Universal Dependency Treebank for Xibe

He Zhou, Juyeon Chung, Sandra Kübler, Francis M. Tyers
Indiana University
{hzh1,juychung,skuebler,ftyers}@iu.edu

Abstract

We present our work of constructing the first treebank for the Xibe language following the Universal Dependencies (UD) annotation scheme. Xibe is a low-resourced and severely endangered Tungusic language spoken by the Xibe minority living in the Xinjiang Uygur Autonomous Region of China. We collected 810 sentences so far, including 544 sentences from a grammar book on written Xibe and 266 sentences from *Cabcal News*. We annotated those sentences manually from scratch. In this paper, we report the procedure of building this treebank and analyze several important annotation issues of our treebank. More specifically, we look at loanwords from Chinese, at the attributive function of the case marker *i*, at the topic marker *oci*, and at relative and adverbial clauses. Finally, we propose our plans for future work.

1 Introduction

The Xibe language (ISO 693-3:sjo) is a Tungusic language spoken by members of the Xibe minority group of China. Based on the 2010 population census of China, the population of the Xibe minority is no more than 200,000[1]. Xibe people are mainly distributed in northeastern China, including Heilongjiang, Jilin and Liaoning, and northwestern Xinjiang Uygur Autonomous Region. However, active native speakers mainly live in Cabcal Xibe Autonomous County and adjacent regions in Xinjiang. The number of native Xibe speakers has dropped below 40,000 and continues to decrease. Therefore, the Xibe language is considered a severely endangered language by UNESCO[2].

There is a limited amount of linguistic studies pertinent to the Xibe language. Gu (2016) provides a survey on Xibe language research since the 1970s. Most of the previous studies are either theoretical description of this language or comparative studies with other languages, including Chinese, Manchu, and Mongolian. However, there is no corpus or any computational tool available for this language so far. In Cabcal Xibe Autonomous County, there is a single newspaper written in Xibe, *Cabcal Serkin* 'Cabcal News', which provides an invaluable resource for linguistic research. Therefore, to start the process of building NLP applications for this low resourced language, we first aim to create a syntactically annotated treebank based on texts from this newspaper.

We choose the Universal Dependencies (UD) framework (McDonald et al., 2013) to create a dependency treebank for the Xibe language. The UD project has been developed for consistently constructing treebanks for many different languages cross-linguistically, aiming to capture similarities as well as idiosyncrasies among typologically different languages (Nivre et al., 2016). The existing universal guidelines[3] have been widely used for a wide range of typologically different languages. Thus, we expect that they will be usable for Xibe without much adaptation. Xibe is an agglutinative language with rich morphological inflections. We decided that annotating word features as detailed as possible will allow us to make available as much syntactic and semantic information as possible.

The remainder of this paper is organized as follows: In Section 2, we provide a comparison of Xibe and Manchu, and explain the differences between written and spoken Xibe. We introduce details of

[1] http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm
[2] http://unesco.org/languages-atlas/index.php
[3] https://universaldependencies.org/guidelines.html

Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020), pages 204–214

the corpus, including transliteration and pre-processing, in Section 3. In Section 4, we discuss several important annotation issues in part-of-speech and syntax. We summarize our work in Section 5.

2 Background

2.1 Xibe and Manchu

The Xibe minority used to reside in northeastern China and had a close relationship with Manchurian and Mongolian in both lifestyle and language. The Xibe people used to be one of the Manchu Eight Banners; therefore, they were considered a part of Manchu. Around 1764, the Xibe troops and their families left their hometown of Mukden (now Shenyang, China) and headed west towards the Ili Valley in Xinjiang to strengthen the border under the decree of Emperor Qianlong. Since their settlement there, they have continued using their own language and there still exists an active language community now.

Since the Xibe language is highly similar to Manchu, the question whether Xibe is an independent language or a Manchu dialect has been the focus of a controversial discussion among historians and linguists. In 1947, the Xibe minority conducted a language reform and determined the modern Xibe writing system, which is based on Manchu with slight modifications. However, modern Xibe has developed characteristics that set it apart from Manchu, as a result of language contact with adjacent languages such as Uygur, Kazakh, Russian, and Chinese. Most of the changes originate from Xibe absorbing a large amount of new words in the political domain from Chinese.

We mainly used *Xiboyu Yufa Tonglun* (General Introduction to Xibe Grammar) by Setuken (2009), a comprehensive description of written Xibe grammar, to guide our annotation decisions. Additionally, although Manchu is rarely used in daily life, there are many more accessible reference works for Manchu than for Xibe. Because of the similarity between the two languages, we have also consulted Manchu materials, such as *the Comprehensive Manchu-Chinese Dictionary* (Hu, 1994) and *Manchu Grammar* (Gorelova, 2002), when making annotation decisions.

2.2 Written and Spoken Xibe

Spoken Xibe is a collection of dialects, and there is no standard. Thus, spoken Xibe differs to a certain point from the written form. Most previous studies are concerned with documenting Xibe dialectal variation or studying spoken Xibe phonology or morphology (Norman, 1974; Li, 1979; Li, 1982; Li, 1985; Li, 1988; Jang, 2008; Zikmundová, 2013). The language data in those works are not written in Xibe script but are collected by recording native speakers' pronunciation and transcribed with IPA or transliterated in Roman alphabet since the goal is documenting the dialectal differences. Considering the variation of spoken language among the Xibe communities and the difference between the written and spoken language, we take written Xibe as the research object, and our data are mainly based on Xibe language publications, that is, a newspaper and a grammar.

3 Corpus

3.1 Data Collection

In our present work, we have collected written Xibe sentences from two data sources. The first part originates from *Xiboyu Yufa Tonglun* (General Introduction to Xibe Grammar) (Setuken, 2009). We extracted all 544 example sentences from the grammar, only excluding examples from poetic language. Using sentences from a grammar has the advantage that they comprehensively cover all grammatical constructions. The 544 sentences contain a total of 5,773 tokens, and the average sentence length is 10.6 tokens per sentence.

The second part was collected from *Cabcal News*. Each issue of the newspaper has four pages, the first two pages are news, the remaining two pages are essays or poems written by native speakers. To keep the genre consistent, we only extracted news. We collected 266 sentences from 9 issues, including 9,716 tokens. The longest sentence has 92 tokens and the average sentence length is 36.5 tokens per sentence. After combining the two parts, our complete treebank consists of 810 sentences, or 15,489 tokens.

vowels (5)	a[a]	e[ə]	i [i]	o [o]	u[u]
consonants (19)	n[n]	k [q]/[k]	g [ɢ]/[g]	h [x]/[χ]	b [p]
	p[p^h]	s [s]	š[ʂ]	t [t^h]	d [t]
	l[l]	m[m]	c[$tʂ^h$]	j[tʂ]	y[j]
	r[r]	f[f]	w[v]	ng[ŋ]	
foreign letters (10)	ck[k^h]	cg[g^h]	ch[x^h]	z[z̧]	ts[ts^h]
	dz[ts]	sy[sz]	tsy[ts^hz]	cy[$tʂ^hz$̧]	jy[tʂz̧]
Manchu vowel(1)	v [u]				

Table 1: Xibe alphabet, with transliterations and IPA.

3.2 Pre-processing

Before converting each sentence into CoNLL-U format, we Latinized each Xibe sentence and translated it into English. We manually transliterated the first 544 grammar book sentences and automatically transliterated the news data using a python script. After the conversion to the CoNLL-U format, each sentence has its original text written in Xibe script, the transliteration, and the English translation. Tokenization assumes that all words are separated by spaces or punctuation. We annotated each word with its lemma, UTS part of speech tag, morphological features, and dependency annotation.

For the first 544 sentences, the annotation work was carried out by two annotators. The first annotator annotated 464 sentences, and the second annotator annotated 80 sentences. The 80 sentences by the second annotator were checked by the first annotator to keep the annotation consistent. As for the second part of the data, the annotation was performed by the first annotator. We used UD Annotatrix (Tyers et al., 2017) to facilitate our annotation.

3.3 Transliteration

The writing system of Xibe is untypical in that its writing direction is from top to bottom, from left to right. The Xibe script is based on Manchu script with slight modifications, which uses traditional Mongolian letters. Xibe letters have different forms: Most of the letters have three forms at initial, medial, or final position, but some letters just have one or two forms. In Table 1, all letters but *ng* ᠩ are the initial forms. For *ng* ᠩ, we show the final form since it cannot occur in initial position.

Modern written Xibe has 5 vowels, 19 consonants, and 10 foreign letters, shown in Table 1 (Setuken, 2009; Xinjiang Ethnic Language Work Committee, 1992). Additionally, the 10 foreign letters are constructed on the basis of elements from which the letters of the Xibe alphabet are formed. They are only used for foreign words, mostly Chinese loanwords. The additional vowel ᠸ listed in the table is one of the Manchu vowels, it is not part of the official Xibe script. However, since Xibe and Manchu have a large amount of words in common, this letter is frequently used in Xibe texts. It has a similar pronunciation to the Xibe ᠤ *u*, thus to differentiate the two, we used *v* to transliterate this vowel.

4 Annotation Issues

Xibe is one of the Tungusic languages. Like other Tungusic languages, it has agglutinative morphology. Xibe morphology mainly focuses on verbs in that verbs are marked for tense, aspect, mood, and voice, but also for converbs and participles. Xibe phrases are head-final, both on the phrasal and the clausal level. The canonical word order is Subject-Object-Verb (SOV) (but see Section 4.3), and arguments are marked for case.

Figure 1[4] shows a Xibe sentence in canonical order, in which *muse* 'I' is the subject, *mini juwe gala* 'our two hands', in instrumental case (marked by *i*), is an adjunct to the main verb, *ice usin tokso* 'new

[4]Because of space limitations, we show short example trees in vertical form with Xibe text and long example trees in horizontal form with transliteration only.

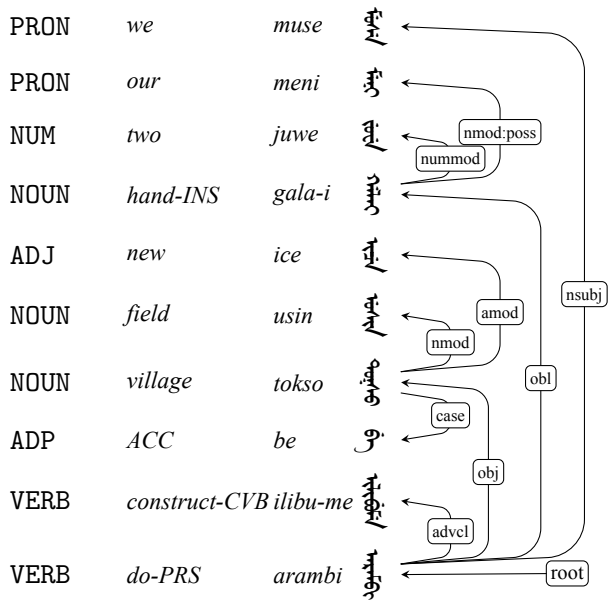

Figure 1: Dependency Tree for 'We construct the new countryside with our two hands'.

```
18-20   ᡤᡠᠩᡥᡝᠴᡤᠣ  _   _     _   _   _   _       _   _
18      ᡤᡠᠩ       ᡤᡠᠩ X     _   _   22  nsubj   _   Translit=gung
19      ᡥᡝ        ᡥᡝ  X     _   _   18  flat    _   Translit=he
20      ᠴᡤᠣ       ᠴᡤᠣ NOUN  _   _   19  flat    _   Translit=cgo
```

Figure 2: Fragment of the multi-word expression ᡤᡠᠩᡥᡝᠴᡤᠣ, *gung he cgo*, 'republic' in CoNLL-U format.

countryside', marked for accusative case *be*, is the direct object, and the verb phrase *ilibume arambi* 'construct' is the main verb.

In the following sections, we will discuss several language phenomena in Xibe, with a focus on the annotation decisions we have made for these phenomena.

4.1 Loanwords from Chinese

In UD annotation, the smallest unit is defined to be the syntactic word rather than morphemes or constituents smaller than words; morphological features can only be encoded as properties of words. However, an issue arises because of the frequency of Chinese loanwords: Xibe adapts Chinese loanwords in different ways, but in most cases, they are handled as phonemic loanwords, i.e., Chinese syllables are transliterated into Xibe letters with similar pronunciation. As a consequence, each Chinese character is written as a separate word in Xibe. Thus, it is necessary to combine these elements into a single syntactic word in the UD annotation scheme.

Figure 2 shows an example. Here the sequence ᡤᡠᠩᡥᡝᠴᡤᠣ , *gung he cgo*, 'republic' corresponds to the tri-syllabic Chinese word 'gòng hé guó'. The three syllables are directly transliterated into three Xibe syllables and written separately as three tokens. In Chinese, the first two tokens are bound morphemes while the third token is a free morpheme. Therefore, we could not find proper part-of-speech tags for the bound morphemes. Here our strategy is to assign `X` to the bound morphemes and the proper part-of-speech to the free morpheme according to the syntactic function of the complete word. In the example, the first line depicts the whole word. The following lines represent the syllables: ᡤᡠᠩ *gung* and ᡥᡝ *he* are the bound morphemes, and ᠴᡤᠣ *cgo* the free morpheme. On the syntactic level, we treat this type of structure as a multiword expression (MWE) and annotate it with a flat internal structure.

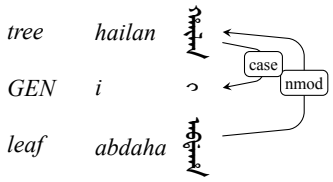

Figure 3: Dependency Tree for noun phrase ‘leaf of tree’.

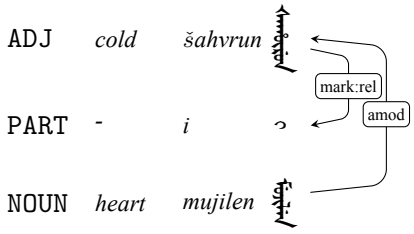

Figure 4: Dependency tree for ‘cold heart’.

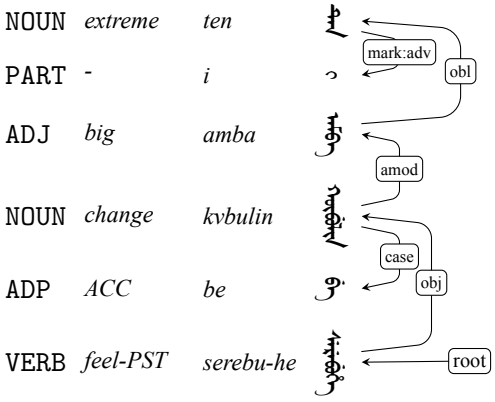

Figure 5: Dependency tree for ‘(someone) experienced extremely big changes’.

4.2 Attributive Function of *i*

ᠢ *i* is one of the case markers in Xibe, and its primary syntactic function is to express genitive and instrumental case. Figure 1 shows an example of the usage of *i* as instrumental case marker: *i* is attached to the head of the noun phrase *meni juwe gala* ‘our two hands’. Figure 3 shows an example of *i* as genitive case maker, which connects *hailan* ‘tree’ and *abdaha* ‘leaf’. Besides these two case, *i* is also used in other syntactic contexts, as an attributive function. We consider this usage a homograph of the case marker, assuming that it is influenced by Chinese. I.e., *i* is a modifier particle `PART` instead of a case marker `ADP`. The attributive function of *i* occurs frequently in *Cabcal News*.

There are two types of attributive functions: In the first case, *i* marks adjectival modifiers. In Figure 4, *šahvrun* ‘cold’ is an adjective and directly modifies *mujilen* ‘heart’, but there is an *i* without obvious function. We assume that this is borrowed from Chinese. We follow the Chinese-HK UD treebank (Leung et al., 2016; Wong et al., 2017) and annotate the adjective as the head of the particle *i*. The particle *i* is treated as a `mark:rel` dependent of the adjective.

In the second case, *i* marks adverbial modifiers. In Figure 5, *ten* is a noun, meaning ‘pole, extreme’. The following particle *i* marks it to be an adverbial modifier of the adjective *amba* ‘big’, describing the degree of the adjective. Thus *ten* is an adjunct depending on the adjective with the relation `obl`, and *i* depends on *ten* with the relation `mark:adv` indicating that the noun functions as an adverbial modifier.

4.3 Topic Marker *oci*

Xibe uses the canonical word order of subject-object-verb (SOV). Rearranging the word order is possible to a certain degree, the syntactic functions and semantics of the sentence are still clear because of the government by case markers. The topic of a sentence tends to occupy sentence-initial position and is marked via topic markers. In written Xibe, *oci* is one of these topic markers, but it shows signs of changing to a copula, influenced by Chinese. *oci* derives from the verb *ombi* ‘to become’ in its conditional converb form, and it literally means ‘if becoming somebody or something’. As a topic marker, *oci* is similar to the Japanese topic marker は *wa* and the Korean topic marker 은 *eun*/는 *neun*. We consider it an ADP, and it assigns nominal case to the subject, but it has the function of topicalization in terms of information

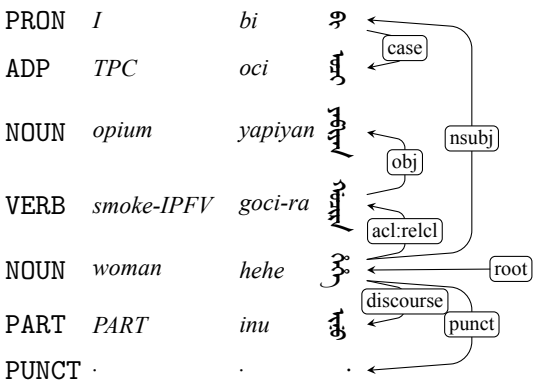

Figure 6: Dependency tree for 'I am an opium smoking woman'.

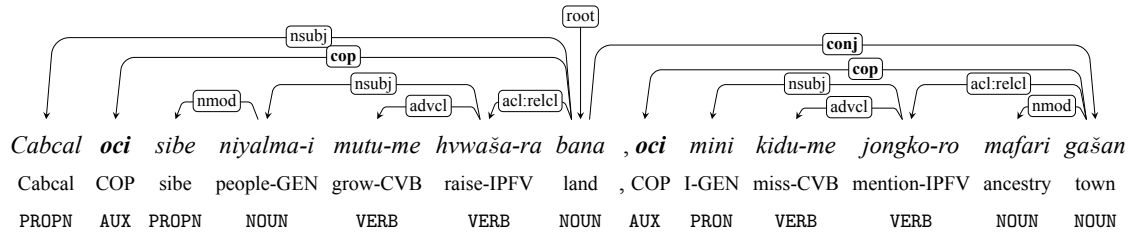

Figure 7: Dependency tree for 'Cabcal is the place where Xibe people grow up and the hometown that I miss'.

structure. In addition, when a topic marker is used, a modal particle *inu* is optionally added at the end of the sentence denoting modality.

In Figure 6, *oci* follows the subject *bi* 'I', functioning as topicalizer for the subject. *inu* – located at the end of the sentence – functions as a modal particle and indicates that the sentence is declarative. The modal particle is a dependent of the head of the sentence, marked as `discourse`, following the Classical Chinese UD Treebank (Yasuoka, 2019).

It is worth noting that *oci* seems to be in the process of changing from a topic marker to a copula, which we assume to be influenced by Chinese since it literally corresponds to the Chinese copula 是 shì . In this usage, *oci* is frequently found in *Cabcal News*, typically in equational constructions. In Figure 7[5], *oci* is used in an equational construction where it functions as a link between the main subject *Cabcal* to each nominal phrase introduced by *oci*. The head of the first conjunct in the coordinating construction is the `root`, and the head in the second conjunct depends on it via the `conj` relation. In each conjunct, *oci* as a copula depends on the nominal head via the `cop` relation. This is the only case we find so far that counters the typical SOV structure in Xibe, and we assume that it is highly influenced by Chinese.

4.4 Relative Clauses

Similar to many Tungusic languages, relative clauses in Xibe are pre-nominal, and there is no relative pronoun. The main device to render relative clauses in Xibe is via the predicate verb in a relative clause, which takes participle form and modifies the following noun or noun phrase. Xibe has imperfect and perfect participles, which express the temporal meanings present or past. The imperfect participle has the suffix *-ra/re/ro*, and the perfect participle has the suffix *-ha/he/ho*.

Under UD guidelines, a relative clause is an instance of an adjectival clause, which is characterized by finiteness and omission of the modified noun in the embedded clause. Therefore the modified noun should be an argument in the clause. In other word, there should be a gap in the clause which the head noun or noun phrase can fill in. Based on this criterion, Xibe has two types of relative clauses, subject-gap

[5]Please note that subjects in relative clauses take genitive case and are not marked for topic. Thus, *sibe niyalma-i* 'Xibe people-GEN' and *mini* 'I-GEN' are the subjects of the relative clauses.

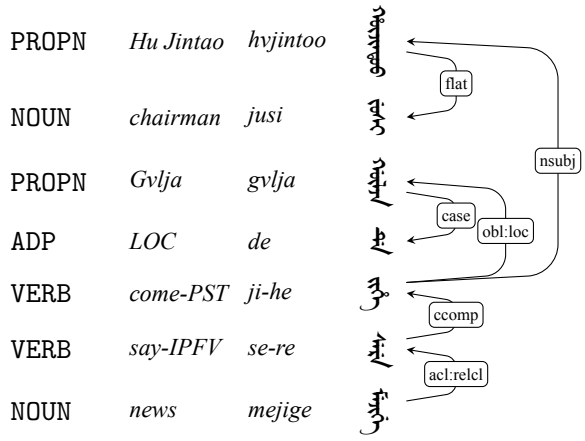

Figure 8: Dependency tree for 'the news that Chairman Hu Jintao came to Gvlja'.

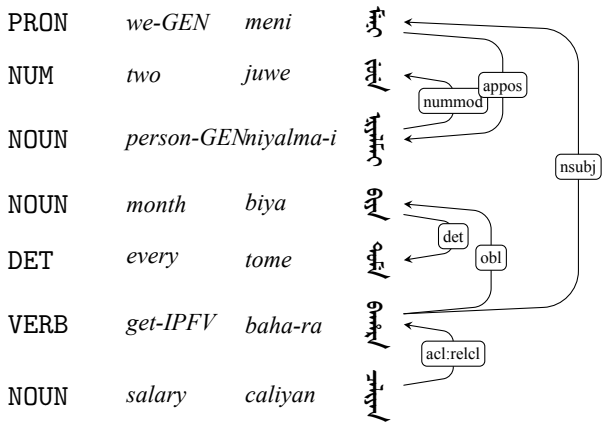

Figure 9: Dependency tree for 'the salary that we two people get every month'.

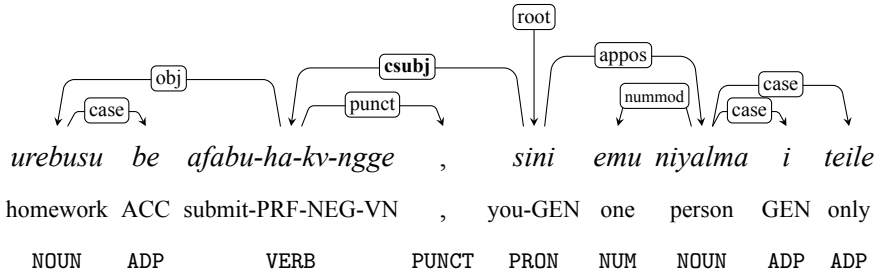

Figure 10: Sentence with headless relative clause 'The only person who did not submit the homework is you'.

and object-gap relative clauses. For a relative clause, the participle is a dependent of the modified noun or noun phrase, and their relation is `acl:relcl`.

In subject-gap relative clauses, the head noun should be able to be filled into the subject position of the clause. In Figure 8, *sere* is the imperfect participial form of *sembi* 'to say', whose object is the sentence complement prior to it. This phrase literally means 'the news which is saying that Chairman Hu Jintao came to Gvlja'. *mejige* 'news' is the subject of *sere* in relative clause.

In object-gap relative clauses, the head noun can be filled in the object gap. In Xibe, the subject noun of the relative clause must take genitive case, which puts the subject noun and the participle morphologically in a possessive relation, but semantically it is the agent. In Figure 9, *bahara* is the imperfect participle of *bahambi* 'to get' and requires two arguments. The subject consists of an appositive phrase, and both of the appositive constituents take genitive case.

In addition to these two basic types, there is a construction that can be considered a special form of relative clause. Instead of modifying a nominal constituent, the participle in the relative clause adds suffix *-ngge*, converting the participle to a verbal noun (VN). *-ngge* semantically denotes an abstract concept of an action, or an object to which the action is applied, or a person (Gorelova, 2002). In Figure 10, '*urebusu be afabuhakvngge*' is such a construction. *afabuhakv* is the negated perfective participle of verb *afabumbi* 'to submit'. By adding the suffix *-ngge*, it changes to a verbal noun and refers to a person according to the context, meaning 'the person who did not submit'. It functions as a clausal subject of the sentence, and the verbal noun '*afabuhakvngge*' depends on the nominal predicate with relation `csubj`.

Xibe also has adjectival clauses that are not considered relative clauses. The modified nominal constituent cannot be filled into either subject or object position of the clause. The participle is dependent on the head noun, and their relation is `acl`. In Figure 11, *toksimbi* 'to knock' requires two arguments in which the subject typically has the semantic feature of animacy. The head noun of this phrase *asuki*

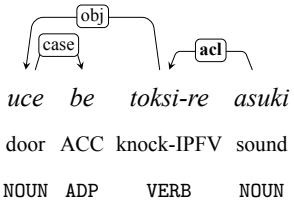

Figure 11: Dependency Tree for noun phrase ‘the sound of knocking on doors’.

-me	imperfect converb, denoting simultaneity of subordinate and main actions
-fi/-pi	perfect converb, indicating the reason for performing the main action
-ci	conditional converb, indicating the subordinate action precedes the principal action in time, meaning ‘if’
-cibe	concessive converb, usually collocates with adverb *udu* ‘although’
-tala/tele/tolo	terminal converb, indicating the main action continues until the final completion of the subordinate action, meaning ‘until...’
-nggala/nggele/ nggolo	denote the subordinate action before which the main action takes place, meaning ‘before...’
-tai/tei	denote an extreme degree of an action
-hai/hei/hoi	denote the action that is durative and intermittent, meaning ‘continually, constantly’

Table 2: Xibe converb suffixes.

‘sound’ does not meet the semantic criterion, it is the result of the action. We treat such case as `acl` as shown in the example in Figure 11.

When an adjectival clause modifies nouns such as *turgun* ‘reason’, *ba* ‘place’, *erin* ‘time’, or *fon* ‘period’, they function as adverbials, that is, causal, locative, and temporal, by adding dative case markers. We will explain these cases in the next section.

4.5 Adverbial Clauses

There are two main devices to express adverbial clauses: converbs and a certain type of adjectival clauses. Converb is a separate subclass of verbal forms, and they function as the means of subordination of one verb to another. Converbs cannot serve as predicates of a simple sentence but can function as adverbs or predicates of adverbial clauses (Gorelova, 2002). A Xibe converb is formed by the verb root and one of the eight types of converb suffixes listed in Table 2. The converb is the predicate of the adverbial clause, and it is dependent on the main predicate, their relation is `advcl`. For example, in Figure 12, *wajinggala* is the converb form of *wajimbi* ‘to finish’, meaning ‘before finishing something’. It serves as predicate of the adverbial clause and modifies the main predicate *yabuha* ‘left’.

However, converbs cannot explicitly express adverbial clause types such as locality, time, or causality. Such meanings are expressed by specific constructions, as described in Section 4.4. The constructions are syntactically adjectival, but semantically express adverbials. The construction is formed by an adjectival clause modifying a noun and a case marker, mostly dative case *de*. The participle is the predicate of the adjectival clause, modifying nouns including *ba* ‘place’, *erin* ‘time’, *fon* ‘period’, *turgun* ‘reason’. In the dative case, these constructions express locative, temporal, and causal relations with the main predicate. Therefore, the cased nouns *bade*, *erinde*/*fonde*, *turgunde* tend to serve as the corresponding subordinating conjunctions. In Figure 13, the imperfective participle *yabure* modifies *erin*, *erin* then takes the dative case, which turns the modified noun phrase into an adverbial attached to the verb, with relation `obl`. The phrase literally translates as ‘at the time of going on the road’.

Postpositions in Xibe are uninflected words denoting syntactic relationships between nouns or a noun and a verb. Postpositions govern nouns, pronouns that they follow. However, such words can also function as a subordinate conjunct when they follows a participle, and the participle serves as the predicate of the clause and is dependent on the main predicate with relation `advcl`. The subordinate conjunct func-

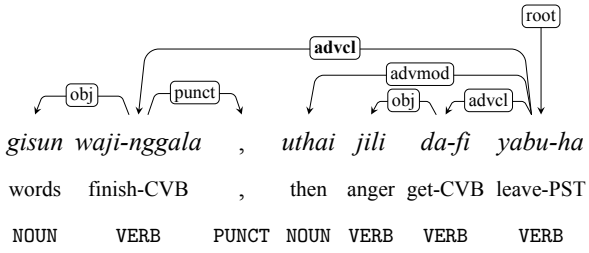

Figure 12: Sentence with adverbial cl.: 'Before (someone) finished the words, (he) got angry and left'.

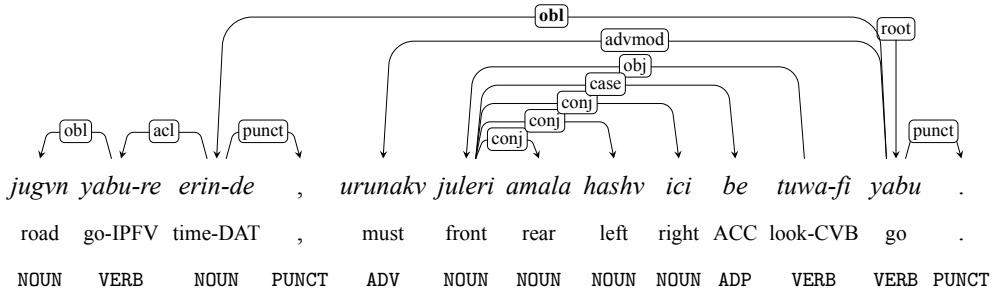

Figure 13: Dependency tree for 'When you go on the road, you must look around your surroundings'.

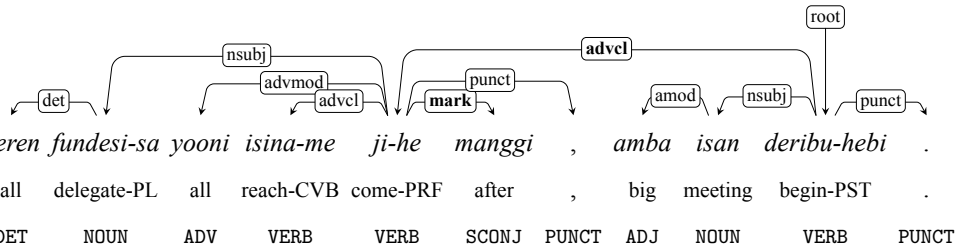

Figure 14: Dependency tree for 'After all the delegates arrived, the conference began'.

tions as a marker, and it is dependent on the clausal head. In Figure 14, *manggi* 'after' is a subordinate conjunct and follows an adjectival clause with the perfective participle *jihe* as head. *jihe* depends on the main predicate *deribuhebi*, having relation `advcl`.

5 Conclusion

In this paper, we have shown the procedure of building the first Xibe treebank using Universal Dependencies. This is an important step towards the documentation of Xibe, for which little previous research exists due to its low number of language resources. Along with the treebank construction, we document several language specific phenomena. For future work, we will continue collecting and annotating sentences from *Cabcal News* and Xibe elementary textbooks to expand our treebank. At the same time, we will also conduct corpus-based linguistic research on the language, and we will start investigating parsing approaches that will work for such a low resource language.

Acknowledgments

We are grateful to Jonathan North Washington, He Ma, and several native Xibe speakers (who want to remain anonymous) for their discussion and assistance. We would also thank the two anonymous reviewers for their helpful comments. He Zhou is supported by China Scholarship Council.

References

Liliya M Gorelova. 2002. *Manchu Grammar*. Brill.

Songjie Gu. 2016. A literature review on Sibe language. *Manchu Studies*, 2(2):83–87.

Zengyi Hu. 1994. *A Comprehensive Manchu-Chinese Dictionary*. Xinjiang People's Publishing House, Urumqi, Xinjiang, China.

Taeho Jang. 2008. *Sibe Grammar*. The Nationalities Publishing House of Yunnan, Kunming, Yunnan, China.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing universal dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 20–29, Osaka, Japan.

Shulan Li. 1979. A survey on the Sibe language. *Minority Languages of China*, 6(3):221–232.

Shulan Li. 1982. Possession category in Sibe. *Minority Languages of China*, 6(5):50–57.

Shulan Li. 1985. Adverbials in Sibe. *Minority Languages of China*, 6(5):12–25.

Shulan Li. 1988. Auxilaries in sibe. *Minority Languages of China*, 6(6):27–32.

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*, pages 92–97, Sofia, Bulgaria.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 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.

Jerry Norman. 1974. A sketch of Sibe morphology. *Central Asiatic Journal*, 18(3):159–174.

Setuken. 2009. *General Introduction to Xibe Grammar*. Xinjiang People's Publishing House, Urumqi, Xinjiang, China.

Francis Tyers, Mariya Sheyanova, and Jonathan Washington. 2017. UD Annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 266−–275, Pisa, Italy.

Xinjiang Ethnic Language Work Committee. 1992. *nei fon sibe šu tacin gisun i arara kooli (Modern Literary Xibe Orthography)*. Xinjiang People's Publishing House, Urumqi, Xinjiang, China.

Koichi Yasuoka. 2019. Universal dependencies treebank of the four books in Classical Chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28, Taipei, Taiwan.

Veronika Zikmundová. 2013. *Spoken Sibe: Morphology of the Inflected Parts of Speech*. Karolinum Press.

A Tagset, Relations and Features

This treebank uses 17 Universal POS Tags, 30 Universal Dependency relations, relation subtypes, and 20 features.

A.1 Universal POS Tags

ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM
PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	

A.2 Universal Dependency Relations and Subtypes

A.2.1 Universal Dependency Relations

`acl`	`advcl`	`amod`	`advmod`	`appos`	`aux`	`case`	`cc`	`ccomp`
`clf`	`compound`	`conj`	`cop`	`csubj`	`det`	`discourse`	`fixed`	`flat`
`iobj`	`mark`	`nmod`	`nsubj`	`nummod`	`obj`	`obl`	`parataxis`	`punct`
`root`	`vocative`	`xcomp`						

A.2.2 Relation Subtypes

`acl:relcl`	`flat:name`	`mark:adv`	`mark:plur`	`mark:rel`
`nmod:poss`	`nmod:range`	`nsubj:pass`	`obl:loc`	`obl:tmod`

A.3 Features

feature	**value**	**feature**	**value**
Abbr	Yes	Polarity	Neg
Aspect	Imp, Perf, Prog	Polite	Elev
Case	Abl, Acc, Cmp, Com, Dat Gen, Ins, Lat, Loc,Nom	PronType	Dem, Ind, Int, Prs, Tot
Clusivity	Ex, In	Poss	Yes
Degree	Cmp, Pos	Reflex	Yes
Foreign	Yes	Tense	Fut, Past, Pres
Mood	Cnd, Imp, Ind, Sub	Typo	Yes
Number	Plur, Sing	VerbForm	Conv, Fin, Inf, Part, Vnoun
NumType	Card, Frac, Mult, Ord, Sets	Voice	Act, Cau, Pass, Rcp
Person	1, 2, 3		