjongke. When automatizing the process of PoS tag annotation, the presence of the "+" operand is non-trivial and any corpora building process must take that into consideration because the "+" operation is highly productive in Drenjongke.

3.3.5 Difficulty of evaluation

Language resource papers are often presented with metrics such as inter-annotator agreement to prove their qualities. However, it is difficult to obtain those metrics for low-resource languages that have a limited number of researchers. Moreover, difficulty lies in the fact that there is no ground truth for PoS tagging.

4 Discussion

As far as we know, no Drenjongke corpus has been available until now. Building a Drenjongke corpus was not a simple task and various considerations were discussed in the process. We present four points: (a) domain specificity of the corpus, (b) social impact of machine-readable resources, (c) language revitalization, and (d) annotation challenges.

4.1 Domain specificity

The Drenjongke corpus is based on a conversational textbook for Drenjongke beginners, targeting young Drenjongke people who do not always speak Drenjongke in their everyday life. A problem with this specific purpose of the text has pros

[11] see the fine-grained tag section of the following corpus for an example: https://github.com/UniversalDependencies/UD_Korean-Kaist/

in that the corpus has sentences relevant to average Drenjongke speakers. Cons of this specific text is the lack of diversity of genres in the data set. In order to address this issue, we are in the process of augmenting the data with texts from other genres such as folk songs and philosophical texts. The soon-to-be-expanded corpora will have annotations using syntactic relation tags adapted from the Universal Dependencies (UD) scheme, which has gained recognition as an optimal tool for a cross-lingual annotation scheme. Tibeto-Burman languages are still absent in the UD version 2.6 (Zeman et al., 2020).

4.2 Phonological Drenjongke

Drenjongke written in the Tibetan script follows the spelling norm of Classical Tibetan resulting in the inclusion of characters that are not pronounced. Written forms following Classical Tibetan are helpful for tracking etymological information, but they add difficulties in reading and writing, especially if one is not familiar with the written Tibetan forms.

The romanization of Tibetan scripts is based on the "Wylie system," which is a grapheme-based transliteration (Wylie, 1959). Among the community members, there is an increasing call for a romanization system that reflects the actual Drenjongke pronunciation. A recent invention of such a system with collaborative work with the Drenjongke community is the "Phonological Drenjongke" proposed in van Driem (2016). The romanization system in our corpus follows "Phonological Drenjongke" added by a native speaker researcher. A larger corpus in the future is expected to aid the development of an automatic transliteration system between Tibetan script and phonological Drenjongke.

4.3 Language documentation and revitalization

Building a Drenjongke corpus may serve as an effective way of documenting the language. One of the greatest challenges in creating this corpus was the lack of detailed morphological information on Drenjongke. The whole process needed constant communication among the three authors, one of whom is a native speaker of Drenjongke. This collaboration is essential in creating corpora that reflect the current Drenjongke language.

Spoken corpora of multiple Tibeto-Burman languages exist, but text-based corpora of endangered Tibeto-Burman languages are rare. The process of building this Drenjongke corpus emphasized the need for an active collaboration between native speaker researchers and NLP researchers, a challenge that is not easily addressable. The current team benefited from a linguistic project by the second author, which may be a way to move forward when interdisciplinary collaboration is needed.

5 Conclusion and Future work

This paper reports the building of a Drenjongke PoS-tagged corpus containing 275 sentences with English translation and romanization in "Phonological Drenjongke." Our future work lies in establishing a reliable word segmentation method as well as in optimizing the PoS tagging process while referring to the existing tools for Tibeto-Burman languages. Additionally, the development of the language-specific documentation for the Universal PoS tagset, as well as dependency relation labelling, remains for future work.

Acknowledgement

The authors are grateful to Jin-Dong Kim and three anonymous reviewers for their feedback on the paper, Mamoru Komachi for insightful discussions regarding the annotation process, Arseny Tolmachev for the post-acceptance mentorship, and the AACL-IJCNLP SRW committee members for providing support of various kinds. Of course, all remaining errors are of our own. Thanks also go to Jigmee Wangchuk Bhutia and Lopen Karma Gyaltsen Drenjongpo for allowing us to edit and publish the contents of the phrase book.

References

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 Resources*, pages 103–110.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) Morphological Analysis: Syllable-Based Tokenization and Part-of-Speech Tagging. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(5):1–34.

Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) Treebank: Guideline and Analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(40):1–13.

Lopen Karma Gyaltsen Drenjongpo. 2017. *Bhutia learning book for young beginners [’lo sar zhön bo tsu lo phen be lho ke jöng d’eb]*. Phen-De Ga-Tshal Ling Dharma Centre, Sikkim, India.

George van Driem. 2016. The phonology of Dränjoke: Experimental development of Roman Dränjoke and Phonological Dränjoke. *Manuscript. University of Bern*.

Nathan W Hill and Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. In *Yearbook of the Poznan Linguistic Meeting*, volume 3.1, pages 47–76. Sciendo.

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.

Huidan Liu and Long Congjun. 2018. CTTC: A Collection of Tibetan Text Corpora . In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 13–19.

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.

Turrell Wylie. 1959. A standard system of tibetan transcription. *Harvard journal of Asiatic studies*, 22:261–267.

Juha Yliniemi. 2019. *A descriptive grammar of Denjongke (Sikkimese Bhutia)*. Ph.D. thesis, University of Helsinki.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2020. Universal Dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.