Clausal complementation in Nepal Bhasa

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Borui Zhang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Claire Halpert

November, 2021

© Borui Zhang 2021
ALL RIGHTS RESERVED

Acknowledgements

I gratefully acknowledge the Institute of Linguistics for the support of my thesis project. I would especially like to express earnest and heartfelt appreciation to my committee members: Claire Halpert, Brian Reese, Diti Bhadra, and Maria Gini.

I am most deeply indebted to my advisor Claire Halpert for her tremendous support and encouragement throughout my graduate program years. I started the course of Field Methods studying Somali with Claire, as my first opportunity to use what I had learned from the theoretical linguistic courses to explore endangered languages. The amazing experience of doing fieldwork with Claire created a foundational interest for me to do fieldwork later for my dissertation project in Nepal Bhasa. This project has flourished under her guidance at each stage. At times, when I encountered difficulties in the process, she was always there to help me: clarifying my thoughts of designing linguistic diagnostics, interpreting odd data patterns, providing executable feedback, and discussing in great detail any small questions that I brought up. She supported me when I was exploring new methods for interdisciplinary research where I could use a combination of linguistic knowledge and computational tools in my fieldwork. It was an invaluable experience of learning from safe failures by trying out new methods as a graduate student. I cannot adequately express my gratitude for her time and efforts to help both me and my ideas grow and develop. I owe Claire.

I learned a great amount of knowledge from Brian Reese on semantics and the computational linguistics side of my project. I enjoyed the wide-ranging discussions with him, from aspectual frameworks and theories to computational representations of complementation. He provided detailed assistance on clarifying my arguments and inspecting the code for my machine learning models. I also thank Diti Bhadra for her helpful insights and solid knowledge to help me shape my dissertation. I learned a wide

range of topics from her, not limited to, wh-movements, complementations, volitionality, and typological features of ergative languages. I would like to especially thank Maria Gini from my minor field, the College of Science and Engineering for her amazing support on interdisciplinary research between linguistics and computer science. She provided me with valuable opportunities (Women in Engineering at UMN, MnDRIVE program, Grace Hopper Conference, MinneWIC, etc.) to share my linguistic research work with the people in Engineering. Those have been inspiring me to take different perspectives to look at and tackle problems in my research.

I would like to extend my deepest appreciation to my Nepal Bhasa language consultants, Tijala Chitrakar, Sujata Bajracharya, Juju Nakarmi, Baadal Chitrakar, Animish Sthapit, Yajuman Manandhar, Aayush Tuladhar, and Arbindra Bajracharya.Their support, time, and language expertise were invaluable to this project.

I would also like to thank Dustin Alfonso Chacón for working with me on Nepal Bhasa wh-scope strategies in the early years. I also thank Jason Overfelt for his seminar course on sluicing, which sparked a new way for me to investigate Nepal Bhasa dependent clauses. I thank Hooi Ling Soh for her insights and the great conversations on syntax and Mandarin Chinese. I would also like to thank Abe Kazemzadeh for the many interesting conversations on NLP algorithms, coding, Python libraries, and Chinese languages. I would like to thank Jean-Philippe Marcotte for the opportunity of discussing Nepal Bhasa aspect with undergraduate students in his Introduction to Linguistics class. I thank Keir Moulton for his workshop on complementation at GLOW in Asia XII and the extended conversations on Nepal Bhasa complementation patterns.

I am also grateful to previous my teachers, Ettien Koffi, who brought me into the Linguistics world; Ed Sadrai, with whom I had my first computational linguistic conversation; Tim Hunter, who brought me into the field of syntax and showed me first how corpora help on theoretical research.

Last but not least, I want to thank my fellow graduate students, faculty, and friends, who have offered their support. In particular, I would like to especially thank Maria Heath, Mskwaankwad Rice, Mitchell Klein, Samantha Hamilton, Jon Coltz, Lu He, and Jon Cotner for their academic and personal help.

Finally, I thank my parents for their support over the years.

Dedication

This dissertation is dedicated to my parents, and to all those who helped make it possible.

Abstract

This dissertation examines the syntax and lexical semantics of finite verbal dependent clauses in Nepal Bhasa through fieldwork and by creating a shallow parsing model and corpus-based search to test descriptive generalizations. Nepal Bhasa deploys two main different syntactic complementation strategies: head-final pre-verbal CPs, which I argue are true complements and head-initial post-verbal CPs, which I argue are parataxis. Complementation additionally introduces certain syntactic and morphological constraints. Inchoative and perfective morphemes appear in free alternation in some mono-clausal environments, whereas in embedding structures, an embedding predicate with the inchoative suffix is restricted. By annotating a small dataset from open-source Nepal Bhasa data, I train a chunking model by adopting the technique of transfer learning in machine learning, with fine-tuning the pre-trained mBERT language model. The preliminary test results show the potential usefulness of using NLP tools to effectively build a corpus for research in low-resource languages. In particular, this method corroborates my descriptive generalization that inchoative is restricted on embedding predicates in Nepal Bhasa. Additional search over the structural treebank corpora of typologically related languages adds evidence to a cross-linguistic generalization on embedding verb restrictions.

Contents

List of Tables

List of Figures

List of Abbreviations

Acronym	Meaning
ABS	Absolutive
ADN	Adnominal
ADV	Adverb/adverbial
AUX	Auxiliaries
C	Complementizer
CM	Covert Movement
CP	Complement phrase
DAT	Dative
DOM	Differential object marking
E	Event time
ERG	Ergative
FA	Focus Alternatives
GEN	Genitive
INCH	Inchoative
IP	Inflectional phrase
IPFV	Imperfective
LF	Logical Form
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
NEG	Negation
NLP	Natural language processing
NMLZ	Nominalizer
NONPST	Non-Past
NP	Noun phrase
OSCAR	Open Super-large Crawled Aggregated coRpus
PFV	Perfective
PL	Plural
PP	Postposition

Continued on next page

Table 1 – Continued from previous page

Acronym	Meaning
PST	Past
PTCP	Participle
pro-drop	Pronoun drop
QP	Question phrase
R	Reference time
RC	Relative clauses
SLC	Sluicing-like constructions
VP	Verb phrase

Chapter 1

Introduction

This thesis focuses on finite verbal dependent CPs in Nepal Bhasa. The term *dependent*, I refer to the kind of clauses that look like the complement of verbs on the linear surface, however, it may or may not be a genuine verbal complement syntactically. A number of studies have investigated complement phrases/clauses (CP) cross-linguistically in terms of the grammatical functions (Bresnan 1972, Stowell 1981), clausal features and requirements (Bayer 1996, Davison 2007, Dayal and Mahajan 2007, Moulton 2009).

Nepal Bhasa (also known as Newari or Newar language [1]) is a Tibeto-Burman language of Nepal, spoken by the Newar people from Kathmandu Valley and surrounding regions. Multiple varieties of the language are described in the literature: Kathmandu Newari (Hale 1980, Malla 1985), Dolakha Newar (Genetti 2009), and classical Newar language. Nepal Bhasa language is an endangered language with about 860,000 native speakers, according to the Nepal (2011) census, and the number keeps declining in recent years. My language consultants in this project are mainly in the United States and grew up in Lalitpur and Kantipur, speaking the Kathmandu dialect of Nepal Bhasa at home. Their ages range from 22 to 35. Some differences in linguistic judgments are observed among the speakers.

The research methodology I use to analyze complementation in Nepal Bhasa in this project brings together three different avenues of linguistic research to build a rich empirical and analytical picture. The main data source for my theoretical analysis comes from my fieldwork. Open-source corpora and pre-trained natural language processing

[1] The term 'Newari' is dispreferred by some of my native speaker consultants.

(NLP) deep learning tools fill out the empirical generalizations based on my fieldwork. Both of these techniques create an empirical basis for my syntactic analysis of aspect and complementation. In fieldwork, I meet my language consultants individually, either in person or online, to directly collect language data from them. Fieldwork offers the most reliable source of data, with the ability to precisely control context, get negative evidence, and capture the most updated language use of the oral language style.

Indirect data collection methods, taking examples from books or articles, may not be able to reflect the dynamic language status. However, data collection from fieldwork can be expensive and time-consuming, and also limited in getting a larger range of vocabulary or a broader variety of sentence structures. As rich support to fieldwork, I used small open-source corpora to supplement my understanding of the thesis research topics. The structured corpora, treebanks, provide syntactic pattern retrieval for drawing cross-linguistic inferences. The unstructured raw authentic Nepal Bhasa data shows its efficiency in the model training process as an example of working on an NLP task for a low-resource language.

Language models as a core application are used to make theoretical or applied predictions in the field of NLP or computational linguistics. With endangered languages, it can be difficult to build good language models due to a lack of data. Multi-language models (Devlin et al. 2018) started to emerge in recent years, and are reported to achieve higher accuracy from the training strategy of fine-tuning a pre-trained language model using the target low-resource data (Cruz and Cheng 2019, Kjeldgaard and Nielsen 2021). I adopt this method and use a customized limited Nepal Bhasa dataset to train a chunking model for identifying embedded CPs.

With this thesis, I hope to add more cross-linguistic variety to the general understanding of complementation in languages by examining the relevant theories in Nepal Bhasa. I also hope to show the possibility of expanding research methodology for low-resource language by using corpora and natural language processing tools.

1.1 Aims and Objectives

On the surface, Nepal Bhasa verbal complementation can be categorized into four types, as shown in (1), headed by four different complementizers (*dhakā*, *dhayā*, *ki*, and *null*).

(1) a. Sitā-na [CP Rām-na oṃ nala **dhakā**] dhā-u.
Sita-ERG Ram-ERG mango eat.NA DHAKA say-U

b. Sitā-na [CP Rām-na oṃ nala **dhayā**] dhā-u.
Sita-ERG Ram-ERG mango eat.NA DHAYA say-U

c. Sitā-na [CP Rām-na oṃ nala Ø] dhā-u.
Sita-ERG Ram-ERG mango eat.NA say-U

d. Sitā-na dhā-u [CP **ki** Rām-na oṃ nala].
Sita-ERG say-U KI Ram-ERG mango eat.NA
'Sita said that Ram ate mangos.'

These four sentences seemingly express the same meaning regardless of the complementizers. The thesis demonstrates that these strategies are not all equivalent complementations. By finding the CP external (the interactions with the main clause) and internal (the elements inside of the CP) factors to unveil the underlying structures and the complementation strategies that are used to derive the surface representations.

Another notable aspect of complementation in Nepal Bhasa is that it triggers surprising restrictions on the aspectual morphology – the optionality of aspect on the matrix verb forms is restricted in certain complementation matrix clauses. For example, -NA and -U are interchangeable as a part of the progressive verbal morpheme as shown in (2). However, the alternation is blocked in (3) by complementation, the -U form is only preferred in the matrix predicate.

(2) Rām-na sāāt bāje TV swa-i chwoṃ**-u**/**-na**.
Ram-ERG seven time TV watch-IPFV remain-U/-NA
'Ram was watching TV at seven.'

(3) Ram-na sāāt bāje [CP Sita-na aṃ na-i chwoṃ-u/-na] swa-i
Ram-ERG seven time Sita-ERG mango eat-IPFV remain-U/-NA watch.IPFV
chwoṃ**-u**/***-na**.
remain-U/NA
Lit: 'At seven, Ram was watching that Sita was eating a mango.'

I will argue that both of the verbal suffixes -U and -NA are aspectual morphemes in Nepal Bhasa. In addition to appearing on the progressive auxiliary, they occur in both simple and complex morphological structures. At a glance, both forms can appear in contexts that are compatible with the past tense in (4).

(4) a. Rām-na oṃ na-la.
Ram-ERG mango eat-NA

b. Rām-na oṃ na-u.
Ram-ERG mango eat-U
'Ram ate a mango.'

Although the previous literature has mentioned some usage of -U and -NA, no discussions about them are found in terms of grammatical aspect. -NA has been described as a past tense disjunct form and -U the stative by Malla (1985). I provide a variety of evidence against a tense account, including that -NA yields an immediate future reading under certain conditions instead of a past reading.

In favor of -NA as inchoative among other evidence, we can compare the -NA in (5) to the one in (4a): the morpheme does not contribute the same temporal information in both sentences. The former one is compatible with the future tense, and the latter one with the past tense. On the other hand, in (6), the morpheme -U does not contribute any future readings even given the same condition as -NA in (5). The thesis further demonstrates that this behavior of -NA is part of a cross-linguistic phenomenon of immediate future readings from an inchoative aspect, which has a direct counterpart in Chinese.

(5) Ji chaen **wa-na.**
1st.SG home go-NA
✓'I'm about to go home.'
* 'I've left.'

(6) Ji chaen **wa-u**.
1st.SG home go-U
*'I'm about to go home.'
✓'I've left.'

A complete understanding of the aspectual properties of -U and -NA is essential to answer the question of why the inchoative aspect is restricted on embedding predicates in complementation. Therefore, knowing how temporal information (tense and aspect) is encoded in Nepal Bhasa is a necessary step. I investigate the distribution of the grammatical aspect (perfective versus imperfective) and lexical aspect (inchoative versus

stative) of the language, both of which exist in the matrix CP and the dependent CP in the complementation structure.

Returning to the variation in complementation strategies, I show that quantifier scoping patterns and CP positioning variation reveal differences between complementation strategies in Nepal Bhasa. Certain complementation types allow embedded quantifiers to take the sentential scope, while others do not. The embedded *wh*-phrase in the DHAKA headed CP in (7) can take the sentential scope (Zhang and Chacón 2018), which turns it to an interrogative sentence, whereas the KI headed CP in (8) disallows the *wh*-phrase taking the sentential scope.

(7) Sitā-ṃ [CP Rām-a **chu** na-u dhakā] dhā-u
Sita-ERG Ram-ERG what eat.PRF DHAKA say-PRF
'Sita said what Ram ate.'
'What did Sita say that Ram ate?'

(8) Sitā-ṃ dhā-u [CP ki Rām-a **chu** na-u]
Sita-ERG say-PRF KI Ram-ERG what eat.PRF
'Sita said what Ram ate.'
*'What did Sita say that Ram ate?'

While a number of diagnostics like the one illustrated in (7) and (8) allow us to distinguish head initial from head-final CPs, the aspectual restriction on embedding predicates appears to cut across dependent clause strategies. I argue that this contrast, combined with other differences, indicates that head-final CPs Are true complements, while head-initial CPs are actually high-adjoined parataxis.

In order to test my hypothesis that the aspectual property of inchoative from -NA that may potentially be the factor of restricting the optionality in the aspectual forms in (3), I conduct a pilot corpus search-based study to test whether inchoative marking in Mandarin and Cantonese is also restricted in the complementation environment.

The thesis also explores ways to make the most of language data for the theoretical research of complementation of endangered languages. I use the open-source Nepal Bhasa raw dataset written in Devanagari collected through OSCAR web crawling (Ortiz Suárez et al. 2019) to perform a computational linguistic task of training a chunking neural network model, for predicting Nepal Bhasa complement CPs, using natural language processing Python library tools.

1.2 Methodology

The research methodology in the thesis includes direct elicitation with speakers, syntactic and semantic analysis built on these empirical findings, and corpora-based searching and modeling. The analysis starts with figuring out the headedness and the grammatical aspect paradigm of Nepal Bhasa. These temporal cancellation tests (Mayol and Castroviejo 2013, Grice 1989) and compatibility tests based on viewpoint semantic aspect (Reichenbach 1949, 2005) to describe the temporal properties of the -NA and -U. Complementation types (preverbal, post-verbal, and sentential CPs) are examined through various syntactic-semantic diagnostics such as island effect test(Bayer 2006) and intervention effect (Beck 2006) for discovering the clausal properties and underlined complementation structures. Aspectual compatibility between the matrix clause and the dependent clause is tested, among the aspectual types of habitual, perfective, inchoative, perfect, and progressive.

A corpus of Nepal Bhasa crawled data is scripted in Devanagari from the Open Super-large Crawled Aggregated coRpus (OSCAR) is used to train a Nepal Bhasa chunking NLP model. I designed a data labeling instruction for the native speakers to identify the dependent CPs in the minimally selected Devanagari scripts. The labeled data get pre-processed and trained with a neural model on GPU with Python libraries. I adopted the method of fine-tuning the pre-trained multi-language Bidirectional Encoder Representations from Transformers (mBERT) cased language model (Devlin et al. 2018) and used NERDA library (Kjeldgaard and Nielsen 2021) for the training part. The model performance is evaluated by F-1, precision, and recall scoring with providing the confusion table for the actual counts. The cross-linguistic corpus-based searching approach is based on the finding of the cross-linguistic behaviors of the inchoative elements shared in Nepal Bhasa and Mandarin. The corpus search uses Tregex to find the target complement patterns in the structured Universal Dependencies treebank data in Mandarin and Cantonese. Its aim is to check the existence of the inchoative element in these languages to further examine the cross-linguistic generalization of complementation.

1.3 Chapter structures

In Chapter 2, I focus on the basic syntax and morphology of the language before getting into the complexity of complementation from reviewing Malla (1985), Hale (1980) and Hale and Shresthacharya (1973) and extend to what I found about the lexical aspectual morphology and word orders in my fieldwork: covering its basic clausal word order as SOV (section 2.2), to headedness in NPs (section 2.4), and non-NPs (section 2.3), to *wh*-in-situ word orders in questions (section 2.5). Nepal Bhasa is head-final, with an independently developed classifier system (Hale and Shresthacharya 1973) which is head-initial. It is an ergative language (Malla 1985) and the ergative case dropping is related to the pragmatic intention/agentivity. Animacy is also involved in grammatical case marking (Tuladhar 1985). Nepal Bhasa also displays differential object marking.

Chapter 3 demonstrates that Nepal Bhasa does not have grammatical tense marking; instead, I argue that verbs are inflected only with aspect (section 3.1). Temporal features of -NA and -U are described using Reichenbach's viewpoints framework (Reichenbach 1949) (section 3.2). I show compatibility tests of aspects in different tense environments to test the temporal features of these morphemes. Additional immediate future reading and other linguistic behaviors found by -NA. Similar elements are also found in Mandarin LE. Literature suggested conjunct/disjunct agreements (Malla 1985, Hale 1980) are tested, and new agreement variations are found in my language consultants' speech (section 3.3) The aspectual morphology has different levels of complexity, from single morphemes to complex auxiliary combinations. Single inchoative and perfective morphemes are used in forming other complex aspectual morphemes. I show the tests of the compatibility of the aspects in the matrix CPs and dependent CPs, in both simple and complex morphological forms.

In Chapter 4, I discuss the four complementation types in terms of the of dependent CPs conditions (section: 4.1). I show that *wh*-operators in Nepal Bhasa may use different scope-taking strategies in matrix clauses and dependent clauses (Zhang and Chacón 2018), as an examination of island effects and intervention effects demonstrates. The test results suggest that the head-less preverbal CPs are likely to be a type of head-final CP. Syntactic-semantic tests suggest that the head-final preverbal CPs are the genuine verbal complement. Only head-final CPs may scramble to the sentence-initial position.

The embedded in-situ *wh*-operators in these clauses can take the local scope or the sentential scope. The head-initial post-verbal CPs (*ki*-clauses) are not complements, but paratactic. Post-verbal is the only allowed position for this kind of CP (section 4.2).

In Chapter 5, I describe two pilot experiments in using open source corpora to study relevant theoretical subjects to assist fieldwork. In the first experiment (section 5.1) I discuss the NLP resources for endangered languages. I show how I adopt the idea of fine-tuning pre-trained a multiple language model to a Nepal Bhasa NER chunking neural model. I pre-processed the open-source Nepal Bhasa raw data and instructed native speakers to label a limited number of complement CPs using the IOB labels. Model training is performed on Google Colab GPU using Python libraries and followed by the evaluation and discussion. The second one is a corpus-based approach cross-linguistically searching features for embedding verbs (section 5.2). This method finds similar inchoative element behavior in Mandarin and Cantonese corpora: as I have observed for Nepal Bhasa that inchoative marking does not exist with the embedding predicates. This finding further supports my cross-linguistic generalization of inchoative behaviors. I discuss the different ways to capture linguistic components in complementation using structured treebank data.

Chapter 2

Basic Syntax

2.1 Introduction

As I showed in Chapter 1, The four types of Nepal Bhasa dependent CP clauses that all look like verbal complements on the surface, as shown in (9) with the abstracted syntactic pattern of each in (10) respectively. I propose in this dissertation that these dependent clauses have different syntactic structures and distinctive linguistic properties. Some are true complements and some are not. As we will see the details in Chapter 3 and Chapter 4, these clauses exhibit a variety of different linguistic behaviors that allow us to distinguish them on syntactic grounds.

(9) a. ✓ Sitā-na [CP Rām-na oṃ na-la **dhakā**] dhā-u.
Sita-ERG Ram-ERG mango.ABS eat-INCH DHAKA say-PFV

b. ✓ Sitā-na [CP Rām-na oṃ na-la **dhayā**] dhā-u
Sita-ERG Ram-ERG mango.ABS eat-INCH DHAYA say-PFV

c. ✓ Sitā-na dhā-u [CP **ki** Rām-na oṃ na-la]
Sita-ERG say-PFV KI Ram-ERG mango.ABS eat-INCH

d. ✓ Sitā-na [CP Rām-na oṃ na-la] dhā-u
Sita-ERG Ram-ERG mango.ABS eat-INCH say-PFV

'Sita said that Ram ate mangos.' [1]

[1] A note on my verb glossing: the verb's temporal morphology is crucial for the discussion in this thesis. I gloss verbal morphology differently from existing literature (cf. Malla (1985) using NONPST and PST for tense information) I will discuss why I think these aspectual verbal morphemes should be glossed in this way precisely in Chapter 3.

(10) a. Preverbal complementizer *dhakā*:
SUBJ [CP ... ... ***dhakā***] V

b. Preverbal complementizer *dhayā*:
SUBJ [CP ... ... ***dhayā***] V

c. Post-verbal complementizer *ki*:
SUBJ V [CP ***ki*** ... ...]

d. Preverbal null-complementizer:
SUBJ [CP ... ...] V [2]

In order to appreciate these differences, this chapter aims to provide some basic syntactic background in Nepal Bhasa. It will not cover a full grammatical sketch of this language but will focus on issues crucial to understanding the topic of complementation.

The organization of the chapter is as follows: section 2.2 begins with Nepal Bhasa word orders in main clauses with NP objects and with CP objects. We will see other types of phrases organized into two subgroups: the ones associated with non-nominal phrases in section 2.3, and those that are associated with nominal phrases in section 2.4. I will discuss some particular derived structures in section 2.5, such as *pro*-drop, *wh*-question, yes-no question, and scrambling.

2.1.1 A note on sources

There is little linguistic work on Nepal Bhasa, as discussed in Chapter 1. There are a few grammatical sketches of Nepal Bhasa that I draw on (Malla 1985, Hargreaves 1986, 1991, Hale and Shresthacharya 1973, Hale 1980, Genetti 1988, Tuladhar 1985), in conjunction with my own fieldwork with Nepal Bhasa speakers. My own research has revealed some dynamic morphological changes between the native speakers of the younger generation and what is reported in older sources. I note these relevant differences throughout the dissertation.

[2]The post-verbal null-complementizer was not acceptable to some of my consultants, so I use the term "null-complementizer" to refer the preverbal embedded CPs. Other movements involving derived sequence will be discussed in Chapter 4.

2.2 Basic clause-level word order

Understanding basic word order patterns is crucial to making sense of the variability in phenomena such as CP positioning and scope-taking in this language. The canonical word order of Nepal Bhasa is SOV (11a).

(11) a. Sitaṃ Ram-ta hala.
Sita.ERG Ram-DAT blame.INCH
S O V
'Sita blamed Ram.'

b. * Sitaṃ hala Ram-ta.
S V O

SVO (11b) is ungrammatical in simple mono-clausal sentences.[3] Some scrambling is possible in question sentences and I will discuss those cases in Section 2.5.

A basic SOV word order suggests that Nepal Bhasa is a head-final language, with the head of a phrase to the right of its complement. In (11a), the verb *hala* is the head of the VP and the NP *Ram* is the complement. Like the NP object in (11a), CPs also appear pre-verbally as in (12).

(12) Sitā-na [CP Rām-na oṃ na-la (dhakā)] dhā-u
Sita-ERG Ram-ERG mango.ABS eat-INCH C say-PFV
'Sita said that Ram ate the mango.'

While NP objects must be preverbal (11b), some embedded CPs can also appear post-verbally on the surface as in (13).

(13) Sitā-na dhā-u [CP ki Rām-na oṃ na-la]
Sita-ERG say-PFV C Ram-ERG mango.ABS eat-INCH
'Sita said that Ram ate the mango.'

The possible CP positions in (12) and (13) only show the surface word order flexibility. The existence of both orders does not necessarily mean that both positions are genuine verbal complement positions.

A key point of difference is that the sentences in (12) and (13) are headed by different complementizers. As this is a main topic of the thesis I will talk more about

[3] I note that the SVO order is not acceptable to my consultants in (11b), but some literature suggests that this order is acceptable in some cases (Tuladhar 1985).

how different complementizers work and how different complementation structures are generated and derived in Chapter 4. Setting aside, for now, the issue of Nepal Bhasa having many different complementizers, a fair question to ask here is which position is the base position for a true verbal complement CP. Zhang and Chacón (2018) argue that CPs may either be generated pre-verbally or post-verbally in Nepal Bhasa. I present new evidence that all complement CPs are generated pre-verbally in Chapter 4.

2.3 Fine-grained clausal word order

In this subsection, I will talk about the word order at the VP/TP level: adverbials (ADV, PP), sentential particles (PTCP), verbs (V), and auxiliaries (Aux). A simple word order of the VP level is shown in (14).

(14) ADV + OBJ-NP + V + AUX + PTCP

As this basic order shows, we see more evidence for a general head-final word order in the language.

2.3.1 Adverbials

I observe that AdvP and PP are left-adjoined to VP in Nepal Bhasa. An example of the temporal adverbial phrase left adjoining to the verb phrase is shown in (15); it can also move to sentence-initial position as in (15b).

(15) a. jiṃ **maṃglbār** kitāb bwanu.
1SG.ERG Tuesday book read.IPFV

b. **maṃglbār** jiṃ kitāb bwanu.
Tuesday 1SG.ERG book read.IPFV
'I read books on Tuesdays.'

A directional postposition phrase (PP) in (16) shows that the P head is on the right-side and the NP complement is on the left of the head. This is also consistent with the head-final feature of the language.

(16) chæṃ likka
home near
'near the house'

PP adjunct phrases, left-adjoin to VP, as the example (17) shows.

(17) Trisala [PP Sayal nāpāṃ] wan-i
Trisala Sayal with go-IPFV
"Trisala goes with Sayal"

Clausal adverbials also left-adjoin to the VP, as the example in (18) shows. The clausal modifier is non-finite and head-final.

(18) Jiṃ [CP ne bāle] tivi soyā
1SG.ERG eat.INF while TV watch-IPFV
"I watch TV while eating."

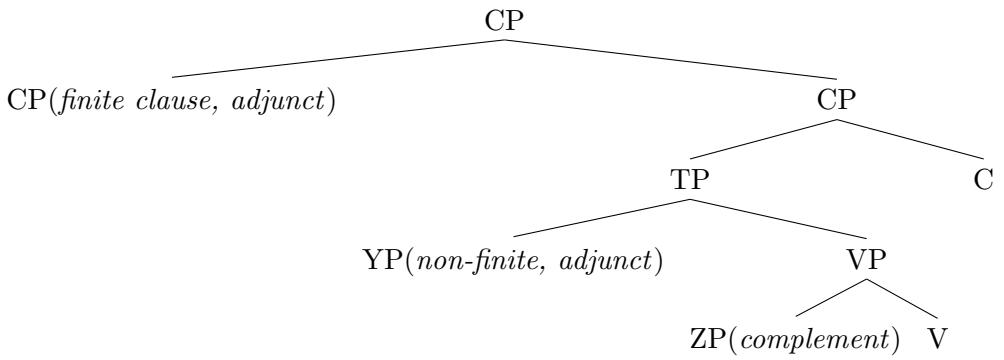
Finite adverbial clauses also left-adjoin to the main clause as the examples in (19) shows. The adjunct clause headed by the word *because* is head-final with a finite clause complement appearing to the left of the head. The adjunct clause as a whole left-adjoins to the main clause, as (19a) shows. It cannot appear in a post-verbal position, as the ungrammaticality of (19b) shows.

(19) a. [CP Jiṃ skul ma waṃu **liṃ**] ji fel ju-la.
1SG.ERG school NEG go.IPFV because 1SG.ABS fail happen-INCH

b. * ji fel ju-la [CP jiṃ skul ma waṃu **liṃ**].
1SG.ABS fail happen-INCH 1SG.ERG school NEG go.IPFV because
"Because I did not go to school, so I failed tests."

So far we have seen non-finite adjuncts left-adjoining to VP, and finite clauses left-adjoining to the matrix clause, as the tree diagram below in (20) shows:

(20)



Clausal adjoining strategies are not limited to these three ways (preverbal, sentence-initial, complement). I will show a similar left-adjoined pattern for the nominal classifier in Section 2.4.1. These will support the suggestion in Hale and Shresthacharya (1973) that Nepal Bhasa has an independently developed classifier system that often left-adjoin nodes in a head-final language. In Chapter 4, I will show another clausal adjoining strategy in Nepal Bhasa – paratacticity, which unlike any of the strategies we see in this section, allows the clause to linearly appear post-verbally (on the right side of the matrix verb.)

2.3.2 Nepal Bhasa verbs

In this subsection, I will focus on the types of verbs in Nepal Bhasa: Malla (1985) categorizes verbs in Nepal Bhasa into two groups: principal verbs and auxiliary verbs. The term principal verbs can be roughly understood as lexical verbs, which are the majority of the VP examples we have seen so far.

The auxiliary verbs in Nepal Bhasa are mainly re-purposed principle verbs (See Malla 1985:58 for the list of these repurposed words.) Malla suggests there are different ways of forming auxiliaries,[4] but he did not provide examples of the formations. I will show two kinds of auxiliaries observed in my data.

The first is modal auxiliaries. As (21) shows, the AuxP is head-final (headed by the modal auxiliary *fu* ('can') with its complement VP appearing on the right.

(21) Sayal-lāṃ [TP aṃ ne] fu
Sayal-ERG mango eat.IPFV can
'Sayal can eat mangos.'

According to Malla (1985), all auxiliaries are derived from lexical verbs, and *fu* itself has lexical meanings of 'to be well' or 'to be able.' So it is likely that the 'auxiliary' *fu* meaning 'be able,' and takes an infinitival complement. But for this surface form I do not have clear evidence to claim either that this is a bare-form verb, or that it gets inflected with the perfective marker *-(g)u*.[5] I will discuss perfective and other aspectual concepts and restriction patterns in Chapter 3.

[4]Malla suggested that one way is modifying with the non-finite form of the verb predicate, and the other one is modifying with the verb with some post-verbal particles, but he did not specify which.

[5]If a verb stem ends with *-u*, the inflected form will be the same as the surface form.

Besides individual auxiliary words, Nepal Bhasa also has auxiliary phrases. For example, the progressive form in (22), is a more complex morphological form than simple modals like *fu* ('can').

(22) Sayal-lāṃ la na-i chwoṃ-na/-u
Sayal-ERG meat eat-IPFV remain-INCH/-PFV
'Sayal is eating meat.'

In this example, *chwoṃ-na* is the progressive auxiliary, which contains a root word *chwoṃ* 'remain,' an aspectual-looking suffix *-na/-u*, and an aspect-inflected main verb *na-e* as the complement, as the linear order shows in (23). Note that both the progressive and the main verb are head-final.

(23) V-IPFV + *chwoṃ*-INCH/PFV

There are several factors that determine how the progressive auxiliary is inflected (the choice between -NA and *u*), including the person feature of the subject, the time of the event happening, and the embedding structure of the sentence. I will discuss them in detail in Chapter 3. Even as the verbal morphology gets more complex, the headedness of this language remains consistently final.

2.3.3 Sentential particles

Nepal Bhasa has many sentence-final particles to mark questions, quotes, hortatories, etc., according to Malla (1985). The particles that I have seen most frequently are the question particle *lā*, and the negative particle *ma*.

Question particle

The sentential final marker *lā* serves as a yes-no question marker (24).[6]

[6]I also observed that another way is placing a rising pitch intonation to the end of a statement sentence. Note that there must not be a rising pitch at the end of the sentence when using the other way to form a polar question.

(1) aṃ caṃ-ta ya (* lā)
mango you-DAT like Q
Do you like mangos? (with a rising pitch on *ya*)

(24) aṃ caṃ-ta ya **lā**
mango you-DAT like Q
Do you like mangos?'

With the particle being sentence-final, (24) is more evidence that CP is head-final in Nepal Bhasa. An unusual case of the question particle *lā* is shown in (25). There is no obvious matrix verb of the sentence, and the appearance of *lā* is pragmatically equivalent to the question verb 'ask.' The entire sentence cannot have a matrix question interpretation, only the declarative sentence shown in the example is available.

(25) Rama Trisālā-ta na Sitāṃ sāri niāna **lā**
Ram.ERG Trisala.DAT either Sita.ERG sari buy.INCH Q
'Ram asked Trisala if Sita bought a sari.'
* 'Did Ram ask Trisala if Sita bought a sari.'

In addition, the embedded clause has the correlative conjunction *na* ('either') on the left edge. If these elements serve as the complementizer 'if,' this head-initial conditional clause seemingly contradicts what we have seen so far about headedness of Nepal Bhasa. However, head-initial CPs have unique complementation properties, which I will discuss in Chapter 4.

Negation particle

The negation particle behaves similar to an adverb, appearing pre-verbally (26), pre-copula (27), and pre-modally (28). The position of the negation is fixed, which is unlike adverbs and cannot move around in a sentence.

(26) a. Rama aṃ **ma** na-la.
Ram.ERG mango NEG eat-INCH

b. * Rama **ma** aṃ na-la.
Ram.ERG NEG mango eat-INCH
'Ram didn't eat a mango.'

(27) a. Rama aṃ ne **ma** khu
Ram.ERG mango eat.INF NEG COP

b. * Rama aṃ **ma** ne khu
Ram.ERG mango NEG eat.INF COP
'It's not true that Ram eats mangos.'

(28) a. Rama aṃ ne **ma** fu
Ram.ERG mango eat.INF NEG can

b. * Rama aṃ **ma** ne fu
Ram.ERG mango NEG eat.INF can
'Ram cannot eat mangos.'

In this subsection, we have seen two examples of particles: the question particle and the negation particle. The headedness of the language still holds based on what I have found so far. [7]

2.3.4 Summary

In this section, I discussed the Nepal Bhasa verbs, auxiliaries, adverbials, and some particles. Besides adverbials, which seem to adjoin the VP from the left, all the rest of the elements are head-final despite some complex morphology. All the elements discussed in this section are non-nominals. I will discuss the nominal domain in the next section.

2.4 Nominals

As we have seen word order at the clausal level is consistently head-final. The next step is to look at word order patterns in the nominal phrases. The basic order of elements in a Nepal Bhasa nominal phrase is shown in (29) (Hale and Shresthacharya 1973, Hale 1985, Malla 1985):

(29) POSS + NUM + CL + ADJ + N + PL

This section is organized starting with the elements that are closest to the nominal. We will see the plural system first, then numerals and adjectives, and finally possessives.

[7] According to Malla (1985), there are more particles in Nepal Bhasa, such as emphatic, persuasive, expletive, etc., so future investigation should consider how they fit into this picture.

2.4.1 Number

Plural

The plural markers are suffixal morphemes: *-ta* and *-piṃ* as in (30) and (31). The morpheme *-piṃ* is reported to mark respect or kinship terms (Malla 1985), though I have also observed that my consultants use either suffix for the kinship terms.

(30) a. misā
woman
'woman'

b. misā-ta
woman-PL
'women'

(31) a. juju
king
'king'

b. juju-piṃ
king-PL
'kings'

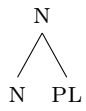
Malla (1985) also reported that the plural suffixes mark animate nouns, but my consultants use *-ta* for marking inanimate nouns too, as shown in (32).

(32) chaeṃ-ta
house-PL
'houses'

(33) kitāb-ta
book-PL
'books'

Nepal Bhasa plural (PL) marking is sensitive to the kinship and the animacy of the head noun. Given that these morphemes exist as bound morphemes (suffixes) of the noun, it is likely that the plural head is the closest functional head to the head noun, without other heads getting in between, as the syntactic structure illustrates in (34).

(34) *Plural is the closest to the N head:*



Numerals and classifiers

Nepal Bhasa has a rich classifier system. Previous studies have attempted to develop theories to capture how classifiers work in this language (e.g., Hale and Shresthacharya 1973) based on semantic values such as animacy and idiomaticity, rather than focusing

on their syntax. The Nepal Bhasa classifier system has some commonalities with many other south Asian languages such as Burmese, Vietnamese, and Thai (Jones 1970): in order to quantify nouns with numerals, classifiers (CL) are required in those languages. Some Nepal Bhasa CLs are optional but when they appear with the numerals, they left-adjoin to NP, as shown in (35). [8]

(35) ni-sāā (-mhā) sā
200 CL cow
'two hundred cows'

Classifiers are often glossed as suffixes of the numerals (Hale and Shresthacharya 1973, Malla 1985, Bhaskararao and Joshi 1985). It still may be under debate whether the classifiers are true suffixes to the numerals, but a classifier usually appears with numerals in Nepal Bhasa. [9]

As example (35) shows, the linear order for my consultants is NUM+CL+N. However, Shakya (1997) suggests that the post-nominal order, N+NUM+CL, is the underlying order in Nepal Bhasa, and the language also accepts some variations of the pre-nominal order: NUM+CL+N. I have not observed any post-nominal numerals in my consultant's speech. The absence of post-nominal numerals may be a potential point of language change or variation.[10]

In another point of contradiction between my speakers' judgments and the existing literature, while classifiers are characterized as obligatory when a numeral is present by Bhaskararao and Joshi (1985), my speakers permit it to be dropped, as (35) shows.

Other linear orders of numeral phrases are suggested. One report from (Bhaskararao and Joshi 1985) gives the example in (36), in which the numeral classifier appears in the beginning of the sentence and away from the head noun that it classifies. They suggest that this order is allowed under the condition that there is no second occurrence of a numeral in the same sentence.

[8] A very simple and clear classifier in Nepal Bhasa is the animate classifier, *-mhā*, as in (35), marking all animate nouns. The rest of the classifiers in this language mark inanimate nouns. For discussion of inanimate classifier selection and patterns, see Bhaskararao and Joshi (1985), Dryer et al. (2008), Hale and Shresthacharya (1973), Shakya (1997).

[9] Time units like 'a day' and 'a year' are the exceptions, for which no classifiers are required.

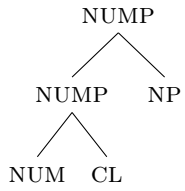
[10] I do not attempt to suggest an underlying order for Nepal Bhasa numeral phrases among the permutations since it seems to be an independent issue from the main concern of the paper of finite CP complementation structures.

(36) Ni-pā Rām-na māri ma-la
two-CL Ram-ERG flat-bread eat-INCH
'Ram ate two pieces of flat bread.' (Bhaskararao and Joshi 1985)

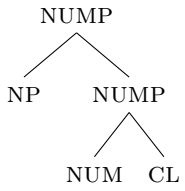
The data yields a gapped surface linear order of NUM+CL+...+N with the subject in between. Though the paper does not point out where the numeral moved from, it is certain that it cannot be the base order of numeral phrases.

Based on the data from the literature and my own fieldwork, we can structure the following two possibilities for Nepal Bhasa numeral phrases. As (37) and (38) show, NUM and CL can attach to an NP.

(37)



(38)



Demonstratives and Adjectives

Like numerals, demonstratives (39) and adjectives (40) are also head-initial in Nepal Bhasa.

(39) a. tho bhaw
this cat
'this cat'

b. wo bhaw
that cat
'that cat'

(40) wo naicicomo bhaw
that soft cat
'that soft cat'

Adjectives also seem to be inflected by the plural, like the example shown in (41). In this example, the plural morpheme is the one for marking the kin terms *-piṃ*, which we saw previously in (31), even though the head noun *misā* ('girl') is not a kin term.

(41) txha-ma-nhi-ma ralahka **piṃ** misā ta
one-CL-two-CL tall PL girl PL
'some tall girls'

Malla (1985) suggests that this is adjective concord, where the adjectives need to agree with the number and gender (animacy) of the head noun. Kiryu (2011), on the other hand, suggests that the pronominal *-piṃ* is an adnominal marker (ADN) for the adjective instead of a plural marker, and the true plural marker in this example is the morpheme *-ta*. According to Kiryu (2011), the 'adnominal marker' works as marking the adjectives that are left-joined to an NP. Adjectives can occur on the right side of the NP (Kiryu 2011, Tuladhar 1985), as (42) shows, but those must not be marked with an 'adnominal marker.'

(42) a. * txha-ma-nhi-ma misā ta ralahka **piṃ**
one-CL-two-CL girl PL tall PL
Intended: 'some tall girls'

b. txha-ma-nhi-ma misā ta ralahka
one-CL-two-CL girl PL tall

I follow Malla's description and treat *-piṃ* as the plural marker for adjectives. Without observing more left-adjoining markers in this language, I treat the absence of *-piṃ* when an adjective modifies a noun from the right as a simple omission.

Possessive/adjective pronouns are formed by adding the suffix *-gu* (allomorph *-yu*) to the bare pronoun form as in (43). As (44) shows, this morpheme is the genitive case.

(43) a. ji-gu nā
1SG-GEN name
'my name'

b. con-gu mā
2SG-GEN mom
'your mother'

c. u-yu kitāb
3SG-GEN book
'his(her) book'

(44) a. Ram-yu lha
Ram-GEN arm
'Ram's arm'

b. Sita-yu kitab
Sita-GEN book
'Sita's book'

The possessives appear as head-initial in these examples. Dryer et al. (2008) suggests that the classifier structural system is developed independently, and the'left-adjoining' of classifiers to nouns is not uncommon in some subgroups of Tibeto-Burman languages. Thus, typologically the Nepal Bhasa classifier system is likely developed independently as head-initial, in a head-final language, but holding its own order in the classifier domain.

Based on what we have seen so far, the basic linear order of the nominal domain will roughly look like the following:

(45) The linear order of nominal domain:

POSS + NUM+ CL/DEM+ ADJ + N+ PL

Cross-linguistically, it is not uncommon in a language that the nominal domain is head-initial, and other domains are head-final.

2.4.2 Cases

We have seen possessive and genitive examples in subsection 2.4.1. They are assigned within the extended nominal domain, rather than at the clause level. In this subsection, we will see Cases in Nepal Bhasa which is closely related to the nominal domain.

Malla (1985) and Genetti (1988) reported that Nepal Bhasa has the following cases: nominative (unmarked), ergative, instrumental, ablative, dative, commitative, locative, and genitive. But the literature does not explicitly demonstrate how each case is assigned.

Subject and object cases

Nepal Bhasa is an ergative language (Hargreaves 1991, Tuladhar 1985, Malla 1985). Transitive verbs that have animate arguments obligatorily mark animate subjects as ergative (ERG) and animate objects as dative (DAT), as the example shows in (46).

(46) Akāsaṃ Tara-ta mhaxi
Akas.ERG Tara-DAT recognize.IPFV
'Akas will recognize Tara.'

In contrast to (46), inanimate objects are unmarked, as shown in (47), and the animate subject in (48) is also unmarked.

(47) Akāsaṃ bol twa-la
Akas.ERG ball kick.INCH
'Akas kicked the ball.'

(48) Tara tinhu-la
Tara jump-INCH
'Tara jumped.'

The pattern of only a subgroup of objects displaying case marking follows the cross-linguistic phenomenon of differential object marking (DOM). In the case of Nepal Bhasa, the DOM system marks the animate objects with dative, and leaves the inanimate objects unmarked.

Unmarked objects are often how absolutive case appears in ergative languages. By definition, absolutive is the case that marks both the subject of an intransitive verb and the object of a transitive verb. For Nepal Bhasa, some literature (Tuladhar 1985) glosses such inanimate NPs with the absolutive case (ABS) (also see Hargreaves 2005, Malla 1985). I choose instead to simply consider absolutive, as in (47) and (48), the word *bal* ('ball') and the word *tara* (name 'Tara'), to be true absence of morphological case.

In the case of ditransitive verbs, the direct object of a ditransitive verb is unmarked, as shown in (49).

(49) sayalaṃ ji-ta inglish coṃ
Sayal.ERG me-DAT English teach.PFV
'Sayal taught me English.'

Hargreaves (2005) uses an absolutive gloss to mark the direct object of a ditransitive verb, dative to the indirect object, and the direct mark inanimate object is still consistently unmarked. Again, in this dissertation, I choose to gloss the 'unmarked' NPs as caseless, but it is a reflection of morphology, and not intended to argue against the existence of the absolutive case in the language.

Additionally, as (50) shows, the animate subject of a transitive experiencer verb (i.e., *ya* 'like') is marked with a dative case. It also seems to follow another cross-linguistically robust case variation: 'quirky' experiencer with DAT case on the subjects (cf. Legate (2012), who argues that they are derived from indirect objects).

(50) Rām-ta Sita ya
Ram-DAT Sita like.IPFV
'Ram likes Sita.'

Seeing all of the instances where DAT is used in different syntactic environments, it is clear that another factor that is related to DAT case assignment is the semantics of predicates. Earlier, we saw in (46) that the transitive verb ('kick') assigns ERG to the subject and DAT to the object. But not all transitive verbs do the same. As the example shown in (51), the verb *napla-* 'meet' is transitive and cannot assign DAT to the animate object. It is likely due to the fact that the verb is a collective predicate and inherently reciprocal. This could be a semantic constraint of case assignment in this language.

(51) Akāsaṃ Ram-*(ta) naplai
Akas.ERG Ram-DAT meet.IPFV
'Akas will meet Ram.'

Another semantic factor influencing case assignment is shown in sentences in (52), where the subject in (52a) is marked with the ergative case while (52b) is not. The two sentences have different meanings.

(52) a. **Rām-na** kitāb bwo-na
Ram-ERG book read.INCH
'Ram (intentionally) read the book.'

b. **Rām** kitāb bwo-na
Ram book read-INCH
'Ram (un-intentionally) read the book.'

Legate (2012) suggests that ergative case is inherently assigned, and non-ergative-marked subjects are derived. Hargreaves (2005) claims that the optional ergative case is only used with some, but not all, verbs in Nepal Bhasa. The case-marked subject has an agent focus (Hargreaves 2005, Genetti 1988), and the non-case-marked one has an event focus. (Also see Searle et al. 1983 for similar phenomenon described as "intentionality".)

Locatives

The locative case, the suffix *-e*, marks a noun phrase.

(53) tebil-e
table-LOC
'at the table'

A locative case can correspond to a variety of spatial interpretations, so depending on the context, a locative can mean 'at' (53), 'on', or 'in' as in (54).

(54) ji lakha-e duna
1SG water-LOC dip.INCH
'I dipped into the water' (Hargreaves 2005)

Genetti (2009) suggests that in Dolakha Newar, another Nepal Bhasa dialect, all cases are clitics, even though the glossing style of the element is much like a suffix. She did not provide the reasoning of why they should be clitics or the methods to differentiate them.

Given the current data I have collected in my study, it is possible to assume that locative is a clitic instead of a suffix, especially given the fact that other case-marking morphemes have a certain number of allomorphs, but locative has only one, so it could well be a stand-alone lexical element.

Treating the suffixes as clitics could make Nepal Bhasa fit into the typological system of other Tibeto-Burman languages. But without other testing methods, it is too strong to claim that all case suffixes are clitics in Nepal Bhasa.

In this subsection, I showed the syntactic elements that are closely associated with nominals, how they are ordered in an NP, and how cases work in the basic settings. Numeral classifiers, adjectives, demonstratives as modifiers precede N. Locatives are likely to be clitics. Nepal Bhasa's case assignment relies on the interplay of several factors in this language. First, DOM marks animate objects and indirect objects with DAT, but leaves inanimate object unmarked. Second, a 'quirky experiencer' subject is assigned DAT. Third, in the domain of transitive verbs, collective predicates do not assign DAT due to their inherent reciprocal feature. At but not least, ERG marks animate, intentional agents of transitive verbs. Agentivity/intentionality can be expressed in syntax with the optional ERG.

2.5 Questions and other structures

In this section, I discuss some constructions in Nepal Bhasa that deviate from a surface SOV word order.

2.5.1 Omissions

Pro-drop (both of subject-*pro* and object-*pro*) can occur in Nepal Bhasa matrix clauses, as the examples in (56) and (57) show, compared to the overt pronoun case in (55).

(55) Wa wana.
3.SG go.INCH
'He/She went somewhere.' (No *pro*-drop)

(56) Pasa-le wana.
store-LOC go.INCH
'Someone went to the store.' (Subject *pro*-drop)

(57) Gana wana?
where go.INCH
'Where to go?' (Subject *pro*-drop in questions)

The copula in Nepal Bhasa is *khā*. It does not have any other inflected forms. The copula cannot be dropped when the predicate is nominal, as in (58).

(58) #(Wa) bitiārti *(khā).
3.SG student COP
'He/She is a student.'

In contrast, copula-drop is allowed in adjective predicates as shown in (59).

(59) (Wa) sāā (khā).
3.SG tasty COP
'It is tasty.'

Pro-drop is less preferred in copular clauses, unless an antecedent can be found in the same sentence[11]. However, copula can be dropped in questions, as shown in (60).

(60) (Wa) su (khā)?
(3.SG) who (COP)
'Who is that?'

[11]Tuladhar (1985) explains that Nepal Bhasa *pro*-dropping is due to the implication of pronouns from the agreement marking on verbs. Since copular clauses do not show subject-verb agreement, *pro*-drop is less preferred.

2.5.2 *Wh*-in-situ

Nepal Bhasa is *wh*-in-situ. In questions, the *wh*-phrases appear where their counterparts in a declarative environment, as the examples in (61) and (62) show.

(61) Rām-na chu na-la
Ram-ERG what eat-INCH
'What did Ram eat?'

(62) Su-na aṃ na-la
Who-ERG mango eat-INCH
'Who ate mangos?'

Wh-phrases cannot be scrambled in matrix clauses, as shown in (63) and (153b).

(63) *chu Rām-na ____ na-la
what Ram-ERG eat-INCH
Intended 'What did Ram eat?'

(64) *Rām-na ____ na-la chu
Ram-ERG eat-INCH what
Intended 'What did Ram eat?'

In sentences like (65), with a focus-marked subject, an object-*wh* can optionally scramble to the left of the focused element.

(65) Chu Rām-na-caka ____ na-la?
what Ram.ERG-only eat-INCH
'What did only Ram eat?'

As we suggest in Zhang and Chacón (2018), a sometimes-covert movement of *wh*-phrase to the left of the focus intervenor *caka* must apply for the *wh*-element to take scope at the CP level. Chapter 4 further discusses covert movement and other accounts for *wh*-phrases scoping in Nepal Bhasa complementation.

2.6 Summary

In this chapter, I discussed some properties of Nepal Bhasa basic syntax to provide an understanding of the grammatical properties that are relevant to investigating complementation strategies in this language. I first talked about basic word order in the clausal

level in section 2.2. VP and CP are head-final with the exception that adverbs are not, among the non-nominal domains discussed in section 2.3. In section 2.4, I discussed the nominal domains and case assigning system and some open issues of certain cases such as intentional/unintentional. I discussed possible *wh*-in-situ word orders in questions, and impossible scramblings in mono-clausal sentences in section 2.5.

I did not go in-depth in exploring the inflectional phrase (IP) projection in this chapter, but I will discuss it in Chapter 3 with topics related to temporality in Nepal Bhasa. We have already seen some semantic properties encoded in syntax in this chapter, such as in case assignment and lexical verbs being used as aspectual elements. In Chapter 3, I will provide some background on lexical aspects and eventuality as they are encoded in the syntax of this language.

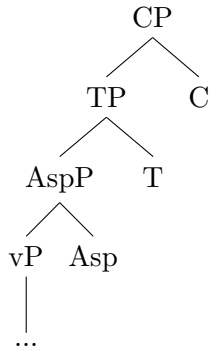
Chapter 3

Temporality in Nepal Bhasa

Temporality is a key component of understanding complementation in Nepal Bhasa. Certain verbal aspectual forms are limited in complementation structures. Before the discussion of why the forms are restricted, we need to first understand how aspect works in this language and how temporal information is encoded through aspectual marking.

Temporality, including the tense and aspectual information, is structurally located between CP and VP, as in the simplified structure in (66) adopted from Bengali (Dayal and Mahajan 2007)[1]. This chapter investigates this domain in Nepal Bhasa, exploring the ways in which temporality is morphosyntactically encoded.

(66)



I propose that Nepal Bhasa only marks grammatical aspect, similar to languages

[1]Bengali is another South Asian head-final language. Tense and aspect are combined to contribute to the temporality of a sentence/utterance and in this chapter, we will look at how temporality is conveyed in Nepal Bhasa. There are various suggested models to account for the aspectual and inflectional domain, and I do not intend to claim this particular hierarchical order presents Nepal Bhasa syntax, but only to give a general structural background of where aspect should normally appear in the structure.

like Mandarin Chinese (Huang 1998, Li 1990, Lin 2006), instead of grammatical tense. Previous literature (Malla 1985) claims that Nepal Bhasa marks the grammatical past tense and non-past tense. In this chapter, I show that Nepal Bhasa does not mark tense grammatically, and the tense information is interpreted via context. I adopt different aspectual tests and suggest aspectual distinctions in Nepal Bhasa using viewpoint semantic representations.

The formation of aspectual morphology in Nepal Bhasa has different degrees of complexity. I categorize it into two groups: the simple verb formation (verb+affix) and the complex verb formation (verb+affix auxiliary+affix). The latter one involves some unusual uses of inchoative and perfective forms.

3.1 Aspectual markers in Nepal Bhasa

The most common grammatical aspectual distinctions are imperfective and perfective (Comrie 1976), but different languages vary in the subcategory distinctions, meaning that not all of the possible grammatical aspect distinctions are in every language. Nepal Bhasa roughly follows this pattern of grammatical classification as in Figure 3.1. The single suffix morphemes *-i/-e* for habitual, *-u* for perfective, and the complex morphemes for progressive (which involves both verbal predicate and the auxiliary *-chwoṃ* and its inflection as part of the formation) are observed in Nepal Bhasa. I discuss Nepal Bhasa's aspectual morphology complexity in section 3.2.5.

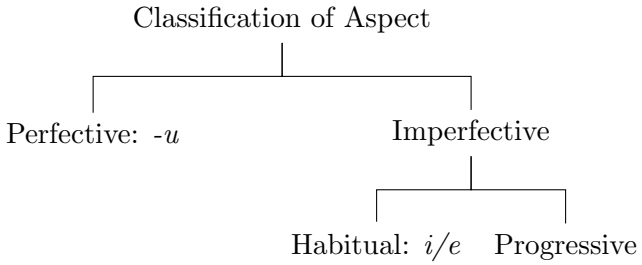


Figure 3.1: Grammatical aspectual category in Nepal Bhasa

Another suffixal morpheme -NA[2] is also found in addition to these three grammatical

[2]I use -NA to represent all the allomorphs, such as -NA, *-la*, *-ka*, etc. collectively. Literature (Malla 1985, Hargreaves 1986) uses the vowel part [*a*] only to describe the same morpheme. I gloss the

aspectual types in Nepal Bhasa. It sometimes appears to be compatible with perfective reading but does not have the same semantics as the perfective *-u*, which I will show in the next section.

3.1.1 Imperfective: Habitual

The habitual suffix *-e/-i* as shown in (67) is the only simple imperfective morpheme I observed in my data. *-e/-i* can also be involved in the complex morphological expression of other aspect, which I discuss in section 3.2.5.

(67) Rām-na oṃ na-e/-i
Ram-ERG mango eat-E/-I
✓'Ram eats mangos.'
✓'Ram will eat mangos.'
*'Ram ate mangos.'

In a past habitual environment, as in (68), *-e/-i* is prohibited. The verb *na* 'eat' must be in its bare form instead of being inflected by *-e/-i*.

(68) Rām-na oṃ na(-*e/*i), taro ā ma na
Ram-ERG mango eat-IPFV, but now NEG eat
'Ram used to eat mangos, but not anymore.'

Malla (1985) treated *-e/-i* as the non-past tense in originally, and later Hargreaves (1986) labels them as imperfective, but neither explicitly explain whether the terms are equivalent in terms of grammatical temporality. The data in (67 and 68) cannot determine whether the morpheme is definitely tense or aspect. The fact that these sentences have habitual readings suggests a trigger of imperfective marking.

I adopt Hargreaves's way (1986) of considering *-e/-i* as the imperfective aspect marker. The sentence in (69) shows that *-e/-i* can appear in past contexts: it is compatible with past progressive. If the morpheme *-e/-i* contributed non-past semantics, its appearance in a past progressive environment would be surprising.

(69) Rām-na oṃ na-i chwoṃ-na
Ram-ERG mango eat-IPFV remain-NA
'Ran was eating mangoes.'

morpheme with the preceding consonant and vowel together.

Beyond arguing that *-e/-i* is a habitual (imperfective) marker in Nepal Bhasa, which is different from what Malla (1985) suggested in earlier literature, I do not find consistent agreement in my data to differentiate *-e/-i*. This finding also differs from Malla (1985), which claims a conjunct-disjunct agreement between a subject and verb that exists in Nepal Bhasa, as shown in Table 3.1.

Person	Morpheme
1st (Conjunct)	-e /-i (phonological variations)
2nd and 3rd (Disjunct)	-i/ -i:

Table 3.1: Nepal Bhasa non-past conjunct-disjunct agreement (Malla 1985)

Malla (1985) claimed that in a declarative sentence, the form of imperfective (non-past in Malla's terms) agrees with the subject: a conjunct agreement requires the subject to be the first person, and a disjunct agreement subject to be the non-first person. According to the table, the morphemes *-e/-i* alternate based on the person of the subject. The suffix *-i* is an ambiguous morpheme here, being able to appear in both conditions. The suffix *-e* appears with the first person subject only. This contradicts with my data in (67), which shows full optionality between the morphemes. This might be revealing that this particular agreement phenomenon is disappearing in the speech of different age groups or regions. I do find other aspectual agreement patterns existing in the complex morphemes such as progressive in complex clausal structures, which is not discussed in any Nepal Bhasa literature. I will discuss those agreement issues in section 3.3.

A last note for the morphemes *-e/-i* is that the free optionality of the two that we see in simple declarative sentences as in (67) disappears when the sentence is under the scope of negation, as shown in (70). Only *-i* is acceptable, and the speakers report a preference for the future reading to be more natural than the habitual reading.

(70) Rām-na oṃ ma na-i/(-*e)
Ram-ERG mango NEG eat-I/-E
'Ram does not eat mangos.'
'Ram might not eat mangos.'

Besides habitual, I also observed grammatical progressive (69) in Nepal Bhasa as a part of the imperfective category in Figure (3.1). Progressive contains more complex

morphology, which involves an auxiliary (in this case: *chwoṃ* 'remain'). I will discuss progressive in section 3.2.5.

3.1.2 Non-imperfective aspect

We continue the discussion of the grammatical categories after viewing the habitual morpheme. I observe two more grammatical suffixes in Nepal Bhasa with simple morphology: -NA and -*u*, as shown in (71). From this section, I gloss them as the inchoative (INCH) and perfective (PFV) respectively. The later sections detail the supporting arguments.

(71) a. Rām-na oṃ na-u
Ram-ERG mango eat-PFV
'Ram ate the mango.'

b. Rām-na oṃ na-la
Ram-ERG mango eat-INCH
'Ram ate the mango.'

On the surface, the two morphemes are compatible with past tense and yield perfective readings. Hargreaves (2005) compares the two morphemes in adjective settings, which yield different aspectual patterns than the event verbs inflected with the morpheme do in my data. One conclusion we can make at this point is that they are both temporal markers in Nepal Bhasa. But this is not enough evidence to claim both aspectual markers are the same.

As the examples in (71) show the same surface meaning, it is reasonable to discuss -NA and -*u* together, but Malla (1985) discusses them in different categories. He considers -NA as a finite conjunct past tense, and -*u* as the stative aspect. A verb inflected with -*u* can be distinguished from the eventive aspect, though Malla (1985) provides no example sentences for the two types of aspect. Another mention of -*u* in (Malla 1985) describes it as a form of the 'quality adjectives' that usually ends with -*u* suffix, but the book did not address whether it is associated with the stative feature. Hargreaves (2005) considers -*u* as the imperfective disjunct marker for lexical verb like *siu* ('know').

Malla (1985) referred to -NA as *non-first person (disjunct) past tense* as shown in Table 3.2, but did not explicitly define the term 'past tense'. Hargreaves (1986) referred to -NA as the *perfective disjunct form.*

Person	Non-past	Past
Conjunct	-e/i (vowel variation)	-ā
Disjunct	-i/i:	-a

Table 3.2: Malla's verb agreement table

We have discussed the column of 'non-past' in Table 3.2 in the previous section. For a declarative sentence, in the second column, a past conjunct agreement involves the first-person subject and verbal suffix -*ā*, which is pronounced as a long low vowel, while past disjunct agreement involves non-first person subject and verbal suffix -*a*, which pronounced as a schwa.

I did not observe systematic uses of -*nā* for the first person subjects in declarative sentences in my data. Instead, my consultants choose to use -NA for most of the cases. On the one hand, this might suggest that the perfective conjunct-disjunct agreement is disappearing in the speech of different age groups or regions. I will discuss the issue with more data in section 3.3.

So far, my data suggests that -*e*, -*i*, -*u*, and -NA are grammatical aspectual morphemes in Nepal Bhasa. But the observations of their syntactic functions and agreement patterns do not align with Malla's generalization. In the next section, I focus on examining the semantics of -NA and -*u*. I will show that -NA and -*u* have different semantics, and propose that -NA may be grammatical inchoative and -*u* may be perfective in Nepal Bhasa.

3.2 The semantics of aspectual morphemes -*u* and -*na*

We come to this section of understanding the semantics of -*u* and -NA in Nepal Bhasa, a key part of investigating the complementation structure in Nepal Bhasa. Certain verbal inflections are restricted in complementation but not in other constructions, as the examples shown in (72 and 73). Both -*u* and -NA are acceptable as aspectual suffixes to the progressive auxiliary verb in a simple sentence; however, for an embedding verb in complementation, the optionality of affixation on the auxiliary disappears. Before identifying the possible factors that may cause the restriction, we need to first understand what the aspectual differences are between -*u* and -NA.

(72) Rām-na sāāt bāje TV swa-i chwoṃ**-u**/**-na**.
Ram-ERG seven time TV watch-IPFV remain-U/-NA
'Ram was watching TV at seven.'

(73) Ram-na swai **chwoṃ-u** /***chwoṃ-na** sāāt bāje [CP Sita aṃ
Ram-ERG watch.IPFV remain-U remain-NA seven time Sita mango
nai chwoṃ-u /chwoṃ-na]
eat.IPFV remain-U /remain-NA
Lit: 'At 7 am, Ram was watching that Sita was eating a mango.'

As discussed in the previous section, -NA is a grammatical aspect in Nepal Bhasa. In this section, I suggest that it may contribute to the inchoative aspectual meaning by examining the semantics of the morphemes.

3.2.1 Lexical Aspect

The aspectual term 'inchoative' is not commonly discussed in terms of grammatical aspect. Instead, it is discussed as lexical aspect, where inchoative refers to a situation that has a change of state from the framework (e.g., Dowty 1979b, 2012, Vendler 1967, Smith 1997. Dowty (1979a) suggests (modified in Smith 1997) some categories of event classifications, States, Activities, Accomplishments, Achievements, and Semelfactive as in Table 3.3. The situation types are determined based on the combined binary values of the temporal properties. The values are arbitrarily assigned from the meaning of a given verb.

Stative	Durative	Telic	Aspectual Classes	Examples
+	+	-	State	*know, own*
-	+	-	Activity	*sing, run*
-	+	+	Accomplishment	*build a house, sing a song*
-	-	+	Achievement	*win a race, reach the top*
-	-	-	Semelfactive (punctual)	*sneeze, hiccup*

Table 3.3: Event types classification

Although inchoative is not introduced as one of the temporal properties in the table, it is loosely defined as atelic, which indicates an ending point of an event. But the inchoative also requires a start point of an event, which is not relevant to the 'atelic' value.

The situation aspect framework is different from the grammatical classification (Comrie 1976), as the aspectual feature values rely on the lexical meaning of a verb in the former one and the latter one focuses on the grammaticality. However, they are not entirely separated either. For example, the perfective grammatical aspect indicates an end of an event, which is how 'telic' is used to describe a verb predicate without inflection. With the same analogy, it is possible to see an inchoative aspect, which is typically be seen as an inherited value of a verb, as being a grammatical aspect.

Additionally, the lexical aspect can be changed by applying grammatical aspectual marking, which is called aspectual class shift or event-type shift (Zucchi 1998). Nepal Bhasa shows that grammatical inchoative making can shift a non-inchoative lexical verbal predicate to inchoative by attaching the inflectional morpheme to the verb stem. So far we conclude that neither of the aspectual frameworks can efficiently fit the inchoative aspect in the system.

3.2.2 Reichenbach's viewpoints semantics

Both grammatical aspect classification and lexical aspect classification are meant to account for aspectual issues only, but the viewpoint framework can be used to represent both tense and aspect types in general. Some fundamental work on tense are Reichenbach (1949) on reference time, Prior (1967) on tense logic, and Montague (1970) on tense and modality.

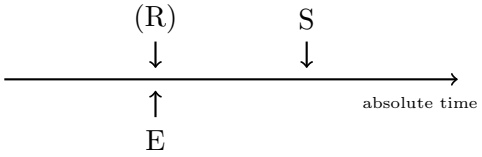
I adopt the general idea from Reichenbach to represent Nepal Bhasa's aspectual markers -*u* and -NA. Reichenbach (1949, 2005) suggests a three-way timepoint tense logic (Speech Time, Reference Time, and Event Time) of a grammatical sentence defined below. In this framework, different tense and aspect types can be represented by identifying the ordering of the relevant viewpoints in the absolute timeline.

1. Speech Time (S): the time at which the current utterance takes place
2. Reference Time (R): the time point the speaker is referring to
3. Event Time (E): the time point at which the predicate takes place

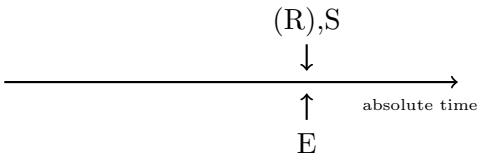
As the timelines shown in (74) is an example of how three simple tenses (past, present, and future) are presented in a naive way (Toews 2015), as the literature suggests

slightly different representations. The three time points are located and their orders are shown in each timeline for each tense type. The relation between S and E is crucial in identifying tenses, but not R. R is the crucial element when identifying aspect types.

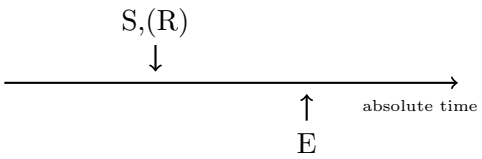
(74) a. Simple past: E ≪ S



b. Simple present: E = S



c. Simple future: S ≪ E



In (74a), the simple past tense is represented as S preceding E (E ≪ S); in (74b), E and S overlapping represents simple present (E = S); S preceding E (S ≪ E) in (74c) represents the simple future.

R becomes crucial in defining aspectual types where the relations between R and E are relevant, and not S. A simple example of the perfect aspect can be represented as E ≪ R, regardless of S. With this said, every language has its own tense-aspect types, so when necessary, all R, E, and S may be considered to describe more specific situations. It does not matter that some languages like English encode grammatical tenses with S and E relation, while some other languages like Nepal Bhasa encode grammatical aspect with R and E relation. I will discuss the representations of Nepal Bhasa -NA and *u* in this section.

3.2.3 Examining perfective and inchoative in Nepal Bhasa

Recall the minimal pair sentences in (71) repeated in (114) below, the two grammatical aspectual morphemes *-u* and -NA could be treated as being identical as they both are compatible with past perfective reading. They have an initial viewpoint representation (Kratzer and Heim 1998, Smith 1997) for past perfective as shown in (76). E $\subseteq$ R represents that the event time is included in the reference time.

(75) a. Rām-na oṃ na-u
Ram-ERG mango eat-U
'Ram ate the mango.'

b. Rām-na oṃ na-la
Ram-ERG mango eat-NA
'Ram ate the mango.'

(76) Past perfective: E $\subseteq$ R $\ll$ S

In the rest of the section, I will show that the two morphemes have different semantics and are not interchangeable in certain temporal contexts. I adopt the cancellability test to help to locate the ordering of the time points.

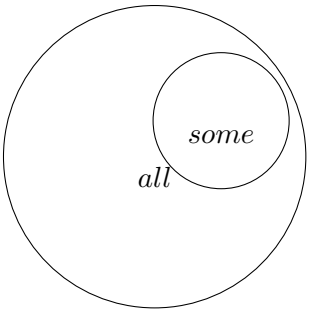
The cancellability test

The cancellability test is usually used in identifying conversational implicatures and it also works well on scalar implicatures (Mayol and Castroviejo 2013, Grice 1989). As the example in (77) shows that the implicature that the quantifier *some* is associated in (77a) -'John did not pass all the exams'- is cancellable by the second half of the sentence. The scales of the quantifiers *all* and *some* are ordered by entailment, as the set displayed in (78): *all* is stronger than *some*. The use of a weak scalar item implicates the negation of any stronger scalar item.

(77) a. ✓John passed some exams. In fact he passed all of them.

b. # John passed all the exams. In fact he passed some exams.

(78) ✓ *some* $\rightarrow$ *all*
* *all* $\rightarrow$*some*



Similarly, the implicature cancellability test also can be used in checking the domain of time span. Some studies like (Iatridou et al. 2001) have used this method to examine different aspectual readings. I use the test to examine the temporal aspectual situations of *-u* and -NA in Nepal Bhasa as shown in (79) and (80). The test is to scale the ending point of the event indicated by each morpheme. As (79) shows, E marked both with -NA and *-u* should precede R.

(79) a. ✓ Rām-na oṃ na-**u**. Neu ee sida-la.
Ram-ERG mango eat-PFV Eating time finish-INCH
'Ram ate the mango. The eating event is finished.'

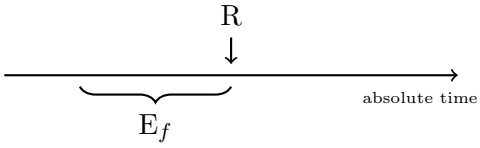
b. ✓ Rām-na oṃ na-**la**. Neu ee sida-la.
Ram-ERG mango eat-INCH Eating time finish-INCH
'Ram ate the mango. The eating event is finished.'

However, in (80a), the event time marked by *-u* cannot be canceled by an ongoing 'eating' event. It means that *-u* marks an E must end before (or at) the R, as shown in 81a (event final point marked as E_f) where fits the viewpoint semantics of the perfective aspect. On the other hand, the event time marked by -NA can be canceled by an ongoing event in (80b). This means the final point of the event does not need to precede R; instead, the final point can be left undefined.

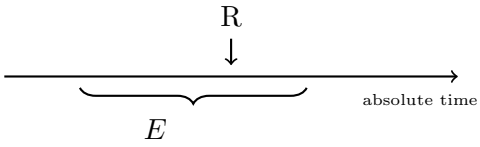
(80) a. # Rām-na oṃ na-**u**. Woṃ oṃ na-i cwona.
Ram-ERG mango eat-PFV 3rd.ERG mango eat-IPFV remain
'Ram ate the mango. He is (still) eating.'

b. ✓ Rām-na oṃ na-**la**. Woṃ oṃ na-i cwona.
Ram-ERG mango eat-INCH 3rd.ERG mango eat-IPFV remain
'Ram ate the mango. He is (still) eating.'

(81) a. -*u* must end before R



b. -NA does not define E_f, can surpass R.



The cancellability test result suggests that -*u* and -NA have different semantics even though they yield the same surface reading in (71). In particular, we see the semantics of -NA makes it compatible with perfective reading, but it is not an actual perfective marker, given its E_f can surpass R.

The cancellability test in (80) helps to find the E_f, but we also need to consider an event starting time (E_s) to complete the event descriptions of the viewpoint models of both morphemes. Otherwise, if E_s is not defined, the aspectual meaning will change to stative or imperfective, which is not how -NA behaves. With a specified E_s and an unspecified E_f -NA, it will fit the description of the inchoative aspect. I summarize -NA and -*u* into the model in (3.4) based on what have been tested so far.

Morpheme	Viewpoints
-NA	$E_s \ll \mathrm{R} \ll E_f$ or $E_s \ll E_f \ll \mathrm{R}$
-*u*	$E \subseteq R$, expands to $E_s \ll E_f \ll \mathrm{R}$

Table 3.4: Viewpoint semantics of aspect marker -*u* and -NA

Next, to further ensure that -NA and -*u* are grammatical aspect and not tense, I test their compatibility in different tense environments.

Temporal compatibility test

Sentences from (82) to (84) show basic compatibility between Nepal Bhasa tense makers and the default tense readings (past, present, and future).

(82) Wi-ta wa kitab ma-**u**
3SG-DAT that book need-U
✓ (S)he needed that book. (past)
* (S)he needs that book. (present)
* (S)he will need that book. (future)

(83) Wi-ta wa kitab ma-**la**
3SG-DAT that book need-NA
✓ (S)he needed that book. (past)
* (S)he needs that book. (present)
* (S)he will need that book. (future)

(84) Wi-ta wa kitab ma-**li**
3SG-DAT that book need-IMPV
* (S)he needed that book. (past)
✓ (S)he needs that book. (present)
✓ (S)he will need that book. (future)

However, if the condition of R changes from default to time span, the tense pattern will change in the sentences where the verbs are marked with -NA and *-u*.

In the examples from (85) to (87), an overt time span *ek mahina-yu lagi* ('for a month') is added on to the sentences in (82) to (84) respectively. -NA shows different compatibility patterns with tenses when R is a time span instead of a time point, whereas *-u* is consistently compatible with past time and consistently incompatible with the present.

(85) Wi-ta wa kitab ek mahina-yu lagi ma-**u**
3SG-DAT that book one month-GEN for need-PFV
✓ (S)he needed that book for a month. (past)
* (S)he needs that book for a month. (present)
✓ (S)he will need that book for a month. (future)

(86) Wi-ta wa kitab ek mahina-yu lagi ma-**la**
3SG-DAT that book one month-GEN for need-NA
* (S)he needed that book for a month. (past)
✓ (S)he needs that book for a month. (present)
✓ (S)he will need that book for a month. (future)

(87) Wi-ta wa kitab ek mahina-yu lagi ma-**li**
3SG-DAT that book one month-GEN for need-IMPV
* (S)he needed that book for a month. (past)
✓ (S)he needs that book for a month. (present)
✓ (S)he will need that book for a month. (future)

Additionally, it is compatible with the future tense when there is a time span.

The same conditions on the verb *bwo* ('read') are shown in (88) and (89).

(88) a. Rām-na wa bahkā bwo-**u**.
Ram-ERG that story read-PFV
✓'Ram read that story.' (past)
* 'Ram reads that book.' (present)
* 'Ram will read that book.' (future)

b. Rām-na wa bahkā bwo-**na**.
Ram-ERG that story read-INCH
✓'Ram read that story.' (past)
* 'Ram reads that book.' (present)
* 'Ram will read that book.' (future)

c. Rām-na wa bahkā bwo-**ne**.
Ram-ERG that story read-IPFV
* 'Ram read that story.' (past)
✓'Ram reads that book.' (present)
✓'Ram will read that book.' (future)

(89) a. Rām-na wa bahkā ek ghanta taka bwo-**u**.
Ram-ERG that story one hour for read-PFV
✓'Ram read that story for an hour.' (past)
* 'Ram reads that book for an hour.' (present)
✓'Ram will read that book for an hour.' (future)

b. Rām-na wa bahkā ek ghanta taka bwo-**na**.
Ram-ERG that story one hour for read-INCH
* 'Ram read that story for an hour.' (past)
✓'Ram reads that book for an hour.' (present)
* 'Ram will read that book for an hour.' (future)

c. Rām-na wa bahkā ek ghanta taka bwo-**ne**.
Ram-ERG that story one hour for read-IPFV
* 'Ram read that story for an hour.' (past)
✓'Ram reads that book for an hour.' (present)
✓'Ram will read that book for an hour.' (future)

The compatibility patterns of the differently formed verb 'read' are the same as the ones of the verb 'need', when there is no additional time span. Adding a time span to the verb 'read' changes the pattern. The reading results of both verbs are summarized in Table 3.5.

Lexicon	Morpheme	R condition	Tenses		
		Time span	Past	Present	Future
Need	-NA	No	✓	*	*
Read	-NA	No	✓	*	*
Need	*-u*	No	✓	*	*
Read	*-u*	No	✓	*	*
Need	*-li*	No	*	✓	✓
Read	*-ne*	No	*	✓	✓
Need	-NA	Yes	*	✓	✓
Read	-NA	Yes	*	✓	*
Need	*-u*	Yes	✓	*	✓
Read	*-u*	Yes	✓	*	✓
Need	*-li*	Yes	*	✓	✓
Read	*-ne*	Yes	*	✓	✓

Table 3.5: Compatibility patterns of *-u* and -NA and tenses conditioning with timespan

First, these is the evidence that both -NA and *-u* are grammatical aspectual markers. Otherwise, the tense reading pattern should be consistent regardless. Second, when R is a time span, the pattern changes to 1) -NA is incompatible with the past, whereas -U is; 2) pattern flipped from 1) for the present; 3) they all seem to be compatible with the future, except the combination of 'read' + NA. These new patterns suggest that time span plays a role in determining temporality. Furthermore, timespan-affected new patterns are strongly associated with the aspectual morpheme rather than the lexical entries. R being a time span suggests more details about the semantics of -NA and *-u*, as the starting and ending points of R (R_s and R_f) are involved in time point ordering.

The patterns of -NA in the table suggests that it is an inchoative aspectual marker.

Emenanjo (1991), quoted in (Chacon 2009), claims that in Igbo, the inchoative suffix portrays an event as "fully started". I interpret this as a requirement of -NA as the following: $E_s \subseteq$ R.

The inchoative aspect has a history of being mistreated as past tense. For instance, Chinese *le* had gone down such a path: studies (Chan 1980, Li and Thompson 1981) find that Chinese *le* is inchoative instead of past tense [3]. There are two crucial parts of the inchoative aspect: that the endpoint of an event is unidentified; and that the starting point of the event proceeds the reference time (Li and Thompson 1981), (Christensen 1990), (Chan 1980).

We can relate it to the case in Nepal Bhasa where the past tense with an R time span does not necessarily indicate this requirement, thus -NA becomes incompatible with the past tense when R is a time span. In the previous case when R is a single time point, the requirement is met.

On the other hand, for the suffix *-u*, which Malla (1985) describes as the stative marking, I suggest it is perfective, because the only pattern that changes when R is a time span is with the future tense interpretation. The overt time span is a typical sign of a stative. Then, the next question will be why a time point R disallows the future interpretation for *-u*. The imperfective suffixes in these examples, *-li* and *-ne* show consistent compatibility patterns with the tense environments regardless of the time span.[4]

Adjectivals suffixed with *-u* and -na

Beside what we have seen so far the compatibility pattern regular verbs with *-u* and -NA suffixing, they can also suffix adjective predicates, as in the examples shown in (90) and (91), where the morphemes attach to the auxiliary *soṃ* ('become'). (90) has a stative reading and (91) an inchoative reading.

(90) Ball wachu soṃ-u / *soṃ-na
Ball blue become-PFV / become-INCH
'The ball is/was blue.'

[3] *le* has other independent function such as resultative verbs, which I will not discuss.

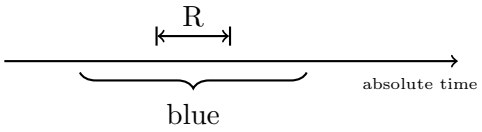
[4] I gloss imperfective with the consonant and the following vowel altogether. Therefore, *-li* and *-ne* are allomorphs in this case.

(91) Ball wachu *soṃ-u / soṃ-na
Ball blue become-PFV / become-INCH
'The ball becomes?/became blue.'

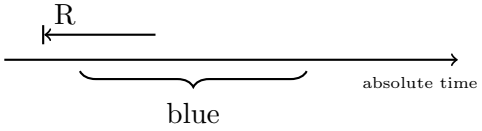
The morpheme *-u* in (90) shows an unchanging status of the color of the ball, so there should not be any state-changing during R as shown in (92a). Contrastively, (91) shows a process of an event of 'color changing' state so the state changing point must be included in R as shown in (92b).

The examples reveal a special property for *-u* to be compatible with both present and the past tense, in contrast to the sentences where *-u* is not compatible with present tense with its suffix regular verbs.

(92) a. *-u* yields stative reading in adjectival predicate



b. -NA yields inchoative reading in adjectival predicate



This again proves that the two morphemes are different aspectual morphemes and they can change the time of an event in different temporality conditions.

Additionally, a resultative reading can only be achieved by -NA but not *-u*. This is not surprising since a resultative aspect includes a change of state with which the inchoative -NA can cope and stative/perfective *-u* does not.

(93) Rām-na khu-ta wachu **soṃ-ka** /***soṃ-u** da-la.
Ram-ERG thief-DAT blue become-INCH become-PFV beat-INCH
'The thief was beaten to blue.'

Soṃ ('become') is not the only auxiliary for adjective predicates. Other auxiliaries are used to forming adjectival predicates in Nepal Bhasa, as the examples of *ju* ('happen') are shown in (94) and (95). Regardless of the chosen auxiliaries, the same pattern of *-u* being compatible with both the present and the past tenses remains.

(94) Sita birami ju / *ju-la
Sita sick happen.PFV / *happen-INCH
'Sita was/is sick.' Implication: 'Sita is not sick now.' (*Stative reading*)

(95) Sita birami *ju / ju-la
Sita sick *happen.PFV / happen-INCH
'Sita got sick.' Implication: 'Sita can be sick or well now.' (*Inchoative reading*)

The event of 'being sick' is stative in (94), which naturally has an E_f is included in R. On the other hand, the event of 'being sick' is inchoative in (95), where the change of state from 'not sick' to 'being sick' must be included in R.

An additional restriction that distinguishes the morphemes apart is shown in (96).

(96) a. Sita birami ma ju
Sita sick NEG happen.PFV
'Sita was not sick.'

b. * Sita birami ma ju-la
Sita sick NEG happen-INCH
Intended:'Sita did not get sick.'

c. Sita birami ju maku
Sita sick happen.PFV NEG.BE
'Sita did not get sick.'

The morpheme *-u* is compatible with negation, whereas -NA is not, as shown in (96a) and (96b). This may indicate that negation is stative thus conflicting with 'changing state', therefore -NA is limited. The compatibility patterns we found so far with *-u* and -NA are summarized as in Table 3.6.

Conditions	Patterns					
	Past		Present		Future	
	-NA	-U	-NA	-U	-NA	-U
− time span	✓	✓	*	*	*	*
+ time span	*	✓	✓	*	✓	✓
adjectival inchoative	✓	*	✓	*	NA	NA
adjectival stative	*	✓	*	✓	NA	NA
resultative	✓	*	✓	*	NA	NA
negation	*	✓	*	✓	*	*

Table 3.6: Summary of *-u* and -NA compatibility pattern with different conditions

As the pattern suggests, -NA is likely to be inchoative and -*u* to be stative/perfective, which supports the predicted viewpoint model repeated in Table 3.7.

Morpheme	Viewpoints
-NA inchoative	$E_s \ll \mathrm{R} \ll$, E_f must undefined
-*u* perfective	$E \subseteq R$

Table 3.7: Viewpoint semantics of perfective and inchoative

To summarize, I examined the semantics of -*u* and -NA with first discussing the term uses of perfective and inchoative in lexical aspect and grammatical aspect, I propose the semantics of the morphemes using viewpoint semantics. I test these semantics with temporal cancellability tests with different event conditions. The results suggest that the semantics of -NA is likely to be inchoative and -*u* to be perfective/stative.

3.2.4 Immediate future interpretation with inchoatives

Here I describe a special property of the inchoative aspect -NA: it yields an immediate future reading in certain situation as the example shown in (97). In contrast, -*u* does not have this property: the future reading is unacceptable in (98).

(97) Ji chaen **wa-na**
1st.SG home go-INCH
✓'I'm about to go home.'
* 'I've left.' (Nepal Bhasa)

(98) Ji chaen **wa-u**
1st.SG home go-PFV
*'I'm about to go home.'
✓'I've left.' (Nepal Bhasa)

In the previous section, we learned that -NA is compatible with the past tense. However, this special case in (97) contradicts the generalization, as only the immediate reading is available and not the past tense reading.

This special property of immediate future reading from inchoative is also found in another Sino-Tibetan language, Mandarin Chinese. The sentence-final particle *le*, which usually has a default past-tense reading, also yields immediate future reading in the same settings, as shown in (99). The difference between Nepal Bhasa and Mandarin

is that while the past tense reading is not available while the immediate future reading is available in Nepal Bhasa, in Mandarin, both readings are available. The sentence is therefore ambiguous Mandarin and unambiguous in Nepal Bhasa.

(99) Wo huijia **le**
1st.SG go.home LE.INCH
'I'm about to go home.' (Mandarin)
'I've left.'

Regardless of the ambiguity, some requirements must be met in both languages for immediate future reading. The first condition is that the subject agent must be the 'knowledge holder' in declarative as in (97) and (99), yielding first-person agent requirement. In questions, the requirement 'flips' to second person, as illustrated by (100) and (101) respectively show that a third-person subject in both Nepal Bhasa and Mandarin does not yield the immediate future reading, in contrast to (97) and (99). Literature (Christensen 1990, Chan 1980, Lin 2003) suggests that Chinese has an inchoative aspect but none of them mention its association with the agent.

(100) Wo chaen **wa-na**
3rd.SG home go-INCH
*Intended: 'He is going home.'
'He left.' (Nepal Bhasa)

(101) Ta huijia **le**
3st.SG go.home INCH
*Intended: 'He is going home.'
'He left.' (Mandarin)

Second, I propose that the immediate future interpretation requires a volitional predicate, like *wa* (leave) above. With non-volitional predicates, no future interpretation is available, even with a first-person subject, as in (102).

(102) Jiṃ zaṃs loma-**na**
1st.SG.ERG exam forget-INCH
*Intended: 'I'm going to forget the exam.'
'I forgot the exam.' (Nepal Bhasa)

Christensen (1990) suggests that in Chinese, the inchoative aspect indicates imminent action. Chan (1980) explains this as "when the inchoative combines with a verb denoting an event, the situation which arises is a state which is the result phrase or aftermath of the event." This means that a verb that fails to have a resultative action will automatically fail to have this kind of reading. For instance, the verb 'arrive' is a resulting state in itself, and there are no results of 'arrive'. This claim is slightly different from what I describe as non-volitional, which emphasizes the event is not controlled by the agent. Therefore, from this perspective, the verb 'arrive' can also be considered as a non-volitional verb and will not have an immediate future reading.

The third restriction is that sequential actions cannot convey immediate future interpretation. As the Mandarin examples shown in (103 and 104) and the Nepal Bhasa example in (105) and (106).

(103) Wo zou le bing shang che le
1SG leave LE and get-on bus LE
*'I'm about to leave and get on to the bus.'
'I left and got on to the bus.' (Mandarin)

(104) Wo shang che le
1SG get-on bus LE
'I'm about to get on to the bus.' (Mandarin)

(105) Jɪm̩ wa-na ale chicken ta-na
1.SG.ERG leave-INCH and chicken cut-INCH
* 'I'm about to leave and cut the chicken'
* I left and cut the chicken. (Nepal Bhasa)

(106) Jɪm̩ chicken ta-na.
1.SG.ERG chicken cut-INCH
'I'm about to cut the chicken.' (Nepal Bhasa)

In Mandarin, when a sequence of volitional verbs is conjoined, only the past tense interpretation remains. In contrast, in Nepal Bhasa neither the past tense nor the immediate future survives in the sequenced action setting.[5]

[5]We have seen in (97) where the person agent yields only the immediate future reading with the verb stem *wa* 'go', instead of past reading. Consequently, putting it together with other action verbs, as in (105), still does not ameliorate the past reading context. This may be due to some special aspectual property that is carried by this particular word, since other words such as *ta* ('cut') in (106) do not have

3.2.5 Aspect with complex morphology containing *-u* and -na

There are aspectual meanings that are not directly indicated by a single morpheme in Nepal Bhasa, rather appearing as a complex verb form with auxiliaries (Malla 1985). Hargreaves (1991) discusses the complex morphology for some auxiliaries *ten* ('ready'), *te* ('time to'), and *dhun* ('finish'). I observed the auxiliaries *chwoṃ* ('remain'), *dhun* ('finish'), and *soṃ* ('become') in my fieldwork. In this section, I will discuss these auxilaries' morphological representations, as one way to express progressive, perfect, and resultative aspect respectively. Crucially, I show the two morphemes, -U and -NA that we have previously seen as perfective and inchoative, also appear these complex forms. As (107) shows, in the progressive, the lexical auxiliary element *chwoṃ* can combine with either morphemes to form the progressive aspect.

(107) Rām-na oṃ na-i chwoṃ-na/chwoṃ-u
Ram-ERG mango eat-IPFV remain-NA/remain-U
'Ram is eating a mango.'

The progressive forms are more complex with multiple verbal elements and different aspectual suffixes. As the pair of examples in (108 and 109) show, both sentences have the same meaning.

(108) Jiṃ aṃ na**-i** chwoṃ**-u** /chwoṃ**-na**
1st.ERG mango eat-IPFV remain-U remain-NA
'I am eating the mango.'

(109) Jiṃ aṃ na**-yā** chwoṃ**-u** /chwoṃ**-na**
1st.ERG mango eat-NMLZ remain-U remain-NA
'I am eating the mango.'

The lexical word root *chwoṃ* ('remain') functions as a progressive auxiliary in Nepal Bhasa. Hargreaves (1986) called it an auxiliary, as when a regular verb is borrowed to function as an aspect. These morpheme combinations are acceptable regardless of the person of the subject, as the third person as the examples in (110) and (111) show.

(110) Akas-aṃ aṃ na**-i** chwoṃ**-u** /chwoṃ**-na**
Akas-ERG mango eat-IPFV remain-U remain-NA
'Akas is eating the mango.'

(111) Aksa-aṃ aṃ na-**yā** chwoṃ-**u** /chwoṃ-**na**
Akas-ERG mango eat-NMLZ remain-U remain-NA
'Akas is eating the mango.'

The progressive aspect in Nepal Bhasa is formed by a sequence of verb (main-verb + *chwoṃ*) morphologically, with each component bearing one of the two possible suffixes: -*i* versus -*yā* and -*u* versus -NA respectively. So theoretically, there can be four possible ways to form the progressive aspect as Table 3.8 shows.

Morpheme combination	Examples
V-*i* + *chwoṃ-na*	(108) and (110)
V-*i* + *chwoṃ-u*	(108) and (110)
V-*yā* + *chwoṃ-na*	(109) and (111)
V-*yā* + *chwoṃ-u*	(109) and (111)

Table 3.8: Possible morpheme structures of Nepal Bhasa progressives

.

The four types of morpheme structures of the progressive are all available with first and third-person subjects. The data, therefore, suggest that there are no subject-verb agreement concerns found in progressives so far. We have not seen any agreement restrictions in the morphological forms, which differs from the conjunct/disjunct agreement pattern from the literature.

The first pair is the main verb suffixes, -*i* and -*yā*. It is interesting to see that -*i*, the imperfective marker we just discussed, is also involved in progressives morphologically. Typologically, it could be an economical strategy to form an aspect with existing aspectual morphemes if a language does not have a variety of single morphemes for aspects in general.

The other morpheme *yā*, Malla (1985) describes it as 'gerundive' as it usually turns a verb to a noun, and I tentatively gloss it as a nominalizer (NMLZ) as shown in (112).

(112) Mi-**yā** wonā chwoṃ-na
sell-NMLZ go-INCH remain-NA
'went on selling' (Malla 1985)

In Malla's example, the main verb root *mi* ('sell') is suffixed with -*yā* and yields a gerundive meaning of selling.

A study by Genetti (2005) on Dolakhā Newar (another closely related Newar dialect) suggests the suffix *-an*, which looks like the counterpart of *yā* as they appear in similar contexts, is a non-finite participial. The Dolakhā Newar sentence below in (113) has the verbal structure the same as those in (108) and (109). The two dialects also share the same lexical root *coṃ* ('remain') for the progressive aspect, though suffixal morphemes differ.

(113) kehē ho cicā-uri dāi coṃ-**an** coṃ-hin.
y.sister and small-top e.brother stay-PART stay-3P.PST3
'The sister and brother continued to stay (where they were).' (Genetti 2005)

It is reasonable to suggest that *-yā* may be non-finite in Nepal Bhasa, given that as a non-finite element, the morpheme cannot stand alone in the matrix clause, and I have not seen a case of *yā* standing alone in a matrix sentence.

I summarize the morphological patterns of Nepal Bhasa imperfective, perfect and inchoative, progressive, and resultative in Table 3.9.[6]

Aspect	Person	Structure	Morpheme Annotation
Imperfective	1 st	*V-(e/i)*	IPFV
Imperfective	3 rd	*V-(e/i)*	IPFV
Inchoative	1 st	*V-na*	INCH
Inchoative	3 rd	*V-na*	INCH
Progressives	1 st	*V-i/yā + choṃ-na*	IPFV + PROG +INCH
Progressives	1 st	*V-i/yā + choṃ-u*	IPFV + PROG + PFV
Progressives	3 rd	*V-i/yā + choṃ-na*	IPFV + PROG + INCH
Progressives	3 rd		
Perfective	3 rd	*V-u + du*	PFV +BE
Perfective	3 rd	*V-e + dhuahka + la*	IPFV + FINISH + INCH
Perfective	1st		
Resultatives	3rd	*soṃ-ka + V-la*	BECOME-INCH + V-INCH
Resultatives	1st		

Table 3.9: Nepal Bhasa aspects that involve inchoative and perfect markers

The table shows various morphological combinations of aspectual markers in Nepal Bhasa, which all related to the morphemes *-u* and -NA. (Hargreaves 2018) discusses

[6]These structures are based on my current data collection and do not intend to say every Nepal Bhasa word will follow the patterns. Hargreaves (1991) suggests the causative case, *-k* or *-ka*, is in perfect aspect. Since the causative is independent, not occupying the morphological position where *-u* or *-na* are located, I consider the causative as part of the auxiliary word.

alternations of the auxiliary *dhun-* that are egophoricity encoded by conjunct/disjunct agreement. Those alternations are different from the alternation dataset that I discuss in my fieldwork.

Unlike *-yā* and *-i*, the other pair of suffixes in (108) and (109) with *chwoṃ* ('remain'), *-u* and -NA, can appear independently in these sentences, so they seem to be finite. There remains work to be done to connect tense/aspect with finiteness before claiming that the inchoative and perfect are in finite environments. The Nepal Bhasa progressive aspect has complex morphological forms, which involve lexical verb, habitual forms, and auxiliaries, I will not simply gloss such complex form as PROG; instead, I mark each component in the form.

3.3 Morphological restriction in complex clauses

The chapter so far has demonstrated the behaviors of aspectual morpheme *-u* and -NA in terms of their compatibility of tense, similarities, and differences in meanings when attaching to different verb classes (e.g., lexical verbs, adjective, and auxiliaries). In this section, we will see the morphological behaviors of the aspect in complex clauses.

3.3.1 Embedding verb restrictions in complement CPs

We have seen *-u* and -NA are optional in the examples like in (114) in section 3.2.3.

(114) a. Rām-na oṃ na-**u**
Ram-ERG mango eat-PFV
'Ram ate the mango.'

b. Rām-na oṃ na-**la**
Ram-ERG mango eat-INCH
'Ram ate the mango.'

However, the optionality of *-u* and -NA is restricted in complementation. As the example in (115) shows, the embedding verb limits the inchoative form, whereas this optionality is not affected in the dependent CPs.

(115) Ram-na [CP Sitā-na oṃ na**-la/-u** dhakā] tā**-u/*-la**
Ram-ERG Sita-ERG mango eat-INCH/PFV C hear-PFV/-INCH
'Ram heard that Sita ate the mango.'

Similarly, the optionality of inflected auxiliaries is also restricted in complementation. As the mono-clause sentences in (116) show, both forms are grammatical, but the embedding verb with -NA ending is blocked in complementation as shown in (117).

(116) Rām-na oṃ na-i **chwoṃ-na/chwoṃ-u**
Ram-ERG mango eat-IPFV remain-INCH/remain-PFV
'Ram is eating a mango.'

a. Rām-na sāāt bāje TV swa-i **chwoṃ-na/chwoṃ-u**.
Ram-ERG seven time TV watch-IPFV remain-INCH/remain-PFV
'Ram was watching TV at seven.'

(117) Ram-na swai **chwoṃ-u** /***chwoṃ-na** suthe sāāt bāje [CP Sita
Ram-ERG watch.IPFV remain-PFV remain-INCH when seven time Sita
aṃ nai **chwoṃ-u** /**chwoṃ-na**]
mango eat.IPFV remain-PFV /remain-INCH
'At 7 am, Ram was watching that Sita was eating a mango.'

In (117) the verb in main and dependent predicates are both progressive auxiliaries. But the suffix -NA is unacceptable for the main clause predicate, while the embedded predicate allows either form [7].

The data below show different aspect used in the embedding verb and the embedded verb. Restrictions of the -NA formed auxiliary must not be used in the main clauses [8].

[7] I use the uppercase of -NA to indicate all the allomorphs of this morpheme.

[8] In mono-clausal sentences of the perfect aspect, there is a preference of which morpheme to choose, which is unlike the free choices in progressive. This might be due to the additional causative marking in the perfective (cf. Hargreaves 1991).

(1) a. * Sita-na nigu chitthi che dhuṃ-gu
Sita-ERG 2 letter write.IPFV finish-U

b. Sita-na nigu chitthi che dhuṃka-la
Sita-ERG 2 letter write.IPFV finish-NA
'Sita has written two letters.'

(2) a. Jin nigu chitthi che dhuṃ-gu
1SG-ERG 2 letter write.IPFV finish-U

b. * Jin nigu chitthi che dhuṃka-la
1SG-ERG 2 letter write.IPFV finish-NA
'I have written two letters.'

(118) Ram-na [CP Sita-na nigu chitthi che **dhuṃ-gu/-kala**] swaya
Ram-ERG Sita-ERG 2 letter write.IPFV finish-PFV/-INCH watch.IPFV
chwoṃ-*na/-u
remain-INCH/PFV
Lit: 'Ram has been watching that Sita has written two letters.'

Conjunct/disjunct agreement (Hargreaves 1991, Malla 1985, Hale 1980) suggests an agreement between subject agent and the verb in Nepal Bhasa. This may not be able to account for (115), (117) and (118). First, the current optionality issue is between *-u* and -NA, but the suggested agreement in Malla (1985) (see Table 3.2) does not include *-u*, therefore it cannot account for *-u* directly. Second, -NA is considered as the disjunct past form, which should not co-exist with the first person in a declarative sentence. But I observed the counterexample of a co-existence in (97).

Additionally, my consultants show different judgments on the conjunct/disjunct agreement pattern with imperfectives. Hargreaves (1991) suggests the conj/disj agreement can account for a referencing pattern between clauses, as the example cited in (119) and (120). The data shows that the embedded subject has the same indexing as the main clause subject when the conjunct imperfective form is used as in (119), whereas the subjects in the embedded and main clause in (120) cannot refer to the same person with the disjunct imperfective marking.

(119) Wo-ṃ$_i$ [CP [TP lā **na-e** dhakā]] dhāl-a
He.ERG meat eat-IPFV.CONJ C said
'He$_i$ said that he$_i$ will eat meat.'

(120) Wo-ṃ$_i$ [CP [TP lā **na-i** dhakā]] dhāl-a
He.ERG meat eat-IPFV.DISJ C said
'He$_i$ said that he$_j$ will eat meat.' (Hargreaves 1991)

However, my consultants believe that (120) is ambiguous between the co-reference and different-referencing readings, as the updated judgement shown in (121). They agree with the judgement in (119).

(121) Wo-ṃ$_i$ [CP [TP lā **na-i** dhakā]] dhāl-a
He.ERG meat eat-IPFV.DISJ C said
'He$_i$ said that he$_{i/j}$ will eat meat.'

Additionally, I found the co-indexing is acceptable with using inchoative -NA (the disjunct form in Malla 1985) in the past tense as in (122).

(122) Wo-ṃ$_i$ [CP [TP lā **na-la** dhakā]] dhāl-a
He.ERG meat eat-DISJ.INCH C said
'He$_i$ said that he$_{i/j}$ ate meat.'

I also tested -*u* in the same embedding environment as shown in (123). The result shows that only different references are allowed, which contrasts with the result of -NA.

(123) Wo-ṃ$_i$ [CP [TP lā **na-u** dhakā]] dhāl-a
He.ERG meat eat-PFV C said
'He$_i$ said that he*$_{i/j}$ ate meat.'

In fact, the embedding verbs are mostly in perfect form -U in my fieldwork data collection. Language consultants report a weaker optionality judgment only when the matrix embedding verb stem is *dhā* ('say') as shown in (124).

(124) [CP Rām-na Sitā-ta khan-u dhakā] Rām-na **dhā-u/-la**
Ram-ERG Sita-DAT see-PFV C Ram-ERG say-PFV/-INCH
'Ram said that Ram saw Sita.'

So far, despite the edge case of the embedding verb 'say', embedding verb restriction in Nepal Bhasa is prominent. Declarative sentences do not reflect the conjunct disjunct agreement.[9]

3.3.2 Morphological restriction in adjunct clauses

In addition to the restrictions in complementation, comparing to (125) to (126) the position of an adverbial CP may restrict the choices of -*u* versus -NA.

(125) [CP Ji-ta na-e māsti **wa-u/ *la** liṃ] jiṃ na
1SG-DAT eat-IPFV.CONJ wanting become.PFV/*INCH because 1SG.ERG eat
wan-i
go-IPFV
'I'm hungry, so I will go to eat some food.'('Lit: Because I want to eat, I will go to eat some food.')

[9]Question sentences are very crucial in conj/disj agreement, I currently do not have enough data to suggest if those agreements exist in my consultants' speech.

(126) Jiṃ na wan-i chae-dhā-xini, [CP ji-ta na-e māsti
1SG.ERG eat go-IPFV why-say 1SG-DAT eat-IPFV.CONJ wanting
wa-la/*u]
become.INCH/*PFV
'I'm hungry, so I will go to eat some food.'('Lit: The reason why I will go to eat some food is I want to eat some food.')

Both sentences have an adjunct *because*-clause embedded. The key difference is the order of the embedded clause's appearance in the sentence hierarchically, and this ordering difference dictates which aspectual marker must be used, instead of free alternatives. We have learned previously that in many cases *-u* and -NA are freely alternate in a compatible environment of tense. This is a case of the same tense environment, but different aspects triggered by the ordering of the embedded clauses. Moreover, this adjunct clause contributing special semantic properties is not language-specific to Nepal Bhasa but is also present in other languages like English: Charnavel (2017) has a discussion about the *since*-clause in English. Therefore, the syntactic environment can possibly place restrictions on aspect choices.

In this subsection, I showed the cases of aspect morphology blocks in complex clauses: complement clauses and adjunct clauses. The former one restricts the embedding verb form which is outside of the embedded CP, whereas the latter case restricts the embedded verb form. I also reviewed some conjunct/disjunct agreements in the literature, and how my data contradicts the traditional morphological pattern of the agreement.

3.4 Conclusion and limitations

In this chapter, we looked at how temporality works in Nepal Bhasa, including aspect and tense information. I propose that *inchoative* and *perfective* are the better characterizations for capturing the aspectual patterns in this language than what the literature describes as *past disjunct* and *stative* suffixes respectively (Malla 1985).

I attempted to model the two Nepal Bhasa aspect in the timeline and examine the aspectual morphemes. Further, I suggested different strategies of forming aspects: inchoative and perfective, and they both can be shifted by adding timespan, resulting in different compatibility patterns with tenses. Nepal Bhasa inchoative is featured with immediately future reading under certain conditions, which perfective *-u* is not. I also

showed how aspectual morphological restrictions work in different structures and illustrated their different interpretations of conjunct/disjunct marking in complex sentences. I examined if conjunct/disjunct occurs in complex sentences with complex morphemes in declarative sentences. I showed that the restrictions only exist in complex structures, which behave differently from mono-clausal sentences.

Chapter 4

Syntactic strategies for Nepal Bhasa complementation

In Chapter 2 and Chapter 3, we learned the syntactic and aspectual basics of Nepal Bhasa to prepare us for understanding complementation in this language. Now, we are set to explore complementation strategies in Nepal Bhasa and the linguistic properties that each of them may have [1].

Recall the complementation examples from Chapter 1, as repeated in (127). There are three observed complementizers (C): *dhakā*, *dhayā*, and *ki*. Each of these C heads are able to embed a finite CP, as the presence of aspectual marking and standard case patterns inside the dependent clause indicates.

(127) a. Sitā-na [CP Rām-na oṃ na-la **dhakā**] dhā-u.
Sita-ERG Ram-ERG mango eat-INCH DHAKĀ say-PFV

b. Sitā-na [CP Rām-na oṃ na-la **dhayā**] dhā-u
Sita-ERG Ram-ERG mango eat-INCH DHAYĀ say-PFV

c. Sitā-na dhā-u [CP **ki** Rām-na oṃ na-la]
Sita-ERG say-PFV KI Ram-ERG mango eat-INCH
'Sita said that Ram ate mangos.'

(128) a. Preverbal complementizer *dhakā*:
SUBJ [CP ... ... ***dhakā***] V

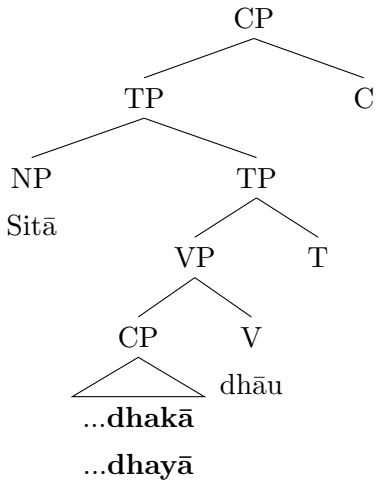
[1] In this dissertation I only focus on the finite verbal CPs, and leave other kinds such as non-finite or NP complement for future discussion.

b. Preverbal complementizer *dhayā*:
SUBJ [CP ... ... ***dhayā***] V

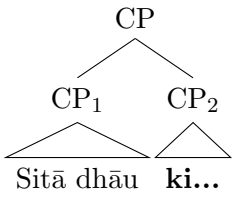
c. Post-verbal complementizer *ki*:
SUBJ V [CP ***ki*** ... ...]

I will discuss Nepal Bhasa complementation strategies in this chapter, reviewing and building on my previous work in Zhang and Chacón (2018) and Zhang (2018). I will demonstrate that post-verbal *ki*-clauses use a distinct complementation strategy from head-final CPs: as the trees in (178) and (179) indicate, I agree that *ki*-clauses adjoin as paratactic adjuncts, while *dhakā*/*dhayā* CPs are true complements to matrix verbs. In the following sections, we will see a variety of evidence that supports this conclusion.

(129) *Nepal Bhasa head-final CP*



(130) *Nepal Bhasa head-initial CP*



4.1 Head-final clauses in Nepal Bhasa

As the example in (127) shows, Nepal Bhasa has head-final complementizers, which follows the headedness in the matrix clause, as we learned from Chapter 2. There are two head-final complementizers, *dhakā* and *dhayā* in Nepal Bhasa. Hargreaves (1991) suggests that *dhakā* is the causative version of the root of the word *dha-* (say). According to Malla (1985), *yā* is genitive marker and nominalizer, so *dhayā* could be the nominalized formed of the 'say' verbal stem.

(131) a. Sitā-na [CP Rām-na oṃ na-la **(dhakā)**] dhā-u.
Sita-ERG Ram-ERG mango eat-INCH C say-PFV

b. Sitā-na [CP Rām-na oṃ na-la **(dhayā)**] dhā-u.
Sita-ERG Ram-ERG mango eat-INCH C say-PFV
'Sita said that Ram ate mangos.'

(132) a. [CP Rām-na oṃ na-la **(dhakā)**] Sitā-na dhā-u.
Ram-ERG mango eat-INCH C Sita-ERG say-PFV

b. [CP Rām-na oṃ na-la **(dhayā)**] Sitā-na dhā-u.
Ram-ERG mango eat-INCH C Sita-ERG say-PFV
'Sita said that Ram ate mangos.'

(133) a. ? Sitā-na dhā-u [CP Rām-na oṃ na-la **dhakā].**
Sita-ERG say-PFV Ram-ERG mango eat-INCH C

b. ? Sitā-na dhā-u [CP Rām-na oṃ na-la **dhayā].**
Sita-ERG say-PFV Ram-ERG mango eat-INCH C
Intended: 'Sita said that Ram ate mangos.'

As the examples from (131) to (133) show, the two Cs seem to be identical in terms of CP positioning. They are optional when the CPs are located in a grammatical position in the matrix clause. Preverbal and sentence initial are the acceptable positions for these CPs; post-verbal head-final CPs are found marginal by some of my language consultants [2]. My consultants use *dhakā* more frequently than *dhayā*, although they are both judged as acceptable in the environments I discuss. I will therefore treat both head-final Cs as interchangeable in single embedding environment, and will not provide data with both versions.

Wh-scope restrictions are often good indicators of what properties a CP may have. Scoping issues are often addressed by looking at the surface syntax and the semantic representations, from general theories on overt syntax and Logical Form (LF) (May 1978, Huang 1982, 1998, Fox 2003, Pesetsky 2000). However, others have argued that *wh*-in-situ can take semantic scope higher than canonical position either through "unselective binding" Pesetsky (1987), or composition (Hamblin 1973, Kotek 2017). In this section, I first discuss the preverbal, head-final Nepal Bhasa CPs and their properties by reviewing

[2]Complementizer dropping is more acceptable when the matrix clause predicate is *dāu* 'say', than other predicates. *dhayā* also tends to occur in the direct quoted sentences with *dāu* 'say' as the embedding predicate

my work on *wh*-scope from (Zhang and Chacón 2018) and (Zhang 2018), including the examinations of island effects, intervention effects in *wh-in-situ*, and sluicing-like constructions (SLC).

4.1.1 Island and intervention effects in CPs (Zhang and Chacón 2018)

We learned in Chapter 2 that Nepal Bhasa is *wh*-in-situ. It can take sentential scope *in-situ* in a mono-clausal sentence, as shown in (134). The in-situ *wh* can also take the sentential scope when it is embedded in pre-verbal CPs, as in (135).

(134) Rām-na **chu** na-la?
Ram-ERG what eat-INCH
'What did Ram eat?'

(135) Sitā-na [CP **su-na** aṃ na-u dhayā] si-u?
Sita-ERG who-ERG mango eat-PFV that know-PFV
'Who did Sita know ate the mango?'

One scope taking strategy for *-in-situ wh* is covert movement. The covert movement (CM) analyses predict that *wh-in-situ* should exhibit the same properties as overt movement, such as island sensitivity. For example, island sensitivity is observed in *wh-in-situ* configurations in Mandarin Chinese. As shown in (136), *in-situ wh*-operators display sensitivity to the Complex NP Constraint (Huang 1982, Bayer 2006, Cheng 2009). The Chinese example shows that the adjunct *wh*-phrase *weishenme* ('why') cannot covertly move out the DP, and thus fails to take sentential scope.

(136) * Qiaofeng xihuang [DP [CP Botong **weishenme** xie de] shu]
Qiaofeng like Botong **why** written DE book
'For what reason x, Qiaofeng likes the book that Botong wrote for x?'
(Mandarin, Huang 1982)

This approach, applied to Nepal Bhasa suggests that the *wh*-operator stays *in-situ* on the surface syntax of (134), as in (137), but moves to the Spec,CP position at LF, as in (138).

(137) Surface syntax: [CP [TP Ram **what** ate]

(138) LF: [CP **what**$_{\text{i}}$ [TP Ram t_{i} ate]]

However, in Zhang and Chacón (2018), we show that *wh-in-situ* is not sensitive to traditional island constraints (e.g., relative clause islands, complex NP islands, comparative clauses, *etc.*) in Nepal Bhasa. For example, the sentence in (139) has a relative clause, but the *wh*-operator takes wide scope regardless. Thus, no island effects are observed in Nepal Bhasa, which is unexpected on a CM analysis.

(139) Rām-na [RC **su-na** dā ma] guru nāplā-u?
Ram-ERG who-ERG hit CL teacher meet-PFV
'Which person x, Ram met the teacher y that x hit y?'

By contrast, when *wh-in-situ* is embedded in a relative clause in a verbal-argument CP, island sensitivity will appear, as shown in (140).

(140) * Ākās-āṃ [CP Rām-na [RC **su-na** dā ma] guru nāplā-u] dhā-u?
Akash-ERG Ram-ERG who-ERG hit CL teacher meet-PFV say-PFV
'Who is the person x, such that Akash said that Ram met the teacher x hit?'

What can account for these unexpected island (in)sensitives in Nepal Bhasa? Focus Alternatives (FA) composition approach suggests that the interpretation of *wh*-phrases does not involve any movement. On this account, *wh*-phrases are focus elements, and are interpreted by computing the semantic focus alternatives of the sentence (Beck 2006, Hamblin 1973, Karttunen 1977). *Wh*-phrases function to"shift" the semantic value into a focus tier until recombining with the Q head in the root clause at LF. Intervention effects are the unacceptability that arises when an *in-situ wh*-operator appears within the scope of a focus-sensitive operator, demonstrated in the sentences in Hindi and Korean in (141) and (142). The ungrammaticality is caused by the *in-situ wh*-operator appearing within the scope of the intervener, 'only'.

(141) * John-*hi* **kyaa** khariide-gaa?
John-only what buy-FUT
'What will only John buy?' (Hindi, Malhotra 2009)

(142) * Minsu-*man* **nuku**-lûl po-ass-ni?
Minsu-only who-ACC see-PST-Q
'Who did only Minsu see?' (Korean, Beck 2006)

In contrast to the Hindi and Korean, Nepal Bhasa matrix clauses do not show intervention effects for *in-situ wh*-operators, as shown in 143. Zhang and Chacón (2018)

suggest that Nepal Bhasa *wh*-operators may take sentential scope over a focus-sensitive operator in matrix clauses.

(143) Rām-na-*caka* **chu** na-u?
Ram-ERG-only what eat-PFV
'What did only Ram eat?'

However, unlike matrix clauses (or non-argument embedded clauses), we do observe intervention effects in dependent CPs. In (144), the sentential scope interpretation of *chu* 'what' is blocked. Instead, *chu* 'what' must be interpreted with embedded low scope. Similar findings are demonstrated in (145) for the adjunct *wh*-operator *chæ* 'why'. We attribute this obligatory low-scope induced by the addition of the focus operator *caka* 'only' to an intervention effect.

(144) Sitā-ṃ [CP Rām-a-*caka* **chu** na-u (dhakā)] dhā-u
Sita-ERG Ram-ERG-only what eat-PFV that say-PFV
'Sita said what only Ram ate.'
* 'What did Sita say that only Ram ate?'

(145) Sitā-ṃ [CP Rām-a-*caka* **chæ** aṃ na-u (dhakā)] dhā-u
Sita-ERG Ram-ERG-only why mango eat-PFV that say-PFV
'Sita said why only Ram ate mango.'
* 'Why did Sita say that only Ram ate mango?'

Overt movement has been independently shown to ameliorate intervention effects Beck (2006). We see this effect in Nepal Bhasa in (146) for argument-*wh* and in (147) for adjunct-*wh*, by overtly scrambling *wh*-phrases above the focus-operator, they can take matrix scope. This amelioration further supports the idea that FA scoping strategy is used in Nepal Bhasa complementation.[3]

(146) Sitā-ṃ [CP **chu** Rām-a-*caka* ___ na-u (dhakā)] dhā-u
Sita-ERG what Ram-ERG-only eat-PFV that say-PFV
#'Sita said what only Ram ate.'
'What did Sita say that only Ram ate?'

[3]Overtly scrambling the *wh*-operator makes embedded low-scope interpretations much more difficult to access. I do not account for this fact for now. However, this may be related to the fact that, in general, scrambling *wh*-operators strongly prefer sentential scope or the low-embedded scope strongly prefer it in the canonical order.

(147) Sitā-ṃ [CP **chæ** Rām-a-*caka* ___ aṃ na-u (dhakā)] dhā-u
Sita-ERG why Ram-ERG-only mango eat-PFV that say-PFV
#'Sita said why only Ram ate mango.'
'Why did Sita say that only Ram ate mango?'

Generally, the CM and FA approaches are understood as alternative analyses for analyzing the scope of *in-situ wh*-operators. However, Nepal Bhasa data shows a different pattern: both island effects and IE occur in pre-verbal CP clauses, whereas neither occurs in matrix sentences.

Furthermore, head-final dependent CPs can be fronted, as in as in (148). The *wh*-operator in this case can take sentential scope. Fronted CPs still show IE, however, as in (149). As with the *in-situ* CP moving the *wh*-phrase over the intervener ameliorates IE, as in (150).

(148) [CP Rām-a chu na-u (dhakā)] Sitā-ṃ dhā-u
Ram-ERG what eat-PFV that Sita-ERG say-PFV
'Sita said what Ram ate.'
'What did Sita say that Ram ate?'

(149) [CP Rām-a-*caka* **chu** na-u (dhakā)] Sitā-ṃ dhā-u
Ram-ERG-only what eat-PFV that Sita-ERG say-PFV
'Sita said what only Ram ate.'
*'What did Sita say that only Ram ate?'

(150) [CP **Chu** Rām-a-*caka* ___ na-u (dhakā)] Sitā-ṃ dhā-u
what Ram-ERG-only eat-PFV that Sita-ERG say-PFV
'Sita said what only Ram ate.'
'What did Sita say that only Ram ate?'

The same result holds for adjunct *wh*-operators, shown in (151). An intervener forces the *wh*-operator *chæ* 'why' to take embedded scope. Fronting the phrase over the intervener permits the sentential scope in (152) .

(151) [CP Rām-a-*caka* **chæ** oṃ na-u (dhakā)] Sitā-ṃ dhā-u
Ram-ERG-only why mango eat-PFV that Sita-ERG say-PFV
'Sita said why only Ram ate mango.'
*'Why did Sita say that only Ram ate mango?'

(152) [CP **Chæ** Rām-a-*caka* ___ oṃ na-u (dhakā)] Sitā-ṃ dhā-u
why Ram-ERG-only mango eat-PFV that Sita-ERG say-PFV
#'Sita said why Ram ate mango.'
'Why did Sita say that Ram ate mango?'

The operation of moving the entire CP clause followed by *wh*-movement is considered as pied-piping structure (Cable 2012) that have been found in other languages. But whether it is a genuine scope-taking strategy in Nepal Bhasa cannot be concluded by this special case with an intervener involved. So far, the data has suggested that Nepal Bhasa employs both FA and CM strategies for *in-situ wh*-operators in CPs, as summarized in Table 4.1 .

Clauses	Structure	Result	Conclusion
Matrix clauses	[M-CP ... *wh* ...]]	[-island effects] [-intervention effects]	FA CM
Dependent clauses	[M-CP V [V-CP ...[Island-CP ... *wh* ...]]]	[+island effects] [+intervention effects]	CM FA

Table 4.1: The (non-)existence of the two effects in Nepal Bhasa CPs

We consistently get IE and island effects for *wh*-operators pre-verbal CPs and fronted CPs, which is unlike matrix clauses. Next, I discuss the sluicing-like construction (SLC) to show another property of genuine sluicing which account on the basis of assuming Nepal Bhasa pre-verbal CPs are true complement (Zhang 2018).

4.1.2 SLC in Nepal Bhasa (Zhang 2018)

Nepal Bhasa matrix sentences do not have flexible word order in general, as discussed in Chapter 2. *Wh*-phrases also cannot scramble in the matrix clauses, as (153) and (154) show. However, we have seen that dependent CPs allow scrambling in ameliorating IE.

(153) a. * **Chu** Rām-na ___ na-la
what Ram-ERG eat-INCH

b. * Rām-na ___ na-la **chu**
Ram-ERG eat-INCH what
Intended: 'What did Ram eat?'

(154) a. * **su** wa ___ kha:?
who 3.SG COP

b. * wa ____ kha: **su**?
3.SG COP who
Intended: Who is that?

I suggest that SLCs are scrambling sluicing in Nepal Bhasa (Zhang 2018). In Nepal Bhasa SLCs like (155), a *wh*-word from the embedded clause survives while the rest of the clause is elided.

(155) Sitā-na su-ita dā-u, tala [CP **su-ita**] jīṃ ma-syu.
Sita-ERG someone-PFV hit-PST, but who-DAT 1SG.ERG NEG-know.PFV
'Sita hit someone, but I don't know whom.'

In this account, the example in (155) is derived from scrambling sluicing, with two steps, *wh*-scrambling into the left periphery of the CP clause, followed by TP ellipsis, as the proposed analysis shown in (156).

(156) Sitā-na su-ita dā-u, tala [CP su-ita ~~[TP Sitā-na ____ dā-u]~~]
Sita-ERG someone-DAT hit-PFV, but who-DAT Sita-ERG hit-PFV
jīṃ ma-syu.
1SG.ERG NEG-know

(Zhang 2018)

Case connectivity has been observed as a property of genuine sluicing (Ross 1969, Merchant et al. 2001): the remnant *wh*-phrase carries the same case marking as in the corresponding non-elided *wh*-question. Nepal Bhasa exhibits full case connectivity: Dative Case, Ergative Case, and Absolutive Case, from (157) to (159).

(157) Sitā-na **su-ita** dā-u, tala [CP **su-ita**$_i$/***su**/***su-na** (Sitā-na t_i
Sita-ERG someone-DAT hit-PFV, but who-DAT/*who/*who-ERG Sita-ERG
dā-u)] jīṃ ma-syu.
hit-PFV 1SG.ERG NEG-know
'Sita hit someone but I don't know who (Sita hit).'

(158) **Su-nā** Sitā-ta yek-i, tara [CP **su-nā/*su/*su-ita** (Sitā-ta
Someone-ERG Sita-DAT like-NON-PST, but who-ERG/*who/*DAT Sita-DAT
yek-i) jīṃ ma-syu].
like-NON-PST 1SG.ERG NEG-know
'Someone likes Sita, but I don't know who (likes Sita).'

(159) **Su** pasa-le wa-na, tara [CP **su/*-ita/*-na** (pasa-le
Someone.ABS store-LOC go-PST, but who-ABS/*DAT/*ERG (store-LOC
wa-na) jīṃ ma-syu].
go.PFV) 1SG.ERG NEG-know
'Someone went to the store, but I don't know who (went to the store).'

Embedded *wh*-phrases must remain in the embedded CP, otherwise, are ungrammatical, in (160) and (161).

(160) * **Su-ita** Rām-na [CP [TP Sitā-na ___ dā-u]] dhā-u?
who-DAT Ram-ERG Sita-ERG hit-PFV say-PFV
Intended: 'Who did Ram say that Sita hit?'
(*Wh-phrase cannot scramble across CP to the matrix clause*)

(161) * Waṃ [CP **Su-ita** Rām-na [CP [TP Sitā-na ___ dā-u]] dhā-u] sy-u?
3.SG who-DAT Ram-ERG Sita-ERG hit-PRF say-PFV know-pfv
Intended: 'Who did he know that Ram said Sita hit?'
(*Wh-phrase cannot scramble across CP to another embedded CP*)

The above data suggest that there is a constraint against long-distance scrambling in Nepal Bhasa, where the *wh*-phrase overtly moves from an embedded CP clause to another CP. However, this is licensed by SLC, as in (162) and the derivation in (163).

(162) [CP [TP Rām-na [CP [TP Sitā-na **su-ita** dā-u] dhā-u], tala [CP
Ram-ERG Sita-ERG someone-DAT hit-PFV say-PFV, but
su-ita] jīṃ ma-syu.
who-DAT 1SG.ERG NEG-know
Ram said that Sita hit someone, but I don't know whom.
Wh-phrase in double-embedded CP

(163) [CP [TP Rām-na [CP [TP Sitā-na **su-ita** dā-u] dhā-u], tala [CP
Ram-ERG Sita-ERG someone-DAT hit-PFV say-PFV, but
su-ita ~~[TP Rām-na [CP [TP Sitā-na ___ dā-u] dhā-u]~~ jīṃ
who-DAT Ram-ERG Sita-ERG ___ hit-PFV say-PFV 1SG.ERG
ma-syu.
NEG-know
Ram said that Sita hit someone, but I don't know who ~~(Ram said Sita hit)~~.
(*Nepal Bhasa SLC licensing long-distance scrambling*)

Nepal Bhasa data suggests that SLC ameliorates the ungrammaticality of overt scrambling across clausal boundary. This analogizes to island repairing by sluicing cross-linguistically (Ross 1969). The ability for SLCs to license long-distance *wh*-scrambling is strong evidence that the elided clauses, therefore their head-final antecedents, are truly subordinated.

4.1.3 Summary

In this section, I discussed the properties of head-final dependent CPs in the pre-verbal and fronted positions. We saw strong evidence of connectivity effects from *wh*-scope taking strategies and scrambling sluicing. Taken together, these effects suggest that head-final CPs are true verbal complements in this language. In the next section,I show that the other type of complement, head-initial *ki*-clause CPs, have none of the patterns we saw in the head-finial CPs.

4.2 Post-verbal head-initial Nepal Bhasa *ki*-clause

Nepal Bhasa is SOV, and complement CPs usually appear pre-verbally, as we saw in the last section. However, it has post-verbal CPs that are headed by *ki*, with the post-verbal *ki*-clauses shown in (164a).

(164) a. Sitā-na dhā-u [CP ki Rām-na oṃ na-la].
Sita-ERG say-PFV C Ram-ERG mango eat-INCH

b. * Sitā-na [CP ki Rām-na oṃ na-la] dhā-u
Sita-ERG C Ram-ERG mango eat-INCH say-PFV

c. * [CP ki Rām-na oṃ na-la] Sitā-na dhā-u
C Ram-ERG mango eat-INCH Sita-ERG say-PFV
Intended: 'Sita said that Ram ate mangos.'

This complementizer may be borrowed from Persian. Many other languages, including Turkish, and North Azeri also have *ki*-clauses (Kesici 2013, Halpert and Griffith 2018).

Two differences between *ki*-clauses and head-final CPs are immediately apparent. First, most of my consultants do not prefer complementizer dropping for *ki*-clauses, in

contrast to the other head-final CPs. In addition, *ki*-clauses cannot be preverbal (164b) or be fronted (164c), a pattern which is also seen in Turkish *ki*-clauses, even though preverbal is the default position for objects in Nepal Bhasa.

These immediate differences suggest that the syntactic properties of post-verbal *ki*-clauses could be different from pre-verbal head-final ones. Some hypotheses we can consider:

(165) Possible syntactic structures for *ki*-clauses:

a. In-*in-situ* complement: *ki*-clauses may be generated low in the post-verbal complement CP. Matrix verbs will be head-initial and take complement CPs on the right side.

b. Extraposition: *ki*-clauses may be generated low but end up as post-verbal under some sort of movement.

c. Parataxis: *ki*-clauses are originally generated high, as parataxis.

With Hypothesis (165a), we should expect *ki*-clauses behave as argument CPs both syntactically and semantically. Hypothesis 165b will predict that matrix verbs will be head-final and take *ki*-complement CPs on the left side. These two hypotheses are similar to each other, with the same proposal that *ki*-clauses are complements. The main difference between the two is how *wh*-phrases get interpreted, either in-situ or moving (if so, possibly via pied-piping) to take scope. Hypothesis (165c), on the other hand, will predict that paratactic *ki*-clauses will behave more like adjunct clauses instead of argument clauses.

Kesici (2013) suggested that Turkish *ki*-clauses are paratactic by examining their scope patterns, anaphor binding, NPI-licensing, semantic pragmatic compatibility of other elements. Halpert and Griffith (2018) examined North Azeri data and suggested that *ki*-clauses in that language are genuine complements. I follow and reproduce some of the diagnostics from (Kesici 2013) and (Halpert and Griffith 2018) to show that Nepal Bhasa *ki*-clauses are paratactic.

4.2.1 Diagnostics on testing Nepal Bhasa *ki*-clauses

The data we have so far finds Hypotheses (165a) and (165b) are less likely to be the case for Nepal Bhasa, but Hypothesis (165c) is promising. First of all, if Hypotheses

(165a) and (165b) are right that *ki* will be the first head-initial phrase that we find in the verbal extended projection in this language, and it will be typologically hard to fit it into the consistent head-finalness we found so far.

Additionally, if the *ki*-clause is a genuine complement, we should expect the same flexibility as for head-final Cs, which may both appear in preverbal and sentence-initial positions. But (164) already shows that *ki*-clauses do not allow multiple positions, in contrast to the head-final Cs. In this section, we will see additional evidence that *ki*-clauses will favor Hypothesis (165c), as being paratactic.

Can *ki*-clauses be questioned?

Testing if *ki*-clauses are restricted in the question environment will help us understand the *wh*-scope and polar question patterns in this language. Recall that *wh*-operators in head-final CPs can take either low (declarative) or high (sentential *wh*-question) scope. In contrast, *wh*-operators in post-verbal *ki*-clauses cannot take sentential scope, as shown in (166).

(166) Sitā-na dhā-u [CP ki Rām-na chu na-la].
Sita-ERG say-PFV C Ram-ERG what eat-INCH
'Sita said what Ram ate.'
* 'What did Sita say that Ram eat?'

This phenomenon has two possible interpretations. First, if *ki*-clauses are generated high as adjunct islands, sentential scope will be blocked according to the island constraint. The second is if we assume that *wh*-phrases in *ki*-clauses need to be overtly moved to take sentential scope, an in-*situ* embedded *wh*-phrase will not be able take sentential scope without moving the entire clause. But such movement cannot happen, as I previously showed that *ki*-clauses in (164b) and (164c) are not allowed to move pre-verbally or be fronted.

A *wh*-phrase in the matrix clause is not affected by presence of a *ki*-clause, as in (167). *Ki* also blocks embedded *wh*-phrases from taking sentential scope in (168).

(167) **Su-na** dhā-u [CP ki Rām-na aṃ na-la].
Who-ERG say-PFV C Ram-ERG mango eat-INCH
'Someone said that Ram ate the mangos.'
'Who said that Ram ate mangos?'

(168) Sitā-na dhā-u [CP ki Rām-na **chu** na-la].
Sita-ERG say-PFV C Ram-ERG what eat-INCH
'Sita said what Ram ate.'
* 'What did Sita say that Ram eat?'

However, when a sentence-final Q particle appears, *ki*-clause block it from taking matrix scope, and the Q must be interpreted in the embedded clause, as in (169).

(169) a. Sitā-na dhā-u ki Rām-na aṃ na-la **lā**
Sita-ERG say-PFV C Ram-ERG mango eat-INCH Q
'Sita said if Ram ate mangos.'
*'Did Sita say that Ram ate mangos?'

b. Sitā-na dhā-u [CP ki Rām-na aṃ na-la **lā**]

c. * Sitā-na dhā-u [CP ki Rām-na aṃ na-la] **lā**

The facts in (168) and (169) strongly suggest that the *ki*-clause is in a high adjunct position: if the *ki*-clause attaches higher than matrix C, then there would be no way for Q in the matrix clause to follow it in (169), or for a *wh*-phrase inside to take scope in the matrix clause in (168).

ki-clause interaction with focus adverbs

The data so far suggests that *ki*-clauses may be paratactic matrix. Kesici (2013) suggests that *ki*-clauses are incompatible with focus adverbs in Turkish. Unlike Turkish, Nepal Bhasa does not seem to show this property, as (170) is grammatical. The focus adverb *caka* 'only' modifies the matrix subject:

(170) Sitā-na **caka** dhā-u [CP ki Rām-na oṃ na-la].
Sita-ERG only say-PFV C Ram-ERG mango eat-INCH
'Only Sita said that Ram ate mangos.'

However, note that when we attempt to position the focus adverb to force it to modify the matrix verb, its meaning changes to 'once'.

(171) Sitā-na dhā-u **caka** [CP ki Rām-na oṃ na-la].
Sita-ERG say-PFV only C Ram-ERG mango eat-INCH
'Sita *once* said that Ram ate mangos.'
*'Sita *only* said that Ram ate mangos.'

This suggests that the focus adverb cannot in fact associate with *ki*-clauses. Therefore, we see no evidence of a c-command relationship with the matrix clause.

Matrix quantifiers binding into *ki*-clauses

The sentence in (172) shows that the quantifier 'everyone' in the matrix clause cannot bind the pronoun in a *ki*-clause. Again, this shows that no c-command relationship or binding exits in Nepal Bhasa *ki*-clause.

(172) dakkaisyaeṃ$_i$ dhā-u [CP ki waṃ*$_{i/j}$ aṃ na-la]
everyone.ERG say-PRF KI 3.SG.ERG mango eat-INCH
*'everyone$_i$ said that he$_i$ ate a mango.'
'everyone$_i$ said that he$_j$ ate a mango.'

ki-clauses as adjunct

Although *ki*-clauses might appear to be complement clauses based the meanings we have seen so far, *ki* can also appear in more obvious adjunct clauses.(173a) is similar to English *not-only... but also* construction. In this case, the *ki*-clause does not appear to be a complement of any matrix verbs. More interestingly, as (173b) shows, the head final complementizer *dhakā* is not allowed, which again implies that this kind of clause may not be a complement clause, but rather that all *ki*-clauses may be parataxis.

(173) a. Tha khānā swe bāle caka baṃla ma-khu (**ki**) ne bāle na sā
this food see while only nice NEG-COP KI eat while also tasty.

b. * Tha khānā swe bāle caka baṃla ma-khu ne bāle na sā **dhakā**
this food see while only nice NEG-COP eat while also tasty C.
'This food does not just have a good shape, but also has good taste.'

Another usage of Nepal Bhasa *ki* is in coordination constructions as shown in (174), with a meaning of *or*. In this case, it is not certain if *ki* is the head of the coordination. While (174) appears to coordinate clauses with inflected verbs, speakers strongly disprefer *ki* in smaller coordinations, as (175) shows (cf. (Malla 1985) Page 95).

(174) Rām [(ki) thaṃuṃ wa-i] [ki kane wa-i]
Ram KI today come-IPFV KI tomorrow come-IPFV
'Ram will come either today or come tomorrow.'

(175) ? Rām [ki thaṃuṃ] [ki kane] wa-i
Ram KI today KI tomorrow come-IPFV
Intended: 'Ram will come either today or tomorrow.'

As from these clearly non-subordinated adverbial clauses headed by *ki*, it is consistent with the findings that *ki*-clauses are not genuine complement CPs.

4.2.2 Summary

The patterns discussed in this section are summarized in Table 4.2.

Tests	Results
wh-embedded in *ki*-clause	embedded scope only
Matrix *wh*-question	sentential scope
Sentence final-Q	embedded scope only
Compatibility with matrix focus adverbs	not compatible
Anaphor binding	no co-indexing

Table 4.2: Syntactic and semantic tests on *ki*-clauses

Bengali These results suggest that *ki*-clauses are paratactic instead of true argument CPs as the head-final *dhakā* and *dhayā* clauses in Nepal Bhasa.

Some South Asian languages like Bengali, which also have both head-initial and head-final CPs show properties of semantic selection (Bayer 2001), as shown by the examples in (176) and (177).

(176) [Ram koljata-y jacche **bole**] {janlam/ bhablam/ Sunlam/
Ram Calcutta-LOC goes C knew-1.SG thought-1.SG heard-1.SG
*dekhlam/ *Osombhob}
saw-1.SG unlikely
'That Ram is going to Calcutta {I knew/ I thought/ I heard/ *I saw/ *is unlikely}.'

(177) {janlam/ bhablam/ Sunlam/ dekhlam/ Osombhob} [**je** Ram
knew-1.SG thought-1.SG heard-1.SG saw-1.SG unlikely C Ram
koljata-y jacche]
Calcutta-LOC goes
'That Ram is going to Calcutta {I knew/ I thought/ I heard/ I saw/ is unlikely}'
(*Bengali*, (Bayer 2001))

In these cases, the matrix verbs *dekhlam* ('saw') and *Osombhob* ('is unlikely') restrict the head-final CPs and only select the head-initial CPs in Bengali, even though both C heads can be interchangeable with other matrix verbs. I have not observed matrix verb semantic restriction in Nepal Bhasa CPs in general. The sentences in (173) could be a marginal case showing that head-final CPs are restricted in a semantic way. More investigation is needed to examine whether matrix verb classes restrict the headedness of dependent CPs in Nepal Bhasa.

4.3 Conclusion

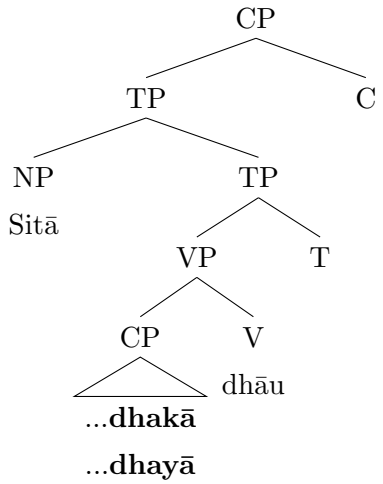
In this chapter, I discussed the Nepal Bhasa head-initial and head-final complementizers, and the CP positioning and the associated *wh*-scope patterns, as the summary table shown below:

	Pre-verbal S [CP *var...*] V		Post-verbal S V [CP*...var...*]		Sentence-initial [CP*...var...*]S V	
C-heads	Acceptability	Scope	Acceptability	Scope	Acceptability	Scope
-dhakā	✓	H&L	?	L	✓	H&L
-dhayā	✓	H&L	?	L	✓	H&L
ki-	*	-	✓	L	*	-

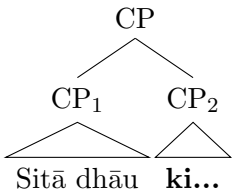
Table 4.3: The summary of the finite dependent CP positions which are headed by different C-heads, and their possible scopes (H: high scope, sentential scope; L: low scope, embedded scope, local scope)

The syntactic-semantic analyses in this chapter suggest that the head-final pre-verbal CPs are the genuine complements where the head-initial *ki*-clause CPs are parataxis. The *null* CP structure is likely to be derived from head-final C head drop. Nepal Bhasa matrix CPs and dependent CPs have a different *wh*-scope pattern. Both FA and CM strategies are used in complementation. Therefore, I propose the following structures for each clause type as the following:

(178) *Nepal Bhasa head-final CP: true complement*



(179) *Nepal Bhasa head-initial CP: parataxis*



These complementation strategies also seem to be compositional, resulting in 'nested' embedding structure. As shown in (180), the head-final *dhayā*-clause is embedded in a head-initial *ki*-clauses.

(180) Sitā-na syu [CP ki Ākāsaṃ [CP Rām-na bal thwa-la dhayā]
Sita-ERG know.PFV C Akas.ERG Ram-ERG ball kick-INCH C
dhā-u].
say-PFV
'Sita knew that Akas watched that Ram kicked the ball.'

I do not have enough data to generalize possible doubling-embedding patterns. Future directions on examining the variations will help better understand how complementation works in Nepal Bhasa. For example, parataxis cannot embed in other parataxis. This should rule out the possibility of double embedded *ki*-clauses. In the next chapter, I will explore some ways we can use computational tools to further investigate the properties of dependent clauses in Nepal Bhasa.

Chapter 5

A corpus-based approach assisting fieldwork in Nepal Bhasa

The syntactic clausal structure and the lexical semantics of the embedding verbs (Moulton 2009, Bresnan 1972) are two major focuses in the work on complementation in general. For complementation in Nepal Bhasa, ideally, we would want to have access to as much language data as possible that contains the CP embedding structures to work with. I mainly use the data collected from my fieldwork to hypothesize and test the linguistic properties in Nepal Bhasa complementation. The speed of data collection and the vocabulary range were some noticeable limitation in my fieldwork data, which may potentially limit the validity of making the theoretical claims. A structured corpus of Nepal Bhasa complementation sentences would be significant for improving the validity of the research results. However, like many other endangered languages, no such structured corpus exists currently for Nepal Bhasa, and building one is costly and time-consuming. In this chapter, I experiment with two ways of using corpora to help my fieldwork research in Nepal Bhasa complementation. The involved techniques in the experiments may assist in speeding up the process of linguistic fieldwork in general.

In the first experiment, I use an unstructured (raw) corpus data set of Nepal Bhasa from Open Super-large Crawled Aggregated coRpus (OSCAR) (Ortiz Suárez et al. 2019) to train a chunking model (or shallow parsing). Chunking is one kind of the

Named Entities Recognition (NER) models, for predicting Nepal Bhasa complementation clauses. I adopt the technique of transfer learning, with fine-tuning a pre-trained neural transformer model - multi-lingual language model (mBERT) - using Python NERDA (Kjeldgaard and Nielsen 2021) training library. Creating chunking models can be the first step of building the structured corpus for Nepal Bhasa. The data format in NER modeling has been highly standardized, which makes it easy to share the corpus data and the trained model for other researchers to use. Training chunking models also present a possible way of speeding up the annotation procedure for linguistic fieldwork documentation.

In the second experiment, I use the available structured corpora of non-Nepal Bhasa, non-endangered languages, Mandarin and Cantonese treebank data (McDonald et al. 2013), to a generalization which has previous observed in the Nepal Bhasa data from my fieldwork – the tendency of inchoative blocking in the embedding verbs. If Mandarin and Cantonese (Sino-tibetan) also show no embedding verb with the inchoative element, it will evidence a cross-linguistic generalization of complementation.

5.1 Shallow parsing of CPs in Nepal Bhasa

Annotated corpora provide naturalistic data and syntactic-semantic annotation labels for setting a foundation that benefits linguistic research and language documentation (Hovy and Lavid 2010, de Marneffe and Potts 2017). Building annotated corpora for endangered languages is particularly beneficial, as the linguistic insights are systematically shown in the data, which are directly reusable and can be improved by adjoined efforts over time. However, there is no annotated public corpus resource of Nepal Bhasa currently found available. A small Nepal Bhasa unstructured corpus comes available from the Open Super-large Crawled Aggregated coRpus (OSCAR)(Ortiz Suárez et al. 2019). With the joined efforts from my language consultants and me we annotated a small amount of complement clauses from the OSCAR data in this experiment. The annotated data is used in training an NLP shallowing parser to predict embedded clauses in Nepal Bhasa. The procedure may be used as a starting step of developing an annotated corpus in general for fieldworkers.

The procedure of training a model with the annotation of embedded CPs belongs

to the shallow parsing (also chunking) technique (Abney 1991). It is one of the Named Entities Recognition (NER) NLP classification tasks, which has been used for capturing linguistic featured labels. Some commonly used applications are in identifying lexical categories in NPs such as location names, brand names, organizations, times, etc. Shallow parsing requires less annotation effort by focusing on task specific labels, compared to the full POS or semantic NER tagging tasks. With the limited available resources, I focused on labeling the embedded clauses, the matrix verbs, and the embedded verbs for a small number of Nepal Bhasa sentences from the OSCAR data for the training purpose.

The accuracy of a trained model prediction determines the effectiveness of the model for practical use. Empirical studies have found that the performance of NLP models suffer from too little training data. Low-resource languages are never able to provide a large amount of training data, which could be one of the reasons of why few NLP tools and models are available for endangered languages. Until lately, NLP/Machine learning methods have been achieving significant advances in many language-related tasks, despite the development still remaining limited in the field of theoretical linguistics. With the help of advanced data crawling techniques, computing power, and complex modeling infrastructure, multiple-language models like mBERT and XLM (Lample and Conneau 2019) have been released. Studies (Meechan-Maddon and Nivre 2019, Lample and Conneau 2019, Ebrahimi et al. 2021, Liu et al. 2020) have shown that using the neural machine learning techniques - transfer learning and pre-train fine-tuning bigger models improves model performance with little to no annotation work. This may be an opportunity for low-resource languages with limited data sources to still be able to train better NLP annotation tools and enrich fieldwork research.

In the experiment, I train a shallow parser to predict complement clauses in Nepal Bhasa by fine-tuning the multiple-language Bidirectional Encoder Representations from Transformers (mBERT) (Devlin et al. 2018) with a small annotated Nepal Bhasa corpus based on the raw data from OSCAR (Ortiz Suárez et al. 2019).

5.1.1 Chunking tasks for embedded CPs

The classic chunking format is the IOB method, as it marks the inside, the outside, and the beginning of a target phrase/chunk. IOB refers to the related phrases as shown in

Table 5.1. I additionally label the verb levels (I-V, O-V, and B-V) for capturing the matrix verbs and the embedded verbs.

Chunking tags	**Meaning**
I_{cp}	inside of a CP
B_{cp}	beginning of a CP
O_{cp}	outside of a CP
I-V	embedded verb
O-V	matrix verb
B-V	embedded verb that begins a CP

Table 5.1: Chunking tagset

The Nepal Bhasa raw data, labeled as '*new*' in OSCAR (Ortiz Suárez et al. 2019) is about 6MB of the entire 20TB dataset. Again, a task of manually capturing all the verbal CPs can be still expensive even when the dataset is not as big as in high-resource languages. Therefore, I minimally labeled the partial data with the help from the native speakers in my fieldwork. Adapting the strategy of fine-tuning the pretrained mBERT model I use NERDA python library (Kjeldgaard and Nielsen 2021) to train and evaluate a chunking model for the Nepal Bhasa verbal CPs.

5.1.2 Annotation and data processing

The Nepal Bhasa raw data source is from OSCAR 2019. The dataset is in Devanagari script. Some pre-processing steps were taken as shown in Table 5.2 before the annotation [1], including removing non-Devanagari characters, aligning one sentence per line, removing sentences that had less three words in one line, resulting in a total of 16603 clean sentence left for CP extraction. There were 684 sentences found that contain *dhaka*, and 2660 sentences found that contain *ki*. The complementizer *dhaka* is a good morphological cue to detect complement sentences in the data, while the morpheme *ki* is ambiguous between being a complementizer or a phrasal conjunction ('and') in this language.

Out of the 3344 sentences (680+2660) that potentially contain CPs, 200 *dhaka*-sentences and 100 *ki*-sentences were randomly selected for the annotation task [2].

[1] data cleaning scripts are available in Github repo.

[2] The complementizer *dhaya* also exists in the corpus. No manual annotations made for it due to limited time. A observation of it only occurs when the matrix verb stem is *dha-* ('say') with quotations.

Nepal Bhasa (*newa*) in Devanagari script	**Sentences**
OSCAR-2019 original file	(5.7 MB)
# remove unreadable chars and empty lines	16694
# remove too short sentences	16603
# complementizer *'dhaka'* captured	684
# *'ki'* captured	2660
Annotated Data	
# total manually annotated	300
# total identified non-embedded	101

Table 5.2: Nepal Bhasa OSCAR corpus status

I worked with two language consultants in the annotation task. Both of them have participated in my fieldwork before. For minimizing the complexity of the annotation instruction for the consultants, one was only given the *dhaka*-sentences, and the other the *ki*-sentences to work on. The working timeline started with meeting them separately to practice labeling with at least one negative example and one positive example. Language consultants were given the task sentences and received the written instructions (See Appendix A), which were covered in the meeting. A meeting was scheduled once every 50 annotated sentences completed by one annotator. Annotation issues such as correcting the labels and clarifying the annotation instruction were discussed during the meeting. Language consultants used square brackets to mark the true embedded CPs, as shown in the first line in Figure 5.1.

[थन धाःसा १८७६य् लुइकूगु धकाः] च्वयातल

लुइकूगु, च्वयातल

Figure 5.1: Annotating verbal CPs in OSCAR Nepal Bhasa corpus

Accumulated work time for the *dhaka*-sentence annotation was about 5 hours including both the annotation and meetings, and 3 hours for the *ki*-sentence annotator. A faster annotating work speed was observed in the later sessions. Within the 300 sentences, 6 true embedded CPs were found in the 100 *ki*-sentences, and more than 190 true embedded CPs in the 200 *dhaka*-sentences.

5.1.3 Model Training

The manually-annotated sentences were converted to the CoNLL-2003 shared NER task format. Two ways of training were designed on the same data sets: in the first one, the model trains both CPs and verbs within one run, and in the second one, CPs and verbs are trained as separated single-task models. The data was divided into training, validating, and testing sets with a ratio of 7:2:1. The average training time is 3 to 4 minutes with GPU.

Training CPs and verbs in a model

The tag level statistic distribution of CPs and verbs is shown in Table 5.3. Label 'I_{cp}' is the most used tag in the annotation with 2133 tokens. Next is Label 'O_{cp}', 1618 tokens, then 'O-V' 315 tokens, 'B_{cp}' 207 tokens, 'I-V' 199 tokens, and 'B-V' only 1 token.

Tag Level Distribution	**Tokens**
#I_{cp}	2133
#O_{cp}	1618
#O-V	315
#B_{cp}	207
#I-V	199
#B-V	1

Table 5.3: Annotation level distribution of CPs and Verbs

The higher number of 'O-V' labels reflects that more complex verbal morphemes are in the matrix clauses than in the embedded CPs. The very low number of 'B-V' means that the embedded verbs rarely occur at the left edge of a CP.

Training CPs and verbs in separate models

A second experiment is designed with tagging the CP clauses and verbs separately in two single trainings, as the tag distributions show in Table 5.4. With a smaller number of labels, each label is used on more tokens in the two single tasks.

The models may potentially increase their performances as more tokens are used in each label. However, we do see that the distribution for VPs is imbalanced: 4334 tokens for outside VPs, only 63 tokens for inside VPs, and 76 tokens for the beginning of a VP.

Tag Level Distribution	**Tokens**
#I_{cp}	2332
#O_{cp}	1933
#B_{cp}	208
#I_{vp}	63
#O_{vp}	4334
#B_{vp}	76

Table 5.4: Annotation level distribution of CPs and verbs separated

This imbalanced distribution of labels on VPs may negatively affect the single task of training the VP chunking model.

5.1.4 Results and Discussion

The performance of the duo-task of training both CPs and verbs in the same run shown as the F1, Precision, and Recall scores summarized in Table 5.5.

Level	F1-score	Precision	Recall
B_{cp}	0.857143	0.800000	0.923077
B-V	0.000000	0.000000	0.000000
I_{cp}	0.848249	0.801471	0.900826
I-V	0.173913	0.285714	0.125000
O-V	0.384615	0.500000	0.312500
AVG-MICRO	0.766467	NaN	NaN
AVG-MACRO	0.452784	NaN	NaN

Table 5.5: Nepal Bhasa CP-verb chunking model performance

Precision quantifies the number of positive predictions that actually belong to the positive levels. Level 'B_{cp}' and 'I_{cp}' has the highest Precision at 80%. Verb levels 'O-V' and 'I-V' get a worse precision at 50% and lower. This may be due to the fact that the number of training labels for verbs are far less than the I-O-B levels. Recall quantifies the number of positive level predictions from all predicted positive examples. Level 'B_{cp}' is with the highest recall 92.3%, and Level 'I_{cp}' at 90%. Similar to their precision scores that both of the recall scores are higher than the verb levels, 'O-V' and 'I-V'. Both precision and recall scores of 'B-V' are 0s the worst among all levels. The F-scores, which are calculated based on the precision and recall, show a balanced

averaged performance of 85% of Level 'B_{cp}' and 'I_{cp}', and 17% and 38% for 'I-V' and 'O-V' respectively. The micro-average of 76% combines the contributions of all levels to compute the average metric, where as the macro-average of 45% computes the metric for each level independently. The confusion matrix of the model in Table 5.6 represents the true labels vertically and the actual taggings horizontally.

	I_{cp}	B_{cp}	O_{cp}	I-V	B-V	O-V
true-I_{cp}	109	1	7	3	0	1
true-B_{cp}	1	12	0	0	0	0
true-O_{cp}	11	2	58	1	0	4
true-I-V	13	0	1	2	0	0
true-B-V	0	0	0	0	0	0
true-O-V	2	0	8	1	0	5

Table 5.6: Nepal Bhasa CP-verb chunking confusion matrix

'I-V' are often predicted as 'I_{cp}' only, and O-V are also more likely to be predicted as 'O_{cp}'. Although they are wrongly predicted, the predicted tags are still in the right clausal categories.

The model is relatively accurate on predicting over the levels of 'I_{cp}' and 'O_{cp}' with high recall values more than 90%. The verb related tags ('I-V', 'B-V', and 'O-V') have lower performance. This may be due to the smaller number of training examples in the limited working data. The 0 performance of 'B-V' is expected due to rare occurrences of a verb found at the clausal beginning. For the same reason, the Macro-average and Micro-average are both *NaN* since B-V level is not found in some sets. The Macro-average F-score especially gets significantly dragged down because of the performance on Level 'B-V'.

Training the CPs and verbs as two separate models achieved a higher performance, as shown in Table 5.7 and Table 5.8.

Level	F1-score	Precision	Recall
I_{cp}	0.964286	0.944056	0.985401
B_{cp}	0.857143	0.800000	0.923077
AVG-MICRO	0.954545	NaN	NaN
AVG-MACRO	0.910714	NaN	NaN

Table 5.7: Nepal Bhasa CP-only chunking performance

Level	F1-score	Precision	Recall
I_{vp}	0.285714	0.333333	0.250000
B_{vp}	0.666667	0.666667	0.666667
AVG-MICRO	0.461538	NaN	NaN
AVG-MACRO	0.476190	NaN	NaN

Table 5.8: Nepal Bhasa verb-only chunking performance

Both 'I_{cp}' and 'B_{cp}' scores were increased in the second training, comparing the numbers in Table 5.7 to Table 5.5. See the output predictions in Appendix B .Even though the verb performance is slightly higher in the single task training, it still remained low in general due to lack of annotated verb tokens.

The hyper-parameter settings 11 epochs, 10 warmups, 7 batches proved to be the best among different trials. A systematic observation is that a larger batch size, 10 for example in this case, significantly lowers the accuracy, which Keskar et al. (2016) suggest in their study of deep learning structures.

5.1.5 Conclusion

The experiments training shallow parsers for Nepal Bhasa complement phrases has shown the potential use of NLP tools in assisting corpus annotation for fieldwork research in endangered languages in general. The models successfully achieve some high model performance with the very limited data source (less than 300 manually sentences, 2% of the entire OSCAR Nepal corpus). The procedure may be used as a starting step in developing more structured corpora for fieldworkers.

Some future next steps are possible to further improve the quality of the entire procedure. Study shows that annotator expertise has a strong influence on the annotation accuracy and speed (Baldridge and Palmer 2009). My language consultants' expertise has grown significantly throughout the experiment. Setting up agreement tests for annotators to review others' annotation work may be helpful to improve future accuracy, although the annotation time might be prolonged and more annotators would be needed.

One of the disadvantages of annotating CPs with shallow parsing is that it may not be able to detect nested embedding structures, i.e. a CP being embedded in another layer of CP. There were few data points like this, and I only label the outermost layer CP for those.

Furthermore, from a theoretical linguistic perspective, using the traditional 'IBO' annotation style for the head-final CP clauses in Nepal Bhasa can be unproductive, since the right boundary of the clause has a stronger cue than the left boundary. An annotation style of 'IEO' ('E': end of the CP clause) may be more appropriate for tagging head-final phrases.

Another possible way to improve the CP embedding general model performance is to have some minimum labeling for the *null*-complementation type. Since *null* CP does not have apparent morphological cues, a larger sentence pool is needed for annotators to identify true CPs initially. A more comprehensive CP chunking model can be trained to detect adjunct CPs such as relative clauses on the basis of verbal CP chunking model.

5.2 Cross-linguistic corpus-based approach

Corpus linguistic approaches are traditionally used for language teaching and learning (Bennett 2010). The method rapidly extends into many research fields that involve natural language research. One goal of applying a corpus-based research method is to find quantitative linguistic evidence, even cross-linguistically, to help propose and test theories. The current study is aimed at capturing linguistic patterns that are associated with complementation structures. This requires the corpora to have particular annotations for capturing the complementation clauses to begin with. The complexity of every corpus varies. The earlier lightly-annotated corpora contain linear sentences with invented part of speech (POS) tags (e.g., Brown Corpus (Francis and Kucera 1979)). Later, more heavily annotated treebanks came out with POS tagged sentences parsed into the hierarchical structures, such as Universal Dependency (UD) treebank (McDonald et al. 2013), Penn Treebank (PTB) (Marcus et al. 1993). Comparing to the linear non-annotated corpora, treebanks corpora often contain some featured annotations based on linguistic frameworks (e.g., Constituency-based, dependency-based, LFG, HPSG). These featured annotations are useful for linguistic research that attempts to capture grammatical patterns.

The fieldwork observation in Chapter 3 states that the aspectual markers are restricted in embedding verb forms, as the example shown in (181 and 182), and are not

restricted in the simple matrix clause sentence (183). If there were structured treebank Nepal Bhasa corpora available, one could verify the generalization over more data points by conducting a corpus search of the complement structures. Although a direct corpus-based analysis is unavailable currently, a linguistically motivated cross-linguistic corpus search may be an alternative to examine the same generalization.

(181) Ram-na [CP Sitā-na oṃ nala dhakā] tā**-u/*-la**
Ram-ERG Sita-ERG mango eat.INCH C hear-PFV/-INCH
'Ram heard that Sita ate the mango.'

(182) Ram-na swai **chwoṃ-u** /***chwoṃ-na** suthe sāāt bāje [CP Sita
Ram-ERG watch.IPFV remain-PRF remain-INCH when seven time Sita
aṃ nai **chwoṃ-u** /**chwoṃ-na**]
mango eat.IPFV remain-PRF /remain-INCH
'At 7 am, Ram was watching that Sita was eating a mango.'

(183) Rām-na oṃ na-i **chwoṃ-na/chwoṃ-u**
Ram-ERG mango eat-IPFV remain-INCH/remain-PRF
'Ram is eating a mango.'

We have learned from Chapter 3 that *na* marks the inchoative aspect and contributes a meaning that is compatible with the past tense, which otherwise is interchangeable with the other morpheme *-u*. Similar inchoative properties, such as immediate future reading, with the sentence-final LE in Mandarin were also judged and accepted by some Chinese native speakers. The shared linguistic behaviors of the languages motivates a cross-linguistic hypothesis that some aspectual properties of inchoative (e.g., telic) may be prohibited in embedding verbs of the complementation structure. Therefore, the tendency can be checked in any language corpus if some potential inchoative morphological properties are found. In the experiment, I tested it on Mandarin and Cantonese Universal Dependencies treebank corpora (McDonald et al. 2013), as the languages are typologically close to Nepal Bhasa, and exhibit possible inchoative behaviors.

5.2.1 Corpus search

The target structural pattern of inchoatives in complementation are illustrated in (184, 185, and 186) for Nepal Bhasa, Mandarin, and Cantonese respectively. If such a pattern is found in the corpora, it will indicate that the proposed generalization is false. Namely,

complementation structure does not restrict the inchoative aspect in the main clause. On the other hand, if the pattern is not found, it will imply that a certain aspect is restricted in complementation structure.

(184) [Subj Verb-INCH [dependent clause]] (Nepal Bhasa)

(185) [[Subj Verb [dependent clause] LE] (Mandarin)

(186) [[Subj Verb [dependent clause] LO] (Cantonese)

The Chinese Treebank Corpus in the Universal Dependencies (McDonald et al. 2013) has a total of 3997 treebank sentences, 1342 sentences that contain the CCOMP label (complement CP) as an indication of complementation clauses. The example in Figure 5.2 shows a sentence containing a CP. Cantonese UD treebank data contains 386 complement clauses in the sentences among 1004 total sentences. As for the structural pattern extraction tool, I used Tregex, one of the treebank community-provided tools.

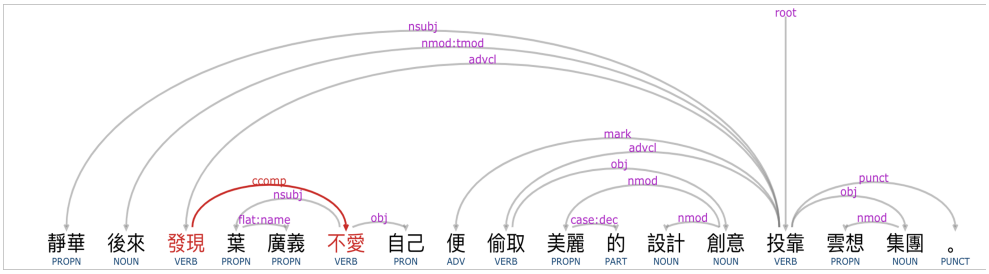


Figure 5.2: Chinese UD Treebank example of Mandarin CPs 'Jinghua later discovered [that YG does not love her], and then she stole ML's design idea and joined Yunxiang Group.'

Any occurrence of Mandarin *le* or Cantonese *lo* in the non-sentence final position, as the example shown in (5.3), is not considered as the inchoative element.

As the corpus research results are shown in Table 5.9, none of the sentences contain the pattern as in (184, 185, and 186). Not finding the target pattern in the corpora supports the hypothesized generalization that inchoative may be prohibited in matrix embedding verbs.

In this section I propose a corpus-based approach for assisting Nepal Bhasa complementation. The experiment relies on the observed shared features in Nepal Bhasa, Mandarin and Cantonese. By searching over the already-existing Mandarin and Cantonese

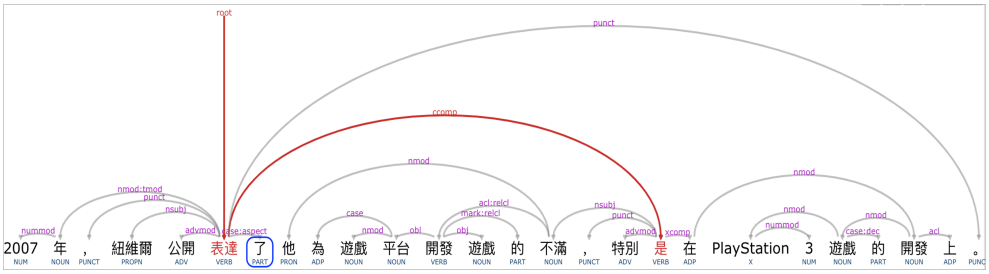


Figure 5.3: Chinese UD Treebank example of non-sentence final *le* (circled) in complementation main clauses: 'In 2007, N publicly expressed his complaining about game platform development, especially on the dev of Play Station3.'

Corpora	total sentences	complements	INCH matrix verbs
Mandarin	3997	1342	0
Cantonese	1004	1386	0

Table 5.9: Corpus Search on inchoative in embedding verbs of Mandarin and Cantonese

structured treebank corpora, we found the results further support a cross-linguistic generalization that the inchoative aspect may be prohibited in embedding verb in complementation. The method avoids necessarily building language corpus before doing any quantitative research, including testing hypotheses that are motivated from linguistic fieldwork. There remains possible debates on whether corpus search results can be directly associated with (un)grammaticality, and the possible human annotation errors that could lead to false positives/negative results.

Building a comprehensive Nepal Bhasa structured corpora will ultimately add stronger evidence to confirm the distributions of embedding verbs and dependent CPs. But the experiment showed that cross-linguistic corpus search is useful for seeking linguistic features when the target data is limited. The method offers the opportunity of helping linguistic research on low-resource, endangered languages by maximizing the use of available data resources from other languages.

I am aware of some of the limitations of corpus-based approaches. For example, the heavier an annotated treebank is, the more linguistic assumptions it has. But chunking tasks may be a good way to start. Additionally, if the treebanks were built on some theories that contradict the theory underlined in the current work, it would be difficult to use such corpora. This also reflects the long-time debate between the ideal grammar

and treebank grammar - treebank grammar does not represent the true grammar, but in a better format for recognition.

5.3 Conclusion

In this chapter, I showed two experiments using corpora to assist in linguistic fieldwork research on Nepal Bhasa complementation. The first experiment shows the possibility of using advanced NLP tools to assist fieldwork on speeding up the annotation process with the joined research methods from fieldwork and NLP tools. The second one takes advantage of the typological similarity of inchoative morpheme behaviors among Nepal Bhasa, Mandarin, and Cantonese to gather a shared cross-linguistic complementation morphological restriction.

Both experiments are rooted in discovering complementation structures from linguistic fieldwork, whereas corpora do not necessarily offer direct linguistic insights until we establish some basic understanding of the language. The fieldwork data analysis from the previous chapters discovers the complementation patterns and the restriction factors. The corpus analysis and the NLP data exploratory experiment help to overcome some of the fieldwork research limitations such as the speed of data collection and analysis. The collaboration effort of the native speakers and the linguistic fieldworker, myself, is highlighted in adapting the combined methodology.

To conclude on using NLP tools for assisting fieldwork on low-resource languages, the following conditions may be satisfied: 1. A small set of natural language corpus. Nepal Bhasa OSCAR corpus is an open source public dataset. 2. Annotation technicians. Annotation requires native speaker knowledge of large vocabulary and script reading sentence comprehension. I do not have advanced language reading ability in Nepal Bhasa, so my main annotation resource is from my native speaker language consultants. Post-annotation meetings are required for ensuring the annotation accuracy. 3. NLP skills. Required computation skills vary for different tasks. My specific experiment requires python programming and some basic understanding of neural deep learning preprocessing, training, evaluation pipelines.

NLP assisted fieldwork may still have a long way to go to become a mature method accepted by professionals across disciplines. Building reusable annotated corpora may

efficiently set a good research foundation. In addition to the current technique of transfer learning with fine-tuning, active learning featured with actively querying annotators for labels, can provide sufficient information to the annotators without being overwhelmed by a mass of data. More high quality training data can be provided under the productive pipe-line.

Chapter 6

Conclusion and Discussion

The study aims to discover clausal complementation in Nepal Bhasa by applying linguistic fieldwork methods and NLP techniques. I start with reviewing and testing the basic syntactic facts in Chapter 2. Chapter 3 moves on to the grammatical aspect through studying the verbal morphology to discover how temporal information is encoded in this language. Based on findings in the Chapter 2 and 3, Chapter 4 displays the examinations of different complementation types and proposes two syntactic complementation strategies that Nepal Bhasa deploys. Chapter 5 shows two experiments of using the open-source data to assist linguistic fieldwork in Nepal Bhasa, which procedure can be adopted in linguistic research for endangered/low-resource languages.

The study suggests that the pre-verbal head-final CPs that are headed by *dhakā* or *dhayā* are true complements and the head-final post-verbal *ki*-clauses are parataxis. The paratactic CPs blocks all matrix scopes including exhibits question particle lā in the matrix clause. Double-embedded *wh*-phrases in a true CP can take the sentential scope resulting in a pied-piping structure (Zhang and Chacón 2018). True complement CPs also allow genuine sluicing via long-distance scrambling (Zhang 2018). The *null*-head preverbal CP is reduced from *dhakā* or *dhayā* clauses, whereas the *null*-head post-verbal CP only marginally acceptable.

I show that Nepal Bhasa uses grammatical and lexical aspect marking systems and propose the inchoative aspect -NA exists in Nepal Bhasa with the event-end point undefined (E_f), where the perfective aspect *-u* requires event time (E) included in R, by applying the temporal compatibility tests. These temporal features result in different

event time entailments of the two aspectual types, even they are both compatible with certain tense environments such as the past tense. The existence of inchoative in Nepal Bhasa is also evidenced by observed cross-linguistic shared immediate future reading in Mandarin Chinese and Cantonese. Inflected auxiliaries in progressive and perfective have multiple forms, and my fieldwork finds the inchoative-like forms are less preferred embedding verb in the matrix clause. A generalization of inchoative-like aspect restricted in embedding verbs in complementation is supported by the data collected from my fieldwork and the OSCAR dataset.

From the research methodology point of view, I explore data from other resources to support the findings in my fieldwork. The first experiment of training the chunking models for predicting Nepal Bhasa CPs shows the effectiveness of applying transfer learning technique on endangered languages. With fine-tuning the large language model (110M parameters) mBERT for the target language, Nepal Bhasa with a smaller dataset (200+ annotated sentences), the models achieve successful performance (85%-90% F1). Furthermore, the entire procedure from data annotation to data-processing to model training took a reasonably short period of time. This shows the possibility of using NLP tools for effectively building corpora for endangered languages. The second experiment of corpus search on Mandarin and Cantonese data merely answers the question of why building structured corpora is helpful: it provides ways to extract natural language sentences with any desired syntactic structural pattern, the clause complementation structures in this case.

The study still remains in many future directions. My experience of working with native speakers on the annotation tasks was fruitful. The annotation speed grows much faster from the first session, a sign of gaining linguistic expertise. This could further advance the speed of building corpora and language models. I had two native speakers helping me read the Devanagri script of Nepal Bhasa raw data, if there were more native speakers, cross-viewing agreement tests of each other annotation work can help reduce human annotation errors. More language treebank corpora can be used for cross-linguistically testing on inchoative behaviors in embedding verb predicates.

I have shown that timespan adverbials can affect the compatibility of inchoative and perfective with tense environments. Testing on more variety of lexical classes is needed to examine whether verb classes are also a factor in changing the default temporality in

Nepal Bhasa. In addition, more inflected auxiliary pattern forms should be discussed for further understanding the roles of the morphemes *-u* and -NA in complex morphology.

The number of data examples of the double-embedding complementation is limited in this work. Further exploring possible double-embeddings formed by the different headed CPs may offer new insights to the understanding of Nepal Bhasa clausal complementation: whether *ki*-clauses can be double embedded and whether there are any hierarchical ordering constraints between the C heads *dhakā* and *dhayā*.

Bibliography

Abney, Steven P. 1991. Parsing by chunks. In *Principle-based parsing*, 257–278. Springer.

Baldridge, Jason, and Alexis Palmer. 2009. How well does active learning actually work? time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 296–305.

Bayer, Josef. 1996. Complementation and the scope of wh in bengali. In *Directionality and logical form*, 251–309. Springer.

Bayer, Josef. 2001. Two grammars in one: Sentential complements and complementizers in bengali and other south asian languages. In *Tokyo Symposium on South Asian Languages: Contact, Convergence and Typology*, 11–36.

Bayer, Josef. 2006. Wh-in-situ.

Beck, Sigrid. 2006. Intervention effects follow from focus interpretation. *Natural Language Semantics* 14:1–56.

Bennett, Gena R. 2010. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.

Bhaskararao, Peri, and Sunder Krishna Joshi. 1985. A study of newari classifiers. *Bulletin of the Deccan College Research Institute* 44:17–31.

Bresnan, Joan W. 1972. Theory of complementation in english syntax. Doctoral Dissertation, Massachusetts Institute of Technology.

Cable, Seth. 2012. Pied-piping: Introducing two recent approaches. *Language and Linguistics Compass* 6:816–832.

Chacon, Thiago Costa. 2009. Lexical and viewpoint aspect in kubeo. In *Conference on Indigenous Languages of Latin America IV, University of Texas*.

Chan, Marjorie. 1980. Temporal reference in mandarin chinese: an analytical-semantic approach to the study of the morphemes le, zai, zhe, and ne. *Journal of the Chinese Language Teachers Association* 15:33–79.

Charnavel, Isabelle. 2017. Non-at-issueness of since-clauses. In *Semantics and Linguistic Theory*, volume 27, 43–58.

Cheng, Lisa Lai-Shen. 2009. Wh-in-situ, from the 1980s to now. *Language and Linguistics Compass* 3:767–791.

Christensen, Matthew Bruce. 1990. The punctual aspect in chinese: a study of the perfective and inchoative aspect markers in mandarin and cantonese. Doctoral Dissertation, The Ohio State University.

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2. Cambridge university press.

Cruz, Jan Christian Blaise, and Charibeth Cheng. 2019. Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409* .

Davison, Alice. 2007. Comp projection. *Linguistic theory and South Asian languages: Essays in honour of KA Jayaseelan* 102:175.

Dayal, Veneeta, and Anoop Mahajan. 2007. *Clause structure in south asian languages*, volume 61. Springer Science & Business Media.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805. URL `http://arxiv.org/abs/1810.04805`.

Dowty, D. 1979a. 1979: Word meaning and montague grammar. dordrecht: Reidel .

Dowty, David R. 1979b. Word meaning and montague grammar. synthese language library, no. 7, dordrecht, boston, london: D.

Dowty, David R. 2012. *Word meaning and montague grammar: The semantics of verbs and times in generative semantics and in montague's ptq*, volume 7. Springer Science & Business Media.

Dryer, Matthew S, et al. 2008. Word order in tibeto-burman languages. *Linguistics of the Tibeto-Burman Area* 31:1.

Ebrahimi, Abteen, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021.

Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726* .

Emenanjo, Emmanuel 'Nolue. 1991. The tense system of igbo. *Afrikanistische Arbeitspapiere (AAP)* 27:129–144.

Fox, Danny. 2003. On logical form. *Minimalist syntax* 82–123.

Francis, W Nelson, and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor* 5:7.

Genetti, Carol. 1988. A contrastive study of the dolakhali and kathmandu newari dialects. *Cahiers de linguistique Asie orientale* 17:161–191.

Genetti, Carol. 2005. The participial construction of dolakhā newar: syntactic implications of an asian converb. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 29:35–87.

Genetti, Carol. 2009. *A grammar of dolakha newar*, volume 40. Walter de Gruyter.

Grice, H Paul. 1989. *Studies in the way of words*. Harvard University Press.

Hale, Austin. 1980. Person markers: Finite conjunct and disjunct verb forms in newari. *Papers in South-East Asian Linguistics* 7:95–106.

Hale, Austin. 1985. Noun phrase form and cohesive function in newari. *Studia Linguistica Diachronica et Synchronica, Berlin: Mouton de Gruyter* .

Hale, Austin, and Iswaranand Shresthacharya. 1973. Is newari a classifier language? *Contributions to Nepalese studies* 1:1–21.

Halpert, Claire, and Carter Griffith. 2018. CPs in North Azeri: New evidence for extraposition. In *10th Workshop on Altaic Formal Linguistics (WAFL10)*, ed. Theodore Levin and Ryo Masuda. Cambridge MA: MITWPL.

Hamblin, Charles L. 1973. Questions in montague english. *Foundations of language* 10:41–53.

Hargreaves, David. 1986. Independent verbs and auxiliary functions in newari. In *Annual Meeting of the Berkeley Linguistics Society*, volume 12, 401–412.

Hargreaves, David. 2005. Agency and intentional action in kathmandu newar. *Himalayan Linguistics* 5.

Hargreaves, David. 2018. *"am i blue?": Privileged access constraints in kathmandu newar*, volume 118. John Benjamins Publishing Company.

Hargreaves, David J. 1991. The concept of intentional action in the grammar of kathmandu newari. Doctoral Dissertation, University of Oregon.

Hovy, Eduard, and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22:13–36.

Huang, C-T James. 1998. *Logical relations in chinese and the theory of grammar*. Taylor & Francis.

Huang, CT James. 1982. Move wh in a language without wh movement. *The linguistic review* 1:369–416.

Iatridou, Sabine, Elena Anagnostopoulou, and Roumyana Izvorski. 2001. Observations about the form and meaning of the perfect. *Current Studies in Linguistics Series* 36:189–238.

Jones, Robert B. 1970. Classifier constructions in southeast asia. *Journal of the American Oriental Society* 1–12.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and philosophy* 1:3–44.

Kesici, Esra. 2013. Ki-clauses in turkish: A paratactic analysis. *Coyote Papers: Working Papers in Linguistics, Linguistic Theory at the University of Arizona* .

Keskar, Nitish Shirish, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836* .

Kiryu, Kazuyuki. 2011. A functional analysis of adjectives in newar. In *Himalayan languages and linguistics*, 99–129. BRILL.

Kjeldgaard, Lars, and Lukas Nielsen. 2021. Nerda. GitHub. URL `https://github.com/ebanalyse/NERDA`.

Kotek, Hadas. 2017. Intervention effects arise from scope-taking over alternatives. In *Proceedings of NELS*, volume 47, 153–166.

Kratzer, Angelika, and Irene Heim. 1998. *Semantics in generative grammar*, volume 1185. Blackwell Oxford.

Lample, Guillaume, and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* .

Legate, Julie Anne. 2012. Types of ergativity. *Lingua* 122:181–191.

Li, Charles, and Sandra Thompson. 1981. A functional reference grammar of mandarin chinese. *Berkeley, CA: University of California Press. Find this author on* .

Li, Yen-Hui. 1990. Audrey (1990) order and constituency in mandarin chinese. *Studies in Natural language and Linguistic Theory. Kluver, Academic Publishers, Dordrecht* .

Lin, Jo-Wang. 2003. Temporal reference in mandarin chinese. *Journal of East Asian Linguistics* 12:259–311.

Lin, Jo-Wang. 2006. Time in a language without tense: The case of chinese. *Journal of semantics* 23:1–53.

Liu, Zihan, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. *arXiv preprint arXiv:2002.05923* .

Malhotra, Shiti. 2009. Intervention effects and wh-movement. *University of Pennsylvania Working Papers in Linguistics* 15:16.

Malla, Kamal Prakash. 1985. *The newari language: A working outline*. 14. Institute for the Study of Languages and Cultures of Asia and Africa.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank .

de Marneffe, Marie-Catherine, and Christopher Potts. 2017. Developing linguistic theories using annotated corpora. In *Handbook of linguistic annotation*, 411–438. Springer.

May, Robert Carlen. 1978. The grammar of quantification. Doctoral Dissertation, Massachusetts Institute of Technology.

Mayol, Laia, and Elena Castroviejo. 2013. How to cancel an implicature. *Journal of Pragmatics* 50:84–104.

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92–97.

Meechan-Maddon, Ailsa, and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 112–120.

Merchant, Jason, et al. 2001. *The syntax of silence: Sluicing, islands, and the theory of ellipsis*. Oxford University Press on Demand.

Montague, Richard. 1970. Universal grammar. *Theoria* 36:373–398.

Moulton, Keir. 2009. Natural selection and the syntax of clausal complementation .

Nepal, CBS. 2011. National population and housing census, national report. *Government of Nepal. Kathmandu* .

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, 9 – 16. Mannheim: Leibniz-Institut für Deutsche Sprache. URL `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215`.

Pesetsky, David. 1987. Binding problems with experiencer verbs. *Linguistic Inquiry* 18:126–140.

Pesetsky, David. 2000. *Phrasal movement and its kin*. MIT press.

Prior, Arthur N. 1967. *Past, present and future*, volume 154. Clarendon Press Oxford.

Reichenbach, Hans. 1949. Elements of symbolic logic .

Reichenbach, Hans. 2005. The tenses of verbs. *The language of time: A reader* 71–78.

Ross, John Robert. 1969. Guess who? In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 5, 252–286. Chicago Linguistic Society.

Searle, John R, S Willis, et al. 1983. *Intentionality: An essay in the philosophy of mind*. Cambridge university press.

Shakya, Daya R. 1997. Classifiers and their syntactic functions in nepal bhasa. *HIMALAYA, the Journal of the Association for Nepal and Himalayan Studies* 17:4.

Smith, Carlota S. 1997. *The parameter of aspect*, volume 43. Kluwer Academic Publishers.

Stowell, Tim. 1981. Complementizers and the empty category principle. In *North East Linguistics Society*, volume 11, 24.

Toews, Carmela Irene Penner. 2015. Topics in siamou tense and aspect. Doctoral Dissertation, University of British Columbia.

Tuladhar, Jyoti. 1985. Constituency and negation in newari. Doctoral Dissertation, Verlag nicht ermittelbar.

Vendler, Zeno. 1967. Facts and events. *Linguistics in philosophy* 122–146.

Zhang, Borui. 2018. Sluicing-like construction in kathmandu newari. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 54, 617–631. Chicago Linguistic Society.

Zhang, Borui, and Dustin Alfonso Chacón. 2018. Embedding, covert movement, and intervention in Kathmandu Newari. *Proceedings of the Linguistic Society of America* 3:69–1.

Zucchi, Sandro. 1998. Aspect shift. In *Events and grammar*, 349–370. Springer.

Appendix A

Annotation Instruction

Three components to capture in each sentence: a.) the embedded clause, marking it with "[ ]"; b.) the "main verb" (aka. the verb outside of the embedded clause); c.) the "embedded verb" (aka. the verb inside of the embedded clause)

Annotation example in English:

Line 1: original sentence

Sam almost didn't believe that the thief might have left the shop without any money.

Line 2: bracketing CP

Sam almost didn't believe [that the thief might have left the shop without any money].

Line 3: Picking out verbs: *[believe, might have left]*

```
00. थन धाःसा १८७६य् लुइकूगु धकाः च्वयातल
---> Leave the original sentence untouched.
00. [थन धाःसा १८७६य् लुइकूगु धकाः] च्वयातल
Step 1: Copy the original sentence and paste the copy right below it. Edit the
copied sentence.
[ लुइकूगु, च्वयातल ]
---> Step 2: find the "main verb" (ZZZ) and put it to the left of the "[ZZZ ,
]";
find the "embedded verb" (YYY) and put it to the right of the "[ZZZ,YYY]".

Please notice that the order of the verbs (which goes to the left and right)
matter in the bracket.

If you are not sure about which verb to put, you can replace it with UNK. For
example,
[लुइकूगु, UNK] or [UNK ,च्वयातल].

If the sentence does not have an embedded clause, add "**" in front of the
sentence. For example,
00. **थुकिया फुक्कं कासा लिंकबेल्ट ओभल रंगशालाय् जुइ
```

Appendix B

CP chunking model predictions

0:

```
true label:['B', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]
```

1:

```
true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
prediction:[['B', 'B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]
```

2:

```
true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']]
```

3:

```
true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O',
'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]
```

4:

```
true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I',
'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O',
```

'O', 'O', 'O']
prediction:[['B', 'O', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]

5:

true label:['O', 'O', 'O', 'B', 'I', 'I', 'I', 'I', 'I', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']]

6:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']]

7:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']]

8:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O', 'O']]

9:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']]

10:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O']]

11:

true label:['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I']
prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I',

'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I']]

12:

true label:['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O']

prediction:[['B', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'I', 'O', 'O', 'O', 'O', 'O']]