Universal Dependencies for Suansu

Jessica K. Ivani[1,2,3] **Kira Tulchynska**[4]

[1]University of Tübingen [2]University of California Santa Barbara [3]University of Zurich
[4]Hebrew University of Jerusalem

Correspondence: jessica.ivani@uni-tuebingen.de

Abstract

This contribution presents the Naga-Suansu Universal Dependencies (UD) treebank, the first resource of this kind for Suansu, an endangered and underdocumented Tibeto-Burman language spoken in Northeast India. This treebank follows the UD annotation framework. We describe the corpus composition, data sources, and annotation process, outlining the general structure of the treebank. In addition, we highlight morphosyntactic challenges where Suansu grammar does not fit neatly into the UD annotation schema and propose adaptations to better capture its structural properties. As the first Tibeto-Burman language included in the UD project, the Naga-Suansu treebank serves several purposes: it contributes to the documentation and preservation of endangered languages, enables the understanding of cross-linguistic variation, and supports future research efforts in refining UD annotation practices for South and Southeast Asian languages.

1 Introduction

Universal Dependencies (UD) is a framework aiming to provide a unified grammatical annotation standard for a wide range of typologically diverse languages (McDonald et al., 2013; Nivre et al., 2016; de Marneffe et al., 2021). Thanks to its rich annotation capabilities and unified approach, UD has become an increasingly popular resource for cross-lingual NLP research, particularly for syntactic parsing and multilingual language modeling tasks. In addition, large UD corpora have been used in descriptive linguistics to explore internal syntactic variation (Kiss and Thomas, 2019) as well as in cross-linguistic and typologically oriented studies (Naranjo and Becker, 2018; Levshina, 2019).

Although the number of languages and treebanks has increased in the last decade, several language families and regions, including low-resource and endangered languages, are still underrepresented in UD.

This contribution introduces the Universal Dependencies (UD) treebank for Suansu (treebank name: Naga-Suansu[1]), an endangered and under-described language spoken along the Indo-Myanmar border. To our knowledge, this represents both the first treebank corpus for Suansu and the first UD contribution for the Tibeto-Burman language subfamily. Developed in close collaboration with the Suansu-speaking community, this resource aims to facilitate the inclusion of additional genealogically and geographically related languages into UD, enhance the typological diversity of UD corpora, and promote greater visibility and representation for under-resourced languages in language technology research.

The paper is organized as follows. First, we introduce Suansu language, outline its most prominent linguistic features, and clarify the naming conventions used for the treebank (Section 2). Section 3 describes the corpus and data sources, while in Section 4 we detail the data preprocessing steps, describing the orthographic conventions (4.1) and lemmatization process (4.2). In Section 4.3 we present the part-of-speech tags used in the treebank, and Section 4.4 outlines the features in the treebank, with a more detailed discussion on converbs (4.5) and alignment (4.6). In Section 5 we illustrate the dependency relations, with examples from copula clauses (5.1) and verb compounds (5.2). We conclude (6) with a summary and some directions for future research.

2 Suansu

Suansu is a Tibeto-Burman language spoken by approximately 2,000 people living in a small cluster of villages located in Ukhrul district, Manipur state, India, near the Indo-Myanmar border. It

[1]The treebank has been released as UD_Naga-Suansu, an official UD treebank, available at https://universaldependencies.org/treebanks/nmf_suansu/ under the CC BY-SA 4.0 license.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 30–38

is an agglutinative language with a strong preference for suffixation and a strict verb-final word order, traits that are typical of languages in this region (Post and Burling, 2017). Prominent morphological processes include verb compounding and serialization. Suansu morphosyntactic alignment does not fit neatly into the common alignment patterns described in the typological literature, a trait it shares with other regional languages such as Meithei (Chelliah, 1997), Poumai Naga (Veikho, 2021), and Mongsen Ao (Coupe, 2007). Suansu argument marking is primarily influenced by semantic and pragmatic factors rather than syntactic ones. Our annotation schema to account for these distributions, developed per the Universal Dependencies guidelines, is described in Section 4.6.

Suansu is an endangered language, and its documentation is ongoing (Ivani, 2023, 2024; Ivani and Zakharko, 2024): cross-village communication, education, and media predominantly use Tangkhul Naga, the regional *lingua franca*. Thus, Suansu lacks a standardized orthography. This paper offers a first orthography for the language, developed jointly with Suansu community members and detailed in Section 4.1.

The decision to name the UD treebank "Naga Suansu" was motivated primarily by logistical considerations. Suansu currently lacks an ISO code (`Glottocode: suan1234`), a requirement for registering UD treebanks. Consequently, the treebank was assigned the ISO code [`nmf`], originally designated for Tangkhul Naga. To clearly differentiate Suansu from Tangkhul within this shared coding scheme, we chose a term that could unambiguously encompass both languages in the future. After consultations with members of the Suansu-speaking community, the term "Naga" was proposed, as it better reflects shared ethnic affiliation while foregrounding linguistic identity. This choice aligns with established naming conventions among nearby languages (e.g., Mao Naga, Poumai Naga). However, readers should note that "Naga" is an ethnic term without inherent linguistic implications. In this paper, we use Naga-Suansu to indicate the UD treebank, and Suansu when we refer to the language.

3 Data

The Naga-Suansu treebank comprises four distinct data sources, each translated into Suansu by native speakers and subsequently manually glossed and annotated. The corpus contains 584 sentences and 3123 words across the four sources. A summary description of the corpus is available in Table 1.

Text	Sentences	Words
film_Bridge	332	1990
grammar_BivalTyp	131	536
grammar_Cairo	20	159
grammar_ValPal	101	438

Table 1: Corpus information

The most significant portion of the treebank consists of the first 300 lines of subtitles from the film *Bridge of Spies* (2015), forming part of a broader parallel subtitles corpus project (Ebert et al., 2023).

The second source is the ValPal dataset (Hartmann et al., 2013), comprising 101 sentences designed to investigate valency patterns across languages. The last two sources of treebank include 131 sentences from the BivalTyp questionnaire (Say, 2020) and 20 sentences from the Cairo CICLing Corpus[2]. All materials were translated directly from English into Suansu by native speakers, followed by manual interlinear morphological glossing and annotation.

The choice of these sources is motivated by practical considerations, such as the availability of analyzed and glossed materials. At present, the corpus is skewed toward written data. We plan to include conversational data and additional spoken materials in future releases.

4 Annotation

The data was divided into small batches, each containing fewer than 100 sentences. The first batch was converted from interlinear glosses into a CoNLL-U file with tokenized sentences without annotated features and dependencies. Morphosyntactic annotation for this initial batch was carried out entirely manually.

Subsequent batches were processed using the UDPipe 1 pipeline (Straka and Straková, 2017), trained on the manually annotated data from previous batches. The resulting automatic parses were then reviewed and corrected manually.

Throughout the process, any ambiguous or previously unencountered structures were discussed collaboratively by the authors and native speakers to establish a consensus on the appropriate annotation strategy.

[2] https://github.com/UniversalDependencies/cairo/tree/master

Phonemes	Orthographic representations
/ɬ/	*lh*
/θ/, /tʰ/	*th*
/x/	*hr*
/χ/	*hw*
/ə/	*ā*
nasalized V	*Vhn*

Table 2: Some examples of orthographic representation of Suansu phonemes. Nasalised vowels are phonemic in Suansu.

4.1 Tokenization and Word Segmentation

In the treebank, whitespace is used to mark word boundaries. Lexical units are identified based on phonological independence, which is determined by the presence of primary stress. The orthography employed in this treebank is a preliminary system developed in collaboration with members of the Suansu-speaking community. Based on the Latin script, it draws partially from the Tangkhul Naga orthographic tradition. This Suansu orthography was designed primarily by and for the community, with several key goals in mind: to serve as an initial step toward a standardized writing system; to be accessible and easily typable, facilitating intuitive use in everyday communication, including messaging and chat platforms; and to accurately represent Suansu-specific linguistic features, such as the lateral fricative /ɬ/, that are not present in Tangkhul.

Additionally, the writing system was designed with future refinements in mind, allowing for straightforward updates as needed. While certain phonemic distinctions in Suansu are currently represented by the same grapheme, for instance, *th* is used for both /θ/ and /tʰ/, this overlap does not impede intelligibility among speakers. Table 2 presents illustrative examples of phonemes, some of which are unique to Suansu in contrast to Tangkhul, along with their corresponding orthographic representations in the treebank. Future orthographic developments will address additional features such as diphthongs and tone marking.

4.2 Lemmatization

Certain morphological processes are not explicitly marked in lemma forms. For instance, lexical compounds are treated as single units in the lemma, rather than being segmented into their individual components. Suansu exhibits several productive derivational processes, with compounding and reduplication being particularly prominent. Examples include *themok.nui* ‘milk’ (literally *themok* ‘cow’ + *nui* ‘breast’) and *sui.sui* ‘follow’ (cf. *sui* ‘back’). Such forms are preserved in their derived state in the lemma, including systematic patterns like full reduplication of verb stems or auxiliaries, which is used to express interrogative mood. Similarly, inherently possessed kinship terms remain unsegmented in the lemma, as in *a.thi* ‘mother’ (literally ‘my mother’).

Class	UPOS	Total	%
Open	ADJ	51	1.63%
	ADV	191	6.11%
	INTJ	82	2.62%
	NOUN	669	21.41%
	PROPN	260	8.32%
	VERB	635	20.33%
Closed	ADP	28	0.90%
	AUX	176	5.63%
	CCONJ	23	0.74%
	DET	136	4.35%
	NUM	47	1.50%
	PART	50	1.60%
	PRON	311	9.96%
	SCONJ	72	2.30%
Other	PUNCT	387	12.39%
	X	6	0.19%

Table 3: POS tags and their frequencies in the Naga-Suansu treebank.

4.3 Universal Part-of-Speech tags

The Universal Dependencies (UD) framework defines 17 universal part-of-speech (POS) tags. Of these, all except SYM (symbol) are attested in the Naga-Suansu treebank. Table 3 summarizes the POS tags used in the treebank along with their respective frequencies. In the following sections, we provide a brief overview of selected parts of speech as they appear in the treebank, highlighting their key properties.

4.3.1 NOUN

Suansu nouns take case (1) and number (2) markers, although the number formative can also be found on verbs, where it functions as a nominalizer (cf. below). In pronominal possession, the pronominal form precedes the possessed noun (3).

(1) *Peter lairak-**di** lua-te*
Peter book-**TOP** take-COMPL
‘Peter took a book.’

(2) *Tye baneo-**pha** lapui-di chuhn-e*
DET boy-**PL** road-TOP cross-PST
'The boys crossed the road.'

(3) ***A** miszu garhe*
1**SG** person NEG.COP
'Not my guy.'

4.3.2 VERB

Suansu verbs are subject to several inflectional and derivational processes. There is no person or number agreement. Suansu verbs take TAM markers (see above and 4), and derivational morphology includes adverbials (5) and directionals (6), among others, and are further discussed in Section 5.2. Nominalization is extensive in Suansu and encompasses several markers, including plural forms (7). Serialization is highly productive in the language (8).

(4) *Tye duh makh-**e***
DET old.person cough-**PST**
'The old man coughed.'

(5) *A phethe **rai** ga re rung*
1SG eat **COLL** HORT like.that say
wi-le sir
have-PRS sir
'I have a dinner date, sir.'

(6) *Nue-ganan la-di rung **kai**,*
ask-CVB AUX-NMLZ say **UP**
Lieutenant
lieutenant
'Just answer the question, Lieutenant.'

(7) *A-va client ba-nan makhwa-da*
1SG-GEN client 3SG-ERG early-ABL
*rung-ha-**pha**-di the-ma ga reha*
say-PST-**NMLZ**-TOP do-NEG EVD REP
lala?
AUX
'...My client is not honoring the claim?'

(8) *Atra thahn kanahn li **la** **ve**-e*
hot place where also **bring** **go**-PST
'I went somewhere hot.'

4.3.3 ADJ and ADV

In Suansu, adjectives and adverbs do not form distinct classes; rather, they are derived from verbs and retain verbal properties (9). Modifiers with adjectival (10) and adverbial (11) functions are typically formed through nominalization, derivational processes, and case marking. In the treebank, such lexemes are annotated as adjectives or adverbs based on their syntactic behavior. However, not all noun modifiers undergo nominalization (12).

(9) *neo-di **szu-e***
child-TOP **be.good-PST**
'The child was good.'

(10) ***ka-szu-e** neo-di hanahn*
NMLZ-be.good-NMLZ child-TOP here
lai-le
be-PRS
'The good child is here.'

(11) *mazohn szuka **szu-nan** thai-le*
all stuff **be.good-ERG** see-PRS
ba-byahn
3SG-BEN
'Everything goes well for him.'

(12) *Peter mobile phone **katha** ska manung-le*
Peter mobile phone **new** one desire-PRS
'Peter wants a new mobile phone.'

4.3.4 PRON

The set of independent personal pronouns in Suansu, as represented in the Naga-Suansu treebank, is shown in Table 4. In addition to personal pronouns, the treebank includes demonstrative pronouns (e.g., *hadi* 'this', *didi* 'that'), quantifiers such as *mazohn* 'all', interrogative pronouns like *thuza* 'who' and *mwe* 'what', as well as indefinite pronouns (e.g., *chatha* 'others'), which frequently function as determiners.

	Singular	**Plural**
1st person	*a*	*ha*
2nd person	*nahn*	*na*
3rd person	*ba*	*bu*

Table 4: Independent personal pronouns in the Naga-Suansu treebank.

4.3.5 DET

Determiners in Suansu precede the noun and include modifiers that express definiteness and deixis, such as *hai* 'this' (13) and *tye* 'that' (14).

(13) ***hai*** *lairak-di* *lua* *dai*
DET book-TOP take JUS
'Take this book.'

(14) ***tye*** *lairak-di* *lua* *dai*
DET book-TOP take JUS
'Take that book.'

4.3.6 AUX

In the Naga-Suansu treebank, the most prominent auxiliary is the existential verb *la*, which frequently bears TAM inflection in periphrastic constructions, such as those expressing progressive aspect (cf. Section 4.5). We also tag as AUX most mood formatives that occur as independent phonological units, with the exception of the imperative suffix *-a*, which is treated as a verbal suffix. Modal forms are treated similarly: the phonologically independent obligative modal *geraha* is annotated as AUX (15), whereas abilitive and permissive modals, which are tightly bound to the verb, are treated as verbal suffixes (15). The copula *e* is also tagged as AUX and linked via the cop relation (see Section 5.1).

(15) *A-nan* *Lynn-di,* *hai* *me* *re*
1SG-ERG lynn-TOP PRO no REP
*rung-**gam**-ma* *rahn* *re,* *nahn* *li*
say-**MOD**-NEG IRR like.that 2SG also
me *re* *rung* ***geraha*** *matikza* *thok*
no REP say **MOD** very difficult
rahn
IRR
'I told Lynn, the firm can't say no, and you'd have a tough time saying no, too.'

4.3.7 PART

Forms that do not meet the criteria for other part-of-speech categories are annotated as particles. In the Naga-Suansu treebank, these include the reportative marker *re* (when it appears after quotes), the copular negator *garhe*, and various discourse particles.

4.3.8 SCONJ

We use the tag SCONJ for nominalizers, including the clause nominalizer *di*, which also functions as a relativizer. This formal overlap between relativization and nominalization is commonly observed in Tibeto-Burman languages. The tag also applies to forms that introduce subordinate clauses, such as the purposive *gase* (16) and the temporal adverbial *ganan*, which additionally serves as a converb in certain aspectual contexts (see Section 4.5).

(16) *Peter* *miaowi-di* *chokla* ***gase*** *the-ganan*
Peter cat-TOP catch **PURP** do-CVB
la-le
be-PRS
'Peter is trying to catch the cat.'

4.3.9 INTJ

Expressions primarily used as exclamations or parts of exclamatory utterances are annotated as INTJ. This category also includes fillers, such as the numeral *ska* 'one' when used in this function, as well as forms like *min* 'right' and *ay* 'yeah' in backchanneling contexts.

4.4 Features

The morphological features annotated in the Naga-Suansu treebank are listed in Table 5. Most align with existing Universal Dependencies feature definitions, though a few additional features and values have been introduced. In the following section, we briefly discuss these newly added features, along with key annotation decisions involving converbs and alignment.

4.4.1 Modal=Abil, Obl, Perm

The Naga-Suansu treebank introduces the feature Modal, which groups together modality-related values as follows: Abil for abilitive constructions (e.g., expressing physical or cognitive ability, comparable to English *can* or *be able to*), Obl for expressions of necessity and obligation, and Perm for constructions encoding permission (e.g., *be allowed to* in English). In Suansu, the verbal suffix *-ngam-* is used to express ability, while *-szu-* encodes permission. Both forms occur immediately after the verb stem, and before any TAM marker. Strong and weak necessity are marked by the phonologically independent form *geraha*, which typically appears in clause-final position. We refer to the discussion of AUX in 4.3 for the POS annotation of these forms.

Features	Values	Count	%
Abbr	Yes	2	0.04%
Aspect	Imp, Perf, Prog	473	9.56%
Case	Abl, Ben, Dat, DatErg†, Erg, ErgTop†, Gen, GenAbl†, GenTop†, Loc, LocTop†, Top	504	10.18%
Degree	Cmp, Pos	45	0.91%
Deixis	Prox, Remt	151	3.05%
Evidentiality	Fh, Nfh	9	0.18%
ExtPos	ADV, VERB	12	0.24%
Foreign	Yes	6	0.12%
Modal†	Abil†, Obl†, Perm†	14	0.28%
Mood	Des, Hort†, Imp, Ind, Int, Irr, Jus	451	9.11%
Number	Plur, Sing	1218	24.61%
NumForm	Digit, Word	46	0.93%
NumType	Card	47	0.95%
Person	1, 2, 3	263	5.31%
Polarity	Neg, Pos	131	2.65%
PronType	Dem, Ind, Int, Prs, Tot	467	9.44%
Reflex	Yes	2	0.04%
Tense	Past, Pqp, Pres	365	7.38%
VerbForm	Conv, Fin, Inf, Vnoun	743	15.01%

Table 5: Features in the Naga-Suansu treebank. † marks new features and values introduced in the treebank.

4.4.2 `Mood=Hort`

Suansu exhibits several strategies for expressing speaker attitude toward actions, including a fine-grained encoding of orders, requests, and encouragements. In the Naga-Suansu treebank, we annotate a range of mood categories alongside modal expressions, including imperative, desiderative, and jussive moods. To account for polite forms of encouragement, we introduce the hortative mood (`Mood=Hort`). This is typically expressed by the clause-final form *ga*, which we tag as an auxiliary. An example illustrating the hortative mood is provided below (17), alongside a jussive example (18), the latter used for polite directives, to highlight the functional and contextual differences between the two.

(17) *Mazwehn, Doug, hai-da phethe* ***ga***
alright doug 1PL-ABL eat HORT
'Alright Doug, you can join us for dinner.'

(18) *Kapiu matha* ***dai***
search start **JUS**
'Start searching the place.'

4.4.3 `Case=DatErg, ErgTop, GenAbl, GenTop, LocTop`

The Naga-Suansu treebank includes instances of case stacking, or *Suffixaufnahme* (Plank, 1995), where two case suffixes co-occur on a single noun phrase. The case combinations attested include genitive + ablative (`Case=GenAbl`, 19), dative + ergative (`Case=DatErg`, 20), ergative + topic (subject) (`Case=ErgTop`), and locative + topic (object) (`Case=LocTop`). Dependency relations are annotated based on the syntactic context and the argument structure of the associated verb root.

(19) *Hai letter-di Peter-**va-da***
DET letter-TOP Peter-**GEN-ABL**
'This letter is from Peter.'

(20) *A-**la-nan** phabta-ma*
1SG-**DAT-ERG** understand-NEG
'It's not clear to me.'

4.5 Converbs

Suansu employs several converbs to express a range of functions, including adverbial subordination, as widely attested in the typological literature (Haspelmath, 1995). One such form, *ganan*, is used to indicate simultaneous or immediately consecutive actions or states. It follows an uninflected verb, while the final verb in the clause, which denotes the subsequent action or state, carries TAM marking (21). Additionally, *ganan* functions in a periphrastic construction to express progressive aspect, where it attaches to the main verb and is followed by the existential auxiliary *la*, which is inflected for tense (22). These distinct uses are differentiated in our Features annotation. Similar annotation strategies have been adopted in the Tatar (Taguchi et al., 2016) and Uyghur (Eli et al., 2016) Universal Dependencies treebanks, where comparable converb forms serve analogous functions.

(21) *Neo pha wiza-va rung-e*
child PL teacher-GEN say-NMLZ
*chu-**ganan** mazohnna-e*
hear-**CVB** sit.down-PST
'The children obeyed the teacher and sat down.'
`VerbForm=Conv`

(22) *Peter guitar khohn-**ganan** la-le*
Peter guitar play-**CVB** be-PRS
'Peter is playing the guitar.'
`Aspect=Prog|VerbForm=Conv`

4.6 Alignment

Suansu alignment system is complex; it is motivated by semantic rather than morpho-syntactic factors and does not fit neatly into the typological classification proposed in the literature. Semantic Agents are generally marked with the ergative marker *nan*, while Patients and Experiencers are often marked with *di*. However, the presence of these markers is not obligatory and is heavily influenced by pragmatic factors. Syntactic subjects of intransitive sentences are typically unmarked.

In our annotation scheme, we label the syntactic subjects of transitive clauses with the ergative case marker *nan* (23). The marker *di*, on the other hand, is annotated as a topic marker, and its syntactic role is determined contextually (23, 24). Previous work (Ivani, 2023, 2024; Ivani and Zakharko, 2024) has shown that *di* is strongly associated with discourse pragmatics; it serves to highlight or single out a referent, drawing the listener's attention to it within the discourse. This function aligns more closely with topicalization than with core argument marking. By tagging *di* as `Topic`, we reflect its pragmatic role in discourse, while syntactic dependencies are annotated separately based on each construction. This approach allows us to remain consistent with UD guidelines without resorting to misleading labels like 'absolutive' or 'accusative', while also preserving the unique alignment properties of Suansu.

(23) *Hui-nan Peter-**di** maka-j*
dog-ERG Peter-**TOP** bite-PST
'A dog bit Peter.'
`Case=Top|DepRel=obj`

(24) *Peter-**di** Maria thrwa-le*
Peter-**TOP** Maria be.similar-PRS
'Peter resembles Maria.'
`Case=Top|DepRel=nsubj`

5 Dependency relations

The Universal Dependency v2 lists 37 syntactic relations. Of these 37 dependency relation types, thirty are used in the Naga-Suansu treebank[3]. Dependencies and related counts are reported in Table 6. We briefly discuss syntactic relations pertaining to copula clauses (5.1) and verb compounding (5.2).

Dependency	**Count**	**%**
root	584	18.69%
nsubj	427	13.67%
punct	387	12.39%
obj	289	9.25%
advmod	159	5.09%
aux	156	4.99%
obl	153	4.90%
det	137	4.39%
nmod:poss	95	3.04%
mark	84	2.69%
xcomp	52	1.66%
discourse	46	1.47%
advcl	44	1.41%
flat	42	1.34%
conj	41	1.31%
nummod	37	1.18%
nmod	32	1.02%
amod	32	1.02%
vocative	32	1.02%
···		
other	295	9.44%

Table 6: Dependency relations used in the Naga-Suansu treebank.

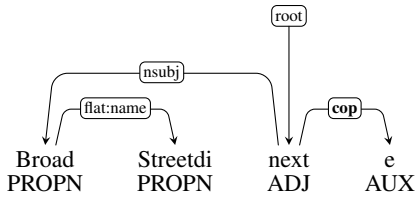
5.1 Copula clauses

Identity clauses in Suansu are expressed via the copula form *e*, which functions as a linker in these structures and does not carry tense or aspect information. It is tagged in the treebank as `AUX` and annotated via a `cop` relation, as shown by the example in Figure 1).

5.2 Verb compounding

Verb compounding processes are abundant in Suansu. In the treebank, we distinguish `compound:prt` for compounds involving verbs and

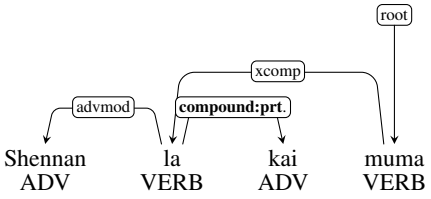
[3] Syntactic dependencies absent in the current Naga-Suansu treebank are `clf`, `dep`, `dislocated`, `expl`, `fixed`, `goeswith` and `list`.



'Broad Street is next'

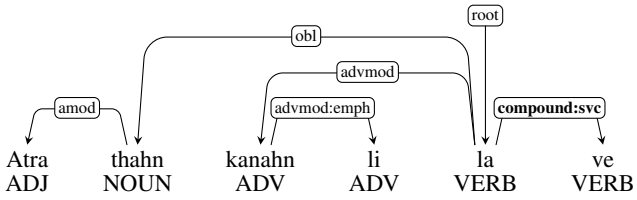
Figure 1: Annotation of copula clauses.

directional particles, and `compound:svc` for serial verb constructions. The syntactic relation `compound:prt` is illustrated in Figure 2a, while the syntactic relation `compound:svc` is exemplified in Figure 2b. These dependency relations reflect both morphological and syntactic integration: directional particles in `compound:prt` are tightly bound to the verb, forming a single semantic unit, whereas in serial verb constructions (`compound:svc`), multiple verbs occur in sequence and jointly express a complex event. Our annotation captures this distinction by analyzing the first verb in the sequence as dependent on the second, more semantically central verb.



'I'm not sure I want to pick it up'

(a) Annotation of verb compounds with directional particles.



'I went somewhere hot'

(b) Annotation of serial verb constructions.

6 Conclusion

We have presented the first treebank for Naga-Suansu, the first resource of its kind for Suansu, an endangered language, and the first known contribution to Universal Dependencies from a language within the Tibeto-Burman subfamily. Alongside the treebank, we introduced our approach to orthographic standardisation and detailed the part-of-speech tags, morphological features, and dependency relations used in annotation, most of which align closely with the UD guidelines. Where necessary, we proposed extensions, such as the addition of `Mood=Hort` and modality features, and offered a proposal for representing morphosyntactic alignment specific to Suansu. We also addressed dependency relations for copular clauses and compounding.

Our future goals include expanding the dataset, by completing the UD annotation of the movie *Bridge of Spies*, which currently covers the first 300 sentences. We also plan to include annotated conversational data, further broadening the range of linguistic contexts represented in the treebank. Finally, we plan on incorporating code mixing and multilingual influences, such as borrowings from the regional lingua franca Tangkhul Naga. Through this work, we hope to encourage the development of additional UD treebanks for other Tibeto-Burman languages and to support a growing body of resources for endangered languages within the UD ecosystem.

Acknowledgments

This research was funded by the Global Strategy and Partnerships Funding Scheme of the University of Zurich, project *Event Packaging in Language*, which we gratefully acknowledge here. We extend our gratitude to the Suansu-speaking community for their invaluable contributions, and especially to Jason M. Vashum for his generous assistance with translation and annotation. We also thank the three anonymous reviewers for their constructive comments on the draft of this paper, and Taras Zakharko for insightful comments and formatting support.

References

Shobhana Lakshmi Chelliah. 1997. *A Grammar of Meithei*, volume 17 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin.

Alexander Robertson Coupe. 2007. *A Grammar of Mongsen Ao*, volume 39 of *Mouton Grammar Library*. Mouton de Gruyter, Berlin.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Christian Ebert, Natalia Levshina, and Paul Widmer. 2023. Partree - parallel treebanks: A multilingual corpus of movie subtitle.

M. Eli, W. Mushajiang, T. Yibulayin, K. Abiderexiti, and Y. Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig (ValPaL) Online Database*. Max Planck Institute for Evolutionary Anthropology. Accessed: 2025-04-16.

Martin Haspelmath. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath and Ekkehard König, editors, *Converbs in Cross-Linguistic Perspective: Structure and Meaning of Adverbial Verb Forms – Adverbial Participles, Gerunds*, pages 1–56. Mouton de Gruyter, Berlin.

Jessica K. Ivani. 2023. Suansu language from northeastern India: A field report. *Linguistics of the Tibeto-Burman Area*, 46(1):138–163.

Jessica K. Ivani. 2024. Caritive expression in Suansu. *Himalayan Linguistics*, 23(3):28–40.

Jessica K. Ivani and Taras Zakharko. 2024. Phasal polarity in Suansu. *Linguistics of the Tibeto-Burman Area*, 47(2):318–342.

Angelika Kiss and Guillaume Thomas. 2019. Word order variation in Mbyá Guaraní. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 121–129, Paris, France. Association for Computational Linguistics.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

M. Guzmán Naranjo and Laura Becker. 2018. Quantitative Word Order Typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 91–104, Oslo, Norway. Linköping University Electronic Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Frans Plank. 1995. *Double Case: Agreement by Suffixaufnahme*. Oxford University Press, New York.

Mark W. Post and Robbins Burling. 2017. The Tibeto-Burman languages of Northeastern India. In Graham Thurgood and Randy J. LaPolla, editors, *The Sino-Tibetan Languages*, pages 213–242. Routledge, London.

Sergey Say. 2020. Bivaltyp: Typological database of bivalent verbs and their encoding frames. Accessed: 2025-04-16.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2016. Universal dependencies treebank for Tatar: Incorporating intra-word code-switching information. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Sahiinii Lemaina Veikho. 2021. *Grammar of Poumai Naga (Poula): A Trans-Himalayan Language of North-East India*. Brill, Leiden.