Object-conditioned differential marking in Chintang and Nepali

Abhandlung
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Universität Zürich

vorgelegt von
Robert Schikowski

Angenommen im Frühjahrssemester 2013
auf Antrag der Promotionskommission:
Prof. Dr. Balthasar Bickel (hauptverantwortliche Betreuungsperson)
Prof. Dr. Fernando Zûñiga

Zürich 2013

Contents

Acknowledgements **v**

Conventions **vii**
0.1 Transcription . . . vii
0.2 Interlinearisation . . . x
0.3 Abbreviations . . . xi
0.4 Sources of data . . . xiii

1 Preliminary considerations **1**
1.1 A quick introduction to the problem . . . 1
1.2 Some basic definitions . . . 2
1.2.1 Valency . . . 2
1.2.2 Roles . . . 4
1.2.3 Grammatical relation . . . 7
1.2.4 Verb class and frames . . . 8
1.3 Object-conditioned differential marking . . . 10
1.3.1 Differential marking in general . . . 10
1.3.2 Differential marking of and conditioned by arguments . . . 11
1.3.3 Differential marking conditioned by objects . . . 13
1.4 Analytical questions . . . 15
1.4.1 Description versus explanation . . . 15
1.4.2 Functions versus conditions . . . 16
1.4.3 Modelling grammatical decisions . . . 17

2 Chintang: S/A detransitivisation **21**
2.1 Language background . . . 21
2.2 Overview of relevant morphology . . . 23
2.2.1 Parts of speech . . . 23
2.2.2 Nominal morphology . . . 24
2.2.3 Verbal morphology . . . 26
2.3 Overview of relevant syntax . . . 28
2.3.1 Valency and basic frames . . . 28
2.3.2 Word order . . . 29
2.3.3 Frames and classes . . . 30
2.3.4 Differential marking . . . 32
2.3.5 Raising of case and agreement . . . 35
2.4 Formal properties of S/A detransitivisation . . . 38
2.4.1 S/A detransitivisation as differential framing . . . 38
2.4.2 Arguments selected by S/A detransitivisation . . . 41
2.4.3 Syntactic independence of detransitivised objects . . . 45
2.4.4 S/A detransitivisation in complex sentences . . . 49
2.5 Functional preliminaries: Identifying referents . . . 57

2.5.1 Introduction . . . 57
2.5.2 The identification process . . . 58
2.5.3 Definiteness . . . 62
2.5.4 Specificity . . . 64
2.5.5 The basis of unique identifiability . . . 67
2.6 Functional properties of S/A detransitivisation . . . 68
2.6.1 Quantifiability . . . 68
2.6.2 Specificity and arbitrary reference . . . 69
2.6.3 Quantifiability in detail . . . 74
2.6.4 Interaction with other factors . . . 84
2.6.5 Conventionalisation . . . 91
2.6.6 Some irrelevant variables . . . 101
2.7 Quantitative analysis based on corpus data . . . 105
2.7.1 Introduction . . . 105
2.7.2 Syntactic annotation and primary variables . . . 105
2.7.3 The centrality of quantifiability . . . 107
2.7.4 The role of exceptions . . . 109
2.8 Summary . . . 110
2.9 S/A detransitivisation in other Kiranti languages . . . 111
2.9.1 Overview . . . 111
2.9.2 Limbu . . . 113
2.9.3 Yakkha . . . 115
2.9.4 Athpare . . . 115
2.9.5 Belhare . . . 116
2.9.6 Chiling . . . 116
2.9.7 Bantawa . . . 117
2.9.8 Puma . . . 117
2.9.9 Summary . . . 119

3 Nepali: Differential A and O marking 121
3.1 Language background . . . 121
3.2 Overview of relevant morphology . . . 122
3.2.1 Parts of speech . . . 122
3.2.2 Nominal morphology . . . 122
3.2.3 Verbal morphology . . . 124
3.3 Overview of relevant syntax . . . 125
3.3.1 Word order . . . 125
3.3.2 Frames and classes . . . 125
3.3.3 Differential marking and valency manipulation . . . 130
3.4 Formal properties of DOM . . . 132
3.4.1 DOM as an isolated pattern . . . 132
3.4.2 Arguments selected by DOM . . . 133
3.4.3 Position of the marker . . . 135
3.4.4 Double datives . . . 136
3.4.5 The question of incorporation . . . 138
3.4.6 DOM and agreement . . . 140
3.4.7 DOM in complex predicates . . . 141
3.5 Functional properties of DOM . . . 144
3.5.1 Uses of the dative . . . 144
3.5.2 Literature review . . . 146
3.5.3 Animacy . . . 148
3.5.4 Specificity . . . 150
3.5.5 Quantifiability . . . 153
3.5.6 Interplay of animacy and specificity . . . 154

3.5.7 Topicality . . . 156
3.5.8 DOM with demonstratives and pronouns . . . 159
3.5.9 DOM with proper nouns . . . 161
3.5.10 Modification . . . 162
3.5.11 Unexpectedness . . . 163
3.5.12 Disambiguation . . . 168
3.5.13 Affectedness . . . 170
3.5.14 Some irrelevant variables . . . 171
3.5.15 One form, one function? . . . 173
3.6 Quantitative analysis based on corpus data . . . 174
3.6.1 Introduction . . . 174
3.6.2 Syntactic annotation and primary variables . . . 175
3.6.3 Calculation of secondary variables . . . 178
3.6.4 Impact of individual variables . . . 179
3.6.5 Interplay of variables . . . 197
3.6.6 Predicting DOM by probability and rules . . . 203
3.7 Some notes on the history of DOM . . . 208
3.7.1 The appearance of *-lai* . . . 208
3.7.2 An etymology of *-lai* . . . 211
3.8 Summary . . . 213
3.9 DOM in other Indo-Aryan languages . . . 215
3.9.1 Overview . . . 215
3.9.2 Panjabi . . . 218
3.9.3 Kumauni . . . 219
3.9.4 Maithili . . . 220
3.9.5 Bhojpuri . . . 222
3.9.6 Hindi-Urdu . . . 223
3.9.7 Bengali . . . 226
3.9.8 Gujarati . . . 227
3.9.9 Oriya . . . 229
3.9.10 Marathi . . . 230
3.9.11 Sinhala . . . 231
3.9.12 Summary . . . 233

4 Conclusions **235**
4.1 Chintang vs Nepali . . . 235
4.1.1 Commonalities . . . 235
4.1.2 Differences . . . 235
4.1.3 Mutual influence? . . . 237
4.2 Repercussions for general linguistics . . . 238
4.2.1 Differential marking . . . 238
4.2.2 Identifiability and quantifiability . . . 239
4.2.3 Non-reductionist explanation . . . 239
4.2.4 Theories of DOM . . . 240

A Annotation guidelines Chintang **241**
A.1 How to tag files . . . 241
A.2 Variables, carriers, and values: overview . . . 241
A.3 Variables, carriers, and values: details . . . 243
A.4 Additional helps for assessing values . . . 250

B Annotation guidelines Nepali **257**
B.1 How to tag files . . . 257
B.2 Variables, carriers, and values: overview . . . 259
B.3 Variables, carriers, and values: details . . . 261
B.4 Problems in complex sentences . . . 271

C Scripts **275**
C.1 sad-parse.pl . . . 275
C.2 sad-consistency.pl . . . 281
C.3 sad-analysis.R . . . 288
C.4 dom-parse.pl . . . 290
C.5 dom-consistency.pl . . . 302
C.6 dom-analysis.R . . . 307
C.7 dom-regression.R . . . 310

D Verb paradigms **317**
D.1 Chintang . . . 317
D.2 Nepali . . . 326

Lebenslauf **345**

Acknowledgements

Writing this thesis has been fun and a fight. Thanks go to all the people who helped me.

- De Balthasar Bickel hät mer d glägehait gää, im CLRP z schaffe und hät mi mit filne mänsche und technologie zämepracht. Är hät sich immer wider ziit gnaa zum mini arbet mit mir diskutiere und isch debii immer sachlich und fair gsii. Danken au an Fernando für di super zwaitbetroiig.

- मेरो सहकर्मी नेत्रजीलाई पनि धेरै धन्यवाद दिन चाहान्छु । वहाँ बिना म नेपालमा हराउँथें र वहाँको राजनैतिक कौशल बिना म छिन्ताङमा काम गर्न सक्दिनँथें ।

- Baiʔ Rʌbiniŋ upʌriwarceniŋ yaŋ dhʌnnebad pimace konnoʔ. Chintaŋbeʔ hunikhimbeʔ yuŋma upidaŋnihẽ, sapphi limlok kok uthuktaŋbidaŋnihẽ, kina oliolikha kamacebeʔ uphadaŋnihẽ. Rʌbiniŋ unisa Sarmilaniŋ thekthekthekwa khasiŋsiciŋ nusayaŋ huncirek maikaiʔmata ucektaktace. Hiccibaŋŋa ta narek tuklok somma Nʌspati Kʌtha ukhemsace. Rʌbiniŋ Kathmandubeʔyaŋ tubaktacehẽ kina paĩyaŋ kam likhi luŋaʔã.

- Suryʌŋa yaŋ baddhe uphadehẽ. Chintaŋge, Bantawa, Paniriŋko miʔmikha sʌbdʌce somma jamma nisoko kha raicha. Bamu pode numde haŋ bhasabʌigyanik lise phe hola. Teĩbe huŋkhiya micciniŋ toŋnogo maʔmi chittuhẽ pache aniŋwa tiye.

- Sarah Schneider half mir bei der Annotation des Chintang Language Corpus. Danke!

- मलाई NNC को टिप्पणी गर्न र DOM को बनावट बुझ्न सघाउनुभएकोमा बलराम प्रसाईंजी र कृष्ण पौडेलजीलाई धेरै धन्यवाद दिन चाहन्छु ।

- त्यसैगरी मेरा नेपाली भाषाका गुरु लक्ष्मी नाथ श्रेष्ठजीलाई पनि धेरै धन्यवाद दिन चाहन्छु । वहाँले मलाई नेपाली भाषा नसिकाउनुभएको भए मैले छिन्ताङमा मान्छेहरूसँग कसरी कुरा गर्न सक्थें ?

- Meine Kollegen Taras Zakharko, Alena Witzlack-Makarevich und Giorgio Iemmolo boten mir ein angenehmes Umfeld und hatten stets ein offenes Ohr für meine merkwürdigen Ideen. Ohne euch wäre ich nur halb so gern ins Büro gekommen! Besonderer Dank geht an Taras für seine Verrücktheit, an Alena für leckeres Essen und an Giorgio für ausgedehnte Kaffeepausen, sowie an alle drei für Kommentare zu dieser Arbeit. Спасиба, дзякуй, grazii.

- I would also like to thank all people working in the CLRP: transcribers, translators, and glossers. Your work has been the base for the Chintang Corpus of which my work has greatly profited. So thanks to all the Rais: रिखी माया, जानकी आन्ने, रेनुका, दुर्गा बहादुर, गणेश, दयाराम, दुर्गा कुमारी, अनिता, चन्द्र कुमारी, शान्ति माया. Thanks to Diana Schackow and सन्तोष घिमिरे, who became a friend while I was in Nepal, as well as to all glossers in Germany and Switzerland: Sebastian Sauppe, Krissi Labs, Claudia Polkau, Kodjo Vissiennon, Saskia Wunder, Anne Wienholz, Joel Prokopchuk, Rachel Weymuth, Marcel Sanjuan, Shirin Hegetschweiler, and Michael Erlach.

- Rechd sche dànga mechd i a no maina familje und am Bernd. Oafach fias dòsai.

The research for this thesis has been financed by the European Science Foundation under the EUROCORES programme EuroBABEL (CRP "Referential Hierarchies in Morphosyntax", IP "Differential agreement in Chintang and differential case marking in Nepali"; DFG/ESF Grant No. BI 799/5-1, 2009-2013).

Conventions

0.1 Transcription

Three main writing systems are used for the object languages in this book. The default is simplified, phonological IPA for both Chintang and Nepali. A further simplified, ASCII-compatible alphabet is used for a few frequent proper names such as *Chintang* and *Nepali*, which it would be cumbersome to write differently. Finally, the International Alphabet of Sanskrit Transliteration (IAST) is used for writing proper names outside examples and for historical materials. IAST can capture all graphemic distinctions made in Devanagari, the native script of Nepali (and of Chintang as far as it is written), which is important for proper names because orthography is distinctive there and for historical materials because we can only speculate about its pronunciation. In addition to the three main systems, phonetic IPA and Devanagari will occasionally be used to illustrate the precise pronunciation and graphic representation of words, respectively.

Table 1 shows the spelling of the words *Chintang* and *Nepali* in all five writing systems. The detailed conventions for writing each language are described below.

Phonetic IPA	Phonological IPA	English Roman	Devanagari	IAST
[ˈʈʂʰiɳʈaŋ]	/chintaŋ/	Chintang	छिन्ताङ	Chintāṅa
[n̪eˈpaːli]	/nepali/	Nepali	नेपाली	Nepālī

Table 1: Spelling the words *Chintang* and *Nepali*

Chintang was not written at all before its linguistic description. The only exception were proper names used by the Nepali administration and therefore written in Devanagari. The version of the IPA I use here is the one developed by the Chintang and Puma Documentation Project (CPDP, Volkswagenstiftung DoBeS programme, grant no. II/79 092, 2004-2008). It is shown in Table 2. The recent Chintang-Nepali-English dictionary by Rāī et al. (2011) uses Devanagari with some special conventions. This system is likely to be taken over by the Chintāṅa Bhāṣā Saṃskṛti tathā Sahitya Pariṣada ('Chintang Language Culture and Literature Council'), which is planning to introduce Chintang classes in schools. It is also shown below for the sake of interest.

Phonetic IPA	Phonological IPA	English Roman	Devanagari
p	**p**	p	प
pʰ	**ph**	ph	फ
b	**b**	b	ब
bʱ	**bh**	bh	भ
m	**m**	m	म
w	**w**	w	व
ʈ	**t**	t	ट
ʈʰ	**th**	th	ठ
ɖ	**d**	d	ड

Table 2: Writing Chintang

Phonetic IPA	Phonological IPA	English Roman	Devanagari
ɖʱ	**dh**	dh	ढ
ɳ	**n**	n	न
s	**s**	s	स
l	**l**	l	ल
lʱ	**lh**	lh	ल
r	**r**	r	र
t̪s̪	**c**	c	च
t̪s̪ʰ	**ch**	ch	छ
d̪z̪	**j**	j	ज
d̪z̪ʱ	**jh**	jh	झ
j	**y**	y	य
k	**k**	k	क
kʰ	**kh**	kh	ख
g	**g**	g	ग
gʱ	**gh**	gh	घ
ŋ	**ŋ**	ng	ङ
ʔ	**ʔ**	'	ः
ɦ	**h**	h	ह
i	**i**	i	इ
ɨ	**ɨ**	i	ऽ
u	**u**	u	उ
e	**e**	e	ए
o	**o**	o	ओ
a	**a**	a	आ
~	~	-	ँ

Table 2: Writing Chintang

Nepali has a long tradition as a written language, the earliest inscriptions stemming from the 13th century AD (Hutt 1988:79). Its traditional script is Devanagari, which is also used for a couple of other South Asian languages such as Hindi and Marathi. Presently there is no generally accepted Devanagari standard orthography for Nepali but a multitude of competing conventions. One big difficulty is that many words display a great degree of pronunciational variation, mainly reflecting the education of the speaker and his knowledge of Sanskrit. For instance, the word *vyavasthā* 'handling' is a loanword from Sanskrit. Its pronunciations range from the most learned variant [ʊyʌʊʌsˈt̪ʰaː] to the most informal [bɛbəsˈt̪aː]. Accordingly, it can be spelt <व्यवस्था/*vyavasthā*>, <ब्यवस्था/*byabasthā*>, <बेबस्था/*bebasthā*>, or <बेबस्ता/*bebastā*>.

I will assume that the most informal variant is basic for speakers and will therefore use it for the transcription of written examples. There are two reasons for this. One is that the pronunciational variation found in Nepali partially seems to be induced by orthography. It can often be observed that a speaker pronounces a word using the most informal pronunciation without caring too much until he sees the word for the first time in a spelling close to Sanskrit, after which he will struggle to adjust his pronunciation to the spelling. Second, "struggle" is to be taken literally here: most speakers have great difficulties with pronouncing words the way suggested by more conservative orthographic conventions. For instance, I have never met a Nepali speaker who could distinguish between <श>/*ś* (originally [ɕ]), <ष>/*ṣ* (originally [ʂ]), and <स>/*s* (originally [s]) – in the basic variant, there is a single phoneme /s/ which is freely realised as [s] or [ʂ]. Nevertheless, speakers who want to present themselves as educated try to come up with some difference in words like <शासन>/*śāsana* 'rule', pronouncing once [ˈʂaːsən], then again [ˈsaːʂən].

Table 3 shows the correspondences between Devanagari letters (in the most conservative available spelling) and the various writing systems used in this work (most importantly, phonological

IPA). Table 4 summarises the differences between the most conservative orthography/pronunciation complex (given in IAST) and the pronunciation that is assumed to be basic here (given in phonological IPA). The latter table also includes clusters and precise conditions for pronunciation.

Devanagari	Phonetic IPA	Phonological IPA	English Roman	IAST
प	p	**p**	p	p
फ	pʰ	**ph**	ph	ph
ब	b	**b**	b	b
भ	bʱ	**bh/b**	bh	bh
म	m	**m**	m	m
व	b/w	**b/w**	b/w	v
त	t̪	**t**	t	t
थ	t̪h	**th**	th	th
द	d̪	**d**	d	d
ध	d̪h	**dh/d**	dh	dh
न	n̪	**n**	n	n
ट	ʈ	**ʈ**	t	ṭ
ठ	ʈʰ	**ʈh**	th	ṭh
ड	ɖ	**ɖ**	d	ḍ
ढ	ɖʱ	**ɖh/ɖ**	dh	ḍh
ण	ɖ	**ɖ**	n	ṇ
स	s	**s**	s	s
श	s	**s**	s	ś
ष	s	**s**	s	ṣ
ल	l	**l**	l	l
र	r	**r**	r	r
च	t̪s̪	**c**	c	c
छ	t̪s̪ʰ	**ch**	ch	ch
ज	d̪z̪	**j**	j	j
झ	d̪z̪ʱ	**jh/j**	jh	jh
य	j	**y**	y	y
क	k	**k**	k	k
ख	kʰ	**kh**	kh	kh
ग	g	**g**	g	g
घ	gʱ	**gh/g**	gh	gh
ङ	ŋ	**ŋ**	ng	ṅ
ह	ɦ	**h**	h	h
ः	-	-	-	ḥ
इ	i	**i**	i	i
ई	i	**i**	i	ī
उ	u	**u**	u	u
ऊ	u	**u**	u	ū
ए	e	**e**	e	e
ऐ	ʌi	**ʌi**	e	ai
ओ	o	**o**	o	o
औ	ʌu	**ʌu**	o	au
अ	ʌ	**ʌ**	a	a
आ	a	**a**	a	ā
ऋ	ri	**ri**	ri	ṛ
ॠ	ri	**ri**	ri	ṝ
ँ/ं	~	~	-	ṃ

Table 3: Writing Nepali

Conservative orthography	Conditions	Basic pronunciation
bh, dh, ḍh, jh, gh	all except #_, _'	b, d, ɖ, j, g
th	s_	t
Vṇ	all	Ṽɖ
v	{ʌ, a} _{ʌ, a}, C_	w
v	rest	b
ś, ṣ, s	all	s
sC	#_	isC
kṣ	#_	ch
kṣ	V_V	cch
kṣa	all	che
Cya, Cva	#_	Ce, Co
Cya, Cva	rest	CCe, CCo
i, ī	all	i
u, ū	all	u
ṛ, ṝ	all	ri
h	V_V	-
ḥ	all	-
ṃ	_{k, kh, g, gh}	ŋ

Table 4: "Basic" pronunciation of Nepali

0.2 Interlinearisation

Glossing generally follows the Leipzig Glossing Rules (www.eva.mpg.de/lingua/resources/glossing-rules.php, last accessed on 18 February 2011).

While glossing Nepali is generally unproblematic, glossing Chintang is often difficult. As in most Kiranti languages, Chintang verb forms are largely agglutinative and therefore easily segmentable but at the same time grossly violate the morphological ideal of 1:1 correspondences between segments and functions. Not only can one portmanteau affix mark several functions, often a single function is marked by several affixes, too. On the one hand, the absence of an affix can mark a function; on the other hand, a visible affix often marks nothing at all (i.e. marking may be redundant). There are complex dependencies between markers so that one affix may mark notably different functions in combination with different affixes, and one function may be expressed by different combinations of markers depending on other functions to be expressed.

There are two options for rendering this complexity in glosses. One is to take a paradigmatic approach, that is, to take every segment and gloss what is common to all paradigm cells it occurs in. For instance, the marker *-i* appears in the scenarios 1piS, 1peS, 2pS, and 2pO, so its paradigmatic function is [1/2pS/O].[1] By contrast, a syntagmatic approach looks at the function of a concrete verb form as a whole and glosses what each marker contributes to it. Depending on the verb form, *-i* might then, for instance, be glossed as [1piS] (without further agreement affixes) or [p] (in combination with the prefix *a-* [2S]). This approach also glosses functions that are not represented by any segment but are marked by the form as a whole.

The present work uses syntagmatic glosses because they generally make it easier to understand the meaning of a verb form. Information that is not part of the paradigmatic function of a marker or marked by the form as a whole is added in square brackets. Redundantly marked information is left away when it can easily be gathered from another marking locus.

Below are two examples for constructed Chintang verb forms glossed in two different ways. The paradigmatic approach is illutrated by (1a) and (2a), the syntagmatic approch by (1b) and (2b).

[1] *-i* cannot be viewed as a general marker of plural since it doesn't occur in all plural cells.

(1) a. *copt-a-n-u-mh-a*
look.at-IMP-2/3p-3O-1/2nsA-IMP
'look at him'
b. *copt-a-n-u-mh-a*
look.at-IMP-2p-3[s]O-2pA-IMP
'look at him'

(2) a. *tham*
fall
'he might fall'
b. *tham*
fall[.NPST.SUBJ.3sS]
'he might fall'

0.3 Abbreviations

The following abbreviations are used:

1	first person	CONJ.PTCP	conjunctive participle
2	second person	CONT	continuous
3	third person	COP	copula
A	transitive agent	CTOP	contrastive topic
ABESS	abessive	CVB	converb
ABL	ablative	CVB.BGR	backgrounding converb
ACC	accusative	CVB.FGR	foregrounding converb
ACCESS	accessible	DAGR	differential agreement
ACROSS	horizontal movement	DAM	differential agent marking
ACT.PTCP	active participle	DAT	dative
ADD	additive	DEF	definite
ADESS	adessive	DEM	demonstrative
ADJ	adjective	DIR	direction; direct case
ADV	adverb	DIST	distal
ADVZ	adverbialiser	DISTR	distributional
AFF	affirmative	DOI	differential object indexing
AGR	agreement	DOM	differential object marking
AMOUNT	amount	DOWN	vertical movement down
ANTE	antessive	DYSF	dysfunctional
ARG	argument	e	exclusive
ASS	assertive	EQU	equative
ATTN	attentional	ERG	ergative
AUX	auxiliary	EXP	experiencer/experiential
AWAY	metaphorical movement AWAY	EXT	extensional
BEN	benefactive	EXTRA	extraessive
CAUS	causative	F	feminine
CHAR.PTCP	characteristic participle	FILLER	filler
CIT	citation particle	FIN	final case
CIT.ADN	adnominal citation	FOC	focus
CLF	classifier	FUT	future
COM	comitative	G	ditransitive goal (≈ recipient)
COMP	comparative	GEN	genitive
COMPL	completive	HAB	habitual
CON	conative	H	honorific
CONCS	concessive	HH	high honorific
COND	conditional	HON	honorific

HUM.CLF	human classifier
i	inclusive
IA	Indo-Aryan
IDF	indefinite
IMP	imperative
IN	metaphorical movement IN
IND	indicative
INF	infinitive
INSIST	insistive
INST	instrumental
INTENS	intensification
INTRA	intraessive
IPFV	imperfective
IRR	irrealis
ITR	intransitive
LH	low honorific
LNK	linker
LOC	locative
MED	medial
METHOD	method
MH	mid honorific
MIA	Middle Indo-Aryan
MIR	mirative
MOD	modalis
n	neuter
N	noun
N.EXP	experiential noun
NA	not applicable
NAME.NTVZ	name nativiser
NEXT	next step
NFUT	non-future
NIA	New Indo-Aryan
NMLZ	nominaliser
NNOM	non-nominative
NOM	nominative
NONF	non-finite
NP	noun phrase
NPST	nonpast
NSAP	non-speech act participant
NTVZ	nativiser
NVOL	non-volitional
O	object (as grammatical relation)
OBL	oblique case
OIA	Old Indo-Aryan
OPT	optative
ORD	ordinal number
OUT	metaphorical movement OUT
p	plural
P	transitive patient
PASS	passive
PERL	perlative
PL	plural
POR	possessor
POSS	possessive
POST	postessive
PRF	perfect
PRFV	perfective
PROB.FUT	probable future
PROG	progressive
PROX	proximate
PRS	present tense
PRS.PRF	present perfect
PST	past tense
PTCP	participle
PURP	purposive
PVB	preverb
Q	interrogative stem
QTAG	question tag
RECNF	reconfirmative
RECP	reciprocal
REF	referential
REFL	reflexive
REL	relative
REP	reportative
RESTR	restrictive
RETRV	retrieval instruction
s	singular
S	intransitive subject
SAP	speech act participant
SEQ	sequentialiser
SG	singular
SORT	sortal
SUB	subessive
SUBJ	subjunctive
SUPER	superessive
SURP	surprise
T	ditransitive theme
TEL	telic
TERM	terminative
TMA	tense-mood-aspect
TMP	temporal
TR	transitive
TRANS	translative
UP	vertical movement up
V	verb
V.NTVZ	verbal nativiser
Σ	verb stem

0.4 Sources of data

All examples in this text contain a source indication. This may be a corpus or my own elicitation data and field notes. In the case of corpus data, the name of the source file and the record or sentence number from which the example was taken are given. Examples from elicitation contain a code for the informant and the time of the interview. Field notes only show the time when they were taken. Linguistic statements on Chintang and Nepali that do not contain a source indication are based on my own work.

All corpus examples for Chintang are taken from the Chintang Language Corpus (CLC). The compilation of this corpus started during the Chintang and Puma Documentation Project (CPDP, Volkswagenstiftung DoBeS programme, grant no. II/79 092, 2004-2008) and is still being continued with financing from several collaborative projects together referred to as Chintang Language Research Programme (www.spw.uzh.ch/clrp). Part of the Chintang Corpus has been made publicly accessible at the Language Archive of the Max-Planck institute Nijmegen (www.mpi.nl/research/research-projects/the-language-archive). The complete corpus is stored at the Department of General Linguistics at the University of Zurich (www.spw.uzh.ch).

All corpus examples for Nepali are from a modified version of the Nepali National Corpus (NNC). The NNC was originally compiled by the Bhasha Sanchar project (www.bhashasanchar.org) with funding from several sources. Since the original version of the NNC contained various file formats and encodings, all files had to be converted to the format used for most files (XML under the XCES standard, www.xces.org) and to UTF-8. The original folder structure was converted so that all genres (core sample prose, books, newspapers, webtext, spoken text) were located on the same level. The modified version of the NNC is likewise stored at the Department of General Linguistics at the University of Zurich.

Subcorpora of CLC and NNC have been annotated for various variables relevant for the present work in order to back up the qualitative discussion with quantitative data. Details on the annotation as well as on the linguistic results can be found in the dedicated sections 2.7 and 3.6. The full annotation guidelines are given in the Appendices A and B.

Both corpora are sociolinguistically diverse, featuring speakers of various ages, genders, social backgrounds etc. A peculiarity of the CLC is that it contains large amounts of child speech (marked by "CL" in the session names) due to the focus on language acquisition during CPDP. The inclusion of these data is, however, unproblematic, because the morphosyntax of Chintang child speech is not different from that of adults' speech – the main areas of divergence are phonology and in the lexicon.

Chapter 1

Preliminary considerations

1.1 A quick introduction to the problem

If the only task of morphosyntax was to indicate the oft-cited "who does what to whom", one would expect that knowing the participants of an event and their roles would be sufficient for determining the basic makeup of a sentence. In reality, however, role is tangled up with properties of the referent occupying the role and of other constituents in a vaste number of languages. The role that is probably best known for this is P, which is in this context mostly talked about in terms of the corresponding grammatical relation, i.e. as object. The best-known pattern involving P is DIFFERENTIAL OBJECT MARKING (shortly DOM). In this pattern, the role of P is marked by different cases depending on properties of its referent such as animacy or topicality. (1) shows a pair of examples from Nepali, an Indo-Aryan language spoken in Nepal. The animate P *manche* 'person' in (1a) is marked by the dative suffix *-lai*, whereas the inanimate P *bhat* 'rice' is in the nominative (zero) in (1b).

(1) a. *Raches-hʌru-le manche-lai kha-e.*
ogre-PL-ERG person-DAT eat-PST.3p
'The ogres ate somebody.'
b. *Raches-hʌru-le bhat kha-e.*
ogre-PL-ERG rice eat-PST.3p
'The ogres had rice.' (elicitation NP 2012)

This pattern is widespread, but it is not the only one. For instance, another possible pattern is differential object indexing, where P is indexed or not according to its referential properties. Yet another, less well-known possibility is what may be called differential framing – patterns in which properties of P affect several marking loci at once. This is illustrated by the sentences in (2), which are a translation of (1) into Chintang, a Tibeto-Burman language also spoken in Nepal. While in (2a) the A *rakkasace* 'ogres' is marked by the ergative and both A and P are indexed on the verb, A is in the zero-marked nominative in (2b) and is the only argument linked to agreement, which looks as if it had been triggered by an S. This pattern is called S/A DETRANSITIVISATION in the present work:

(2) a. *Rakkas-a-ce-ŋa ma?mi u-c-o-he.*
ogre-NTVZ-ns-ERG person 3pA-eat-3[s]O-IND.PST
'The ogres ate somebody.'
b. *Rakkas-a-ce kok u-ci-e.*
ogre-NTVZ-ns rice 3pS-eat-IND.NPST
'The ogres had rice.' (elicitation RBK 2012)

Although the two patterns illustrated in (2) and (1) look superficially similar, they are actually very different in many respects. While Nepali DOM is formally rather simple, S/A detransitivisation

in Chintang is complex even in its most basic variant and has links to all areas of grammar. On the other hand, the latter pattern is functionally simple in that it is basically governed by a single functional variable. By contrast, Nepali DOM involves a multitude of functional factors that call for new ways of predicting grammatical phenomena in general. Further, whereas the central variable for Chintang is akin to specificity, most important factors in Nepali DOM come from the area of topicality and topicworthiness.

These differences are the topic of the present study, which seeks to make contributions in two areas. First, it describes in detail the form and function of the mentioned phenomena, DOM in Nepali and S/A detransitivisation in Chintang. Second, it is hoped that the individual descriptions together with a comparison of the phenomena in question can provide new insights into the nature of object-conditioned differential marking patterns in general. There are several reasons why Chintang and Nepali are highly relevant for this:

- S/A detransitivisation is not a well-known pattern, so its description may widen the horizon of descriptions of differential marking in general.
- Nepali DOM is functionally complex and that is a property that is shared by many other DOM systems (and maybe differential marking systems in general), so a sophisticated model of its function may be of use for the description of other languages and phenomena, too.
- Nepali represents a frequent but Chintang a highly marginal pattern, so the comparison of these two languages may provide hints as to which extent the typology of object-conditioned differential marking patterns has so far been biased towards DOM and how it is possible to integrate all such patterns into a single framework.

The advantages of the comparison of two languages are that the phenomena in question can be studied in depth and that certain characteristics of each phenomenon become better visible in contrast with the other language. On the other hand, the method also has its drawbacks. In particular, this study does not provide any insights into the general typology of object-conditioned differential marking – rather, it makes a small contribution to such a typology by exploring some general possibilities in a nutshell.

1.2 Some basic definitions

This section provides definitions for a few basic concepts that will be used throughout this work. Since all concepts mentioned below have been described elsewhere at book-length and this is a descriptive rather than a theoretical study, I will keep my own thoughts rather short and just try to make clear what I mean by each term and briefly explain why this definition is useful. The only concept that I will discusse at some length is the one which is most important for this study, viz. object-conditioned differential marking. There is a dedicated section (section 1.3) for this. The general approach I take is typological, that is, I will try to define concepts in a way that maximises comparability across languages and will adapt the description of the languages in question to this goal.

1.2.1 Valency

The term VALENCY was borrowed into linguistics from chemistry, and is still easiest to gain an intuitive understanding of it by looking at the chemical definition: the valency of an atom is the number of bonds it can have to other atoms. In the case where there is only a single bond between each pair of atoms in a larger complex, the valency of an atom equals the number of other atoms bonded to it. This latter situation is the base for the linguistic metaphor, where the atom whose valency is described is a predicate (usually a verb in terms of morphosyntax) and the other atoms are referents (nouns). Different predicates intuitively have different valencies: for instance, while it is possible to say *He saw the children*, **He slept the children* is impossible. Thus, in these examples *see* has a higher valency than *sleep*.

Linguistic valency is, obviously, not as straightforward as the chemical one – it is linked to many theoretical problems; for instance, how the valency of a predicate is to be determined, how one can ascertain whether a referent is an ARGUMENT (i.e. "bonded" by the predicate) or not, or whether bonding isn't rather a gradual than a binary property. These questions are important for the present study because objects, its central topic, can only be defined with recourse to valency. For instance, we would like to make statements such as that the single argument of a monovalent verb is never an object. But for this we first need to have a clear notion of what that means.

A formal definition is out of the question because languages are often incomparable with respect to formal criteria. For instance, we could easily say that in Chintang everything is an argument that can trigger agreement. However, if we applied the same definition to Nepali, that would make all verbs that are bivalent in Chintang monovalent in Nepali, because Chintang has bipersonal agreement but Nepali has not. So what we need is a functional definition that captures the intuition that some referents have stronger "bonds" with a predicate than others. Here is the proposal:

> Let there be a clause containing a predicate and at least one referent. Both predicate and referents may be covert, but it must be possible to mention them overtly in order to view them as contained in the clause.[1] Let each referent occupy a role expressing what it does in the state of affairs coded by the clause.
> Then any referent is an argument of the predicate if its role can only be determined with reference to properties of that predicate, and the valency of the predicate is the number of its arguments.

For instance, in *Mary met Peter on the plane*, the roles of *Mary* and *Peter* (let's call them "agent" and "patient" for the moment) can only be determined if one knows that the relevant predicate is *meet* – otherwise these NPs could occupy quite different roles (cf. *Mary turned around*, *Mary gave Peter a turtle*, *Mary took Peter to the haunted house*). By contrast, *on the plane* could be interpreted in any clause because (almost) all states of affairs have a place.

It would be possible to view argument status as gradual based on this definition, because one often has to know more or less about the predicate in order to determine the role of a referent. For instance, consider the following (constructed) Chintang sentence:

(3) *Kapp-e-ŋa* *phakcilek khorek-be? yuŋs-o-ŋs-e.*
Kalpana-NAME.NTVZ-ERG piglet sty-LOC$_1$ put-3[s]O-PRF-IND.PST[.3sA]
'Kalpana put the piglet into the sty.' or 'In the sty, Kalpana put the piglet down.'

The suffix *-ŋa* is polyfunctional in Chintang – it can not only mark agents, but also instruments and causes. *Kappeŋa* is therefore a prototypical case for a referential expression whose role can only be determined with respect to the pertaining predicate, *yuŋs-*. The locative *-be?*, on the other hand, is also polyfunctional, but its functions are much less widely dispersed: it can either mark the destination of a movement (cf. the first translation) or the place where something happens (cf. the second translation). Both these functions could be summarised as "places", and accordingly *-be?* tells a hearer much more about the role of a referent than *-ŋa*. Although this is an interesting possibility, we will not pursue it further here for the practical reason that binary variables are easier to deal with. So whenever one needs to know *something* about a predicate in order to determine the role of a referent we will say that that referent is an argument, no matter whether one needs to more more or less.

Another interesting point is that although predicate properties are a precondition for determining argument roles, they are not always sufficient for that. (3) is a good example – the role of *khorekbe?* stays ambiguous even at the end of the utterance. What *yuŋs-* does tell the hearer, though, is that a destination reading is also *possible* (because *yuŋs-* codes a movement).

[1] This addition is necessary to handle various cases such as diatheses which completely remove an argument, or arguments which are semantically present but can never be overt. For instance, the Chintang passive participle *-mayaŋ* does not allow the overt realisation of an agent, so it would be assumed to mark a monovalent predicate here. Similarly, many Chintang verbs with a petrified applicative suffix *-t* express that an action is done for somebody, but that person cannot be expressed overtly unless one additionally uses the productive benefactive *-bid*. A verb like *chitt-* 'wash (for somebody)' would therefore assumed to be bivalent here.

1.2.2 Roles

Any syntactic description must make reference to ROLES in a wide sense when it comes to describing phenomena like case and agreement. But what exactly is a role? Again, the history of the word gives a good first impression. According to Weekley (1921:1246), the term "role" was borrowed into English from French and originally referred to a roll of paper on which an actor's text was written. From there it got extended to what we call an actor's role today and to the more general meaning found in phrases like *the role of sugarcane in Hawaii's economy*. The actors in a linguistic utterance are its referents (which are not by coincidence often called actants), and the roles played by them are semantic relations holding between them and the predicate they are associated with as well as among themselves.

One important question is to which degree semantic roles are abstract and stereotyped. For instance, one may describe the role of an actor in a play as *Hamlet*, but one may also simply say that he is the hero. Similarly, one may describe the role of Giorgio in *Giorgio repaired the coffee machine* alternatively as coffee-machine-repairer, repairer, agent, or doer. Obviously the most concrete descriptions fit the role most closely but are at the same time not very interesting because they don't contain any generalisation. On the other hand, a role like "doer" is very general but runs the risk of being so abstract that it becomes hard to define and says almost nothing. The more abstractly one defines roles, the less roles one gets altogether: whereas there is an infinite number of roles on the maximally fine-grained level of coffee-machine-repairer (coffee-machine-destroyer, car-repairer etc.), an abstract role like agent already covers a large portion of the possibility space, so that the number of other roles on the same level is naturally limited.

Most typologists today operate with a small set of roles (cf. Haspelmath 2011) whose definitions are closely related to semantic transitivity in two respects: first, all polyvalent predicates are assumed to feature an agent. Second, a special role symbol is used for monovalent predicates, where the only argument role cannot be easily related to transitivity. Further, it is usual to restrict the scope to argument roles, whose behaviour tends to be most idiosyncratic. The present study also subscribes to an approach of this kind because it has proven to be well usable for language description and comparison. More precisely, the role system used here is based on Dowty (1991), Primus (1999), and Bickel (2011) and would be classified as a "Bickelian approach" by Haspelmath 2011. I will briefly explain this role system below.

Dowty defines only two roles, which he calls "proto-agent" and "proto-patient". Both roles are clusters of properties. The arguments of a "predicate with grammatical subject and object" are checked for these properties, and the one with more proto-agent properties will become the subject and the one with more proto-patient properties the (direct) object (p. 576). The same mechanism is assumed to be at work in trivalent predicates, where proto-agent is mapped to subject, proto-patient to direct object, and whatever remains to "oblique or prepositional object". In principle it is also possible to apply proto-roles to monovalent predicates in order to find out whether the single argument is more agent- or more patient-like. Table 1.1 shows the properties associated with each role.

	proto-agent	**proto-patient**
cause and effect	volitional	causally affected
change of state	causes it	undergoes it
movement relative to other participant	moving	stationary
experiencer properties	sentient	-
independent existence	yes	no
aspectual properties	-	"incremental theme"[2]

Table 1.1: Proto-A/P in Dowty (1991:572-573)

[2] An "incremental theme" is defined by Dowty as a referent that has parts that can be mapped to the parts of an action (as in *I'm reading the book*, where each progress in the action of reading can be mapped to a portion of the book).

Primus (1999) makes a couple of useful suggestions for improvements of Dowty's approach that I would like to take up here:

- p. 62ff:: The notions "subject" and "(direct) object" used by Dowty are not well-defined crosslinguistically and are especially problematic in languages with ergative traits. Primus instead speaks of the two highest-ranking coding categories of a language (where "coding category" is a compound of case and agreement) and emphasises that these categories are assigned based on proto-roles. She does not specify which proto-role will be mapped to which category since this depends on alignment. The central function of proto-roles thus is two keep two arguments apart.

- p. 37: Volition is only one aspect of the characteristics of the agent in a transitive event, besides the ability to start, stop, and accomplish the event and the responsibility for this. All these can be summarised under the more general concept of control.

- p. 38: Movement in the present form is not very informative, since both agents and patients frequently move. When the movement of a referent is induced by another referent it can be viewed as a change of state. When it is induced by itself it can be replaced by another property which Primus calls (somewhat vaguely) "autonomous activity".

Note that I do not take over one of Primus' more radical ideas, viz. that proto-roles can ultimately be distinguished solely on the base of "their relative structural position in the thematic structure of a verb or sentence" (Primus 1999:60). Elegant as this reduction is, it is not worked out well by Primus herself and brings with it the practical problem that in order to determine roles one first has to determine thematic structure, which is by far not as trivial as checking proto-role properties. This is also why I do not use Primus' definition of Proto-Recipient, which is based precisely on this idea (Primus 1999:55).

Some further important additions to the definition of roles used in this work are found in Bickel (2011) and Bickel et al. (2010). Bickel (2011) brings together the Dowtyan idea of roles as property clusters with the labels S, A, and O, which have been in wide use in syntactic typology ever since Comrie (1978) and Dixon (1979), although with slightly different content: Comrie defined them on the base of prototypically intransitive and transitive verbs, and for Dixon they represented complex "semantico-syntactic" notions. Both approaches are less widely applicable than Bickel's: while Comrie's approach is similar to it in making use of prototypes, it excludes less prototypical cases from crosslinguistic comparison and thus deprives itself of a bulk of evidence.[3] Dixon's approach fails to distinguish between semantic and syntactic properties, which makes it likewise less useful, in this case both for language comparison (where syntactic properties are seldom universal) and description (where there are often mismatches between the two levels). Bickel takes over Dowty's list of proto-properties but dismisses incremental theme as based on *aktionsart* and therefore being a property of the predicate rather than of a referent. I will adopt this suggestion here.

While A and O are thus defined on purely semantic grounds (A = proto-agent, O = proto-patient), S does not have a semantic content but is a convenient label for the single argument of a monovalent predicate. I will take over these labels here but use P instead of O because it emphasises the relation between this role and the concept of proto-**p**atient and because I reserve O as an abbreviation for the grammatical relation of object (see section 1.3.3 below).

One last point we have to regard before proceeding to a summary is another critique by Haspelmath (2011), who says on p. 18 that the Dowtyan proto-role properties were not intended for

[3]Haspelmath (2011) argues in the opposite direction by saying that meaningful typological generalisations are only possible in the Comrian framework, which is simply not true – generalised roles defined on semantics allow exactly the same and in fact more statements than roles based on a notion of core transitivity. The difference is that some statements must be relativised. For instance, instead of saying "in all languages A and O get (respective) identical marking across verbs" one has to say "every language has a major verb class within which all A and all O are marked identically". Haspelmath also overlooks the importance of prototypicality in all three approaches: he criticises Dixon for basing his definitions of A and O on the prototypical and therefore fuzzy notion of transitivity, but at the same time the concepts of "agent" and "patient" used by Comrie are prototypical and fuzzy as well, just as the Dowtyan proto-agent and proto-patient – with the difference that for the latter it is clear which properties determine membership.

language comparison but for the description of English and that the selection of precisely those properties for typology and the description of other languages is not motivated. While the latter argument is irrelevant – any definition is information-free and has the sole use of capturing data patterns, which this definition is obviously able to do – it is true that the list of proto-properties is somewhat arbitrary and could be enriched by what is known about the typology of transitivity. However, most descriptions of prototypical transitivity such as Hopper and Thompson (1980), Kittilä (2002), Næss (2007) use rather similar criteria to the list found in Dowty (1991), at least as far as the properties listed there pertain to referents and not to the predicate.

Two candidates for additional properties from Hopper and Thompson (1980) are individuation and agency. The first is problematic, as discussed in Iemmolo (2011:29): clauses containing a P that is weakly individuated should be highly transitive according to Hopper and Thompson (1980), but in fact they tend to be formally lowly transitive or even intransitive across languages. The reverse conclusion (typical P should be highly individuated) is likewise not very useful since it doesn't contribute to distinguishing A and P.

The other property, agency, seems, however, a good candidate. I will not use this term here in the wide sense of Hopper and Thompson, who do not make explicit what they mean by it but seem to view it as a complex of person, mode of reference, and animacy (cf. the connection they draw to the Silverstein hierarchy on p. 273). Instead I will define it simply as the potential of a referent to affect other referents. For instance, human beings are generally highly agentive because they have the ability to affect a wide variety of other referents in deep ways. Note that agency is not the same as animacy: while a rock will almost never be conceptualised as animate, it may easily become highly agentive when it comes rolling down a slope.

The revised list of proto-properties is shown in Table 1.2.

	proto-agent	**proto-patient**
agency	highly agentive	lowly agentive
cause and effect	in control	under control
change of state	causes it	undergoes it
experiencer properties	sentient	-
independent existence	yes	no

Table 1.2: Proto-A/P in the present work

Bickel (2011) also extends the idea of proto-roles to trivalent predicates and uses the labels T (theme) and G (goal) (first introduced by Croft 1990 according to Haspelmath 2011) for the two additional roles and A_2 for the agent. This set is described in greater detail in Bickel et al. (2010), where it is assumed that in trivalent predicates A_2 is determined first. After that G is determined based on the following list of proto-goal properties (p. 384), and T is the argument that is left:

- undergoing a change of state or in experience
- causally affected by another participant
- stationary relative to movement of another participant

I do not use these properties in the present work because they are too specific to capture patterns across trivalent verbs. This is true even for Chintang, for which this system was originally designed. As explained in Bickel et al. (2010), Chintang has three major classes of trivalent verbs that all have a clear A argument and two other arguments that can be marked as T-NOM/G-NOM, T-NOM/G-LOC, or T-ERG/G-NOM (see section 2.3.3 for details). But when one takes a closer look at the verbs in these classes, it turns out that not all their "G" match Bickel's definition of proto-goal. For instance, *paŋs-* 'send (somebody somewhere)' is classified as a T-NOM/G-LOC verb, but actually the NOM-marked argument has more proto-goal properties than the LOC-marked one: only this argument undergoes a change of state and is directly causally affected, so it should actually be classified as G (T-LOC/G-NOM). It seems to me that Bickel implicitly gives greater importance to the third property, relative movement. This would place his definition closer to more standard

approaches to ditransitivity such as Malchukov et al. (2010), where the concept of transfer plays an important role. I will therefore stick to Bickel's approach of applying one set of roles (A-G-T) to all trivalent predicates but will only take over movement as a proto-property. In addition, I propose two more properties as useful, size (G is bigger than T) and affectedness (T is directly affected, G indirectly via T). The proto-properties for T and G are summarised in Table 1.3.

	proto-theme	**proto-goal**
size	relatively small	relatively big
affectedness	direct	indirect
movement	moves relative to G	stationary relative to T

Table 1.3: Proto-T/G in the present work

Since cases where A_1 and A_2 are distinguished are exceedingly rare (Bickel and Nichols 2009 admit that they are only aware of a single language where this regularly happens) and Nepali and Chintang do not contain any constructions where this is the case, I will simply use A to cover both.

A final note on the role system used here concerns points of divergence from other common role systems. The one property that is above all responsible for such divergences is the restriction to a simple role set, based on the assumption that all predicates that have the same valency can be described with the same role set. For instance, *go* is usually bivalent and thus gets the same role set as more prototypically transitive verbs such as *kill*: the mover is A and the destination is P. Similarly, *cut* is trivalent and thus gets the same role set as other trivalent verbs such as *give* or *put*: the cutter is A, the thing cut is G, and the instrument is T. Experiencer predicates also do not require a special role set. For instance, the experiencer and stimulus of *like* are mapped to A and P, respectively, based on the properties of agency, sentience, and independent existence. Some other divergences are created by the strictly semantic base of the role system used here. For instance, the experiencer and stimulus of *please* are mapped to A and P just like those of *like*, even though their morphosyntactic encoding points to the exact opposite.

1.2.3 Grammatical relation

If there is any syntactic term that is used with yet less consistency than valency and role, it is certainly this one. Even most works that are dedicated to GRAMMATICAL RELATIONS are not very clear about what they mean with it. For instance, Bossong (2001) and Farrell (2005) do not make a clear distinction between grammatical relations and roles, Croft (1991), Müller-Gotama (1994), and Givón (1997) do make a distinction but do not explain it, and Palmer (1994) starts off with a definition based on roles that is actually similar to the one that will be used here but then lets seep in syntactic criteria.

The implicit received understanding seems to be something like this:

> A grammatical relation is a recurrent dependency between exponents of syntactic relations that is formally tied to an argument and functionally related to roles (in the sense applied above), though not necessarily in a straightforward way.

The grammatical relation par excellence is the subject, which can also be used to get a more concrete understanding of what is meant by this term. For instance, in English there is a strong (if not absolute) dependency between agreement, case marking, and word order, in that the argument that triggers agreement is always marked by the nominative and is mostly placed before all other arguments and non-arguments. Since these three factors so often go together, it is convenient to summarise them under one label and call the agreement-triggering, NOM-marked, first argument of a clause its subject. This label is of great descriptive use since it helps to describe a lot of grammatical processes in a concise way. For instance, it makes it possible to say that it is subjects that can be passivised. The roles that the English subject is related to are S and A.

Alas, when it comes to defining the subject or any other grammatical relation in a meaningful way *across* languages, it turns out that they are as problematic as they are convenient – a detailed discussion of this is found in Witzlack-Makarevich (2011). When one looks at the loose definition we just gave of subject in English in the last paragraph, two simple but deep problems become apparent: first, not all languages have agreement, case marking, or a fixed word order, so formal criteria for establishing a cross-linguistic notion of subject are not very useful in general. Second, the internal structure of grammatical relations varies across languages. Thus, for instance, while probably all languages have *some* grammatical relation that is related to S and A, its precise extension is not always the same. The most well-known case of this is morphosyntactic ergativity, where precisely those marking criteria that group S and A in English (case, agreement, word order) group S and P in other languages or subsystems of theirs. But there might still be other syntactic dependencies (for instance, in nominalisation or clause chaining) that are related to S and A.

An elegant way of dealing with these problems while keeping the descriptive advantage of grammatical relations is offered by the approach introduced in Bickel and Nichols (2009) and Bickel (2011) and elaborated in Witzlack-Makarevich (2011). Their idea of grammatical relations is as follows. Since formal criteria are not suitable for establishing cross-linguistic notions, the functional component is given priority. Further, since the grouping of roles depends on the language and the kind of exponent of syntactic relations one looks at (e.g. morphosyntactic marking vs behavioural properties), this is simply admitted into the definition. An important notion in this context is *argument selector*. The term is introduced in Witzlack-Makarevich (2011) (although Bickel 2011 already speaks of argument *selection*) and refers to any minimal exponent of syntactic relations that treats some argument roles differently from others. All exponents we have mentioned so far – case, agreement, word order, nominalisation, clause chaining – and many others may function as argument selectors. Argument selectors are minimal because separate selectors are used whenever it is possible to formally keep apart two exponents. For instance, case and agreement are treated as separate argument selectors because although there usually is a strong correlation between the two they do not fully depend on each other.

A grammatical relation, then, is a set of argument roles defined by an argument selector. For instance, the term "subject" is a convenient label for the grammatical relation S/A (or {S A} in Bickel's and Witzlack-Makarevich's notation). The precise extension of this category varies depending on language and argument selectors.

The grammatical relation that is of greatest interest for the present work is the object. The easiest way to define this is as non-subject, that is, P/T/G. However, for the purpose of this study it will be convenient to choose another, more narrow definition – see section 1.3.3 below for details.

1.2.4 Verb class and frames

A VERB CLASS may be loosely defined as a set of verbal lexemes that behave similarly. Depending on which aspect one looks at, verbs may be grouped quite differently. For instance, temporal-aspectual properties need not coincide with morphological properties. The aspects of verbal behaviour that are most relevant for the present study are the ones that were the subject of the preceding sections: valency, roles, and grammatical relations. In our definition of grammatical relation we didn't make a difference between argument selectors involving marking (case, agreement, word order) and behavioural argument selectors. When looking at verb classes, however, we will ignore all behavioural argument selectors and also word order for the reason that for the languages under investigation these are either completely irrelevant or the classes defined by them coincide with those defined by morphological marking.

For instance, the Chintang converb *-saŋa* must share its S or A with an associated finite verb form regardless of the verb it attaches to, and there is no verb that allows anything else. The coreferentiality constraint of this form is thus irrelevant for verb classes. Similarly, verbs do differ with respect to which of their arguments can be bound by a reflexive, but since this is simply any NOM-marked, non-S/A argument, the classes defined by these are identical to those defined by valency and case.

An important term that I will frequently use in this context is FRAME. A frame is a construct that contains all semantic and morphosyntactic information that is of interest here (valency, roles, case, agreement) in relation to a concrete verb form. One verbal lexeme may be used with many different frames. In order to take down frames in a concise form, I will use the formalism defined in the database of the Leipzig Valency Classes project (Hartmann et al. 2013) and elaborated in Schikowski et al. (forthcoming):

- A verb form with a set of argument roles X, Y is given as {X Y V} (e.g. {S V} for an intransitive frame). In this work the order of X, Y, and V also reflects the most frequent word order in the described languages.
- A role X marked by case C is given as X-C (e.g. P-NOM: P marked by the nominative).
- When there is only a single agreement slot as in Nepali, a role X linked to agreement is shown as V-X (e.g. V-A: the verb agrees with A).
- When there are several agreement slots as in Chintang, a different strategy is needed in order to show which role is linked to which agreement slot. Since there are no standardised terms for agreement (parallel to e.g. nominative in the domain of case), we will refer to agreement slots via the role they are most frequently linked to. A link of role X to agreement slot Y is then given as V-y(X) (e.g. V-s(S): the verb agrees with S in the way it usually does, or V-s(A): the verb agrees with A as if it was S).
- Potential coreferentiality across frames can be indicated by indices where necessary (e.g. A_1, S_1: A in one frame is potentially coreferential with S in another).

Here are two examples for complete frames that will turn out to be central for Chintang:

- {A-ERG P-NOM V-a(A).o(P)}: A is marked by the ergative and has A-AGR (that is, it triggers the agreement pattern that is most usual for A), P is marked by the nominative (zero) and has O-AGR.
- {A-NOM P-NOM V-s(A)}: Both A and P are marked by the nominative, and A has S-AGR (that is, it triggers the agreement pattern that is normally associated with S).

In principle it is also possible to underspecify frames. For instance, {A X-NOM V-a(X)} would refer to a frame with an A marked by any case and at least one more argument role X linked to A-AGR. Such underspecified or ABSTRACT FRAMES are often useful for making generalisations.

Based on the notion of frame, we can now define verb class in a stricter way:

> A verb class is a set of verbal lexemes that take identical sets of frames.

Since one verbal lexeme is often associated with a range of frames, the last part of this definition is important: two verbs are only considered to be in the same class if the complete set of frames is identical between the two.

Just as there are abstract frames, abstract verb classes can also be defined when of use. For instance, the abstract frame {X-NOM V-X} (at least one argument role marked by nominative and linked to the only agreement slot) could be used in English to define an abstract class of verbs with an unambiguous subject.

In order to describe a verb class in the most economical way, only those of its frames have to be specified which distinguish it from at least one other class and which cannot be derived from other frames. This means that frames generated by differential marking patterns and alternations do not generally form part of what defines a verb class unless they depend on verb class. For instance, Nepali has a passive that can be formed from almost all verbs, no matter whether they are transitive (e.g. *mar-i-y-o* [kill-PASS-PST-3s] 's/he was killed') or intransitive (e.g. *mʌr-i-y-o* [die-PASS-PST-3s] 'somebody died', lit. 'it was died'). The relevant alternative frames (e.g. {A_1-ERG P_2-NOM V-A} vs {S_2-NOM V-S}) thus do not have to be separately specified for every verb class. Similarly, Nepali

features differential agent marking (A-ERG/NOM). Since all A that can be marked by ERG can also be marked by NOM (if not the other way round), A-NOM can be easily predicted from all frames containing A-ERG and does not have to be specified. Where differential marking patterns are characteristic for a frame or where extra explicitness is required, such patterns may be indicated within a single frame (e.g. {A-ERG/NOM P-NOM V-A}).

1.3 Object-conditioned differential marking

1.3.1 Differential marking in general

The history of the term DIFFERENTIAL MARKING starts with Bossong's (1982, 1985) work on Sardinian and on New Iranian languages, where he introduces the term DIFFERENTIAL OBJECT MARKING ("differenzielle Objektmarkierung" in the German original). Recent years have seen a boom in research on differential case marking so that parallel terms were formed for other roles and grammatical relations: Hoop and Swart (2008) seem to be the first to speak systematically of differential subject marking (DSM), Fauconnier (2011) coins the term differential agent marking (DAM), and Kittilä (2008) even speaks of differential goal marking. Iemmolo (2011) treats agreement analogous to case marking and consequently speaks of differential object indexing (DOI) in cases where objects can be indexed or not. Together with these extensions, differential marking has started to gain the status of an independent typological concept.

Note, though, that the idea has been around for a long time. For instance, Kellogg (1875 [1972]:101) notes the following about case marking in Hindi:

> "The accusative appears in Hindi under two forms, the one identical with the nominative, the other consisting of the noun in its oblique form with the appended postposition को. In this last case, when the accusative is the object of a transitive verb, को is incapable of translation, and merely gives a certain definiteness of the noun. (...) को is also used as the postposition of the dative, when it is always rendered 'to.'"

Since this is a grammar written in the Graeco-Roman tradition, the concepts of case and role are not fully separated yet. The case names used by Kellogg are rather similar to modern roles: his accusative corresponds to P and his dative to G. If one takes this into account, the quotation above is clearly one of the first descriptions of DOM in Hindi. Differential marking is thus not a radically new concept – the idea that one and the same "thing" can be marked in different ways is an old one.

One very basic question that has to be asked at this point is why one would consider two different forms as referring to the same thing at all. For Kellogg the answer is clear – accusative and dative are part of a universal grammar defined by the classical languages, so Hindi must have them, too. Modern linguistics does not have such restrictions any longer, so a different kind of answer is required.

For typology this is a simple, practical issue: decomposition is a precondition for comparing languages. For instance, if one claimed that Hindi को (*ko*) had a single function, that would make the description of Hindi more concise but would make it at the same time impossible to compare the way "objects" are marked in Hindi with other languages where P and G are always marked differently from each other.

Apart from this, however, there is also a more theoretical reason why it is possible to assume that differential marking patterns indeed involve one and the same thing marked by different forms. In many cases, it is simply not possible to find a single condition that is both necessary and sufficient for the occurrence of the forms involved in differential marking. The New Indo-Aryan case markers including the Nepali dative *-lai* are a good example for this: so far no serious grammarian has been able to give a unified characteristic of all arguments marked by NOM on the one hand and all marked by DAT on the other, so it is still easiest to classify them on the base of roles, even if those roles are not consistently linked to a single case. Role may then be said to be one of several conditions on case in these languages.

A situation where one function corresponds to various forms is a necessary constituent of differential marking. However, it is not yet sufficient – at least not if the term is to be of any descriptive use. For instance, definite and specific indefinite objects in English are marked by the definite and indefinite articles, respectively, yet nobody would say that English has differential object marking. The reason for this is not that the function of the article is relatively easy to identify – there are, for instance, DOM systems exclusively based on the similarly easy to recognise factor of animacy (Malchukov 2007). Rather, it is that the English definite articles are not restricted to objects but can be used on all NPs. Thus, we will only identify a pattern as differential marking when it is restricted to certain conditioning values without on the other hand being fully determined by them.

There is yet another thing that needs to be added to a satisfactory definition of differential marking. In German, only singular masculine nouns have an accusative – all other nouns are marked by the nominative in P (or, as school grammar would put it, their accusative equals the nominative). Thus, there is an alternation of forms (NOM/ACC) that is linked to conditions (role, gender) and restricted by one of them (role must be P) – yet German is not usually recognised to have DOM. This is because the second condition, gender, is not what is normally called a function but a lexical parameter that is not actively chosen by the speaker but comes packaged with any chosen noun.

The following definition summarises the thoughts from above:

> Let there be a function F, and let two or more values $V_{1...n}$ of F be associated with two or more markers $M_{1...n}$. Let M be found only or at least characteristically with V, however, without V being sufficient for predicting M. Then if one or several additional functional conditions $C_{1...n}$ can improve the prediction of M, F and V will be said to be differentially marked.

For instance in the case of Nepali DOM, F is argument role and V_1 are various object-like roles (mostly P or T). These roles are associated with two markers, *Ø* [NOM] and *-lai* [DAT]. The markers *-Ø/-lai* in this combination are highly characteristic of object-like roles, but no role is sufficient for predicting them. Instead, additional conditions such as animacy or topicality are required to predict the use of *-Ø/-lai*. Therefore, argument role (and more precisely, P and T) may be said to be differentially marked in Nepali.

It should be understood that this definition does not try to capture the "essence" of differential marking – it simply formalises what seem to be some tacit assumptions behind the present use of the term (as in "differential object marking", "differential subject marking" etc.). The definition is hoped to be useful in making it possible to call similar phenomena by a common name.

Another important point about the definition is that it creates a continuum between differential marking and other phenomena. For instance, differential marking is less typical when the set of alternating markers M is not restricted to or less characteristic of V, or when the functional conditions C are less open to active choice.

1.3.2 Differential marking of and conditioned by arguments

So far we have been talking about differential marking in a very general way. We have noted that although the idea of differential marking has been around for a long time, it has gained the status of an independent theoretical concept only recently. The definition of differential marking given in the last section is broad enough to cover all kinds of phenomena – one could, for instance, speak of differential tense marking in cases where tense markers interact with mood, aspect, and polarity. Here, however, we are rather interested in the kind of differential marking patterns that gave rise to the concept in the first place – that is, differential object marking and its extensions to other roles and marking mechanisms. These patterns may be summarised under the term of DIFFERENTIAL ARGUMENT MARKING.

In terms of the definition above, differential argument marking can be viewed as a type of differential marking where F is argument role and $V_{1...n}$ are individual roles (possibly clustered with

non-functional factors such as noun and verb class) that share an alternation in formal marking. Most existing terms mentioned above (DSM, DAM, differential goal marking) specify further which role or set of roles is differentially marked but do not talk about the nature of the markers $M_{1...n}$. Instead, case marking is assumed as the default M for roles. The only exception is Iemmolo (2011), who speaks of DOI (differential object indexing) parallel to DOM, thereby extending the range of M from dependent to head marking.

This is an important point. Ever since Nichols' (1986) groundbreaking paper on head-marking and dependent-marking grammar it has been clear that roles (as well as other functions) can be marked on dependents (NPs occupying a role) as well as on heads (predicates defining a role). Thus, there is no *a priori* reason to ignore differential argument marking on heads.

But one could go even further. The difference between F (the function that is differentially marked) and $C_{1...n}$ (the additional conditions working together with F to determine M) depends on one's viewpoint. From a more abstract perspective, both F and C are nothing but conditions on the form of M. This means that if we take the definition above seriously, we will not only have to include patterns like DOM and DOI under the heading of differential argument marking but any patterns where roles feature among $C_{1...n}$.

This is a rather radical view since it includes patterns where traditionally one wouldn't say that they mark an argument. A case in question are antipassives. If there is an overt marker of diathesis one might say that the antipassive is marked on the verb, but it seems impossible in present terminological tradition to say that the antipassive marks an argument – yet many antipassives have properties of an object such as specificity as their most important condition (Cooreman 1994). The conceptual twist involved here is not trivial. To me the reason why it seems odd to say that an antipassive marks an object seems to be that an antipassive does not indicate which referent is the object. A case marker does so by being adjacent to an NP, and agreement does by indexing properties of the referent such as person, gender, or number. Neither of these can be said of an antipassive.

However, closer inspection reveals that the distinction is not at all clear-cut. While case markers are probably the most watertight method of marking roles, there are cases where several distinct arguments are marked by the same marker or where the interpretation of a marker depends on the verb class. With agreement, the possibility of ambiguity is even more obvious: it arises as soon as several arguments have identical indexed properties, e.g. in the case of a 3s>3s scenario in a language indexing person and number. Compared to this, antipassives do not seem to do a much worse job at pointing out object referents. For instance, the West Greenlandic antipassive has been variably described as being conditioned by the givenness, definiteness, or the scope of the object (Bittner 1987). While in the normal transitive construction A is marked by ERG and O by NOM ("absolutive"), A is marked by NOM and O by INST in the antipassive:

(4) a. *Jaaku-p ujarak tigu-a-a.*
Jaaku-ERG stone take-IND.TR-3s>3s
'Jaaku took the stone.'

b. *Jaaku ujaqqa-mik tigu-si-vo-q.*
Jaaku stone-INST take-AP-IND.ITR-3s
'Jaaku took a stone.' (Bittner 1987:1)

Most verbs require one of several suffixes for antipassivisation. Thus, when the hearer detects one of these suffixes, that helps him to resolve role distribution – otherwise the only means of knowing which role NOM marks are cotext and context. When the argument that is not marked by NOM is covert (*ujarak tiguaa* 'he took the stone', *Jaaku tigusivoq* 'Jaaku took (something)'), the antipassive becomes even more important because it tells the hearer that the present NOM-marked argument can only be P if it has the required semantics (given/definite/wide scope).

Thus, antipassives seem to be functionally similar enough to differential case marking and differential indexing to classify them as another subtype of differential argument marking. The easiest criterion for separating this type from the other two is that it involves markers in several places. We will not assume that the presence of verbal markers or of markers on the conditioning argu-

ment are constitutive for this type. Consider again the Chintang example in (5), repeated from above:

(5) a. *Rakkas-a-ce-ŋa ma?mi u-c-o-he.*
ogre-NTVZ-ns-ERG person 3pA-eat-3[s]O-IND.PST
'The ogres ate somebody.'
b. *Rakkas-a-ce kok u-ci-e.*
ogre-NTVZ-ns rice 3pS-eat-IND.NPST
'The ogres had rice.' (elicitation RBK 2012)

The differential marking pattern found here is very similar to the antipassive in (4) both functionally and structurally. However, the case marking of P is the same (NOM) in (5a) and (5b), and there is no dedicated verbal marker in either case. An intermediate case is found, for instance, in the antipassive of Kalkatungu (Blake 1979, Isaak 1999), where there is also no verbal marker but the case frame alternates between {A-ERG P-NOM} and {A-NOM P-DAT}.

I will refer to patterns like these where M in several loci are bundled as DIFFERENTIAL FRAMING. There is a great number of differential argument framing patterns beside the antipassive – basically, this term covers everything that is more traditionally known as an alternation, and a couple of more phenomena, for instance, all diatheses, ambitransitivity, and reflexivisation, but also noun incorporation, cases of coupled differential case marking and indexing, and the pattern found in Chintang.[4]

So far we have identified differential argument marking as a type of differential marking and have further subdivided this type into differential case marking, differential indexing, and differential framing. At this point, we have to get rid of a terminological problem. "Differential argument marking" is already slightly ambiguous – it would normally be taken to refer to differential case marking only. This problem is more pronounced with "differential object marking", which is exclusively reserved for differential case marking for historical reasons. I will therefore keep this term in its usual meaning and instead use OBJECT-CONDITIONED DIFFERENTIAL MARKING as a cover term for DOM, DOI, and differential object framing. This difficulty also explains the title of the present work.[5]

1.3.3 Differential marking conditioned by objects

Although it is useful to define differential argument marking and its subtypes in a more general frame, the present study is interested in only one type of arguments, namely OBJECTS. We will thus first have to define what we mean by this term and then make some comments on specific properties of object-conditioned differential argument marking.

Although the discussion of objecthood has never been as intensive as that of subjecthood, there nevertheless is a large body of literature on the topic and a great degree of variation in the use of the term. As noted by Plank (1984:vii), nothing much is agreed upon except that objects are not subjects, and that is not much given that subject is a highly controversial category. A lot of basic publications on grammatical relations presuppose a loose understanding of objecthood without defining it at all (see for instance Dowty 1991, Müller-Gotama 1994, Ackerman and Moore 2001, Swart 2007). I will not indulge in searching for the true meaning of the term here but use it in a rather special sense which is most apt for the purposes of the present work:

> Object is a grammatical relation covering one or more semantic roles except S or A which is defined by a specific differential argument marking pattern.

[4] Many of these have been compared before – cf. for instance, Lazard (2001) on parallels between DOM, DOI, incorporation, and antipassives, or Kulikov's (2011) equation of diathesis and ambitransitivity. However, so far no comprehensive typological treatment of differential framing seems to exist (not to speak of an even wider perspective that would include other kinds of transitivity-related alternations such as differential case marking and differential indexing).

[5] An alternative would have been "differential object coding". However, this term also seems awkward with patterns such as antipassives of which in standard terminology one wouldn't say that they code objects.

What makes this definition special is the last clause, since it excludes many arguments that would be called object in standard usage. For instance, *sheriff* in *I shot the sheriff* will not be referred to as an object by default here but only when considering the differential marking patterns it participates in, such as the conative alternation (*I shot at the sheriff*) or the passive (*The sheriff was shot*). This usage is admittedly peculiar but very practical for the present purpose of investigating object-conditioned differential marking patterns since it allows to refer in an easy way to whichever grammatical relation is defined by a pattern.

The two objects that will be mentioned most frequently in the present work are the one defined by S/A detransitivisation in Chintang (see section 2.4.2) and the one defined by Nepali DOM (see section 3.4.2), briefly also "the object in Chintang" and "the object in Nepali". In addition to the term object I will also use the letter O as a shortcut (not to be confused with P, T, G, which may all coincide with O but basically denote roles).

Research on object-conditioned differential marking has concentrated in two areas. Investigations of relevant patterns in individual languages have been focussing on their language-specific conditions $C_{1...n}$ (but not on the meta-question of how to relate these conditions to each other, see section 1.4.3 below). By doing so, they have also increased the typological inventory of potentially relevant functions and made it more precise. This area is relevant to the present study insofar as it provides inspiration – conditions that are relevant elsewhere could also be relevant for the languages investigated here. Works on the syntax of Nepali and other Indo-Aryan languages as well as on Chintang and other Kiranti languages have therefore made an important contribution to this work.

The other area is typological work. The subtype that has attracted most research here is differential object marking, which also started the history of differential marking as an independent concept. Typological research in DOM almost always contains an additional component that asks why DOM is there or what its ultimate function is. An excellent overview of this debate is given in Iemmolo (2011:25ff.), where two main types of approaches are distinguished. "Distinguishing" approaches claim that DOM serves to disambiguate role in cases where otherwise several arguments could be easily interpreted as subject or object, whereas "indexing" approaches view DOM as a means of marking salient properties of objects such as high animacy. For the present study I will not subscribe to either of these in order to keep all analytical possibilities open. What's more, I don't believe that the two functions do necessarily exclude each other. For instance, as will be shown in section 3.5, indexing best summarises the function of DOM in Nepali, but disambiguation is also relevant (see section 3.5.12).

In addition to this, there are several other theoretical decisions that cut across the problem just mentioned and that are also relevant for other types of object-conditioned differential marking. One that has deep consequences is the use of referential hierarchies. Referential hierarchies have featured prominently in linguistics ever since Silverstein's (1976) seminal paper and are also frequently employed in research on object-conditioned differential marking – cf. for instance Bossong's (1998) "dimensions" of "inhérence" and "référence" (corresponding roughly to animacy and identifiability) or the various scales in Aissen 2003. Since the universality of such hierarchies has been called into question (Bickel 2008c), I will not assume any of them prior to the description of the languages that are of interest here.

Another important theoretical decision is whether to put one's focus on abstract functions (such as disambiguation or highlighting) or concrete functions (e.g. specificity, definiteness). Although I must admit that abstract functions are ultimately of greater interest because they offer generalisations, I would like to argue for a "concrete first" approach, especially in the description of language-specific phenomena: any description of the function of a marker should first try to get as close as possible to an ideal situation where the description is both necessary and sufficient, or in other words, where it predicts all instances of the marker without overgeneralising. If this imperative is not followed, one easily gets into situations where one misses patterns in the data or, even worse, confirms a theoretical preconception based on itself. In the present study, the functions of S/A detransitivisation in Chintang (Section section 2.6) and DOM in Nepali (Section section 3.5) will therefore first be described in detail. Summaries are given at the end of the relevant chapters

(section 2.8, section 3.8), and commonalities and differences on an abstract level are given in the conclusions (section 4.1).

1.4 Analytical questions

1.4.1 Description versus explanation

Description and explanation are controversial concepts in linguistics. While modern linguistics started out with the descriptive framework of Structuralism, the next big paradigm, Generative Linguistics, was explicitly explanatory (Dryer 2006). Whereas explanation is the ultimate goal of linguistics in most contemporary theories of language as well as in typology, descriptivism has a strong stance in work on individual languages, especially in grammar writing. What's more, there are influential descriptivist frameworks such as Documentary Linguistics (Himmelmann 1998, Woodbury 2003) or Basic Linguistic Theory (Dixon 2010). The problem is also relevant for the present study because it delimits its possible goals. Are the phenomena in question to be described or should they be explained?

This question presupposes an understanding of what is meant by "describe" and "explain" which I think doesn't exist in linguistics. Although there are a few articles that explicitly address the question of description vs explanation (Frawley and Golinkoff 1995, Haspelmath 2004, Dryer 2006), none of them defines these terms.

A word that often falls when explanation is mentioned is *why*, and this reflects the everyday understanding of the terms, where a description is a mere representation of a state of affairs, whereas an explanation looks at its broader background, too. But what does it mean to ask *why*? An old answer that I still find very convincing comes from David Hume, who states in his Treatise of Human Nature (Hume 1739 [2003]) that the perception of a causal relation between two phenomena requires that they are contiguous in time and space, that the cause take places prior to the effect, and, most importantly, that the effect follows necessarily from the cause. Hume's idea of necessity is based on co-occurrence: he claims that human beings perceive things as linked by necessity when one of them never occurs without the other.

This definition can be taken as the base for a more precise definition of explanation. If an explanation asks why a phenomenon (an effect) is there, we may now say that it tries to discover another phenomenon (a cause) with which it necessarily co-occurs (or, put the other way round, without which it does not occur). A description is then any other approach to understanding a phenomenon that does not look at co-occurrence.

Let's consider a concrete case. As we will see later (section 3.5), DOM in Nepali is rather complex in being based on a whole range of functional factors. Now assume that we want to know something about this pattern: when is the nominative used on objects, and when the dative? The simplest way to answer this question would be to collect all corpus sentences with O-NOM in one place and all sentences with O-DAT in another – or even simpler, to take down the numbers of the relevant sentences in the corpus. That would give us not only an accurate description of the distribution of NOM and DAT on O in the corpus but also an inventory of possible sentences[6] that could be re-used in other contexts. But even though a collection of this type represents the prototype of a description as just defined, most linguists would probably agree that it is rather restricted and does not deserve to be called even a description.

What would be the next step in our analysis? Obviously it is not only desirable to know how O-NOM and O-DAT are distributed in our sample (the corpus) but also in the population (the language). In order to say something about this it is no longer sufficient to list numbers of sentences – we would now like to predict the case of O in an unattested sentence (or actually, in every possible sentence). For this we need to relate the attested to the possible, which in this case can be done via the functional factors correlating with DOM. We will thus try to list these and to analyse their

[6] More precisely we should not speak of sentences but of paragraphs centered around a sentence, because some factors influencing case such as topicality are clearly suprasentential. However, speaking of sentences should suffice in the present hypothetical situation.

interplay. The output would then by many be called a functional description of Nepali DOM. However, according to the definition above this would already clearly count as an explanation, because we have asked *why* NOM or DAT is where it is in general.

Objections to this are easy to imagine: an analysis of this kind does not yet explain why DOM is there at all in Nepali, or why DOM is there at all in the languages of the world, and since these questions contain the greater potential for generalisation, only an account that addresses them should be viewed as truly explanatory. However, obviously there is no straight way to determine which answers are general enough to be considered an explanation and which are so specific that they are "only" descriptions, so I suspect that there often is a hidden criterion for distinguishing between these two, which is personal knowledge and interests: a description contains only the obvious things one already knows, whereas an explanation gives new answers to important questions.

It thus seems hard to establish an objectively motivated cut-off point between description and explanation that at the same time matches our common understanding of these terms. If we use an objective definition like the one presented above, almost everything ends up as explanation except the described option of simply listing all observed phenomena, which is of little practical relevance. Of course it cannot be denied that there is an important difference between questions like "Why are certain O in Nepali marked by the dative?" and "Why do certain O trigger differential marking in many languages?" (as well as between the corresponding answers), but this difference is gradual.

The conclusion for the present study is that there doesn't seem to be great benefit in asking whether an analysis is descriptive or explanatory. Rather, it should be asked what its scope is, or put differently, *to what extent* it is explanatory. This question is easily answered for the present work: it seeks in the first place to explain when S/A detransitivisation in Chintang and DOM in Nepali are used, that is, it is concerned with language-specific phenomena. Apart from that, it may also make a small contribution to the bigger question of how object-conditioned differential marking works in general.

1.4.2 Functions versus conditions

In our definition of differential marking in section 1.3.1, we made a distinction between a differentially marked functional value V and additional conditions C that must both be considered to explain the distribution of a set of markers M. There is the question of how this distinction is motivated.

Superficially V and C look similar – both are ultimately nothing but conditions on the appearance of M. When looking at individual markers, neither of them has to be completely necessary or sufficient. For instance, the Nepali dative is neither found only on P (= one of V) or specific referents (= one of C) nor on all P or all specific referents. One might conjecture that only V is completely necessary for the *alternation* of $M_{1...n}$. For instance, the NOM/DAT alternation in Nepali at first sight looks as if it was only found on P. However, if one takes a closer look it soon turns out that this is not true – NOM/DAT is not associated with the role P but with the grammatical relation O, which is trivial since the very definition of O is based on this alternation. Thus, there seems to be no independent method for determining V.

But this is again not the whole truth. In Nepali, precisely the same arguments that allow the NOM/DAT alternation can also acquire subject-like properties in passives. In Chintang, the same arguments that trigger S/A detransitivisation also trigger O-AGR of various forms. In other words, in both languages in question O is defined by several constructions. Thus, even though O cannot be determined completely independently, it is also not just an arbitrary grouping in the eye of the beholder.

Further, although it is impossible to say that P is necessary for NOM/DAT and information-free to say that O is necessary for NOM/DAT, it is possible to say that either P or T or G is necessary for NOM/DAT, and that is still much more than in the case of any C, where one can only make trivial statements like "either specificity or non-specificity is necessary for NOM/DAT". Put differently, the NOM/DAT alternation (and S/A detransitivisation alike) is much more constrained by role than by anything else.

But role is not only more relevant for the alternations in question but also in the whole language system. In both Chintang and Nepali, roles are important co-determinants of case marking and agreement. By contrast, most of the conditions that co-determine differential marking are completely irrelevant for the rest of the language, that is, it is much easier to assume that they are not marked at all than that they are zero-marked. The few conditions that are relevant elsewhere manifest themselves in different shape. For instance, topicality in Nepali is relevant both for DOM and for word order. However, since these are two very different mechanisms, it is still reasonable to assume that they constitute distinct marking systems. We may thus say that $C_{1...n}$ are only relevant within the frame given by V.

To summarise, there seems to be a base for the intuitive separation of V and C, at least for the two languages in question: V is relevant in the whole language system, comes closest to being necessary for the alternation $M_{1...n}$, and is (as a set) also defined elsewhere in the language.

It is a different question whether there is one dominant condition within C. Most publications on differential marking patterns in individual languages implicitly claim this, e.g. by starting the discussion with one condition to which most space is dedicated. There are two ways in which dominance can be defined here: statistical relations may hold between a condition and the alternation or between a condition and other conditions. For instance, the quantifiability of referents in Chintang is highly relevant for their marking because most quantifiable O are used with the transitive frame (relation to alternation), but also because other distinctions such as specificity are only relevant when quantifiability is given (relation to other conditions).

There is, however, obviously no way of determining when a relation is strong enough for C_1 to be considered dominant. If we assume for a moment that the values of C_1 in an arbitrary system would fully predict M, probably everybody would agree that C_1 is dominant in that system – or rather, that it is the only factor that needs to be described. However, this is only the extreme end of a theoretical continuum whose other end is zero relevance. I will try to show in the language-specific parts of this work that this continuum is also relevant in practice: while it is relatively easy to determine a dominant condition for S/A detransitivisation in Chintang, it is very hard for DOM in Nepali, so that it is better there to simply quantify the importance of the individual factors instead of mapping it to a binary distinction of dominant vs ancillary.

The intuitively appealing concept of dominant conditions also seems to be relevant for an important term in this context, namely MARKING. There is a tendency in the literature to reserve this term for dominant conditions. For instance, the Nepali dative marker *-lai* could be said to mark specificity as one rather dominant C, but it would be odd to say that it marks affectedness, which is very likely to be involved in DOM but marginal as compared to the other conditions. From what was said above about dominance being a continuum, it follows that this specific use of "marking" does not make much sense for the present work. I will therefore use the term here in a simple way: in a concrete utterance, all information is considered marked that is associated with a form. Thus, *-lai* may equally mark role, specificity, and affectedness. What *-lai* marks *in general* is a different question but can be answered on the same base: a piece of information is the more integral to the function of a marker the more often it is associated with it in concrete utterances.

1.4.3 Modelling grammatical decisions

A large part of the core of linguistics is concerned with describing grammar in individual languages and in general. Views on which principles such descriptions should follow and what structure it should have vary greatly across time and theories. A descriptive goal that is abstract enough to be common to all linguistic theories is the goal of observing associations between linguistic *signifiants* and *signifiés*. This rather abstract definition leaves a lot of space for theoretical variation: for instance, *signifiants* may be small and monolithic (morphemes, words) or large and discontinuous (constructions), *signifiés* may be located in the "language system" or in the mind and may be formal (e.g. abstract syntactic structures) or functional (e.g. grammatical semantics), and associations between the two may be anchored in competence or in performance. All these differences are ignored for the moment.

One important question that can be asked at this abstract level is what logical relations hold between an associated *signifiant* and *signifié*. The simplest possible answer is that there is a 1:1 relation where *signifiant* and *signifié* are both necessary and sufficient for each other. This seems to have been the intuition behind Saussure's famous egg-shaped diagram representing the linguistic sign (Figure 1.1).

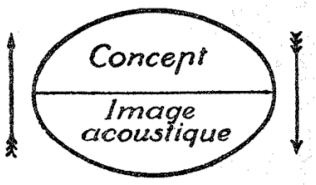

Figure 1.1: Saussure's conception of the linguistic sign (Saussure 1915 [1975]:99)

Now it is a truism that 1:1 relations are an ideal – there are plenty of dedicated terms such as "polysemy", "allomorphy", or "multi-word expression" that describe various well-known cases where one *signifiant* may correspond to several *signifiés* and vice versa. Nevertheless, this ideal hasn't lost attractivity, and most modern linguistic theories seem to strive to maximise the number of 1:1 relations in their descriptions of language and especially grammar.

One subtype of 1:1 relations that is of interest here and that seems to have been rarely questioned so far is the relation between a set of functional conditions and a form in a concrete utterance. It is usually (implicitly) assumed that once a speaker has decided about what he wants to say and once one knows everything about the cotext and the context, the form of the utterance he will make can be fully predicted.[7]

Of course, it is theoretically plausible that a set of conditions given to a speaker should produce a unique output – if that wasn't the case one would have to assume that speakers resort to some sort of random mechanism in order to produce varying outputs in spite of exactly the same input. The point about this subtype of 1:1 model that I would like to criticise here is rather that it assumes that it is in principle possible to *know* all relevant conditions. For one thing, this is theoretically unlikely from an inductive perspective, given all that we know about the history of sciences dealing with complex systems such as language. But practically it is completely impossible for a number of reasons. First, there are way too many variables and values at play to identify all of them even in a single case. Second, many of the relevant variables are rather elusive because they are difficult to measure or quantify or because their values can change depending on the measuring method and even the mind of the observer.

The conclusion from this is simple enough: no description of the function of a linguistic form should consider itself complete, and instead of implicitly claiming that the function predicts the form by 100%, it should make explicit what impact the function has on the form. This is best done by quantifying the impact, which usually results in a probabilistic model of the distribution of the form in question.

The use of statistics is already quite widespread in subfields of linguistics that are in close contact with other disciplines, such as sociolinguistics, psycho- and neurolinguistics, or computational linguistics. However, it hasn't spread so far into the core of language description (Abney 1996). The only publications that I am aware of in this area which make use of probabilistic models are Williams (1994), Wulff (2003), and Bresnan et al. (2007) (who also gives an impressive list of further arguments in favour of such models, see p. 70 ff.). This is a pity given the obvious usefulness of statistics in science in general.

Of course, probabilistic models of grammatical phenomena require greater corpora and more analytical work than absolute (let alone monocausal) analyses. For this reason it is impossible to provide sophisticated probabilistic models for even a few grammatical phenomena in an ordinary

[7] The reverse statement that it is possible to fully predict the function of a given utterance for some speaker is a bit more problematic because there may be ambiguities in the input, but if one assumes that these can in most cases be resolved then that statement is also possible. For the sake of simplicity, however, I will only speak about the former case below.

reference grammar. On the other hand, that's not necessarily what follows from the imperative to specify the impact of a functional variable one uses for describing/explaining the distribution of a form. In many cases simply mentioning that a set of variables does not explain everything and giving a rough, subjective estimate of how much it explains may already help. For instance, an endless row of grammars and articles (e.g. Kleinschmidt 1851 [1968], Kalmár 1979b,a, Johnson 1980, Fortescue 1984, Bittner 1987, Bok-Bennema 1991, Bjørnum 2003, Sadock 2003, Schmidt 2003) have been concerned with the function of the antipassive in Eastern Eskimo, one example of which was given above in (4). One of the reasons why not much progress can be seen in this area is that every proposal in this row views itself as absolute and must therefore reject all others.

The present study is detailed enough to try to incorporate statistics into the offered explanations of S/A detransitivisation in Chintang and DOM in Nepali based on large corpora (see section 0.4, section 3.6, section 2.7).

Chapter 2

Chintang: S/A detransitivisation

2.1 Language background

Chintang [ˈʈʂʰiɳʈaŋ] is a Kiranti language spoken by about 4000 - 5000 speakers in Eastern Nepal (Kośī zone, Dhanakuṭā district, Chintāṅa VDC). The maps in Figure 2.1 and Figure 2.2 show the location of the language area within Nepal and within Dhanakuṭā district.

Figure 2.1: Location of Chintāṅa VDC within Nepal (United Nations Cartographic Section 2007, accessed on 1 November 2012)

The name of the language is derived from Chintāṅa, the name of the Village Development Committee where it is mainly spoken (hence simply "Chintang"). In Nepali it is more commonly referred to as Chintāṅge Bhāṣā, which literally means 'Chintangish language'. Chintāṅge alone is also possible, and both variants are commonly spelt <Chhintang> and <Chhintange> when using Roman letters.[1] The speakers themselves prefer the less technical term *anirɨŋ* 'our language'.

[1] [ʈʂʰ] in Nepalese languages is commonly transcribed <chh> in non-linguistic usage.

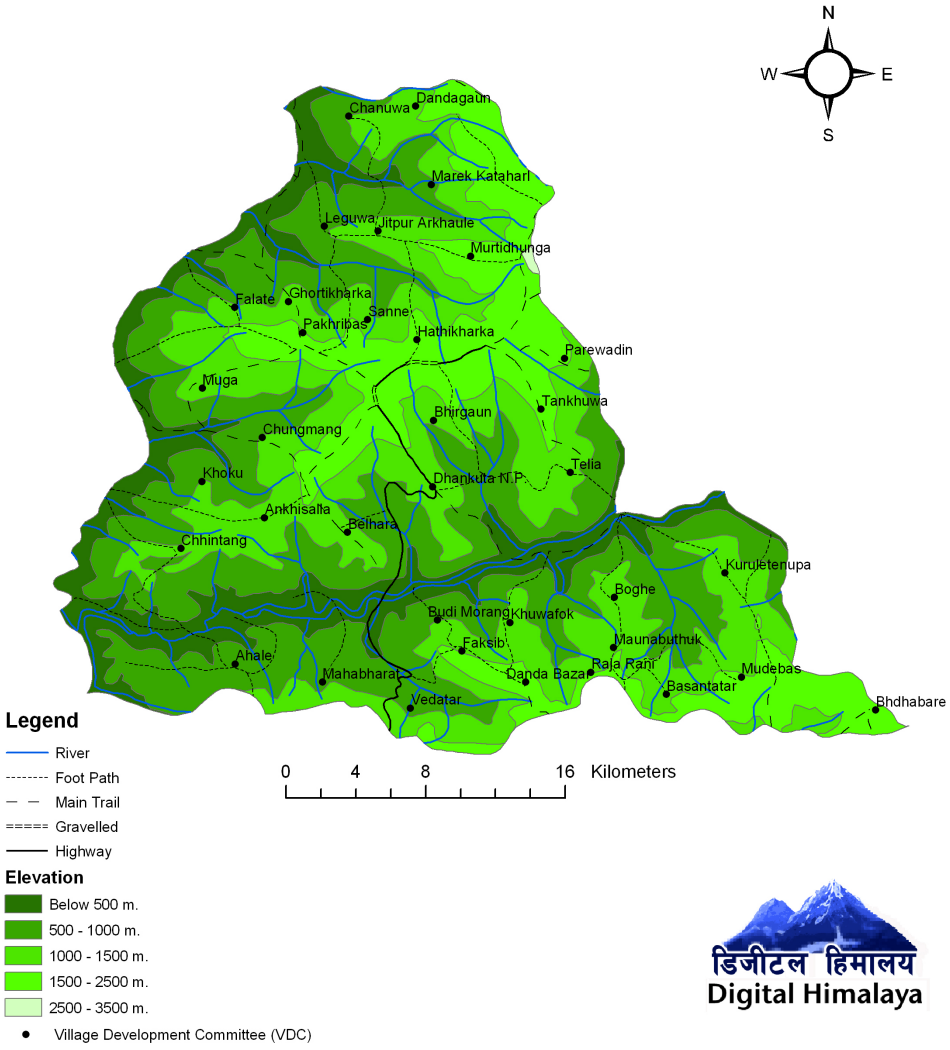

Figure 2.2: Topographical map of Dhankuṭā district with Chintāṅa and Āhāle VDC in the southwest (Joshi 2012, accessed on 1 November 2012)

Within Chintang most speakers are found in wards 1 to 5. Apart from Chintang VDC, the language is also spoken in by a few speakers in the neighbouring VDC Āhāle.

There are no reliable data concerning the number of speakers. The number above is an estimation based on the number of people living in Chintang (about 8000 – 10,000) and the statements of speakers and other researchers working on the language, in particular Netra Paudyal and Balthasar Bickel. Most speakers are bi- or trilingual, with Nepali (Indo-European > Indo-Aryan) as one and Bantawa (Tibeto-Burman > Kiranti > Central Kiranti) as the other additional language. Monolingual speakers can still be found, especially among elderly women.

Using the criteria for describing language endangerment from the UNESCO's Language Vitality and Endangerment framework (UNESCO Ad Hoc Expert Group on Endangered Languages 2003), Chintang gets the average vitality value 2.17 (values range from 0/extinct to 5/safe). Table 2.1 shows the individual values this is composed of.

Genetically Chintang is a Kiranti language. The Kiranti languages are generally accepted to belong to the large Tibeto-Burman family, although their position inside this family is being disputed (cf. Ebert 2003:516). Within Kiranti, Bickel (2008a:3) identifies Chintang as Central-Eastern >

variable	value	comment
language transmission	3	Many but not all children learn the language, and the tendency is going down.
absolute number of speakers	3	The majority of speakers live in a single VDC, and the population is definitively too small to make itself be heard on a national level.
relative number of speakers	3	The majority of the (adult) inhabitants of wards 1 to 5 of Chintang speak the language.
domains of use	3	The language has a strong standing in homes, the traditional economy (agriculture), and religion, but is rarely used in politics and trade and never in education.
new media	0	The language is not used in any media, be it books, newspapers, radio, TV, SMS, or the internet.
education and literacy	1	A practical orthography has been established with the publication of the Chintang dictionary (Rāī et al. 2011), but apart from that no other texts have been published, nor are orthography and grammar taught in school.

Table 2.1: Endangerment of Chintang

Greater Eastern > Eastern > Greater Yakkha. Despite its small size, Chintang itself is not internally homogeneous. Lexical and morphological differences can be observed between varieties spoken uphill and downhill as well as between eastern and western varieties. For instance, *pukt-* 'begin' is only used in higher regions (*puŋs-* being preferred elsewhere), the negative past marker *-t* is used in the western half of the village Mulgāuṃ and in Sāmbugāuṃ, and the imperfective marker *-k* is only used in Sāmbugāuṃ. Speakers tend to identify two major dialects, Mulgāuṃ and Sāmbugāuṃ, but this distinction seems to have an ideological rather than a linguistic base, since in general the influence of Bantawa is stronger in Sāmbugāuṃ and other settlements farther to the west. It is not clear whether the mentioned and other criteria form dialect clusters at all. Fortunately, so far no syntactic differences have been observed, so the question of dialects is only a marginal concern for the present work.

Ethnically the speakers of Chintang identify themselves as Rai, a group that comprises the speakers of most Kiranti languages but excludes the big languages Yakkha and Limbu and also Sunwar. When asked to which group they belong within Rai, speakers usually answer that they are Bantawa. Bantawa is at the same time an ethnonym and the name of the associated language, which is spoken by many people living in Chintāṅa and is even dominant in some western parts of the VDC that lie close to the core language area. However, speakers of Bantawa normally do not refer to Chintang speakers as Bantawa but call them *chendaŋpaci*, literally 'Chintang people'. These days some speakers also refer to themselves as Chintang Rai rather than Bantawa Rai. This may be due to the increased ethnic consciousness in present-day Nepal and the knowledge that Chintang is not a dialect of Bantawa that has kept spreading in the VDC ever since the beginning of the Chintang and Puma Documentation Project.

2.2 Overview of relevant morphology

2.2.1 Parts of speech

Establishing parts of speech in Chintang is comparatively easy due to its wealth of inflectional morphology. Three criteria are sufficient for distinguishing 13 parts of speech:

- dependency: does a form belonging to a part of speech require another, separate form?

- inflection: which inflectional categories does a part of speech have, and how are these realised?
- syntactic use: for which syntactic macrofunction is a part of speech typically employed?

Table 2.2 shows an overview of the part of speech system.

	dependency	inflection	syntactic use
verb	no	S/A/O agreement, TMA, polarity	predicate
noun	no	possession, two numbers, case	referent
adjective	nominaliser	(determined by nominaliser)	qualification of referent
pronoun	no	three numbers, case, clusivity	deixis to SAP
demonstrative	no	two numbers, case, distance from origo	deixis to NSAP
numeral	no	classifier, case	quantification of referent
adverb	no	no	modification of predicate
verboid	no	no	predicate
interjection	no	no	equivalent to clause
particle	any other word	no	grammatical
nominaliser	any other word	two numbers, case	referent
affix	specific p.o.s.	no	grammatical
vector verb	verb	S/A/O agreement, TMA, polarity	grammatical

Table 2.2: Chintang parts of speech

There are two cases where the part of speech labels chosen here deviate slightly from what is commonly understood by them:

- Adjectives in Chintang are peculiar in that they are obligatorily nominalised. This is possible because of the special properties of nominalisation in Chintang (and other Tibeto-Burman languages, see e.g. Matisoff 1972, Genetti 2011), one of which is that all nominalised forms can be directly used to modify other constituents. An example would be *the=go ma?mi* [big=NMLZ$_1$ person] 'big guy' (but also only *the=go* 'big one').
- The label "pronoun" only refers to forms pointing to speech act participants. Third person deixis is functionally similar but morphologically distinct in Chintang so that the corresponding forms must be considered a part of speech of their own ("demonstratives").

Nouns, adjectives, pronouns, demonstratives, and numerals share the important characteristics of being usable as arguments without further marking and of taking case suffixes. They can therefore be subsumed under the label "nominals" where necessary. The parts of speech that are of interest for the present work are verbs and nominals, in particular nouns.

2.2.2 Nominal morphology

There are two nominal inflectional categories that are relevant for the study of S/A detransitivisation. One, case, is shared by all nominals. The other, number, can be marked on all nominals except numerals. Number precedes case marking.

There are two numbers, an unmarked singular and a non-singular marked by *-ce*:

(1) a. *Ba=go cha ghãsa hek-ni?-niŋ.*
PROX=NMLZ$_1$ child grass cut-IND.NPST[.3sS]-NEG
'This child doesn't cut grass.' (CLC:CLLDCh3R08S05.0144)

b. *Ba=go cha-ce=lo aŋ u-num-no? u-yu-ba?.*
PROX=NMLZ$_1$ child-ns=SURP what 3[p]S-do-IND.NPST ACCESS-DEM.ACROSS-LOC$_1$
'What are these children doing over there?' (CLC:CLLDCh1R06S02.0630)

The label "non-singular" is appropriate because other morphological subsystems of the language distinguish three numbers (singular, dual, plural), where the non-singular corresponds to the latter two. One such subsystems are verbs (for which see section 2.2.3 below), the other are pronouns. Pronouns are also special in that they do not make use of *-ce* [ns] at all. Table 2.3 shows the pronominal system.

	s	**di**	**de**	**pi**	**pe**
1	*akka*	*anci*	*ancaŋa*	*ani*	*anaŋa*
2	*hana*		*hanci*		*hani*

Table 2.3: Chintang pronouns

The only nominals that do not mark number are the numerals. Note, however, that two of the three existing native numerals have incorporated what seems to be a cognate of *-ce*: *thitta* 'one', *hicce* 'two', *sumce* 'three'.

-ce can not only mark groups of categorially identical referents but can also be used as an associative plural:

(2) *I-ppa-ce liŋwakha khaŋ-si u-kha?-n-ei.*
2sPOR-father-ns pasture look-PURP 3[p]S-go-[SUBJ.]OPT-ATTN
'Your father and the others should go to take a look at the pasture.' (CLC:CLLDCh2R02S10.143)

Later (section 2.6.3.1) we will see that *-ce* has a special, more narrow semantics when used on objects.

The number of values in the category of case depends on what one counts as case. The working definition used here is that a case is any form which is regularly used to mark a semantic relation between a referent and a predicate. Borrowed markers are included if they occupy a functional niche that did not have a dedicated marker before. The part of speech of the marker is not relevant as long as these conditions are met, so there are both affixes and particles in the class of case markers. The definition also includes markers which are confined to nominal subclasses (*-khi?* [MOD] and its derivates can only be used on demonstratives) and markers which can also be used with verbs (*gari* [TMP.LOC], *khe?ŋa* [TMP.ABL], *likhi* [EQU], *pache* [POST]).

Table 2.4 shows an overview of the 21 cases.

-Ø	NOM	nominative	*-khi?*	MOD	modalis
-(bai)?ni	DIR	directional I	*-lam*	PERL	perlative
-(ba)mu	LOC.DOWN	inferior locative	*-laŋtĩ*	FIN	finalis
-(ba)ndu	LOC.UP	superior locative	*likhi*	EQU	equative
-(ba)yu	LOC.ACROSS	ulterior locative	*-nɨŋ*	COM	comitative
-be?	LOC$_1$	locative I	*-ŋa*	ERG	ergative
gari	TMP.LOC	temporal locative	*pache*	POST	postessive
-i?	LOC$_2$	locative II	*-patti*	LOC4	locative IV
-ko	GEN	genitive	*-sɨrɨŋ*	DIR2	directional II
-kha	LOC$_3$	locative III	*somma*	TERM	terminative
khe?ŋa	TMP.ABL	temporal ablative			

Table 2.4: Chintang case markers

The genitive *-ko* is functionally special in that it only rarely marks relations between referents and predicates but mostly relations between referents. Formally it is special in that a genitive-

marked NP can express a possessum without an overt head following (*ma-ko khim* [woman-GEN house] 'the woman's house', but also *ma-ko* 'the woman's (house)') and that consequently it can be combined with all other case markers (*ma-ko-be?* [woman-GEN-LOC_1] 'in the woman's (house)').

The most important cases for the present study are the nominative and the ergative. The nominative is the default case and is used in too many functions to subsume them under a meaningful label. Some of its most important functions are marking intransitive subjects (S), transitive patients (P) and ditransitives themes (T) or goals (G). The ergative marks transitive and ditransitive agents, instruments, and (combined with locative I or II) sources or objects of comparison.[2] Besides these two cases, the various locative cases (above all locative I and II) are also used to mark argument roles. More detailed information on the distribution of the core cases is given in the overview of syntax in section 2.3.

Cases in Chintang can be stacked in various ways. For instance, locative I and II can be combined with the ergative to mark sources (*khim-be?-ŋa* [house-LOC_1-ERG] 'from the house'), and the altitudinal locatives can be combined with directional I to mark altitudinal directions (*hoŋku-bamu-ʔni* [river-LOC.DOWN-DIR_1] 'down to the river'). Case stacking is not relevant to S/A detransitivisation, so its details can be ignored here.

2.2.3 Verbal morphology

Verbs are characteristically inflected for tense, mood, aspect, polarity, and index person/number/-clusivity of one or two arguments. Agreement is the only relevant category for the treatment of S/A detransitivisation and complex enough, so all others will be ignored here. An overview of the aspectual system is given in section 2.6.4.3 in connection with the question how quantifiability and aspect/aktionsart interact. Inflection paradigms for all finite and non-finite forms can be found in the appendix (section D.1). A sketch of Chintang verbal morphology can also be found in Bickel et al. (2007a), and a more detailed account is given in Schikowski (2011).

The problem of the lack of 1:1 correspondences between form and function that has already been addressed in section 0.2 is especially prominent in the case of agreement affixes. Table 2.5 shows all of them together with their paradigmatic function (i.e. a summary of all functions they carry out in individual paradigm cells) and their slots. Prefixes can be freely ordered (Bickel et al. 2007a) and therefore do not have slot numbers.

a-	2S/A	*-na*$^{+2}$	1s>2
-ce^{+5}	d	*-ni*$^{+5}$	2/3p
-ce^{+9}	3nsO	*-ŋ*$^{+4}$	1sS/O
-i^{+4}	1/2pS/O	*-ŋ*$^{+7}$	1sA
kha-	1nsO	*-ŋa*$^{+2}$	1sS/O
-m^{+7}	1/2nsA	*-ŋa*$^{+10}$	e
ma-	1nseO	*-u*$^{+6}$	3O
mai-	1nsiO	*u-*	3S/A
na-	3>2		

Table 2.5: Chintang agreement markers

Although the paradigmatic functions of markers are often complex and ambiguous, concrete verb forms as a whole almost always code one scenario unambiguously due to the complex interplay of markers. For instance, the combination of *a-* [2S/A], *-u* [3O] and *-m* [1/2nsA] marks the scenario [2p>3s]: the form is bipersonal, so *a-* cannot mark S but must mark A; since the A and O slots are occupied by a 2nd (*a-*) and a 3rd person (*-u*) there is no place left for an additional 1st person and *-m* must mark [2nsA]; finally, the A must be plural because if it was dual *-ce* [d] would have been used, and the O must be singular because otherwise *-ce* [ns] would have been required.

[2]Language-internal reconstruction does not make it clear which of these functions is the primary one. If grammaticalisation procedes from more to less concrete meanings one would have to assume that this case first marked sources.

As the table shows, agreement markers do not define a uniform alignment pattern. For instance, *a-* indexes both S and A of the second person, which is an accusative pattern, *-i* is ergative in indexing S and O, and *-ce* [d] is neutral in being used independently of role. As a consequence, saying that a verb form has S-AGR very rarely means that there is a single marker indexing S. The usual meaning is that the verb form as a whole indicates a single argument. Similarly, saying that a verb form has A- and O-AGR does not mean that there are two markers for A and O but that the form indicates two arguments. The reason why the corresponding patterns are called A- and O-AGR is that they are typically linked to A and another core argument (P, T, or G). Note that this other core argument is also the O selected by S/A detransitivisation (see section 2.4.2 for details). Since A-AGR and O-AGR cannot be observed in isolation we will often simply speak of transitive verb forms, and forms with S-AGR will be called intransitive.

Cases where the agreement patterns are not linked to S, A, and P/T/G arise in less frequent valency classes or in alternations. For instance, in S/A detransitivisation the role of A is linked to S-AGR. A couple of experiencer verbs link A to O-AGR and P to A-AGR. In various deponent frames, the argument indicated by inflection is not linked to an argument in the valency. Using the same labels for roles and agreement positions in spite of such mismatches may seem confusing, but since Chintang does not have any clearly separable sets of markers that could be arbitrarily labelled (e.g. as I, II, III), other options would do even worse.

There is one systematic ambiguity that is also important for S/A detransitivisation and therefore should be mentioned here. Chintang has a morphophonological rule that disallows sequences of vowels in the suffix chain. Therefore, in sequences all vowels but the last are dropped. This rule also affects the marker *-u* [3O], which is dropped when it stands next to *-a* [IMP] or *-e* [IND.PST]. The transitivity of the surface forms can in that case not be determined. For instance, the verb *hatt-* 'wait (for)' can be used transitively or intransitively. The difference is easily visible in the nonpast: /hatt-u-kV/ [wait-3[s]O-IND.NPST[.3sA]] 'he waits for her' yields *hattoko*, and /hatt-no/ [wait-IND.NPST[.3sS]] 'he waits' yields *ha?no*. In the corresponding past forms, however, the difference disappears: both /hatt-a-u-e/ [wait-PST-3[s]O-IND.PST[.3sA]] 'he waited for her' and /hatt-a-e/ [wait-PST-IND.NPST[.3sS]] 'he waited' yield *hatte*. The difference becomes again visible when a suffix intervenes between *-u* and *-e*, as in /hatt-a-u-ŋ-e/ [wait-PST-3[s]O-1sA-IND.PST] 'I waited for him' > *hattuhẽ*, /hatt-a-ŋ-e/ [wait-PST-1sS-IND.PST] 'I waited' > *hattehẽ*.

Chintang has a couple of non-finite forms which express a reduced set of categories as compared to finite forms. However, there are few non-finite forms that cannot express any inflectional categories at all, and some also have means of indexing arguments:

- *-ma* [INF] frequently takes *-ce* [3nsO], especially with deontic semantics and scheduled events:

  (3) *Kattikhera a-tei?-ce wadhap-ma-ce=kha?*
  what.time 1sPOR-clothes-ns wash-INF-3nsO=NMLZ$_2$
  'What time (should I) wash my clothes?' (CLC:CLLDCh3R09S06.052)

- *-saŋa* [CVB.FGR] is compatible with all agreement prefixes (though it rarely takes them):

  (4) *Na-cop-saŋa yuŋ-no.*
  3>2-look.at-CVB.FGR sit-IND.NPST[.3sS]
  'He sits (there) watching you.' (elicitation DKR 2011)

- *-si* [PURP] can index P/T/G using nominal possessor prefixes:

  (5) *Ba-ce-ŋa a-ses-si u-tiy-a-ŋs-e.*
  PROX-ns-ERG 1sPOR-kill-PURP 3[p]S-come-PST-PRF-IND.PST
  'They have come to kill me.' (CLC:INT_JYR.0488)

2.3 Overview of relevant syntax

2.3.1 Valency and basic frames

Chintang has monovalent, bivalent, and trivalent verbs.[3] These terms will only used below where the number of arguments is to be stressed; otherwise the more common terms intransitive, monotransitive, and ditransitive will be used.

For a first overview of morphosyntax, a "basic" frame may be determined for each valency. This has been done below by determining all possible frames for each valency and by choosing the one frame that occurs with most verbs under the least specific conditions. For instance, most bivalent verbs license a group of frames where A can have ERG or NOM and can be linked to S-AGR or A-AGR and where P can be linked to O-AGR or not have agreement at all. Out of this group the frame that occurs under the least specific conditions is {A-ERG P-NOM V-a(A).o(P)}, so this frame is said to be basic for bivalent verbs. Below the basic frames for all three valencies are listed with examples.

- monovalent: {S-NOM V-s(S)}

  (6) *Ama, nunu hap-no.*
  mother baby cry-IND.NPST[.3sS]
  'Mum, the baby is crying.' (CLC:CLDLCh3R01S02.293)

- bivalent: {A-ERG P-NOM V-a(A).o(P)}

  (7) *Dhami-ce-ŋa dokh-a u-loĩs-o-ko.*
  shaman-ns-ERG illness-NTVZ 3[p]A-bring.out-3[s]O-IND.NPST
  'The shamans remove the illness.' (CLC:Jan-Gen.1142)

- trivalent: {A-ERG T-NOM G-LOC V-a(A).o(T)}

  (8) *Sa-ŋa marci huŋ=go-i? yuŋs-o-ŋs-e?*
  who-ERG chilli MED=NMLZ-LOC_2 put-3[s]O-PRF-IND.PST[.3sA]
  'Who put the chilli there?' (CLC:CLLDCh1R05S01.115)

There are two facts about valency in Chintang that will be important for the discussion of S/A detransitivisation. One is that in Chintang, valency is completely independent of the overtness of arguments. Every argument can be covert, no matter whether it corresponds to a known referent or not. For instance, a sentence that is commonly heard in Chintang when people exchange news is:

(9) *Si-ad-e.*
die-AWAY.ITR-IND.PST[.3sS]
'(He/somebody) has died.'

Because arguments are dropped all the time, Chintang has an extremely low referential density, that is, the proportion of argument positions that are occupied by overt NPs is very small compared to other languages (cf. Bickel 2003b, 2006, Stoll and Bickel 2009 on closely related Belhare). For this reason it is hard to find fully expanded frames such as the examples for the basic frames above.

The other important fact is that valency is a relatively fluid concept in Chintang. A large number of verbs have one monovalent and one bivalent sense and accordingly can take both corresponding basic frames. This phenomenon is best known as ambitransitivity or lability in typology and works quite parallel in Chintang to English *the bottle broke* : *he broke the bottle.* However, differently from many languages, ambitransitivity in Chintang is minimally lexicalised and basically fully

[3] The cross-linguistically most common semantic group with zero-valency, atmospheric events, is represented by monovalent verbs in Chintang (e.g. *wei? ta-no* [rain come-IND.NPST[.3sS]] 'it rains'. Tetravalent verbs do not exist in the lexicon but can be derived from trivalent verbs via causativisation, e.g. *hak-mett-* [send-CAUS] 'make somebody send something to somebody'. This valency never occurs in natural speech but only in elicitation and will therefore be ignored here.

transparent and productive. It will therefore be referred to as S/O detranstivisation here. See section 2.3.4.2 for some more details and examples and Schikowski et al. (forthcoming) for a more comprehensive description and corpus counts.

As a result of S/O detransitivisation, it would be imprecise for many verbs to say that they have a fixed valency. For instance, *ot-* 'break' can be used in Chintang as in the English example just given. Accordingly it does not have a valency of 1 or 2 but a *maximal* valency of 2.

2.3.2 Word order

Roles in Chintang are not linked to fixed positions, but there are clear defaults: SV, APV, AGTV. The frequency of these compared to other word orders in fully expanded frames (i.e. with no zeros) in a syntactically annotated part of the CLC is shown in Figure 2.3.

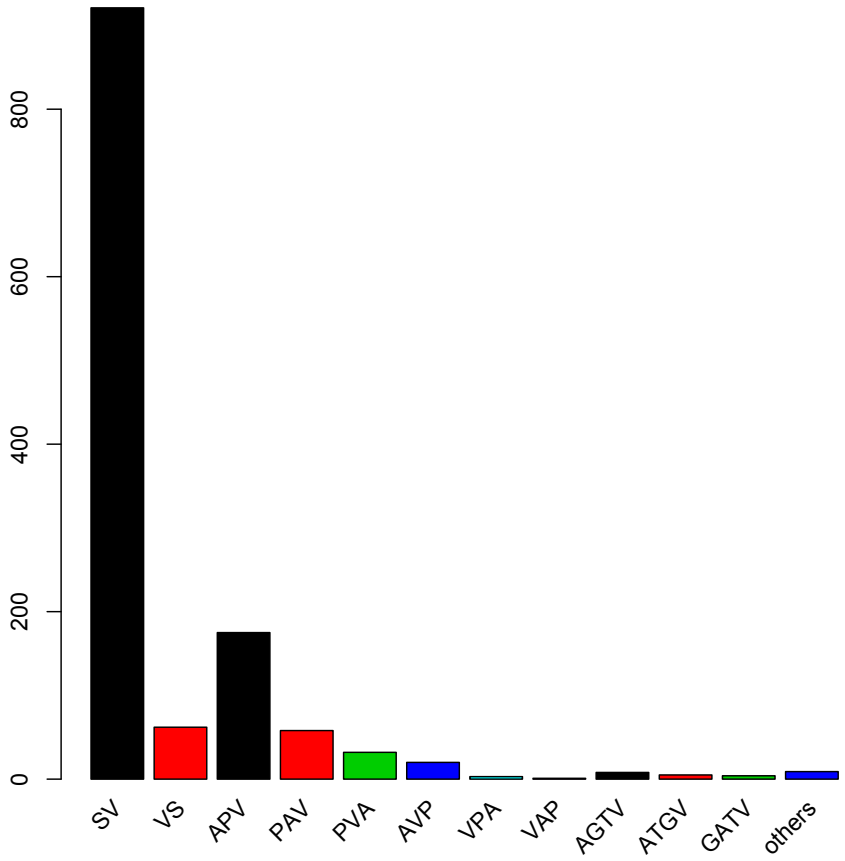

Figure 2.3: Attested word orders in fully expanded frames

The main factor behind role ordering seems to be topicality in the sense of mental presence: the more highly activated a referent, the farther to the left of the verb an overt NP representing it tends to be placed. NPs are mostly placed to the right of the verb when the speaker first thought about dropping them but then changed his mind after he already produced the verb.[4] This happens, for instance, when they are highly activated but their role is not quite clear or when they are deemed to be less important in some way than overt preverbal referents.

[4]The special status of postverbal NPs is also shown by their intonation. Pitch typically reaches its lowest point in a clause after the predicate. When NPs are placed after the predicate, pitch does not rise again but stays low and flat. This also suggests that the sentence as initially planned stopped after the verb and that the intonational contour was fit to this original sentence rather than to the version with the additional postverbal arguments.

The function of NP ordering before the verb and of placing NPs after the verb is illustrated by the three examples below. They all come from a story about a cat and a mouse which used to be friends but broke apart and tried to harm each other later. At the time (10) is uttered, the mouse has been the topic of a couple of sentences (it has been spreading gossip about the cat). It is therefore natural that it occupies the first position in (10) even though the cat now becomes A. In (11), the new status of the cat is already established. The mouse could have been dropped altogether but is overtly mentioned after the verb in order to foreground the cat's rage and background the mouse, which plays a much less active role in the paragraphs to come. Finally, the changed relation is so clear in (12) that the default word order is reinstated.

(10) *Sencak menuwa-ŋa ca-ma puŋs-o=kha=pho.*
mouse cat-ERG eat-INF start-[SUBJ.3sA.]3[s]O=NMLZ$_2$=REP
'Now the cat was about to eat the mouse.' (CLC:story_cat.250)

(11) *Menuwa-ŋa carko=ta kond-o-ko sencak.*
cat-ERG much=FOC search-3[s]O-IND.NPST mouse
'The cat searches a lot for it (the mouse).' (CLC:story_cat.255)

(12) *Menuwa-ŋa sencak khel-a mett-o-ko.*
cat-ERG mouse game-NTVZ do.to-3[s]O-IND.NPST
'Now the cat plays with the mouse.' (CLC:story_cat.260)

2.3.3 Frames and classes

Chintang altogether employs 15 frames (Schikowski et al. forthcoming). The number of verb classes as defined as sets of verbs licensing identical sets of frames, however, is much bigger and amounts to more than 50 (of which, however, only 20 have more than a single member). The biggest classes are all linked to the most frequent frames by simply licensing only one frame (ignoring alternations within frames that are independent of lexical class). Since verb class does not matter to S/A detransitivisation independently of frames, we do not have to talk about this in detail. Also note that when we speak, for instance, of "intransitive verbs" in later sections, that should not be taken to refer to the lexical class of intransitive verbs (i.e. the class of verbs that can only be used with the intransitive frame) but rather to all verbs licensing the intransitive frame (many of which license other frames in addition).

S/A detransitivisation is only possible with frames which are at least bivalent and can have an A-ERG and another NOM-marked argument linked to O-AGR. We will refer to this important abstract frame as the transitive frame (in contrast to the mono- and ditransitive frames; see section 2.4.2 for details). The list below shows all frames that fall under this schema as well as a few other highly frequent frames. One that is especially important is the intransitive frame, which bears bears formal similarities with the detransitivised variant of the transitive frame (and is sometimes hard to distinguish from it, see section 2.6.5.1). See Schikowski et al. (forthcoming) for a list also including marginal classes.

2.3.3.1 **Intransitive frame** {S-NOM V-s(S)}

This frame is the most frequent one in terms of licensing – 45% of all verbs can take it. However, only 20% of verbs take *only* the intransitive frame. Examples are *that-* 'appear', *ma-* 'get lost', *ŋoms-* 'taste buttery', *ims-* 'sleep':

(13) *Ba=go im-nik-niŋ hola.*
PROX=NMLZ$_1$ sleep-IND.NPST[.3sS]-NEG maybe
'Maybe this one won't sleep.' (CLC:CLDLCh3R01S02.152)

2.3.3.2 **Monotransitive frame** {A-ERG P-NOM V-a(A).o(P)}

The monotransitive frame is licensed by 45% of all verbs and is thus equally frequent to the intransitive frame. The corresponding lexical class even is the biggest class, taking up 40% of verbs. This

number only holds if one assumes that S/O detransitivisation is non-lexical, as is done here (see section 2.3.4.2). If one assumes a separate class of S/P ambitransitive verbs instead and takes as monotransitive verbs only those which are never used with the intransitive frame, the proportion shrinks to 30% (which is still clearly above the proportion of intransitive verbs). Examples are *nus-* 'heal', *ca-* 'eat', *putt-* 'pluck', *set-* 'kill':

(14) *Dosi-ko phak=pho thippa-ŋa=ta sed-o-ko.*
Daśaiṃ-GEN pig=REP grandfather-ERG=FOC kill-3[s]O-IND.NPST[.3sA]
'I heard grandpa himself will kill the pig for the Daśaiṃ festival.'
(CLC:CLLDCh1R13S02.1469)

2.3.3.3 Direct object ditransitive frame {A-ERG G-LOC T-NOM V-a(A).o(T)}

The names for this and the other ditransitive frames have been taken from Bickel (2007) and Bickel et al. (2010) and are motivated by their alignment with the monotransitive frame (see Dryer 1986). The direct object ditransitive frame treats T like P (in terms of both case and agreement). It is the most frequent ditransitive frame, being licensed by 18% of all verbs. All these verbs involve caused motion, e.g. *haŋs-* 'send', *bhokt-* 'stick', *thapt-* 'bring over', *tis-* 'put in':

(15) *ʃibanjal ba-sa-ŋa tis-o-ŋs-e ba-iʔ.*
jibanjal PROX-OBL-ERG put.in-3[s]O-PRF-IND.PST[.3sA] PROX-LOC$_2$
'He has put the *jibanjal* (a medicament) in here.' (CLC:CLLDCh4R05S05.754)

2.3.3.4 Primary object ditransitive frame {A-ERG G-NOM T-ERG V-a(A).o(G)}

This frame aligns G with monotransitive P. All verbs using it code physical manipulation of an object (G) with the help of an instrument (T). Examples are *hekt-* 'cut', *thup-* 'sew', *dhekt-* 'block', *bhukt-* 'cover':

(16) *Durga-ŋa u-chau-ce teiʔ-ŋa bhukt-o-ko-ce.*
Durga-ERG 3sPOR-child-ns cloth-ERG cover-3O-IND.NPST-[3sA.]3nsO
'Durga covers her children with a piece of cloth.' (elicitation PRAR 2010)

2.3.3.5 Double object ditransitive frame {A-ERG G-NOM T-NOM V-a(A).o(G)}

This frame treats T and G alike in terms of case marking. Agreement aligns G with P. Although it is not used by a lot of verbs (6% of all), many of them have meanings that are often thought of as prototypically ditransitive in the typological literature (Malchukov et al. 2010). They typically involve an animate recipient in G that benefits from an action. Examples are *hakt-* 'send', *lud-* 'tell', *yukt-* 'keep back for', *pid-* 'give':

(17) *A-pakku-ŋa cha-ce mithai pid-u-c-e.*
1sPOR-younger.uncle-ERG child-ns sweet give-3O-ns-IND.PST[.3sA]
'My uncle gave sweets to the children.' (elicitation PRAR 2010)

2.3.3.6 Transitive experiential frame {A-ERG P-NOM por(A)-N.EXP-NOM V-a(A/3s).o(P)}

This frame is quite different from all other transitive frames in that it contains a noun coding an experience ("N.EXP"). This noun has a possessive prefix indexing the experiencer and must be combined with a light verb in order to make a predicate out of it. Although only four verbs license this frame, one of them (*katt-* 'bring up') is quite productive and can form many experiential idioms such as *laja katt-* 'be ashamed of' (lit. 'bring up one's shame'), *lamma katt-* 'have an appetite for', *remsu katt-* 'be envious of', *rek katt-* 'be angry with':

(18) *Hana-ŋa hun-ce i-rek (a-)katt-u-c-e?*
2s-ERG MED-ns 2sPOR-anger 2[s]A-bring.up-3O-ns-IND.PST
'Are you angry with them?' (elicitation RBK 2011)

As the example shows, A-AGR can be either linked to A (the experiencer) or to a dummy 3s. This differential indexing pattern is unique to this frame. So far I haven't been able to find out what governs it.

2.3.3.7 *som-set(t)-* 'be satisfied, satisfy' {A-ERG P-NOM por(A/P)-N.EXP-NOM V-a(A).o(P)}

This peculiar frame is only used by two etymologically related verbs, *som-set-* and *som-sett-*, which can both mean 'be satisfied with' or 'satisfy'. With the meaning 'be satisfied with', the experiencer is A and the object of satisfaction is P. The experiencer is indexed by a possessive prefix on the experiential noun *som*, a trait shared by this frame with the transitive experiential frame. When the meaning is 'satisfy', the referent that brings about satisfaction is A and the experiencer is P. The experiencer is again indexed by a possessive prefix, even though its role has changed. This alternation is illustrated by (19) (note that *akka* [1s] in (19a) is not marked by ERG because it is a pronoun).

(19) a. *Akka hun-ce a-som sett-u-cu-h-ẽ.*
1s MED-ns 1sPOSS-liver kill.for-3O-3nsO-1sA-IND.PST
'I was satisfied with them.'

b. *Hun-ce-ŋa a-som u-sett-a-ŋs-a-ŋ-ni-h-ẽ.*
MED-ns-ERG 1sPOSS-liver 3A-kill.for-PST-PRF-PST-1sO-3p-1sO-IND.PST
'They have satisfied me.'

c. *Hun-ce-ŋa huni-som u-sett-a-ŋs-a-ŋ-ni-h-ẽ.*
MED-ns-ERG 3pPOSS-liver 3A-kill.for-PST-PRF-PST-1sO-3p-1sO-IND.PST
'They have been satisfied with me.' (elicitation GAR 2010)

When the experiencer is mapped to P (meaning 'satisfy'), it can be marked as the possessor of N.EXP by GEN (e.g. in (19b) *akka a-som* [1s 1sPOR-liver] or *ak-ko…* [1s-GEN]). Because the NOM/-GEN alternation is possible in all possessive NPs (e.g. *akka a-khim* [1s 1sPOR-house] or *ak-ko a-khim*), one may also say that the P experiencer is consistently marked as a possessor by case and indexing whereas the A experiencer (meaning 'be satisfied with') is a hybrid (A case marking, possessor indexing).

Since this frame is so rare, it will not be discussed in great detail in the following sections. In the present context it is only of interest because it can be S/A detransitivised.

2.3.4 Differential marking

Chintang is rich in differential marking patterns of various kinds – there is differential case marking, differential indexing, and differential framing. Most of these patterns are, however, irrelevant for the present study. We will only talk about differential A marking and S/O detransitivisation. The most important pattern, S/A detransitivisation, will be discussed in detail in the following sections.

2.3.4.1 Differential agent marking

The differential marking pattern which is most frequently attested in Chintang is split and fluid A. Until quite recently (e.g. in Bickel et al. 2010) it was thought that ERG was optional on second person pronominal A and impossible on first person pronominal A. In fact, it is optional on pronouns of both persons, as is shown by (20) and (21). All other nominal A including demonstratives require ERG (22).

(20) a. *Akka-ŋa cekt-u-ŋ=go ba-i? lon-n-a?-no.*
1s-ERG speak-3[s]O-1sA=NMLZ$_1$ PROX-LOC$_2$ come.out-LNK-AWAY.ITR-IND.NPST[.3sS]
'What I say comes out here (on the camera).' (CLC:khinci_talk.037)

b. *Akka wa-ce tis-u-ku-ŋ-cu-ŋ ni.*
1s hen-ns put.in-3O-IND.NPST-1sA-ns ASS
'I'll put in the hens.' (CLC:CLLDCh1R02S04.1118)

(21) a. *Aba huŋ=go na hana-ŋa=yaŋ a-ŋis-o-ŋs-e.*
now MED=NMLZ$_1$ CTOP 2s-ERG=ADD 2[s]A-recognise-3[s]O-PRF-IND.PST
'Now you, too, have recognised this.' (CLC:suntala_talk.61)

b. *Hana them a-hekt-o-ko huŋ=go-i??*
2s what 2[s]A-cut-3[s]O-IND.NPST MED=NMLZ$_1$-LOC$_2$
'What are you cutting there?' (CLC:CLLDCh2R14S03.0366)

(22) a. **Huŋ=go aŋgreji pad-e numd-o-ko.*
MED=NMLZ$_1$ English study-V.NTVZ do-3[s]O-IND.NPST[.3sA]
'He's studying English.' (elicitation RBK 2010)

b. *Huĩ-sa-ŋa jamma kob-o-ko=kha?*
MED-OBL-ERG everything pick.up-3[s]O-IND.NPST=NMLZ$_2$
'So it (the camera) picks up everything?' (CLC:CLLDCh2R06S07.441)

The only two lexemes where ERG is completely ungrammatical are the pronouns *ancaŋa* [1de] and *anaŋa* [1pe].[5] Apart from that ERG is very rare on *akka* [1s] (only two instances in the CLC) and unattested with *anci* [1di]. The rareness of *akka-ŋa* seems to be principled. I showed an informant the attested examples and asked what he thought about them. The answer was that he himself wouldn't use them but old people such as his great-uncle might. He also produced further examples and said that these were not wrong but merely old-fashioned. By contrast, the lack of attestations of *anci-ŋa* is simply due to the rareness of overt 1di – 5051 instances of *akka* are opposed to a mere 253 of *anci*. Speakers readily accepted *anci-ŋa* in elicitation. The remaining first person pronoun, *ani* [1pi], is equally frequent with and without *-ŋa* in A function (18 instances for each).

The picture looks similar for the second person. *Hana* [2s] in A is attested 12 times with ERG and 174 times without it. *Hanci* [2d] is unattested with ERG, but since the dual pronoun is again relatively rare (2615 *hana* vs 199 *hanci*), frequency once more explains this – *hanci-ŋa* is accepted in elicitation. *Hani* [2p] is similar to *ani* [1pi] in that marked (7) and unmarked forms (9) are about equally frequent.

The factors governing the presence of ERG on pronouns are yet unknown. Presently it looks like at least three factors favour the marking of ERG: the presence of deontic modality (23), the conservativeness of the language (24), and contrastive focus (25). The following three sentences exemplify these.

(23) *Ã, ani-ŋa ba-ce man-e num-ma-ce=ta kon-no.*
yes 1pi-ERG PROX-ns obey-V.NTVZ do-INF-3nsO=FOC be.necessary-IND.NPST
'Yes, we have to obey them.' (CLC:chintang_now.1314)

(24) *Ani-ŋa ba-khi kha-u-m kina khaŋ-ma=yaŋ*
1pi-ERG PROX-MOD look.at-3[s]O-[SUBJ.]1pA SEQ look.at-INF=ADD
hid-u-m-num.
be.able-3[s]O-1pA-[SUBJ.]NPST.NEG
'When we look at it we can't even overlook it (in its entirety).'
(CLC:origin_myth.558, speaker 70 years old at recording time)

(25) *Hid-u-m-num, ani-ŋa hid-u-m-num.*
be.able-3[s]O-1pA-[SUBJ.]NPST.NEG 1pi-ERG be.able-3[s]O-1pA-[IND.]NPST.NEG
'We won't be able to do it, *we* really won't (but others who have more money may).'
(CLC:ctn_prob_talk 119)

Note that independently of what was said above, A-ERG is only ever possible in fully transitive frames. Under S/A detransitivisation A-NOM is obligatory (see section 2.4.1).

[5]Historically these contain the exclusive suffix *-ŋa*, which is also found in the verb.It is not unimaginable that this suffix is related to *-ŋa* [ERG], the functional bridge being that a transitive action can be viewed as a characteristic achievement of the agent in a similar way any action is characteristic of an exclusive first person plural (precisely because several other referents are excluded). This might seem far-fetched, but the other possible explanation – a phonological constraint against /ŋaŋa/ – is equally weak since there is no evidence for such a constraint except this restricted context. It looks like presently there simply is no good reason for why ERG is strictly impossible only on *ancaŋa* and *anaŋa*.

2.3.4.2 S/O detransitivisation

Many verbs taking one of the transitive frames have an alternative frame where O becomes the only argument and gets linked to S-AGR. This detransitivised variant is used when an event is perceived as happening spontaneously (i.e. without an obvious A) or when its result continues without the participation of an A. (26) and (27) show pairs of examples for the two cases.

(26) a. *Ram-e-ŋa u-tec-ce kosi-be? lums-u-c-e.*
Ram-NAME.NTVZ-ERG 3sPOR-clothes-ns river-LOC$_1$ sink-3O-ns-IND.NPST[.3sA]
'Ram dumped his clothes into the river.'
b. *Kosi-be? ma?mi lums-e.*
river-LOC$_1$ person sink-IND.NPST[.3sS]
'Someone sank in the river.' (elicitation RBK 2010)

(27) a. *Sa-ŋa u-lett-o=kha phuŋ?*
who-ERG 3[p]A-plant-[SUBJ.]3[s]O=NMLZ$_2$ flower
'Who planted the flower?' (CLC:CLLDCh3R07S01.953)
b. *Makkai-ce u-lett-a-ŋs-e.*
maize-ns 3[p]S-plant-PST-PRF-IND.PST
'The maize plants have been planted.' (field notes 2010)

Although it is convenient to call this alternation detransitivisation, it is by no means clear that the transitive frame is in some way basic and the intransitive frame derived. For instance, whereas the transitive frame is by far more frequent than the intransitive one for *lett-* 'plant', it is about equally frequent with the intransitive frame for *lums-* 'sink'. For yet other verbs the intransitive frame is more frequent. For instance, the transitive variant of *ghoŋs-* 'grow big' could so far only be observed in elicitation:

(28) a. *Saĩli, kana-phak na ba-tta=kha ghoŋ haŋ*
third.daughter 1pePOR-pig CTOP PROX-EXT=NMLZ$_2$ grow.big[.SUBJ.NPST.3sS] COND$_2$
na aŋ…
CTOP QTAG
'Saĩli, suppose our pig grew as big as this…' (CLC:CLLDCh1R06S03.0151)
b. *Ba=go phak them-ma ba-tta ghoŋs-o-ŋs-e?*
PROX=NMLZ$_1$ pig what-ERG PROX-EXT grow.big-3[s]O-PRF-IND.NPST[.3sA]
'What has let this pig grow this big?' (elicitation RBK 2010)

Examples such as this one show two things. First, S/O detransitivisation is productive. When I first confronted my informant with the transitive form *ghoŋsoŋse*, he denied that it was possible – most likely because he had never heard it before. However, when I came up with the sentence in (28b), he had to admit that it was well possible in that context. The productivity of S/O detransitivisation also becomes apparent in the lexicon, where about 21% of all transitive vebs are attested with both frames. Second, S/O detransitivisation is non-directional: it can subtract an A from a known transitive frame or add an A to a known intransitive frame.

The relevance of this pattern for our topic, S/A detransitivisation, is indirect. S/O detransitivisation is interesting because it formally is the mirror image of S/A detransitivisation but functionally it is quite different from it. Even when the A of an detransitivised sentence is covert, the two patterns can still be easily distinguished by their semantics:

(29) *Phuŋ nam-no.*
flower smell-IND.NPST[.3sS]
'The flower smells.' or 'It (e.g. a dog) smells at flowers.' (CLC:CLLDCh3R07S01.778)

2.3.5 Raising of case and agreement

2.3.5.1 Introduction

Chintang has numerous constructions where an argument of a subordinate verb leaves morphological traces in the matrix. Such traces can be found both in case and in agreement, and I will refer to both as raising here instead of using different terms such as raising and long distance agreement. Constructions with morphological raising play an important role for the present study because they change or narrow down the possibilities of marking S/A detransitivisation. It should also be noted that most of these constructions are by no means marginal or exotic but highly frequent.

Raising occurs in constructions with two non-finite forms. One is the infinitive *-ma*, which is used with about 15 light verbs expressing a wide range of functions such as ability (e.g. *hid-* 'be able to'), necessity (e.g. *kond-* 'must') or phase semantics (e.g. *puŋs-* 'start to'). The other form is the foregrounding converb *-saŋa*, which is used together with 7 regular verbs that have a special metaphorical meaning in this construction in order to express temporal-aspectual meanings (e.g. *yuŋ-* 'be there' : *-saŋa yuŋ-* 'stay doing'). Both constructions exhibit a high degree of integration in the sense of Raible (1992), that is, their properties place them relatively far away from two juxtaposed independent clauses.

In infinitival subclauses, it is often difficult to determine for these constructions whether an NP belongs to only one predicate or both and by which predicate it is assigned a role. Consider the two examples below. Both are possible with or without the infinitive, and in both the meaning of the two variants is rather similar. The meaning of *hid-* without INF is 'be able to handle, cope with, finish', with an INF it is 'be able to, finish doing'. The meaning of *mund-* without INF is 'forget', with an INF it is 'forget to':

(30) *Marci (ca-ma) hid-u-ku-ŋ-niŋ.*
chilli eat-INF be.able-3[s]O-IND.NPST-1sA-NEG
'I can't (eat) chilli.' (field notes 2010)

(31) *Hana jaileyaŋ yum (ti-ma) a-mund-and-o-ko!*
2s always salt put.in-INF 2[s]A-forget-COMPL$_1$-3[s]O-IND.NPST
'You always forget (to add) salt!' (elicitation RBK 2010)

In (31) it is not quite clear whether *yum* 'salt' is the P of *mund-* 'forget', the T of *tis-* 'put in', or both. Its case does not tell us because *mund-* has P-NOM and *tis-* T-NOM. Agreement is on *mund-* only, but that is not very telling either because INF cannot carry any agreement affixes except *-ce* [3nsO]. Similarly, *hana* [2s] is assigned the same role by both verbs and also functions in that role. The same holds for (30). We will take the simple stance here that in cases such as (31) and (30) the frames of the two participating verbs are superimposed so that *yum* is both P and T and *hana* is both monotransitive and ditransitive A.

Not all complement verbs behave this way. For instance, *lapt-* 'be about to' can only be used with infinitives and thus does neither have an independent frame nor a standard role set. In such cases we will assume that all arguments belong to the embedded predicate and are assigned their roles only by it. Where it is necessary to distinguish this mechanism from frame superimposition we will speak of true raising (because it is only in this case that one can say that an argument is morphosyntactically part of the matrix claus *in spite* of its semantic affiliation). Mostly, however, such a distinction need not be made because the formal result is the same in both cases.

The whole problem is much less pronounced with constructions involving *-saŋa* because all possible matrix verbs acquire a special, abstract meaning in these constructions that makes it clear that they do not have arguments or assign roles any longer. Compare, for instance:

(32) a. *Ba-ce aŋ u-numd-a-ŋs-e mo-ba?*
PROX-ns what 3[p]S-do-PST-PRF-IND.PST DEM.DOWN-LOC$_1$
'What have they done down there?' (CLC:CLLDCh2R02S06.1169)

b. *Wei? ta-saŋa numd-a-ŋs-e acikali.*
rain come-CVB.FGR do-PST-PRF-IND.PST[.3sS] these.days
'It's kept raining over the last days.' (CLC:RM_JK_talk01.189)

2.3.5.2 Constructions with transitive embedded frame

For the study of S/A detransitivisation, the constructions which are of the greatest interest are those which involve transitive embedded frames. There are three options in this case. The default is to raise the complete transitive frame so that the complex sentence as a whole acquires transitive characteristics: A is marked by ERG and there is A+O-AGR. AGR is realised on both predicates with a couple of restrictions: the non-finite embedded forms are only compatible with a few agreement affixes (see section 2.2.3 above) and these are always optional, and intransitive matrix predicates (only found in the *-saŋa* constructions, e.g. *yuŋ-* 'be there') can only have S-AGR. In the case of frame superimposition, the matrix predicate always assigns the same case and agreement to A and links the same NOM-marked referent to O-AGR as the embedded predicate, so the two frames can never clash.

Below are some examples. (33) and (34) show *-ma* [INF] with and without frame superimposition, respectively. (35) show *-saŋa* [CVB.FGR] with and without agreement on the *-saŋa* form.

(33) *U-ko-no-ko-ce sa-ŋa hiŋ-ma hid-u-ku-ce naŋ?*
3[p]S-roam-IND.NPST=NMLZ$_1$-ns who-ERG feed-INF be.able-3O-IND.NPST-[3sA.]3nsO but
'But who can feed the ones wandering around?' (CLC:RM_JK_talk01.073)

(34) *Ma?mi-ce-ŋa the?nuwa thuk-ma na-lapt-i-ŋs-i-hẽ.*
person-ns-ERG saliva spit.at-INF 3>2-be.about.to-2pP-PRF-2pP-IND.PST
'People are about to spit (saliva) at you.' (CLC:CLLDCh3R08S01.1021)

(35) *Cha-ce-ŋa badhe=ta u-ni-saŋa u-thapt-o-ŋs-e.*
child-ns-ERG much=FOC 3nsA-know-CVB.FGR 3[p]A-bring.across-3[s]O-PRF-IND.PST
'The children have come to know a lot.' (CLC:chintang_now.738)

(36) *Ba-khi=ta i-bhog-a ca-saŋa*
PROX-MOD=FOC 2sPOR-sacrificial.meat-NTVZ eat-CVB.FGR
a-khatt-o=kha.
2[s]A-take.away-[SUBJ.]3[s]O=NMLZ$_2$
'You will eat your sacrificial meat like this from now on.' (CLC:CLLDCh1R05S05.0719)

The second and third option for dealing with transitive embedded frames are linked so that one verb can only allow both or none. They only occur with infinitival subclauses. The relevant matrix predicates can have 3sS-AGR (which can be interpreted as indexing the infinitive itself) or raise embedded O-AGR to S-AGR. The pair of examples in (37) shows both options for *kond-* 'must, be necessary':

(37) a. *U-lapthaŋ-ce=yaŋ mi?~mi=kha khok-ma-ce kon-no?.*
3sPOR-wing-ns=ADD small~INTENS=NMLZ$_2$ chop-INF-3nsO be.necessary-IND.NPST[.3sS]
'Its wings also must be cut into tiny pieces.' (CLC:muncurup_numma.29)

b. *Yo a-nne-ce tiyar-a u-lis-e?, pi-ma-ce*
DEM.ACROSS 1sPOR-elder.sister-ns ready-NTVZ 3[p]S-become-IND.PST give-INF-3nsO
u-kon-no?.
3[p]S-be.necessary-IND.NPST
'Those girls are ready, they should be given (rice).' (CLC:CLLDCh2R10S01.359)

Note that with these options, the A of the embedded predicate is marked by ERG independently of AGR:

(38) *ʄamma-ŋa akka cop-ma kon-no/ koĩ-ya-ʔã.*
all-ERG 1s look.at-INF be.necessary-IND.NPST[.3sS] be.necessary-1sS-IND.NPST
'Everybody should look at me.' (elicitation GAR 2011)

2.3.5.3 Constructions with intransitive embedded frame and transitive matrix

Another area of interest are constructions that involve a transitive matrix but no raising. Since transitive embedded frames are always raised or suppressed, such constructions are only found with intransitive embedded frames. There they form a subset of all available complex frames:

- {A-ERG P-[V.NONF] V-a(A).o(V.NONF)}
- {A-ERG/NOM P-[V.NONF] V-a(A).o(V.NONF)}
- {A-NOM P-[V.NONF] V-s(A)}
- {S-NOM S-[V.NONF] V-s(V.NONF)}
- {S-NOM [V.NONF] V-s(S)}

Note that when the embedded predicate is intransitive and the matrix is transitive, role clashes become possible. This is different from the constructions with transitive embedded predicates discussed above, where even the roles of superimposed arguments were always very similar (e.g. monotransitive and ditransitiive A, monotransitive P and direct object ditransitive T). With these constructions, one predicate may assign S and the other A.

This conflict was resolved for the list above as follows. When the matrix predicate is transitive, the argument in question is assumed to be A of the matrix predicate and the non-finite form itself is P. INF can only P become in this group of constructions and not in the raising pattern for embedded transitive frames that we saw above because there the embedded P is more clearly referential and has more proto-patient properties than INF. When the matrix predicate is intransitive or does not assign any roles, the argument in question is assumed to be the S of the embedded predicate.

The constructions that are of interest to us are those with a bivalent matrix – that is, constructions which have an A and a P in their valency, regardless of their case marking and indexing. Below is one example for each of these. Most examples had to be elicited in order to illustrate the fully expanded frame in a single sentence.

{A-ERG P-[V.NONF] V-a(A).o(V.NONF)}

(39) *Ep-ma kond-o-ko ni ba-sa-ŋa.*
stand.up-INF want-3[s]O-IND.NPST[.3sA] ASS PROX-OBL-ERG
'This one wants to stand up.' (CLC:CLLDCh4R02S01.0413)

{A-ERG/NOM P-[V.NONF] V-a(A).o(V.NONF)}

(40) a. *Ba*(-sa-ŋa) chepmu-ma nad-o-ŋs-e,*
PROX-OBL-ERG urinate-INF refuse-3[s]O-PRF-IND.PST[.3sA]
hi-nik-niŋ hola.
be.well-IND.NPST-NEG[.3sS] maybe
'He refuses to pee, maybe he's ill.'
b. *Philim(*-ŋa) thai?-ma nad-o-s-e.*
film-ERG appear-INF refuse-3[s]O-PRF-IND.PST[.3sA]
'The movie just doesn't want to appear.' (elicitation SAR 2011)

{A-NOM P-[V.NONF] V-s(A)}

(41) *Pecce le?le lik-ma hi-no.*
Pecce only go.up-INF be.able-IND.NPST[.3sS]
'Only Pecce can go up.' (CLC:CLLDCh3R06S05.720)

The choice of frame is conditioned by several factors. The most important factor is the matrix predicate, since many verbs allow only a single frame once the embedded clause and its valency are given. For instance, *kond-* in the sense 'want, try' as in (39) and *hid-* 'be able' in (41) both are only grammatical with the complex frames they exemplify. The ERG/NOM alternation in the complex frame represented by *nad-* in (40) is to a great part determined by volitionality: volitional A as in (40a) must be marked by ERG, non-volitional A as in (40b) by NOM. However, there are

intransitive	transitive	meaning with *-saŋa*
khat- 'go'	*khatt-* 'take'	'start doing, do from now on'
thap- 'come across'	*thapt-* 'bring across'	'have been doing, come to do'
yuŋ- 'be there'	*yuŋs-* 'put'	'stay doing, keep doing'

Table 2.6: Etymologically related matrix verbs with *-saŋa* [CVB.FGR]

other factors at work here that are not well understood as yet. The two verbs whose behaviour is so far least understood are *let-* 'stop, abandon' and *latt-* 'stop, have had enough of'. Volitionality probably also plays a role for these, but it can by far not explain all of their uses.

2.3.5.4 Transitivity marked by verb stems

There are a couple of constructions involving *-saŋa* [CVB.FGR] where intransitive and transitive embedded frames are linked in a special way. As shown in Table 2.6, the morphological transitivity of the embedded frame determines the lexical transitivity of the matrix verb. In all cases where an intransitive and a transitive matrix verb are available the two are etymologically related.

In these constructions, an intransitive embedded verb is used with the intransitive matrix variant and a transitive embedded variant with the transitive one. For instance:

(42) a. *Ani=lo naŋ ba-i pop-saŋa pop-saŋa khad-i-ki*
1pi=SURP but PROX-LOC$_2$ degenerate-CVB.FGR degenerate-CVB.FGR go-1p[i]S-IND.NPST
naŋ.
but
'But we just waste away more and more here.' (CLC:INT_MXR.0903)

b. *Akka khem-saŋa khatt-u-ŋ-kh-a-ŋ-ne i-katha.*
1s listen-CVB.FGR take-3[s]O-1sA-CON-3[s]O-1sA-[SUBJ.NPST.]OPT 2sPOR-story
'I will try to listen to your story from now on.' (CLC:kazi_trip_talk.115)

Note, though, that the association between the transitivity of the embedded frame and the matrix verb is not perfect. Although it is true that the intransitive matrix verbs only occur with intransitive embedded frames and that transitive embedded frames are only compatible with the transitive matrix verbs, transitive matrix verbs can (if rarely) also be used with embedded intransitive frames, as in (43):

(43) *Ani toŋ-saŋa=ta khatt-u-m, pa-saŋa=ta*
1pi get.together-CVB.FGR=FOC take-3[s]O-[SUBJ.NPST.]1[pi]A grow-CVB.FGR
khatt-u-m.
take-3[s]O-[SUBJ.NPST.]1[pi]A
'Let's keep working together and growing.' (CLC:Student_life.060)

From the available examples it looks as if this use was once more triggered by volitionality, but this cannot be said with certainty yet.

2.4 Formal properties of S/A detransitivisation

2.4.1 S/A detransitivisation as differential framing

We will now start our investigation of S/A detransitivisation by looking at its formal characteristics. Below is a pair of examples for what we will refer to as the transitive frame (44a) and the S/A detransitivised frame (44b).

(44) a. *Debi-ŋa seu kond-o-ko.*
Debi-ERG apple look.for-3[s]O-IND.NPST[.3sA]
'Debi is looking for the/an apple.'
b. *Debi seu kon-no.*
Debi apple look.for-IND.NPST[.3sS]
'Debi is looking for apples.' (elicitation PRAR 2010)

Both frames are bivalent, i.e. they contain an A (*Debi*) and a P (*seu*). In (44a), the A is marked by ERG and linked to A-AGR and the P is linked to O-AGR. By contrast, A is marked by NOM in (44b) and linked to S-AGR, and P is not indexed at all. The functional properties of S/A detransitivisation are discussed in detail in sections 2.5 and 2.6. For the time being we will use the term specificity where required: the P in (45a) is specific, the P in (45b) is non-specific.

Of the three factors that change between the two frames (A case, A-AGR, O-AGR), only the first is subject to some variation because of the optionality of ERG on pronouns (section 2.3.4.1). Note, though, that it wouldn't be correct to say that the rules for DAM override those for S/A detransitivisation because then we would expect variation in the case of A in both frames. This is not the case: A-ERG is only possible in the transitive frame. Pronouns thus always license NOM but not ERG – in order to know whether the latter case is possible, one has to know the frame into which the pronoun is inserted.

Since A-AGR, O-AGR and the possibility of A-ERG always change together in S/A detransitivisation and cannot be manipulated independently, this pattern can be characterised as differential framing (cf. the definition in section 1.3.2). If one wants to force a reduction to either differential case marking or differential indexing, the latter is the better candidate. One reason is that S/A detransitivisation usually only becomes visible on the verb because arguments are so frequently covert. The other is that the differential case marking component can be theoretically derived from differential indexing but not the other way round.

If we assume that the root of S/A detransitivisation is the presence or absence of O-AGR, we may first say that the change of A-AGR to S-AGR in the detransitivised frame is a corollary: because agreement affixes in Chintang do not have a uniform alignment pattern and because the indexation of A and O are so closely linked, it is usually not possible to just take away the O-AGR affixes and get a meaningful verb form. Instead, the whole pattern has to change to something else, S-AGR being the only remaining alternative. From there it can be argued that S-AGR with an ERG-marked argument is banned in Chintang (it is not attested in a single construction, in contrast to A-AGR with NOM-marked arguments – see section 2.3.5.3) and that therefore the case of the A associated with the changed AGR must change to NOM.

This way of interpreting S/A detransitivisation as a special kind of differential indexing is certainly elegant. However, it also has its weaknesses, and I therefore won't adapt it here. One is that it requires a lot of theoretical assumptions about causal chains in synchronic syntax that I would rather not make. Further, there are some contexts where there is no agreement but S/A detransitivisation can still be expressed via A case (non-finite clauses with overt A, see section 2.4.4.5). This is hard to explain if one assumes that A-NOM is a consequence of changing A+O-AGR to S-AGR.

Note that there are no reasons to assume that one of the frames is more basic than the other. Although the term "detransitivisation" suggests that the detransitivised frame is somehow derived from the transitive frame, this is a mere terminological weakness. In fact, it is possible to derive both frames from each other by simple rules. The only case that is not completely trivial occurs when a detransitivised double object ditransitive frame is to be converted to the corresponding transitive variant. Since there are two NOM-marked arguments, it is not immediately clear which of them should be linked to O-AGR. However, since G-NOM always takes precedence over T-NOM in indexing in Chintang, this is not a real problem either.

It is also not the case that one of the frames is found with more verbs or under more special conditions (the criteria for "basicness" that were used in section 2.3.1). Both frames are possible with every transitive verb, and the condition for both is a relatively simple binary property of the O referent so far approximated as specificity. The only hint to basicness is given by token frequency: the transitive frame is altogether more frequent than the detransitivised frame in the Chintang

corpus in covering about 72% of all relevant observations (see section 2.7 for more quantitative data). However, even this only holds as a general tendency, because individual verbs and object nouns may prefer the detransitivised frame depending on their semantics (section 2.6.3.1).

Now that we have seen the characteristic formal properties of S/A detransitivisation, the question may be asked whether there isn't a better name for it. The pattern resembles many well-known phenomena: it has something of differential agent marking and of differential agent and object indexing, but differently from these phenomena it affects several coding loci at once. It is also reminiscent of noun incorporation in that the object in the detransitivised frame impressionistically is no longer fully referential and the frame as a whole looks rather similar to the intransitive frame (except that the object is still present in the valency).[6] But then again, as we will see shortly (sections 2.4.3.2, 2.4.3.1), detransitivised objects can be moved around freely and can be covert, which is at odds with any conceivable definition of noun incorporation. Another phenomenon that is similar to S/A detransitivisation is S/A ambitransitivity, but this is by default a lexical phenomenon bound by verb class, whereas S/A detransitivisation is possible with any verb in Chintang. Finally, the typological canon that S/A detransitivisation probably resembles most closely is the antipassive. However, in a typical antipassive the downgraded object should be marked by a peripheral case, not by NOM, and the verb should bear a dedicated marker (Cooreman 1994:50, Dixon 1994:146, Primus 1999:237).

A typology of alternations linking fully transitive frames to frames with a demoted O and/or a promoted A should ultimately get rid of prefabricated labels such as "antipassive". A decompositional approach that classifies alternations according to various criteria such as A and O case, A and O agreement, word order etc. seems to be more promising. This is further confirmed by the existence of phenomena that are akin to Chintang S/A detransitivisation but not fully identical to it. I will briefly discuss two examples to illustrate this point.

One group of languages where object-demoting alternations are widespread is the Oceanic branch of the Austronesian family. Margetts (2008, 2011) calls this group of alternations "transitivity discord" but also mentions two other labels that circulate in the literature on Oceanic languages, "semitransitive" and "pseudo noun incorporation". Transitivity discord in Oceanic does not seem to have uniform formal characteristics but depends on the marking mechanisms each language provides. For instance, Niuean has case marking but no agreement. The main factor changing between the two frames is the case of A, which can be ERG or NOM, as shown in (45).

(45) a. *Takafaga tūmau nī e ia a tau ika.*
hunt always EMPH ERG 3s NOM PL fish
'He is always fishing.'
b. *Takafaga ika tūmau nī a ia.*
hunt fish always EMPH NOM 3s
'He is always fishing.' (Seiter 1980:69, cited in Massam 2001:157)

By contrast, Manam has agreement but no case marking. Here, the factor that changes is the presence of separate S/A and O agreement markers in one frame vs the absence of O markers in the other, as shown in (46).

(46) a. *Bóro ŋe u-rere-t-áʔ-idi.*
pig this 1sS/A-like-THC-TR-3pO
'I like these pigs.' (Lichtenberk 1982:272)
b. *Deparóbu u-rére.*
rice 1sS/A-like
'I like rice (in general).' (Lichtenberk 1982:271)

Finally, (47) illustrates the case of a language with case marking and agreement, Sinaugoro. O-AGR is expressed by suffixes, S/A-AGR by preverbal particles. This last case comes very close to S/A detransitivisation in Chintang but still differs from in it in that S and A are have the same

[6] In fact, a related construction in Bantawa has been analysed as noun incorporation by Angdembe (1998). S/A detransitivisation in this and other Kiranti languages is discussed in section 2.9.

agreement markers, so the S/A particle does not change between the frames:

(47) a. *Au-na forara a bubu lausi-a-to.*
1s-ERG sand 1sS/A pour spread-3sO-PRF
'I spilt the sand.'
b. *Au forara a bubu lausi-to.*
1s sand 1sS/A pour spread-PRF
'I spilt sand.' (Tauberschmidt and Bala 1992:184)

According to Margetts (2008), the main factor behind transitivity discord across Oceanic is the "individuation" of the object (where less individuated O are demoted).

Another interesting language family is Algonquian. Algonquian languages have transitive agreement systems whose complexity is comparable to those found in Kiranti languages. In addition to regular agreement, verbs carry a marker that contains information on transitivity and the animacy of S or O. Since Bloomfield (1946), the traditional labels for these markers have been AI (intransitive with animate S), II (intransitive with inanimate S), TA (transitive with animate O), and TI (transitive with inanimate O). A widespread alternation across Algonquian languages links a TA or TI form with A+O-AGR to an AI form where A gets S-AGR. The detransitivised frame is known as "pseudotransitive" since Bloomfield's (1957) treatment of Ojibwe. (48) shows an example from Blackfoot, where the pseudotransitive frame is used with non-specific O (glosses adapted from Ritter and Rosen 2010:134; cf. also Frantz 1991:40):

(48) a. *Na-ow-ats-iw amo mamii.*
PST-eat-TA-3s>4s PROX fish(ANIM)
'S/he ate this fish.'
b. *Na-ow-atoo-m ani akoopis.*
PST-eat-TI-3s>4s MED soup(INAN)
'S/he ate that soup.'
c. *Na-oy-i-w (mamii/akoopis).*
PST-eat-IA-3sS (fish/soup)
'S/he ate (fish/soup).'

The pseudotransitive alternation resembles S/A detransitivisation in Kiranti in that it is not possible to simply take away O-AGR because of the complex interaction of A and O markers. Instead, the complete pattern has to change from A+O-AGR to S-AGR. However, the fused animacy/transitivity markers that are typical for Algonquian are alien to Kiranti. Another major difference is that there is no case marking on arguments in Algonquian, so when A is overt it is zero-marked in both frames.

To conclude, whilst Chintang S/A detransitivisation is more or less similar to many phenomena, it does not correspond to the typical form of any of them. The one phenomenon that it does correspond to does not have an established name yet. The term S/A detransitivisation is meant to be a neutral term that does not only cover the pattern in question but also other differential framing patterns where A can assume S-like properties – for instance, those listed above: noun incorporation, S/A ambitransitivity, and antipassives.

2.4.2 Arguments selected by S/A detransitivisation

The monotransitive frame is not the only frame which is accessible to S/A detransitivisation – all transitive frames that were described in section 2.3.3 are. The correspondences between transitive and detransitivised variants are summarised in Table 2.7.

The transitive frames share two important characteristics: they have an A marked by ERG and linked to A-AGR and a non-A argument marked by NOM and linked to O-AGR. These two features do not only describe what is common to the six frames in Table 2.7 but also separate them from all other frames, including more marginal frames that have not been mentioned in section 2.3.3. We may thus posit an abstract transitive frame of the form {A-ERG O-NOM V-a(A).o(O)}. The

transitive frame	detransitivised frame
{A-ERG P-NOM V-a(A).o(P)}	{A-NOM P-NOM V-s(A)}
{A-ERG T-NOM G-LOC V-a(A).o(T)}	{A-NOM T-NOM G-LOC V-s(A)}
{A-ERG T-ERG G-NOM V-a(A).o(G)}	{A-NOM T-ERG G-NOM V-s(A)}
{A-ERG T-NOM G-NOM V-a(A).o(G)}	{A-NOM T-NOM G-NOM V-s(A)}
{A-ERG P-NOM por(A)-N.EXP V-a(A/3s).o(P)}	{A-NOM P-NOM por(A)-N.EXP V-s(A)}
{A-ERG P-NOM por(A/P)-N.EXP V-a(A).o(P)}	{A-NOM P-NOM por(P)-N.EXP V-s(A)}

Table 2.7: Frames linked by S/A detransitivisation

same procedure can be applied to the detransitivised frames. These frames have an A marked by NOM linked to S-AGR and a non-A argument marked by NOM and not triggering agreement. The abstract detransitivised frame is then {A-NOM O-NOM V-s(A)}.

The use of O (the abbreviation for the grammatical relation selected by a differential marking pattern, see section 1.3.3) in the abstract frames is intended since the argument triggering O-AGR in Chintang is also the one that is central to S/A detransitivisation. Specific O require the transitive frame, non-specific O the detransitivised frame. This is easy to see for the two bivalent frames:

(49) Monotransitive – *khoŋs-* 'play with':

a. *Menuwa-ŋa sencak khoŋs-o-ko.*
cat-ERG mouse play.with-3[s]O-IND.NPST[.3sA]
'The cat plays with the mouse.'

b. *Menuwa sencak khoŋ-no.*
cat mouse play.with-IND.NPST[.3sS]
'The cat plays with mice.' (elicitation PRAR 2010)

(50) Transitive experiential – *rek katt-* 'be angry about':

a. *Ani-ŋa hun-ce ani-rek katt-u-ku-m-ci-m.*
1pi-ERG MED-ns 1piPOR-anger bring.up-3O-IND.NPST-1pA-3nsO-1pA
'We get angry about them.' (elicitation GAR/RBK 2010)

b. *Bamna ani-rek katt-i-niŋ.*
Brahman 1piPOR-anger bring.up-[SUBJ.]1p[i]S-NEG.NPST
'Let's not be angry about the Brahmans (a caste).' (elicitation DKR 2011)

The unique frame used by *som-set(t)-* 'be satisfied with, satisfy' has a special behaviour with respect to S/A detransitivisation. The sense 'be satisfied with', where the experiencer is A, cannot be detransitivised. All detransitivised sentences constructed by me were rejected, and whenever I asked for sentences with non-specific P I got translations with the intransitive equivalent of *som-set(t)-*, *som-si-*, as the one shown in (51d).

(51) *som-set(t)-* 'be satisfied with':

a. *Hun-ce i-som a-sett-u-c-e?*
MED-ns 2sPOR-liver 2sPOR-kill.for-3O-3nsO-IND.PST
'Were you satisfied with them?' (elicitation GAR 2010)

b. **Huŋ=go jo=go=yaŋ u-som se?-no.*
MED=NMLZ$_1$ whoever=NMLZ=ADD 3sPOR-liver kill-IND.NPST
'He's satisfied with anybody.' (elicitation RBK 2012)

c. **Yakkheŋ=le u-som se?-nik-niŋ.*
vegetables=RESTR 3sPOR-liver kill-IND.NPST-NEG
'She's not satisfied with vegetables alone.' (elicitation RBK 2012)

d. *Yakkheŋ-ŋa=le u-som si-nik-niŋ.*
vegetable-ERG=RESTR 3sPOR-liver die-IND.NPST-NEG
'She's not satisfied with vegetables alone.' (elicitation RBK 2012)

This is not completely unexpected. Even in English it is hard to find sentences with a non-specific P referent for 'be satisfied with' because people are usually satisfied with particular things – anything else would seem foolish. In Chintang this tendency seems to be grammaticalised.

Non-specificity is, however, semantically compatible with the other sense of *som-set(t)-* ('satisfy'), and accordingly S/A detransitivisation is possible here. Interestingly this does not depend on whether the P is marked by NOM or GEN, so this frame is the only one that may slightly digress from the abstract frames presented above:

(52) *som-set(t)-* 'satisfy':

a. *Huĩ-sa-ŋa i-som na-sei?.*
MED-OBL-ERG 2sPOR-liver 3>2[s]-kill[.SUBJ.NPST]
'He will satisfy you.' (elicitation GAR 2010)

b. *Cha(-ko) u-som mithai pid-i kina sei?-ma*
child(-GEN) 3sPOR-liver sweets give-[SUBJ.NPST.]1p[i]S SEQ kill-INF
hid-i-ki.
be.able-1p[i]S-IND.NPST
'One can satisfy children by giving sweets to them.' (elicitation RBK 2012)

For the ditransitive frames there is the question whether T, G, or both can trigger detransitivisation. Bickel et al. (2010) argue that the detransitivised variants of these frames imply non-specificity of both T and G and illustrate this first for the double object ditransitive frame with the following examples:

(53) a. *Pi? ghãsa pid-e-h-ẽ.*
cow grass give-PST-1sS-IND.PST
'I gave grass to cows.' (Bickel et al. 2010:388)

b. **Ba=go pi? ghãsa pid-e-h-ẽ.*
PROX=NMLZ$_2$ cow grass give-PST-1sS-IND.PST
'I gave grass to this cow.' (Bickel et al. 2010:389)

c. **Pi? huŋ=go ghãsa pid-e-h-ẽ.*
cow MED=NMLZ$_2$ grass give-PST-1sS-IND.PST
'I gave this grass to cows.' (Bickel et al. 2010:389)

The first example shows that the default reading of the detransitivised double object frame is one where both T and G are non-specific. The second and third examples show that detransitivisation becomes impossible as soon as either T or G is made specific.

This is in conflict with my own elicitation data and also with the data found in the Chintang corpus. The examples in (54), which were elicited by myself, indicate that the detransitivised frame is chosen whenever G is non-specific, no matter whether T is non-specific, too, as in (54a), or whether T is specific, as in (54b). When G is specific as in (54c), the transitive frame is chosen even when T is non-specific. Since G is marked by NOM and linked to O-AGR in the transitive variant of the double object ditransitive frame, this complies with our earlier statement that S/A detransitivisation is generally triggered by O:

(54) Double object ditransitive – *pid-* 'give'

a. *A-pakku cha mithai pi-no.*
1sPOR-younger.uncle child sweets give-IND.NPST[.3sS]
'My uncle gives sweets to children.'

b. *A-pakku ba=go mithai=le cha pi-no.*
1sPOR-younger.uncle PROX=NMLZ sweets=RESTR child give-IND.NPST[.3sS]
'My uncle gives only these sweets to children.'

c. *A-pakku-ŋa ba=go cha-ce=le mithai*
1sPOR-younger.uncle-ERG PROX=NMLZ$_1$ child-ns=RESTR sweets
pid-u-ku-ce.
give-3O-IND.NPST-[3sA.]3nsO

'My uncle gave sweets only to these children.' (elicitation PRAR 2010)

Here are two more examples from the Chintang corpus for specific T and non-specific G with S/A detransitivisation (55b) and for non-specific T and specific G with the transitive frame (56):

(55) a. *Hali, theke khaŋ-a-mett-u-c-e?*
oh why see-2[s]A-CAUS-3O-3nsO-IND.PST
'Why did you show them (your ass)?' (CLC:CLLDCh4R06S03.0950)

b. *Khaŋ-mett-i-niŋ=kha lo!*
see-CAUS-1p[i]S-NEG.[SUBJ.]NPST okay
'Let's not show (your ass) to people, okay?' (CLC:CLLDCh4R06S03.0955)

(56) *Cha-ce ma-khu-th-o-c-a jamma.*
child-ns NEG-bring-NEG-3O-3nsO-IMP[.2sA] all
'Don't bring stuff to all the children.' (CLC:Tel_talk_02.028)

The only sentence in (53) that is in direct contradiction to our claim that S/A detransitivisation is triggered by O is the third, where S/A detransitivisation is impossible in spite of G being non-specific and seemingly because T is specific. However, the ungrammaticalness of this sentence may be explained by another factor. In Chintang, arguments are in general put the farther to the left of the predicate the more topical they are (cf. section 2.3.2). In the sentence in question, the argument that is more likely to be perceived as topical is the specific T *ghãsa*. However, the argument that is *marked* to be more topical by means of word order is the non-specific G *pi?*. Although it is by no means impossible to bring non-specificity and topicality together, this requires a special context that does not necessarily come to the mind of a speaker in an elicitation context. This, and not the fact that T is specific, was probably why *Pi? bago ghãsa pidehẽ* was judged as ungrammatical by Bickel's informant. In the examples in (54) elicited by myself, the specific argument was placed before the non-specific argument in order to avoid clashes between specificity and topicality. See section 2.4.3.1 below for a general discussion of the position of detransitivised O.

Bickel et al. (2010) repeat the claim that detransitivisation is triggered by the non-specificity of T *and* G for the other two ditransitive frames (direct and primary object ditransitive). The examples below are intended to show that both specific T and G are incompatible with S/A detransitivisation.

(57) *(*A-)kam (*a-)khim-be paŋs-e-h-ẽ.*
1sPOR-friend 1sPOR-house-LOC_1 send-PST-1sS-IND.PST
'I sent friends home.'

(58) *(*A-)kam (*a-)gol-ŋa or-e-h-ẽ.*
1sPOR-friend 1sPOR-ball-ERG hit-PST-1sS-IND.PST
'I hit friends with balls.' (Bickel et al. 2010:390)

If it is really O that triggers S/A detransitivisation, the double direct ditransitive verb *paŋs-* 'send' should be detransitivisable even when its G-LOC is specific (*akhimbe* 'to my house'), and the same should be true for the primary object ditransitive verb *or-* 'hit (by throwing)' with a specific T-ERG (*agolŋa* 'with my ball'). However, these examples have the same flaw as those for the double object ditransitive frame – placing O (i.e. the T *kam* for *paŋs-* and the G *kam* for *or-*) to the left of the other ditransitive argument while at the same time using S/A detransitivisation and marking the other argument by a possessive prefix suggests that the referent of O is both more topical than this argument *and* less specific, which is a highly marked constellation and therefore likely to lead to the rejection of these sentences.

The sentences in (59) and (60) below show that S/A detransitivisation is again possible when a clash between specificity and topicality is avoided by placing non-specific arguments closer to the verb. The *a* examples illustrate S/A detransitivisation with both O and the third argument being non-specific, the *b* examples have S/A detransitivisation triggered by a non-specific O and in spite of the third argument being specific, and the *c* examples have the transitive frame triggered by specific O and in spite of the third argument being non-specific.

(59) Direct object ditransitive – *yokt-* 'apply':

a. *Daktar khuwa-be dabai yok-no.*
doctor wound-LOC$_1$ apply-NPST[.3sS]
'The doctor applies medicine to wounds.'

b. *Daktar ba=go khuwa-be khalakhala=kha dabai*
doctor PROX=NMLZ$_1$ wound-LOC$_1$ various=NMLZ$_1$ medicine
yokt-a-s-e.
apply-PST-PRF-IND.PST[.3sS]
'The doctor has applied various medicines to this wound.'

c. *Daktar-ŋa ba=go dabai bibhinnʌ maʔmi-ko khuwa-be*
doctor-ERG PROX=NMLZ$_1$ medicine several person-GEN wound-LOC$_1$
yokt-o-s-e.
apply-3[s]O-PRF-IND.PST[.3sA]
'The doctor has applied this medicine to several people's wounds.'
(elicitation PRAR/RBK 2010)

(60) Primary object ditransitive – *dipt-* 'wrap':

a. *Yo=go kok khali=ta lapho?ã-ŋa dip-no.*
DEM.ACROSS=NMLZ$_1$ rice always=FOC leaf-ERG wrap-IND.NPST[.3sS]
'He always wraps rice with leaves.'

b. *Yo=go kok ba=go lapho?ã-ŋa dip-no.*
DEM.ACROSS=NMLZ$_1$ rice PROX=NMLZ$_1$ leaf-ERG wrap-IND.NPST[.3sS]
'He wraps rice with this leaf.'

c. *Yo-sa-ŋa ba=go kok lapho?ã-ŋa dipt-o-ko.*
DEM.ACROSS-OBL-ERG PROX=NMLZ$_1$ rice leaf-ERG wrap-3[s]O-IND.NPST[.3sA]
'He wraps this rice with leaves.' (elicitation PRAR 2010)

To summarise, the object selected by S/A detransitivisation in Chintang is the NOM-marked argument that is linked to O-AGR in the transitive frame and not indexed at all in the detransitivised frame. This O does not only subsume the formal behaviour of all transitive frames but is also the argument whose referent is responsible for the alternation: specific O require the transitive frame, non-specific O the detransitivised frame. Functionally this puts S/A detransitivisation close to typical differential object indexing systems, where object agreement is a device for tracking referents (Iemmolo 2011:133).

At the end of this section it should be mentioned that S/A detransitivisation is not the only process that defines O. The following argument selectors make reference to precisely the same grammatical relation:

- In S/O detransitivisation, the argument that can become S is always O (section 2.3.4.2).
- The referent of the passive participle *-mayaŋ* is the O of the verb it attaches to.
- Raised agreement in the constructions discussed in section 2.3.5 is always linked to O, be it O-AGR or S-AGR (see sections 2.4.4.1, 2.4.4.2 for details).
- The purposive *-si* can index O and only O by nominal possessive prefixes.

Raised agreement and the purposive also interact with S/A detransitivisation in interesting ways and are therefore discussed in sections 2.4.4.2, 2.4.4.1, and 2.4.4.4 below.

2.4.3 Syntactic independence of detransitivised objects

2.4.3.1 Position

In the last section we have seen that S/A detransitivisation can sometimes interact with word order. Arguments that precede others tend to be interpreted as more topical, and topical referents are much more often specific than non-specific. Therefore, when the non-specific O of a detransitivised

trivalent frame is combined with a specific third argument (T or G, depending on the frame), it may normally not precede that argument since that would imply that it is both more topical and less specific than it.

This could lead us to assume that detransitivised O must always stand next to the verb, which would create another parallel to noun incorporation. This is, however, not the case. While in bivalent frames AOV as in (61) certainly is the default, OAV as in (62) is also possible:

(61) *Khem-e caklet ca-no?*
Khem-NAME.NTVZ sweets eat-IND.NPST[.3sS]
'Does Khem eat sweets?' (CLLDCh1R02S03a.061)

(62) *Nassa akka ca-ŋa-niŋ.*
fish 1s eat-1sS-NEG.[SUBJ.]NPST
'I don't eat fish.' (elicitation PRAR 2010)

Fronting the detransitivised O as in (62) does imply topicality, but as has been mentioned above, this is not impossible in principle. There are two contexts in which it is possible to utter a sentence like (62). Either there could be a fish (or a fish dish) that has been mentioned several times and that is therefore topical. The referent that triggers S/A detransitivisation is then not the fish as a whole but a non-specific subdivision of it – a more appropriate if artificial translation would then be 'I don't eat from the fish'. More about non-specific subamounts can be found in section 2.6 and especially in section 2.6.3.2.

The other option to interpret (62) is to assume contrastive topicality. The topic that justifies fronting *nassa* is then not fish but a larger category such as available dishes ('I would like some soup, but I don't eat fish') or edible animals ('I eat meat but I don't like fish') within which fish is contrasted with one or several other options. This interpretation is more likely than the first possibility because highly topical referents that can be easily inferred are very rarely overtly mentioned in Chintang (cf. section 2.3.1). In the case of contrastive topic it is necessary to mention the contrasting category because only the supercategory (dishes, edible animals) is already accessible.

(63) shows that detransitivised O can not only be separated from the verb by A but also by adjuncts:

(63) *Yum athaba kok=yaŋ car din khe?ŋa le?le ani ca-i-ki.*
salt or rice=ADD four day TMP.ABL only 1pi eat-1p[i]S-IND.NPST
'As for salt and rice, we only eat them four days after (the death of a close relative).' (CLC:LH_Lal.0715)

Detransitivised O can not only occupy the initial position in a clause, they can also be put into the afterthought position after the verb:

(64) *A-ca-no kok?*
2[s]S-eat-IND.NPST rice
'Do you eat rice?' (CLC:CLLDCh2R11S06.243)

In section 2.4.2 above, some examples from Bickel et al. (2010) were discussed. It was speculated that one of these was ungrammatical because of its word order. This example is repeated below for convenience.

(65) **Pi? huŋ=go ghãsa pid-e-h-ẽ.*
cow MED=NMLZ$_2$ grass give-PST-1sS-IND.PST
'I gave this grass to cows.' (Bickel et al. 2010:389)

Pid- 'give' is a double-object ditransitive verb, so O-AGR goes with G and S/A detransitivisation is conditioned by the same argument. The problem in (65) is that the high topicality of *pi?* that is suggested by its being placed before T clashes with the low specificity suggested by its not being indexed on the verb. That this is in fact the case is confirmed by the example in (66), which is structurally parallel but grammatical. The reason is that the *=yaŋ* 'also' on the fronted G invokes a

context of contrastive topicality ('we used to tell this story not only to adults, but also to children'), which reconciles topicality and non-specificity.

(66) *Cha=yaŋ ba katha lud-i-yakt-i-hẽ.*
child=ADD PROX story tell-1p[i]S-IPFV-1p[i]S-IND.PST
'We used to tell this story to children, too.' (elicitation RBK 2012)

Although the placement of O is free in principle, there still might be a statistical association. I tested this for the monotransitive frame, which is the most frequent of the transitive frames. In the syntactically annotated part of the CLC, P directly precedes V (ignoring particles) in 196 out of 737 fully monotransitive frames (27%). The same proportion for the detransitivised monotransitive frame is 134/466 (29%). This difference does not only look small but is also statistically insignificant ($p = 0.32$). It follows that detransitivised P (and probably O in general) are not any more often directly followed by V than other P/O.

To summarise, this section has shown that the position of detransitivised O is flexible and follows in principle the same rules as that of other arguments. This is evidence against an analysis of S/A detransitivisation as noun incorporation.

2.4.3.2 Presence

Differently from S/O detransitivisation and other processes such as noun incorporation, S/A detransitivisation in Chintang does not change the valency of a predicate. We have already seen numerous examples with overt O in various positions that show this. In fact, so far we have not seen any examples with covert detransitivised O, which is unexpected given the general low referential density found in Chintang (section 2.3.1). However, dropping detransitivised O is possible, even though this happens much less frequently than with transitive O or any other arguments. The sentences below show examples of covert O for each major transitive frame.

Transitive – *nek-* 'bite':

(67) *Yaŋ-ce u-nek-no.*
fly-ns 3[p]S-bite-IND.NPST
'Flies bite.' (field notes 2010)

Direct object ditransitive – *tat-* 'bring':

(68) *Hunci-jhani=yaŋ ta?-no naŋ.*
2dPOR-wife=ADD bring-IND.NPST[.3sS] but
'Their wives also bring (that much into marriage).' (CLC:CTN_Fut_Pln.493)

Primary object ditransitive – *ap-* 'hit (by hurling/shooting)':

(69) Ap-no=kha=lo!
hit-IND.NPST[.3sS]=NMLZ$_2$=SURP
'He shoots (at the tangerines).' (CLC:CLLDCh1R06S01.342)

Double object ditransitive – *cind-* 'teach':

(70) *Aba akka=ta cĩ-ya-ʔã=mo kina na aba akka na*
now 1s=FOC teach-1sS-IND.NPST=CIT SEQ CTOP now 1s CTOP
mai-cek-yokt-a-ŋs-e-h-ẽ.
NEG-say-NEG-PST-PRF-PST-1sS-IND.PST
'As for me I didn't say "I will teach".' (CLC:Durga_job.181-182)

Transitive experiential – *rek katt-* 'be angry about':

(71) *Ani-rek katt-i-niŋ.*
1piPOR-anger bring.up-[SUBJ.]1p[i]S-NEG.NPST

'Let's not be angry.' (elicitation RBK 2012)

The conditions for dropping detransitivised O are pragmatic, so an O can be covert whenever it can be reconstructed from the cotext or the context. For instance, (67) was uttered in a very specific context: the speaker was sitting on a veranda on a hot day doing nothing. She said *Yaŋce unekno* while shooing away some flies that were sitting on her daughter's skin. It would have been unnecessary to add *ma?mi* 'people' to the sentence because all other beings that are pestered by flies were irrelevant in that context, anyway.

The other examples with covert O given above can also be explained via pragmatics, too:

- (68) was uttered within a long conversation on the (female) speakers' possibilities and inequalities between women and men. At the time the sentence was uttered it was clear that the only possible T (= O) could be property (men have an advantage over women because the family of their wife has to pay a dowry).
- In (69), an older boy has just given the boy who uttered the sentence an orange. The younger boy sits down to eat it, and the older boy goes away to shoot down more (presumably with a sling, the preferred weapon of boys in Chintang). When the younger boy sees what the other does he says the sentence.
- In (70) the whole preceding paragraph was about teaching students (which are, in a sense, the only possible G (= O) of *cind-* 'teach', anyway).

Another context where S/A detransitivised O can be covert is when they are so unspecific that basically any referent could be inserted. For instance, in one story in the Chintang corpus about a mouse and a cat, the malicious mouse tells all inhabitants of a village that the cat is used to stealing various things such as milk, meat, eggs, and bread (72). When the cat arrives, all villagers know that it steals anything (73) and start to beat it.

(72) *Pempak=yaŋ khut-na-ca-no=kha.*
bread=ADD steal-LNK-eat-IND.NPST[.3sS]=NMLZ$_2$
'"It even steals bread" (, said the mouse.)' (CLC:story_cat.233)

(73) *ʄamma ma?mi-ce-ŋa u-nis-e=pho, menuwa khut-no=kha=mo kina.*
all person-ns-ERG 3pS/A-know-IND.PST=REP cat steal-IND.NPST=NMLZ$_2$=CIT SEQ
'All the people knew that the cat was stealing/was a thief.' (CLC:story_cat.240)

This condition makes dropping O possible even when there is neither cotext nor context. For instance, I was asked the sentence in (74) by a speaker out of context:

(74) *Hana-ko-be u-khu?-ni?-niŋ?*
2s-GEN-LOC$_1$ 3pS-steal-IND.NPST-NEG
'Don't people steal in your country?' (field notes 2011)

Note that dropping O is much less frequently possible in elicitation than in natural data. The reason for this seems to be that elicited sentences usually lack a context, and when the speaker cannot come up with a plausible context himself he will reject the sentence just for that reason. This is especially common in argument dropping, which depends on cotext and context even more than other mechanisms. Elicitation data are thus not very useful here. For instance, O was judged to be obligatorily overt in the elicited example in (75) in spite of the utterance being structurally identical to (67):

(75) *I-phak *(ma?mi) nek-no?*
2sPOR-pig people bite-IND.NPST[.3sS]
'Does your pig bite people?' (elicitation PRAR 2010)

Inspite of the relative freedom speakers have in making detransitivised O overt or covert, it is still remarkable that transitive O are dropped much more often. Out of 1368 fully transitive O in the

annotated part of the CLC, 559 (41%) are overt. The proportion for detransitivised O is notably higher, with 338 out of 544 (62%). This difference is highly significant with $p < 0.01$.

Presently I do not have a straightforward explanation for this. I can offer a motivation, though. Since transitive O are specific, it is possible to track them in discourse. Even if their identity in a greater context is not clear at the time they are first introduced via indexing, it is usually possible to identify them via the things they do or that are done to them within discourse. This possibility bears the promise that the hearer might be able to infer more about the identity of the referent as discourse develops and he learns more and more about it. By contrast, non-specific O cannot be tracked, so it is not possible (or at least rather difficult) for the hearer to learn more about them in the course of a conversation and ultimately identify them. The speaker is thus under greater pressure to identify at least the category of such referents upon their first mention by using an overt NP.

2.4.3.3 NP-hood

If S/A detransitivisation is akin to noun incorporation, one would expect that detransitivised O cannot form the head of fully expanded NPs. This is, however, not the case. Below are examples for detransitivised objects modified by an adjective (76a) and by a numeral (76b).

(76) a. *Akka bajar-be mi=kha bada khed-e-h-ẽ.*
1s market-LOC$_1$ small=NMLZ$_2$ pot buy-PST-1sS-IND.PST
'I bought small pots on the market.'
b. *Akka thitta seu koĩ-ya-ʔã.*
1s one apple search-1sS-IND.NPST
'I'm looking for one (arbitrary) apple.' (elicitation PRAR 2010)

Detransitivised objects are very rarely modified by demonstratives (77a) and possessors (77b). This is due to functional reasons (cf. sections 2.6.4.1, 2.6.4.2).

(77) a. *To cuwa a-thuŋ-no=kha?*
DEM.UP water 2[s]S-drink-IND.NPST=NMLZ$_2$
'Do you drink (from) the water up there?' (CLC:CLLDCh1R03S06.221)
b. *U-kok=ta ca-ŋa-nuŋ.*
3sPOR-rice=FOC eat-1sS-NEG.[SUBJ.]NPST
'I won't eat (from) his rice.' (CLC:CLLDCh1R12S03.428)

Detransitivised O can also be modified by a relative clause (78a) and even form the head of a relative clause themselves (78b):

(78) a. *Akka haŋ-no=go ca-ma le-ŋa-ʔã.*
1s be.hot[.3sS]-IND.NPST=NMLZ$_1$ eat-INF like-1sS-IND.NPST
'I like eating hot stuff.' (elicitation PRAR 2010)
b. *Dhankuta-be tog-i-ki=go kitap caĩ akka ne-ŋa-ʔã-niŋ.*
Dhankuta-LOC$_1$ get-1p[i]S-IND.NPST=NMLZ$_1$ book RETRV 1s read-1sS-IND.NPST-NEG
'I don't read the kind of books one finds in Dhankuta.' (elicitation RBK 2010)

2.4.4 S/A detransitivisation in complex sentences

2.4.4.1 Raising with the infinitive *-ma*

The infinitive *-ma* is one of the two non-finite forms that occur with raising (cf. section 2.3.5). Matrix verbs taking infinitival clauses make use of three raising modes that were already discussed in section 2.3.5.2. They are illustrated once more in (79) with the verb *kond-* 'want, try, must', which occurs with all three modes depending on its semantics. Complete raising (A-ERG and A+O-AGR in the matrix) can be seen in (79a), O to S-AGR (A-ERG and S-AGR with O in the matrix) in (79b), and minimal raising (A-ERG and dummy 3sS-AGR in the matrix) in (79c).

(79) a. *Bhale-ŋa thok-ma na-kon-no goneĩ!*
cock-ERG peck-INF 3>2[s]-want-IND.NPST ATTN
'Watch out, the cock wants to peck you!' (CLC:CLLDCh1R07S02.847)

b. *Master-ce namaskar aphis-be=yaŋ meʔ-ma-ce u-kon-no.*
teacher-ns greeting office-LOC$_1$=ADD do.to-INF-3nsO 3pS-be.necessary-IND.NPST
'We also have to greet the teachers in the office.' (CLC:exp_uni.180)

c. *Huŋ=go cakhaŋ-ce kok-ce na u-mma-ŋa ca-ma=ta kon-no.*
MED=NMLZ$_1$ millet.porridge-ns rice-ns CTOP 3sPOR-mother-ERG eat-INF=FOC be.necessary-IND.NPST
'His mother has to eat those servings of porridge and rice.'
(CLC:phengniba_tale.129c)

The argument that is raised to matrix O- or S-AGR in modes 1 and 2 is the same argument that would trigger O-AGR in an independent matrix, i.e. O. This is illustrated for each of the transitive frames by the sentences below.

(80) Monotransitive frame:

a. *Akka kam-ce tup-ma mai-tok-t-u-cu-h-ẽ.*
1s friend-ns meet-INF NEG-get.to-NEG-3O-3nsO-1sA-IND.PST
'I didn't get to meet (my) friends.'

b. *Akka-ŋa hana khem-ma a-kon-no.*
1s-ERG 2s listen.to-INF 2[s]S-be.necessary-IND.NPST
'I have to listen to you.' (elicitation RBK 2010)

(81) Transitive experiential frame:

a. *Sita-ŋa akka khoiʔ-ma u-ramma u-kaiʔ-ya-ʔã.*
Sita-ERG 1s pester-INF 3sPOR-joy 3[s]A-bring.up-1sO-IND.NPST
'Sita enjoys pestering me.'

b. *Bhaggemani ghatana-ce ani-ramma kaiʔ-ma u-kon-no.*
fortunate event-ns 1piPOR-joy bring.up-INF 3pS-be.necessary-IND.NPST
'One should appreciate fortunate events.' (elicitation SAR 2011)

(82) Direct object ditransitive frame:

a. *Yo-sa-ŋa akka bibhinna des-beʔ khaiʔ-ma u-hi-ya-ʔã.*
DEM.ACROSS-OBL-ERG 1s various country-LOC$_1$ take-INF 3[s]A-be.able-1sO-IND.NPST
'He can take me to various countries.' (elicitation PRAR 2010)

b. *U-mu=ta pok-ma-tha-ma a-kond-e=phe!*
ACCESS-DEM.DOWN=FOC leave-INF-NEXT-INF 2[s]S-be.necessary-IND.NPST=IRR
'One should have left you down there!' (CLC:CLDLCh3R01S04.019)

(83) Primary object ditransitive frame:

a. *Huŋ=go siŋ-ce hana cakku-ŋa dhɨk-ma a-hid-o-ko-ce-niŋ.*
MED=NMLZ$_1$ wood-ns 2s pen.knife-ERG cut-INF 2[s]A-be.able-3O-IND.NPST-ns-NEG
'You can't cut those logs with a pen knife.' (elicitation PRAR 2010)

b. *Hicci-baŋ=ta lauri-ŋa teĩ-ma a-kon-ce-ke.*
two-HUM.CLF=FOC stick-ERG beat-INF 2S-be.necessary-d-IND.NPST
'All two of you should be beaten with a stick.' (elicitation RBK 2010)

(84) Double object ditransitive frame:

a. *Yo-sa-ŋa maʔmi-ce koseli pi-ma nis-o-ko-ce-niŋ.*
DEM.ACROSS-OBL-ERG person-ns present give-INF know-3[s]O-IND.NPST-3nsO-NEG
'He doesn't know to give the people presents.' (elicitation PRAR 2010)

b. *Hun-ce gali pi-ma-ce u-kon-no.*
MED-ns scolding give-INF-3nsO 3pS-be.necessary-IND.NPST
'They should be scolded.' (elicitation RMR 2010)

Bickel et al. (2010:11) claim that with primary object ditransitive verbs both T and G can be raised. However, the examples they provide for this are morphologically ambiguous. They use (85a) to illustrate that G can raise – this is expected, since G is O for primary object ditransitive verbs such as *or-* 'hit (by throwing)'. They then continue with (85b), which allegedly illustrates raising of T with case reassignment (T-ERG > T-NOM). However, the infinitive *oma* can be derived from either *or-* or *os-* 'throw', which is a double object ditransitive verb. If one assumes that (85b) is an instance of *os-*, the example is in accordance with what has been said above: T is O and hance marked by NOM and raisable to S-AGR. No case reassignment has to be assumed.

(85) a. *Gol-ŋa o-ma a-kon-no.*
ball-ERG hit-INF 2[s]S-be.necessary-IND.NPST
'You must be hit with a ball.'
b. *Gol-ce o-ma u-kon-no.*
ball-ns throw-INF 3pS-be.necessary-IND.NPST
'Balls must be thrown.' (Bickel et al. 2010:393)

The specificity of O does not only govern A case and A+O-AGR in simple frames but also in complex frames involving infinitives. A-ERG and A+O-AGR can therefore only be raised when the embedded O is specific. Otherwise S/A detransitivisation is carried out as far as possible under the restrictions posed by the raising mode. In complete raising, this means that the whole complex frame appears in the detransitivised variant. (86) shows an example where (86a) is fully transitive and (86b) is detransitivised.

(86) a. *Pheri maowadi-ce-ŋa maʔmi-ce seiʔ-ma u-lapt-u-ku-ce.*
again maoist-ns-ERG person kill-INF 3pA-be.about.to-3O-IND.NPST-3nsO
'The maoists are about to kill (certain/some) people again.'
b. *Pheri maowadi-ce maʔmi seiʔ-ma u-lap-no.*
again maoist-ns person kill-INF 3pS-be.about.to-IND.NPST
'The maoists are about to kill people again.' (elicitation RBK 2012)

A couple of matrix verbs have A+O-AGR even when the embedded verb is intransitive (section 2.3.5.3). However, this does not in principle constrain their ability to raise S/A detransitivisation. (87) shows an example for *chitt-* 'find the time to'. When the embedded frame is intransitive, *chitt-* has dummy 3sO-AGR (87a). When the embedded frame is transitive, the default is to link embedded O to matrix O-AGR (87b). However, when the embedded O is non-specific, S/A detransitivisation is raised (87c).

(87) a. *Sa-ŋa im-ma chitt-o-ko-niŋ/*
who-ERG sleep-INF find.time.to-3[s]O-IND.NPST[.3sA]-NEG
**chiʔ-nik-niŋ?*
find.time.to-IND.NPST[.3sS]-NEG
'Who doesn't find the time to sleep?' (elicitation SAR 2011)
b. *Phalto maʔmi-ŋa hun-ce cĩ-ma chitt-u-c-e.*
other person-ERG MED-ns teach-INF find.time.to-3O-3nsO-IND.PST[.3sA]
'Somebody else found the time to teach them.' (elicitation RBK 2010)
c. *Sa-lo kam tup-ma chiʔ-nik-niŋ?*
who-NOM friend meet-INF find.time.to-IND.NPST[.3sS]-NEG
'Who doesn't find the time to meet friends?' (elicitation SAR 2011)

(87c) shows that verbs like *chitt-* do not ban intransitive morphology in general but only when the embedded frame is intransitive.

Raising S/A detransitivisation is also possible with the highly flexible and polysemous verb *kond-*, which uses complete raising in the sense 'want, try' but O to S-AGR or dummy 3sS-AGR in the sense 'must'. When the O is specific, the two senses can always be distinguished by the transitive or intransitive inflection of *kond-*. For instance, (88a) can only be taken to code a volitional action. However, when O is non-specific and accordingly the whole frame is detransitivised even with 'want, try', ambiguities as in (88b) can arise.

(88) a. *Debi-ŋa u-kam-ce Kathmandu-be tup-ma kond-u-ku-ce.*
Debi-ERG 3sPOR-friend-ns Kathmandu-LOC$_2$ meet-INF want-3O-IND.NPST-[3sA.]3nsO
'Debi wants to/*must meet her friends in Kathmandu.'
b. *Debi Kathmandu-be nayã kam tup-ma kon-no.*
Debi Kathmandu-LOC$_1$ new friend meet-INF be.necessary-IND.NPST
'Debi wants to/must meet new friends in Kathmandu.' (elicitation RMR 2010)

Two verbs, *let-* 'stop, abandon' and *latt-* 'stop, have had enough of' display a rather special behaviour in that dummy 3sO-AGR is maintained even with non-specific O but S/A detransitivisation still shows up in the case of A:

(89) a. *Abinas-ŋa kok ca-ma led-o-s-e.*
Abinas-ERG rice eat-INF stop-3[s]O-PRF-IND.PST[.3sA]
'Abinas has stopped eating the rice.'
b. *Abinas kok ca-ma led-o-s-e.*
Abinas rice eat-INF stop-3[s]O-PRF-IND.PST[.3sA]
'Abinas has stopped eating rice.' (elicitation SAR 2011)

In raising mode 2 (O to S-AGR), embedded O is linked to matrix S-AGR. We would expect that this is not possible with non-specific O since they also cannot trigger O-AGR. This is hard but not impossible to prove. The problem is that SAP O are always specific, so raising is always possible with them. In order to produce a minimal pair, we have to use NSAP O. 3sO is excluded because the result of raising 3sO to S-AGR is indistinguishable from dummy 3sS-AGR, so the only context where the effect of specificity on this kind of raising can be investigated is with 3nsO. (90) shows an example with a dual. In (90a), which doesn't have raising, both a specific and a non-specific interpretation are possible. (90b), which does have raising, only allows a specific interpretation. Thus, while specificity does not fully determine raising (a specific O is compatible with both raising and dummy 3sS-AGR) it does constrain it (only a specific O can be raised).

(90) a. *Ba teĩ-be hicci-baŋ ni-ma kon-no.*
PROX village-LOC$_1$ two-HUM.CLF know-INF be.necessary-IND.NPST[.3sS]
'In this village you have to know two (specific) people.'
b. *Ba teĩ-be hicci-baŋ ni-ma u-kon-ce-ke.*
PROX village-LOC$_1$ two-HUM.CLF know-INF 3nsS-be.necessary-d-IND.NPST
'In this village there are two people you have to know.' (elicitation RBK 2012)

In mode 3 (dummy 3sS-AGR), where only A case is raised, it is still possible to distinguish between the transitive and the detransitivised frame precisely by this criterion. Compare (91a), where the embedded O is specific and accordingly A is marked by ERG, with (91b), where O is non-specific and A marked by NOM:

(91) a. *Ha=go bha-i?=ko rahansahan lis-e, hun=go*
PROX=NMLZ$_1$ PROX-LOC$_2$=NMLZ$_1$ tradition become-IND.PST[.3sS] MED=NMLZ$_1$
samet kani-ŋa ni-ma kond-a-ŋs-e.
as.well 1pi-ERG know-INF be.necessary-PST-PRF-IND.PST[.3sS]
'That is a tradition of this place, so we have to know that as well.'
(CLC:Student_life.085)

b. *Hun-ce thitta them=yaŋ khei?-ma kon-nik-niŋ.*
MED-ns one what=ADD buy-INF be.necessary-IND.NPST[.3sS]-NEG
'They don't have to buy anything.' (CLC:Durga_job.161)

2.4.4.2 Raising with the converb *-saŋa*

The foregrounding converb *-saŋa* is the other non-finite form that occurs with raising (see section 2.3.5). Since the light verbs used together with *-saŋa* have special grammaticalised meanings in this construction, *-saŋa* always has true raising, that is, the finite verb does not assign roles to the arguments that it indexes (in contrast to the infinitive, where frame superimposition is possible). (92) shows another example for this.

(92) *Pacche jo pujari-ŋa sahuliyat pi-ma par-ne pi-saŋa*
later whatever priest-ERG assistance give-INF be.necessary-IPFV.PTCP give-CVB.FGR
na-khai?.
3>2[s]-take[.SUBJ.NPST]
'Later the priest will start giving you whatever assistance he must give you.' (CLC:kothari_talk.164)

As with the infinitive, the specificity of the O governs raising. When the embedded O is specific, the matrix frame is transitive (93a); when it is non-specific, the matrix frame is S/A detransitivised (93b).

(93) a. *Huĩ-sa-ŋa jagga-be=ko urbarasakti is-saŋa*
MED-OBL-ERG land-LOC$_1$=NMLZ$_1$ fertility destroy-CVB.FGR
khatt-o-ŋs-e.
take-3[s]O-PRF-IND.PST[.3sA]
'But that started to spoil the fertility of the land.' (CLC:exp-wadh_DK.096)

b. *Abo pahila bhonda ani teĩ-be nikkai ta tup-saŋa*
now earlier COMP 1pi village-LOC$_1$ much FOC understand-CVB.FGR
thapt-i-ki.
bring.across-1p[i]S-IND.NPST
'Now we come to understand more than in earlier times in the village.' (CLC:chintang_now.713-714)

A specialty of the *-saŋa* constructions is that in some etymologically related pairs of light verbs a lexically intransitive light verb is used when the embedded frame is intransitive and a lexically transitive light verb when the embedded frame is transitive (section 2.3.5.4). Interestingly, both the intransitive and the transitive variant are possible when the embedded P is non-specific and the whole sentence is therefore S/A detransitivised. This is shown in (94), where *khat-* 'go' and *khatt-* 'take' are alternatively used with a non-specific O:

(94) a. *Huŋ-khi pod-e num-saŋa khatt-i pacchi…*
MED-MOD learn-V.NTVZ do-CVB.FGR take-[SUBJ.]1p[i]S POST
'After we take up studying in that way…' (CLC:Ganesh_talk.131)

b. *ʒatti badde pod-e num-saŋa khad-i…*
however.much much learn-V.NTVZ do-CVB.FGR go-[SUBJ.]1p[i]S
'The more we study…' (CLC:Ganesh_talk.129)

This points to the detransitivised frame taking an intermediate position between the transitive and the intransitive frame: case and agreement justify using an intransitive light verb, whereas the presence of more than two arguments justifies a transitive light verb. So far I have not been able to find a difference between the two variants. A similar phenomenon is found with the transitivity-sensitive aspect marker *-hat(t)* (see section 2.6.4.3).

2.4.4.3 Agreement with the infinitive *-ma*

Many non-finite verbs in Chintang are only non-finite in the sense that they have reduced possibilities for expressing morphological categories. In particular, most non-finite forms can show agreement of some kind. The infinitive is no exception here in that it has optional agreement with 3nsO (cf. section 2.2.3). This is particularly frequent in connection with deontic light verbs (95) or in the independent use of the infinitive, which occurs with events that follow a schedule (96).

(95) *Koni abo pha-ma-ce par-y-o naŋ.*
no.idea now help-INF-3nsO fall-PST-3s but
'I don't know, probably (I) should help them.' (CLC:Gen_talk.065)

(96) *Taŋphekma mei?-ma, jhyal dhoka-ce cup-ma-ce…*
broomstick apply-INF window door-ns close-INF-3nsO
'(I had various chores such as) sweeping the floor, closing the windows and doors…'
(CLC:origin_myth.412)

The verbal suffix *-ce* [3nsO] is obviously related to the nominal suffix *-ce* [ns]. However, synchronically these two should be kept separate. One reason is that they are functionally distinct even on verbs. Several events or several states resulting from events are not the same as a single event with several objects. Compare the following two clauses, where the first contains a nominalising infinitive with *-ce* [ns] and the second contains *-ce* [3nsO]:

(97) a. *U-ca-ma-ce charasta a-pokt-u-m-cu-m.*
3sPOR-eat-INF-ns scattered 2A-leave-3O-2pA-3nsO-[SUBJ.NPST.]2pA
'You might leave his foodstuffs scattered.' (CLC:CLLDCh2R04S04.0494)
b. *Ca-ma-ce kon-no?.*
eat-INF-3nsO be.necessary-IND.NPST[.3sS]
'They should be eaten.' (CLC:phengniba_tale.398)

Another, more important reason why *-ce* on infinitives should be considered as a verbal suffix is also relevant in the context of S/A detransitivisation. *-ce* can only be used with transitive verbs and is not used with any non-singular objects but only with those roles that would also trigger O-AGR with a finite verb form. This is shown in the examples below.

Monotransitive – *khag-* 'look after':

(98) *hani-gor-ce=yaŋ ma-khaŋ-ma-ce=ta*
2pPOR-ox-ns=ADD NEG-look.after-INF-3nsO=FOC
'without looking after your oxen' (CLC:CLLDCh2R04S04.0616)

Transitive experiential frame – *patte katt-* 'trust':

(99) *Ani-guru-ce patte kai?-ma-ce kon-no.*
1piPOR-teacher-ns trust bring.up-INF-3nsO be.necessary-IND.NPST[.3sS]
'We should trust our teachers.' (elicitation SAR 2011)

Direct object ditransitive – *choŋs-* 'take (to), deliver':

(100) *Ka-ŋi-ce caĩ ba-i tai?-ma-ce u-ŋakt-e.*
ACT.PTCP-know-ns RETRV PROX-LOC_2 bring-INF-3nsO 3pS-be.necessary-IND.PST
'All knowing people had to be brought here.' (CLC:origin_myth.412)

Primary object ditransitive – *humd-* 'bury':

(101) *Akka hum-ma-ce hou!*
1s bury-INF-3nsO AFF
'I'm going to bury you two!' (CLC:CLLDCh1R02S04.0218)

Double object ditransitive – *cind-* 'teach':

(102) *Akka cĩ-ma-ce hid-u-ŋ-cu-ŋ-niŋ.*
1s teach-INF-3nsO be.able-3O-1sA-3nsO-1sA-[SUBJ.]NEG.NPST
'I can't teach them.' (CLC:Durga_job.051)

As a logical consequence, *-ma* [INF] can only be followed by *-ce* [3nsO] when its O is specific, i.e. under the same conditions when O-AGR would be expected on a finite verb:

(103) a. *Ana-pic-ce ghãsa pi-ma-ce kon-no.*
1sPOR-cow-ns grass give-INF-3nsO must-IND.NPST[.3sS]
'Our cows should be given grass.'
b. *Maʔmi khem-ma(*-ce) kon-no.*
people listen.to-INF-3nsO must-IND.NPST[.3sS]
'People should be listened to.' (elicitation RBK 2010)

2.4.4.4 Agreement with the purposive *-si*

The purposive *-si* is used with intransitive verbs of motion and with direct object ditransitive verbs to indicate the purpose of a motion. For instance:

(104) a. *U-phuwa-ŋa u-nisa sambok biu bhuk-si*
3sPOR-elder.brother-ERG 3sPOR-younger.brother millet seed sow-PURP
paŋs-e.
send-IND.NPST[.3sS/A]
'The elder brother sent the younger to sow millet.' (CLC:phengniba_tale.027)
b. *Mo kina u-nisa sambok biu bhuk-si khad-e.*
CIT SEQ 3sPOR-younger.brother millet seed sow-PURP go-IND.PST[.3sS]
'And so the younger brother went to sow millet.' (CLC:phengniba_tale.029)

When the O of the verb marked by *-si* is animate it can optionally be indexed by a nominal possessive prefix. This is shown in the following examples.[7]
Monotransitive – *las-* 'fetch':

(105) *J̌ite=lo i-la-si kad-a-ŋs-e.*
Jite=SURP 2sPOR-fetch-PURP come.up-PST-PRF-IND.PST[.3sS]
'Jite has come up to fetch you.' (CLC:CLLDCh2R03S04.0140)

Direct object ditransitive – *choŋs-* 'take (to), deliver':

(106) *U-choŋ-si khac-ce o.*
3sPOR-deliver-PURP go-[SUBJ.NPST.1]d[iS] RECNF
'Let's go to bring him back.' (elicitation SAR 2011)

Primary object ditransitive – *ap-* 'hit (by hurling/shooting)':

(107) *U-ap-si u-kuŋs-a-s-a=kha.*
3sPOR-hit-PURP 3pS-come.down-PST-PRF-PST=$NMLZ_2$
'They have come down to shoot it.' (elicitation SAR 2011)

Double object ditransitive – *pid-* 'give':

(108) *Hana i-saman i-pi-si lond-e-h-ẽ.*
2s 2sPOR-goods 2sPOR-give-PURP set.out-PST-1sS-IND.PST
'I've come to give you your stuff.' (elicitation SAR 2011)

[7] Since it is hard to imagine a context where an emotion is the purpose of a motion I did not add an example for a transitive experiential verb marked by *-si*.

As one would expect, this is not possible when the O of the purposive is non-specific (109a), and when O is indexed it can only be interpreted as specific (109b). Note, though, that the purposive without a possessive prefix can be interpreted as having either a specific or a non-specific O (109c). Thus, while specificity constrains indexation with *-si* it does not fully determine it.

(109) a. *Kam (*u-)tup-si kha?-no.*
friend 3sPOR-meet-PURP go-IND.NPST[3sS]
'He goes to meet friends.'
b. *Kam u-tup-si kha?-no.*
friend 3sPOR-meet-PURP go-IND.NPST[3sS]
'He goes to meet a (specific) friend/*friends.'
c. *Kam tup-si kha?-no.*
friend meet-PURP go-IND.NPST[3sS]
'He goes to meet a friend.' (elicitation RBK 2012)

2.4.4.5 Overt A in non-finite subclauses

There are two contexts where non-finite verb forms can have an overt A that is unambiguously syntactically affiliated with them. One is where there is no coreferentiality constraint operating between the non-finite verb and the associated finite verb, as in (110), where *ani* is only assigned a role by the infinitive *helakhaŋma* and accordingly marked by ERG:

(110) *Ani-ŋa Ɉalpadebi helakhaŋ-ma i?-no=kha.*
1pi-ERG Jālpādevī neglect-INF be.bad-IND.NPST[.3sS]=NMLZ$_2$
'We are not allowed to neglect *Ɉālpādevī* (a goddess).' (CLC:on_jalpadebi.011)

But even when there is a coreferentiality constraint (as in most constructions involving non-finite forms), syntactic affiliation sometimes becomes clear by case. For instance, the purposive *-si* is only used when its own S/A is coreferential with a moving argument (S or double object ditransitive T) in the matrix. In the examples in (111) the A of the *-si* form is coreferential with the S of the non-finite verb. Since A and S require different cases in the relevant frames, it is clear that the overt NP representing S/A has been assigned case by the non-finite verb in (111a) but by the *-si* form in (111b):

(111) a. *U-nna=pho Gauroŋ-ma jethi pha-si*
3sPOR-elder.sister=REP Gaurong-F eldest.daughter help-PURP
khad-a-ŋs-a=kha.
go-PST-PRF-PST[.3sS]=NMLZ$_2$
'I heard her sister has gone to help the eldest of the Gaurong clan.'
(CLC:CLDLCh2R02S02.504)
b. *Ba-ce-ŋa a-ses-si u-tiy-a-ŋs-e.*
PROX-ns-ERG 1sPOR-kill-PURP 3pS-come-PST-PRF-IND.PST
'They have come to kill me.' (CLC:INT_JYR.0488)

Although it may seem unusual that non-finite forms can assign case, this is well attested in Chintang with the purposive and the foregrounding converb *-saŋa*.

If non-finite forms can assign case to an A, it is theoretically possible that they can express S/A detransitivisation even without O-AGR, viz. via A-NOM. The conditions for this are, however, highly restricted: when the finite verb shares its S or T with the non-finite verb's A, it cannot be decided whether an overt NOM-marked NP has been assigned case by the non-finite or by the finite verb. One thus either needs a case where A is shared between the two verbs but their O are different and only the main verb O is specific, or a case where there is no sharing at all (and ERG is not optional, differently from (110), where NOM would also have been possible because the A is a pronoun). These special constellations are not attested in the Chintang corpus. Here is an elicited example that shows that S/A detransitivisation can indeed be expressed via A case in the latter situation:

(112) a. *Cha-ŋa ba arkha thuŋ-ma i?-no.*
child-ERG PROX alcohol drink-INF be.bad-IND.NPST[.3sS]
'A child mustn't drink this schnapps.'
b. *Cha arkha thuŋ-ma i?-no.*
child alcohol drink-INF be.bad-IND.NPST[.3sS]
'Children mustn't drink schnapps.' (elicitation RBK 2012)

2.5 Functional preliminaries: Identifying referents

2.5.1 Introduction

The referential property of O that triggers S/A detransitivisation has so far been approximated as specificity. Below are two examples that give a first impression of what this means.

(113) a. *Debi-ŋa seu kond-o-ko.*
Debi-ERG apple look.for-3[s]O-IND.NPST[.3sA]
'Debi is looking for the/an apple.'
b. *Debi seu kon-no.*
Debi apple look.for-IND.NPST[.3sS]
'Debi is looking for apples.' (elicitation PRAR 2010)

(114) a. *Abo sa tac-c-o.*
now meat bring-[1]d[iA]-[SUBJ.]3[s]O
'Now let's bring the meat.'
b. *Abo sa tac-ce.*
now meat bring-[SUBJ.NPST.1]d[iS]
'Now let's bring (some) meat.' (elicitation PRAR 2010)

The first sentence in each pair is transitive, the second sentence detransitivised. In (113a), Debi is looking for one specific apple, which may or may not be identifiable for the hearer. Similarly, (114a) is about a specific amount of meat. In (113b), Debi is looking for apples in general – there might be one or several. The parallel (114b) is about an non-specific amount of meat – the speaker's group could bring more or less. These cursory characterisations already indicate that the kind of specificity that is relevant for Chintang is closely connected to quantification. This connection is especially striking with mass nouns such as *sa* 'meat' but can, as we will see, also be claimed for count nouns such as *seu* 'apple'.

In most instances of the transitive frame the English translation has an article (definite or indefinite) on the object, whereas in most instances of the detransitivised frame the zero article is used. This shows that the two phenomena revolve around a similar functional variable, or possibly the same variable with somewhat different settings. Cases of divergence are useful for studying the semantics of S/A detransitivisation more precisely. Below are two example sentences for this, which follow each other in the corpus. *Dabaice* 'medicines' in the first sentence has O-AGR but no article in English. The same word in the second sentence is most natural with the definite article in English but triggers S/A detransitivisation in Chintang.

(115) a. *C-o-wakt-u-c-e-ta dabai-ce.*
eat-3O-IPFV-3O-ns-IND.PST[.3sA]-CONT medicine-ns
'He used to take various kinds of medicine.' (CLC:appa_katha_talk.020)
b. *Ca-saŋa=ta numd-a-kt-a-lok ek dini*
eat-CVB.FGR=FOC do-PST-IPFV-[SUBJ.]PST[.3sS]-CVB.BGR one day
a-phe-ce bhai?-ni u-thab-a-ci-e.
1sPOR-elder.brother-ns PROX-DIR$_1$ 3[p]S-come.over-PST-COMPL$_2$-IND.PST
'While he was still taking the medicine (he had got from the hospital), one day my brother's family came over for a visit.'
(CLC:appa_katha_talk.021-022)

In order to explain cases like this one and to get a more precise idea of what specificity means in Chintang, we first need a clear definition that is both strict enough to describe phenomena such as the English articles and flexible enough to account for differences between languages. Below we will not only discuss specificity but also definiteness since that gives us a broader view on the topic and both are closely linked in the literature, anyway.

The categories in question have been among the most intensively discussed in linguistics for more than a century: probably the oldest work which is still of relevance today is Russell (1905),[8] two very recent ones Abbott (2010) and Kibrik (2011). Though there has been some progress in that new and sometimes more powerful concepts have been developed to explain definiteness and specificity, research in this area is still hampered by a couple of flaws:

- There has been what one might call "theoretical wholism": functional concepts such as familiarity and uniqueness have been viewed as monolithic. Accordingly the discussion has mostly been for or against concepts as a whole, thus preventing a more fine-grained understanding of definiteness and specificity.[9]
- Instead of explaining the whole range of definiteness and specificity, many scholars base their discussion on a few very special examples. An early counterexample to this is Hawkins (1978), who has collected and discusses a large amount of diverse examples. However, only recently have there been tendencies to include data from natural language corpora (e.g. Epstein 2002, Kambarov 2008).
- There has been little comparative work. To my knowledge the only large-scale typological work so far is Lyons (1999), who is, however, theoretically superficial and also does not make clear statements about what unites and what distinguishes phenomena in various languages on the functional side. A more recent milestone has been the book by Kibrik (2011), which is very aware of linguistic diversity and proposes a lot of typological parameters but still is not based on a large set of systematically arranged data itself and has too wide a scope to cover all details. Typological work is urgently needed since definiteness and specificity are by no means exotic phenomena. Apart from the well-known Germanic and Romance article languages, many more exotic languages and families feature articles, too (e.g. Insular Celtic, Arabic, Lakhota…). What's more, two widespread phenomena are often linked to these functions, viz. differential argument marking (especially differential object marking, cf. Bossong 1998) and antipassives (Heath 1976, Cooreman 1994).

The third problem is clearly out of scope of the present work – in this section, we will focus on Chintang. The second problem is not a problem for this study since most examples have been taken from the Chintang corpus. As regards the first problem, I would like to briefly sketch below how I believe the discussion can be made more transparent. I will start from an intuitive understanding of definiteness and specificity based on English and will then try to render this understanding more precisely step by step.

2.5.2 The identification process

What are definiteness and specifity about? This question has so far generally been taken to be about the functional factors behind these phenomena. As important as these may be, I believe there is a sense to the question which is prior to them, viz. 'What are they functions *of*?', or, more precisely, 'Which cognitive process motivates paying attention to definiteness and specificity?' This is the process of identifying referents. Speakers mention or imply referents all the time, and hearers have to identify these referents with entities in their own mind. Definiteness and specificity as

[8] If one takes into account less influential work research on definiteness and related topics dates even farther back – cf., for instance, the bibliography in Christophersen (1939).

[9] One counterexample to this practice is the work of Chesterman (1991), who tries to explain the distribution of the English articles and the Finnish partitive by combining three functional concepts, viz. (mental) locatability, inclusiveness, and "extension" (a variable indicating whether a referent is instantiated ("actualised") or not). He then presents a unified theory based on relations between various kind of sets such as (p. 69) the "entity set" and the "referent set". He does not make sufficiently clear, however, how the three concepts presented first are related to this theory.

grammatical functions serve to indicate whether and how this is possible. They are thus about IDENTIFIABILITY in the first place.

The statements above may seem trivial, but they provide a simple base for talking more systematically about definiteness and specificity. If the identification of referents is the process motivating these phenomena, the following components may play a role for their explanation:

- Pointing: The speaker points to a referent.
- Enhancement: The hearer enhances the given information concerning the identity of the referent by all available means.
- Identification: Using the combined information, the hearer tries to identify the referent in a mental space.
- Ability estimation: The speaker estimates whether the hearer can do this.

Several comments are in place at this point.

The first component might be taken to equal giving a referring expression. In fact, this is what many authors have implicitly assumed, especially those coming from the philosophical tradition or focussing on English (e.g. Chesterman 1991, Gundel et al. 1993, Abbott 2010). However, in many of the languages of the world, the default for given referents in argument roles is not to mark them overtly with a referring expression but to leave them covert. This is why I have chosen the term POINTING to cover both overt marking and implication. This issue is highly relevant for Chintang because it is an extreme language with respect to "pro-dropping" (cf. section 2.3.1). It is not only possible and in fact usual to drop any referent that has been previously mentioned in discourse but also to drop referents that have not been mentioned before. For instance, the following sentence is completely normal, whether uttered at the beginning or in the middle of a conversation:

(116) *Pid-o-ŋs-e.*
give-3[s]O-PRF-IND.PST[.3sA]
'(He/somebody) has given (it/something) (to him/somebody).'
(CLC:CLLDCh2R01S01b.1368)

I will use the term POINTER below where it is necessary to cover both referring expressions (= overt pointers) and argument roles not occupied by an overt NP (= covert pointers).

The relevance of the second component in the process of referent identification – the enhancement of information by the hearer – is directly related to that of the first, because overt pointers often (if not mostly) do not provide all the information that is necessary for identifying a referent. The important distinction between pointers and what the hearer makes out of them is not always acknowledged in the literature. For instance, Birner and Ward (1994:93) cite the sentence *"It's hot in here. Could you please open the window?"* (as uttered in a room with three "equally salient" windows) as an example for the use of the English definite article with an "entity" that is not uniquely identifiable. However, even though the NP *the window* does of course not refer to such a unique referent in an arbitrary context, it must do so in the context where this utterance is made, i.e. the hearer has to be able to enhance the given pointer so that one referent emerges as the intended one. If this condition is not given, the definite article becomes infelicitous. If all three windows are closed and there are no other hints to the identity of the window (most importantly the locations of speaker and hearer in the room), *Could you please open the window?* will most likely result in the reply *Which one?* Thus, Birner and Ward's example only demonstrates that pointers marked by *the* need not have a uniquely identifiable referent; however, the example is irrelevant for *enhanced* pointers. This distinction becomes all the more important in a language like Chintang where arguments are covert all the time.

Epstein (2002:337) provides a list of sources of additional information that are widely recognised in the literature:

- previous discourse
- situational context
- common background of speaker/hearer

- world knowledge
- bridging (= association based on world knowledge)

The third component of the identification process – identification proper – is the most important one. Several things need to be said here. First, it is by no means a matter of course that referents are mental entities. Though much of the more recent linguistic discussion of reference (especially in typology) is based on this typically implicit assumption, there are fields (e.g. formal semantics) where it is still highly unusual. What's more, there is also the philosophical tradition where it is one of the defining criteria of referents that they exist in the real world. For instance, Heim (1983) states that not all indefinite and definite expressions refer and introduces the term "file card" for entities that correspond to NPs uttered in discourse but not necessarily to referents in the real world. Chesterman (1991:10) likewise mentions "non-referential" definite expressions and concludes that reference is not important for the description of definiteness. Abbott (2010), which otherwise presents very informed overviews of research into reference and a lot sophisticated discussion, does not even mention the possibility that referents might not be real world entities.

The reason why I chose to define referents as mental entities here is a terminological one. Presently there can be no doubt that the inclusion of mental referents is most useful for the description of definiteness and specificity (and probably of any linguistic phenomena related to reference) – for instance, the use of the English definite article in fiction can hardly be explained if one assumes that its use marks identifiability in the real world. However, among the scholars acknowleding this so far nobody has produced a comprehensive terminology. What is lacking in particular is a common term for the relation between pointers and entities – if "refer" and "reference" are disallowed for talking about mental entities, various non-technical terms have to be resorted to here. To me it seems easier and clearer to understand reference in a broader sense than to try to find a new term. Kinds of referents can be distinguished by adjectives where necessary (e.g. "real world referents" vs "mental referents").

Another comment concerns the notion of mental spaces used in the list above. This term was introduced by Gilles Fauconnier, first as "espaces mentaux" in Fauconnier (1984) and later in the now more commonly known translated form in Fauconnier (1994), and has been popular ever since in cognitive linguistics. In discussions of definiteness and specificity it has been used, for instance, by Epstein (1999, 2002) and Kambarov (2008).[10] Including mental spaces in the description of identification processes does not only make clear that referents are primarily mental entities but also makes it possible to put into words the difference between classical cases such as *"x was the father of Charles II"* (Russell 1905:481) and examples such as *"He had been an academic gypsy ever since the fire"* (Epstein 1999:65, cited from a work of fiction).

In the first example, identification takes place in a mental space that maps the real world. Thus, *the father of Charles II* does not only have a referent in the sense adopted here but also in the philosophical sense. By contrast, in the second example the speaker does not know anything about *the fire*, which is mentioned here for the first time, not even whether it has a counterpart in the real world or not. Its referent is thus only identifiable in the mental space opened by the story.[11] Note that "discourse referents" as they are used in Discourse Representation Theory (Kamp 1981) or File Change Semantics (Heim 1983) are not exactly identical to mental referents. There is at least one important difference between mental spaces and discourse, which is that mental spaces may be multiple, so one referent can exist in several linked spaces simultaneously.

If one takes together all the extensions of identifiability made above, one arrives at a concept which is quite far from the everyday understanding of identifiability: the speaker has to give very

[10] Interestingly, spatial metaphors were already present in the discussion of definiteness before Fauconnier. For instance, Hawkins (1978) generally speaks of "locating" referents.

[11] While it is mostly easily possible to identify which mental space is relevant for the identification of a referent, this is not always the case. For instance, in a sentences like *Anybody could do it, anybody* may not be identifiable in the base space. However, if the relevant space is the one where *it* actually happens, it is possible to refer to *anybody* as to an identifiable referent: *Anybody could do it, and then after that he'd just disappear. Imagine you met that guy.* This problem is known as the problem of donkey anaphora (based on the oft-cited *"Every farmer who has a donkey beats it"*) in the literature (cf. e.g. Roberts 2003:321).

little to no information, the hearer adds whatever he can from his own knowledge, and identification only takes place within mental spaces. A better term might thus be accessibility, as advocated e.g. by Ariel (1988, 1990) and von Heusinger (1997, 2007). I will still stick to the term identifiability because it seems to me that there are important links between the term as it is used here and its more common meaning. Another reason is that proponents of accessibility usually view this as a scalar concept, which to me seems to be a confusion of the notions of possibility and ease of identification.

Finally, the fourth component of referent identification – the assumption of the speaker about the abilities of the hearer – actually comes first in chronological order. Without such an assumption, the speaker could not use linguistic markers of definiteness and specificity. That identifiability is never an objective truth but always filtered by the mind of the speaker becomes clear from cases of mismatches between speaker assumptions and hearer knowledge, e.g. when a speaker asks a hearer *Have you seen the book?* and gets the answer *Which one?* Here, the speaker assumed too much knowledge on the side of the hearer. The book was not identifiable in general but thought to be so by the speaker.

It is important that this component does not only concern the knowledge of the hearer but also his assumed ability to adapt to new situations. This is because a speaker may be more or less challenging when he presents a referent as identifiable. He may do so when the present knowledge of the hearer is already sufficient to identify the referent. However, he may also do so when he well knows that it is not but thinks that the hearer is able to orient himself in a mental space that is such that once one is familiar with it one can identify the referent. Situations of the latter type are discussed as "first-mention definites" in the literature.

An example cited by Abbott (2010:220) is *"The new curling center at MSU, which you probably haven't heard of, is the first of its kind".* Here, the speaker marks *new curling center at MSU* as definite even though he himself acknowledges in the inserted relative clause that the hearer does not have the knowledge to identify it. However, he still assumes that the hearer will be able to construct a mental space where there is only one center that can be talked about. This is evidenced by two facts. One is that he gives some additional hints to the identity of the bowling center – it is new, and it is located at MSU (Michigan State University). Even if a university had several bowling centers, it would be highly unusual if two new ones opened at the same time, so these hints greatly facilitate the construction of a mental space within which the bowling center is identifiable (by contrast, consider how strange *The bowling center, which you probably haven't heard of, is the first of its kind* sounds – such usage would only be possible in combination with a great amount of information enhancement from the side of the hearer, but probably not if there is no such information and the speaker wants to prompt the hearer to construct a new mental space). The second piece of evidence is the minimally contrasting sentence *A new bowling center at MSU, which you probably haven't heard of, is the first of its kind.* The difference to the first example is not that the bowling center is not identifiable in this sentence. It is only *presented* as not identifiable, or in other words, the speaker behaves less challenging here.[12]

Being challenging is conventionalised in many situations. For instance, Fraurud (1996:76) discusses the example *"There is a problem with the carburettor"* (said by a mechanic to a customer) as evidence against identifiability as a major condition for using *the* – an average customer may not even know what a carburettor is, let alone be able to identify it among the parts of a car. However, a gentler mechanic could again have said *There is a part in most cars that is called "carburettor", and there is a problem with that in your car.* The second version is semantically sound but pragmatically strange because the convention in short conversations of this type is for the expert to talk to a customer *as if* to another expert.

To summarise, we have discussed four components which may play a role for the description of identifiability: pointing, enhancement, identification, and ability estimation. The next step in the discussion will be to consider how these components can be used to formulate claims about the function of definiteness and specificity markers more clearly.

[12]Note that these remarks are not what Abbott connects with the example cited from her work. She uses it to argue against the familiarity theory of definiteness (see below).

2.5.3 Definiteness

In this section I will briefly examine a couple of functional concepts that have frequently been used to explain DEFINITENESS (almost always as exemplified by the English article *the*). These concepts are familiarity, identifiability, uniqueness (together with inclusiveness), and determined reference. They will be analysed based on the terms introduced above in order to highlight their characteristics and to arrive at a definition of definiteness that brings together their advantages and is flexible enough to allow adjustments for the description of individual languages.

The first concept, FAMILIARITY, was made popular by Christophersen (1939). Though he was not the first to use the term (he himself dates it back to Brown 1851), his study was certainly the most influential one using it. However, somewhat ironically, the "familiarity theory of definiteness" that is generally criticised for being simplistic and that has been ascribed to him e.g. by Heim (1983) and Lyons (1999) does not correspond to Christophersen's original definition, which is anything else but simple – some scholars may have been misled by the everyday sense of the term familiarity. Still, since the focus of the discussion of the usefulness of familiarity has not been on the technical sense employed by Christophersen but precisely on the everyday sense I will start from that sense, too.

Familiarity may be used to explain the use of the English definite article, which would be used whenever a hearer is already familiar with a referent at the time of its mention. There are many obvious counterexamples to it, e.g. *"The president of Ghana is visiting tomorrow"* (Lyons 1999:5).[13] Using the components described above, we can now state what precisely is wrong about the familiarity theory instead of rejecting it as a whole. One problem is that the theory only accepts a limited set of enhancement methods. Previous mention in discourse certainly makes a referent familiar, as does the personal acquaintance that is implied by a common background and by some types of situations. However, there are also situations where a hearer only becomes familiar with a referent *after* it is mentioned. It is also not clear how to link world knowledge and especially bridging to familiarity. – Another problem is that this approach has a tendency towards viewing the real world as the main space for identification. This is due to the non-technical semantics of the term "familiar": one would probably say someone who has actually seen dodos is more familiar with them than someone who has only read about them in books. However, both sources of knowledge may suffice to identify a dodo in a given context (cf. *The last dodo died in the 17th century*).

Another concept discussed by Lyons is IDENTIFIABILITY. Identifiability in his sense is a much narrower concept than the one laid out above. For instance, Lyons tries to prove that identifiability is not sufficient to explain the use of the English definite article by citing sentences such as *"I've just been to a wedding. The bride wore blue"* (Lyons 1999:7). Thus, what he seems to have in mind when talking about identifiability is identifiability in the real world (that is in our terms, in a mental space corresponding to a hearer's knowledge of the real world). However, there is no obvious reason why identifiability should be restricted in this way. Even if the hearer was not at the wedding and consequently cannot identify the referent called *the bride* in the real world, he can do so within the mental space spanned by the story of the speaker. *The bride wore blue* is noteworthy not because it informs the hearer that some person he knows wore blue but because it is generally more usual to wear white. Thus, it seems imprecise to say that identifiability is insufficient for explaining definiteness – it only is in the narrow sense Lyons attributes to it. Translated into the terms introduced above we can say that restrictions on identification space are of little explanatory use.

One of the most popular concepts in the literature is UNIQUENESS, which according to Abbott (2006) goes back to Russell (1905). Lyons (1999:8) states that a description is unique if "there is just one entity satisfying the description used". A similar definition is probably implied by Farkas (2002), who claims that one of the problems of this concept is to include contextual information (Farkas 2002:8). Her critique depends on a narrow reading of uniqueness as represented by Lyons' definition. However, nowhere does she make clear why such a narrow definition is useful. Instead

[13] Christophersen has, among other things, provided for the possibility of bridging (one aspect of what he calls "implicit contextual basis", cf. Christophersen 1939:29), so this example would not be a counterexample to familiarity in his sense – he even mentions a virtually identical example (*the king* after mentioning *a country*) on p.30.

of linking uniqueness to descriptions, one can simply link it to the combined information gained from pointing and enhancement. This not only preserves the usefulness of uniqueness for the discussion of identifiability but also avoids the problems we have already seen in connection with a fixation on overt pointers.

Another problem Lyons sees with uniqueness occurs with plurals and mass nouns: one can say something like *"The queen gave out the prizes"* (Lyons 1999:11) even though there is no unique entity satisfying *the prizes* but several possible sets, among them the set of all prizes, but also less complete subsets. This is an old but solved problem; Abbott (2010:159) states that the solution provided by Link (1983) is generally accepted in formal linguistics, and Lyons himself cites Hawkins (1978), whose functional concept of INCLUSIVENESS is likely to have originated from a similar intuition to that of Link. According to Hawkins (1978:17), inclusive reference is to "the totality of the objects satisfying the descriptive predicate within the relevant pragmatic set".

Lyons tries to present examples where a definite article is used while reference is not inclusive but is apparently not fully aware of the implications of Hawkins' definition. For instance, he cites the sentence *"Close the door, please"* (as uttered in a room with three doors, Lyons 1999:14) as evidence against the rule of inclusiveness. To be sure, *the door* does not refer to all doors in the room but to only one. However, this is all doors that are found in Hawkins' "relevant pragmatic set" – the most plausible case would be one where there is only one door that is open.

In summary, uniqueness is not as problematic as viewed by some – problems mainly arise under too narrow definitions. Farkas' critique is only relevant if one restricts the information necessary for identifying a referent to what is provided by the speaker – which is problematic, anyway, as has been shown earlier. Lyons' critique is based on a similar flaw, as he overlooks the possibility of narrowing down the set of potential referents by making use of pragmatic information. Thus, of the concepts discussed so far, uniqueness (with the addition of inclusiveness) seems the most useful one. Expressed in the terminological system introduced above, uniqueness (or better "unique identifiability") is one possible value of the ability estimation performed by the speaker before marking definiteness or specificity.

There is one concept left for discussion, which is DETERMINED REFERENCE. In contrast to the older ideas presented above, determined reference is relatively recent and seems to have been introduced by Farkas (2002). According to her definition, an NP has determined reference if the choice of value for the variable introduced by it into the discourse is fixed (Farkas 2002:9). Swart (2006:168), who refers to Farkas, puts this into somewhat more accessible words: "a variable x has determined reference if the value assigned to the discourse referent in the model remains stable across further developments of the discourse" (Swart 2006:168). Farkas uses determined reference to overcome the shortcomings of both familiarity and uniqueness, two approaches she views as opposed to each other. However, as we have already seen, Farkas' view of uniqueness is unnecessarily strict. In fact, Farkas herself says that "determined reference is a special type of uniqueness" (Farkas 2002:9), so instead of introducing another new term we may simply state that uniqueness needs to be taken in a wider sense in order to take into account enhancement, as has already been done above.

I will now summarise the discussion of concepts that have been used in the literature to explain definiteness and specificity. I hope to have shown above that a more fine-grained terminology for the basic components of referent identification makes it possible to draw connections between the most important views and to contrast them more easily. The components that were used for this are pointing, enhancement, identification, and ability estimation. Instead of discussing the usefulness of explanatory concepts as a whole it is often more fruitful to look at which components they focus on and which they neglect. An integrated view of identifiability in a wide sense should ideally consider all components. The concept that has proven to be most useful under this aspect is uniqueness (combined with inclusiveness). If we take uniqueness as one value of identifiability as suggested above, we get to the following preliminary definition of definiteness: *a set (or mass) of referents is definite if the speaker thinks the hearer can uniquely identify all its members (or parts) in the relevant mental space, using and enhancing the information given by himself.*

Note that this is not to say that definiteness is the same in all languages. It is widely known

that it is not – cf. the sentences from well-known European languages below, English vs French in (117), English vs German in (118):

(117) a. I like dogs.
b. *J'-aime l-es chien-s.*
1s-love DEF-PL dog-PL
'I like (the) dogs.'

(118) a. He is at school.
b. *Er ist in d-er Schule.*
3sm be.NPST.3s in DEF-fs.DAT school(f)
'He is at school.'

Variation across languages can be accounted for in two ways under our definition. One is to adjust the components of the identification process. For instance, one language may allow less forms of enhancement than another one, or speakers of one language may be conventionally more challenging when estimating the ability to identify a referent. The other possibility is to formulate language-specific rules. For instance, possessive pronouns may incorporate definiteness (as in English: **the my friend*) or not (as in Italian: *mi-o amico* [1sPOR-sm friend(m)] can be preceded by either *il* [DEF.sm] or *un* [IDF.sm]). Such particularities are usually easier to describe by simple rules than with recourse to function. The flexibility of definiteness built into the definition presented here is an advantage over models which are fixated on English.

2.5.4 Specificity

Linguistic SPECIFICITY has never been as popular a topic as definiteness – maybe because it is not as deeply rooted in philosophy, and maybe also because *the* symbol of definiteness (the English article) is much more frequent than any similar form that could be taken to mark specificity (for instance, it is more than twice as frequent as *a* according to the BNC frequency lists found on www.kilgarriff.co.uk/bnc-readme.html, accessed on 9 February 2011). Nevertheless, specificity is quite a frequent term in grammatical descriptions. This section examines the concept and tries to find out whether it can be rendered more precisely in terms of the identification process framework introduced above.

Definitions of specificity are often surprisingly unspecific. A good example is Enç 1991, who in spite of having written a dedicated, oft-cited article becomes no more precise than saying that specificity means "being a subset of or standing in some recoverable relation to a familiar object" (p. 24). Lyons (1999:35) spells out common sense when he contrasts specificity and definiteness by saying that definiteness depends on two persons (hearer and speaker), whereas specificity only depends on one and is given whenever the speaker has a "particular referent in mind". This puts specificity very close to the definition of definiteness given above – having a particular referent in mind is similar to being able to identify it (in the sense used above, that is, not necessarily in the real world). That would make specificity basically another value of identifiability, but with an additional parameter (the person being able to identify a referent) set to the speaker and with severe consequences for the steps in the identification process: while it is still meaningful to talk of pointers, pointers are no longer used to facilitate identification, and enhancement even becomes completely redundant – the speaker simply uses his own knowledge. The ability estimate changes to simple ability. This is the solution we will choose in the end. Before that, however, some problems have to be discussed.

One possible objection is that viewing specificity in this way is an oversimplification. For instance, Lyons (1999:174) (citing Ioup 1977) mentions that two types of specificity have to be distinguished. One type is specificity in transparent contexts (i.e. where no counterfactual operator is present), the other specificity in opaque contexts (where there is such an operator). A similar distinction is made by Farkas (1994), whose "epistemic specificity" and "scopal specificity" roughly correspond to Lyons' specificity in transparent and opaque contexts, respectively (the main difference being that Farkas' scopal specificity makes reference not only to the scope of counterfactual

operators but also to that of quantifiers).[14] Lyons mentions the following values and examples for these types:

- **referential** := specific in a transparent context. Speaker refers to a particular referent, e.g. *"I haven't started the class yet; I'm missing a student – Mary's always late."* (Lyons 1999:170)
- **non-referential** := non-specific in a transparent context. Speaker referes to no particular referent, e.g. *"I haven't started the class yet; I'm missing a student – there should be fifteen, and I only count forteen."* (Lyons 1999:170)
- **narrow scope** := specific in an opaque context. Counterfactual operator does not take scope over existential operator, e.g. *"Peter intends to marry a merchant banker – even though he doesn't get on at all with her."* (Lyons 1999:167)
- **wide scope** := non-specific in an opaque context. Counterfactual operator does take scope over existential operator, e.g. *"Peter intends to marry a merchant banker – though he hasn't met one yet."* (Lyons 1999:167)

As insightful as this classification is – the terminology obscures the common base of all these phenomena. Why, after all, is it possible to refer to both referential and narrow scope NPs as specific in non-technical language? In order to find out, "opaque context" and "transparent context" should first be replaced by terms connected to the theory of mental spaces. Lyons' counterfactual operators correspond to what is called "space builders" there (Fauconnier 1994, earlier "introducteurs d'espace" in Fauconnier 1984). A space builder is any linguistic form (e.g. a modal verb, a conjunction, or a mood marker) that has the ability to derive spaces from the current base space which are partially or fully independent of it. Opaque contexts are then contexts where a space builder has set up an additional space whereas no such space is present in transparent contexts.

Now the remaining distinctions can be integrated into the framework. The difference between the two *I'm missing a student* examples concerns the knowledge of the speaker, or more precisely, the richness of the information he can access in order to identify a referent. In the first case this information is very detailed – the speaker knows the name of the referent and presumably a couple of other things, too, such as her outer appearance, parts of her behaviour etc. In the second case, by contrast, the information is as poor as can be – the only characteristic of the student is that he is missing. Still, that would usually be enough to identify him if he came in and sat down.

The *Peter intends to marry a merchant banker* sentences show a different distinction but can likewise be easily integrated. This distinction tends to correlate with the knowledge of the speaker but does not necessarily do so – the speaker might know much more about Peter's imagination than about his intentions in the real world. Thus, what really distinguishes the two sentences is the connectedness of the referent to be identified. In one case that referent is located in a derived space but is connected to an entity in the base space. In the other case there is no such connection – identification is only possible within the derived space.

In this way Lyons' fourfold distinction can be broken up into meaningful components. For instance, his "narrow scope use" can be replaced by the more transparent (if lengthy) characterisation "referent is identifiable by the speaker within a derived space connected to the base space".

Another problem for our initial suggestion to put definiteness and specificity on a common base is that it has often been claimed that the two phenomena are independent of each other. For instance, Klages-Kubitzki (1995:32) cites the following examples from Dik (Dik 1989:144) in order to prove that the basic values of definiteness and specificity can be freely combined (note that in her terminology, "generic" is the opposite of "specific"):

- indefinite + specific: *I saw a dog in the garden.*
- indefinite + generic: *A dog is a faithful pet.*
- definite + specific: *I saw the dog in the garden.*
- definite + generic: *The dog is a very faithful pet.*

The specificity Klages-Kubitzki is talking about here corresponds to the token/type distinction: *a dog* and *the dog* correspond to tokens in the first and third examples but to types in the second and

[14] A third type recognised by Farkas, "partitive specificity", is only applicable to plural referents and not relevant here.

fourth examples. However, this distinction is not only independent of definiteness as encoded by the English articles – basically any type of NP can be used to signify a type or token. Chesterman (1991) shows this with examples such as *"Oil floats on water"* (p. 35, type use of zero-article NP) or *"Continued destruction of the rainforest will lead to the extermination of some rare insects"* (p. 37). Frequently a single NP may have both readings, e.g. in *Two whales have already disappeared.* Thus, all the examples above really say is that specificity in the sense of the type/token distinction does not help much to explain the use of the English articles.

Somewhat more watertight arguments for the independency of definiteness and specificity are once more put forward by Lyons. He argues that the distinction between transparent and opaque contexts cannot only be applied to indefinites but also to definites, including the distinctions made within each type. Here are his examples:

- **referential**: *"We can't start the seminar, because the student who's giving the presentation is absent – typical of Bill, he's so unreliable."* (Lyons 1999:172)
- **non-referential**: *"We can't start the seminar, because the student who's giving the presentation is absent – I'd go and find whoever it is, but no-one can remember, and half the class is absent."* (Lyons 1999:172)
- **narrow scope**: *"I'm still searching for the solution to this puzzle – and I think I'm close to finding it."* (Lyons 1999:168)
- **wide scope**: *"I'm still searching for the solution to this puzzle – though John insists it's insoluble and I think he's probably right."* (Lyons 1999:168)

If one assumes that definiteness and specificity as a whole are values of a single variable, examples such as these must indeed be taken as counterevidence. However, under our hypothesis that the two phenomena only have a common base (identifiability) they come as nothing much of a surprise. In fact, they can not only be easily integrated into our framework but are even expected, since the parameters distinguishing them (identification space, richness of knowledge, connection to base space) are independent of the parameter distinguishing definiteness and specificity (i.e. person).

To summarise, putting specificity and definiteness on a common base has the advantages of simplifying terminology and clarifying the relations between the two phenomena. As long as one uses several parameters to characterise each of them there won't be any problems with partial independence.

Before we give the final definitions for definiteness and specificity that will be used henceforth, there is one question that remains to be answered: is specificity entailed by definiteness? So far we have taken definiteness as identifiability on part of the hearer and specificity as identifiability on part of the speaker, so the question may also be asked as: are there cases where the hearer can identify a referent but the speaker can't? The answer depends on how one defines identification. To be sure, there are many cases where the knowledge the hearer can access for identifaction is richer than that available to the speaker. For instance, when phoning a friend who has just given birth one might well ask *How is the baby?* even if one hasn't seen the child so far. However, even though the hearer knows more about the referent in question than the hearer in such cases, it is still true that both can identify it within the relevant mental space (and in fact, the information that is available to the speaker would also suffice for the hearer). We will thus hypothesise that the assumption that the hearer can identify a referent is only possible if the speaker himself can do so. This means that definiteness indeed entails specificity.

One last practical point concerns terminology. There should be a term covering both groups of individual referents and masses. "Set" is inappropriate for this because it is not easily extended to masses, especially when taken in the mathematical sense. Bunt (1979, 1985) uses the term ENSEMBLE to cover referent groups, masses, and singular referents (ATOMIC ENSEMBLE). I will take this over as a practical cover term here, however, without the formal implications made by Bunt.

Here are the final definitions for both specificity and definiteness:

*A referent ensemble is **specific** if the speaker can uniquely identify all its members (or parts) in the relevant mental space.*
*A referent ensemble is **definite** if it is specific and the speaker thinks the hearer can identify it the same way he can, using and enhancing the information given by himself.*

2.5.5 The basis of unique identifiability

In the discussion above, we have relied on the important concept of unique identifiability without defining our intuitions about it precisely. This is maybe excusable since, although the term has been used a lot in the literature, so far I haven't been able to find a definition anywhere.

In order to uniquely identify a single referent, a speaker has to be able to tell it apart from all other potential referents. The addition "potential" is crucial because it is usually impossible to tell a referent apart from all other *existing* referents. In order to be able to tell apart one referent from others, it is necessary to have sufficient knowledge about criteria that distinguish it from them, such as its present location, its name, its colour, and a hundred thousand more. For instance, if somebody tells his child sitting under the christmas tree *Start with the red one!*, the child may use the criterion *red* to distinguish one present from all others that are potentially relevant and thus identify it. If the child is told *Come on, open it!*, things become more complicated but do not change in principle: the child would have to infer that his parent does not consider further description necessary because the intended present is in some way more salient than others – for instance, it might have the child's name on it, or it might be the one he is already holding in his hands. After checking which criteria might be relevant, the present is picked out in a similar fashion as before.

There are two ways in which one could imagine this process to work. One would be to go through all potential referents and check whether they have the necessary criteria or not. The other would be to have an index where it is possible to search for criteria and jump directly to the referents matching them. In either case the identification process creates two groups: a set of "good" referents which have the necessary criteria (containing only one member in the present case) and a set of "bad" referents which do not have it.

Now what about referent groups? Here, instead of telling apart a single referent from others, it is necessary to tell apart several. However, the method to do this is just the same (i.e., checking criteria) – the only difference is that there is more than one referent matching the relevant criteria. We still get the same two sets as a result of the process.

Masses are a little tricky, but again not different in principle: in the easiest case they can be checked as a whole (especially if they are in a container). If a mass does not only have to be distinguished from other referents but also from an adherent mass with similar referential criteria it becomes necessary to subdivide it until one finds out where the boundaries for the relevant criteria are.[15] The result of the criterion check is not a set but again a mass – however, one with fixed boundaries.

The point I would like to make here is that in all cases (single referent, plural count or mass referent), the identification process seems to imply a referent ensemble with fixed quantity. This means that upon closer inspection, inclusiveness is actually not an optional addition to unique identifiability but an inherent characteristic: if one tries to break down identification to simpler notions, it turns out that it has got to do with distinguishing referents from each other using criteria, and this process automatically creates inclusive groups. For instance, recall the example *"The queen gave out the prizes"* cited by Lyons (1999:11) in order to show that inclusiveness is needed in addition to unique identifiability to explain the use of *the*. His argument was that in such sentences *the* is used even though there is no unique group of prizes but a number of subsets for which the predicate is also true. This argument becomes void once we assume identification to work as described above. The pointer *the prizes* tells the hearer that there is a group of referents that can be distinguished from all other potential referents by checking the criterion *prize*. This means that when going

[15] This case is rare but possible. Imagine, for instance, that someone has cooked three pots of rice and put all the rice into a large bowl. It is only after that that somebody else realises the third pot has cooked a little too long and tells him *Maybe we should take the overdone rice out again*

through all referents or through an index of criteria, the hearer cannot stop after he has identified one or a couple of matching referents – otherwise he cannot be sure that the members of the group are really different from all other potential referents (there might still be some left to which they are identical with respect to the relevant criteria).

Only a referent that can be identified in the described way can be tracked in discourse, and all referents that can be tracked must be identifiable. The reason is that if it is not sufficiently clear which or what belong to the "good" ensemble it is not possible to identify two instantiations of the ensemble with each other. For instance if I say *He likes pears*, it is clear that all object referents to which the statement is applicable are pears but not which pears are intended.[16] Even if I use the same pointer again later it will be impossible to tell whether the ensembles are identical. It is therefore grammatical to say *John likes pears, but only the green ones. Susan also likes pears, but only overripe ones.*

2.6 Functional properties of S/A detransitivisation

2.6.1 Quantifiability

We can now reconsider the semantics of S/A detransitivisation before the background of the last section. So far the function of S/A detransitivisation has been approximated as specificity. Above we have argued that specificity has a common base with definiteness in unique identifiability, that unique identifiability entails inclusiveness, and that using it in the identification process creates referent ensembles with fixed quantities.

My claim for Chintang is that the kind of specificity that is relevant for S/A detransitivisation is strongly associated with this last aspect. In order for a referent to be trackable, its quantity has to be in principle determinable. This is because tracking a referent in discourse means to be able to identify intended ensembles with each other across clauses and larger units. This is impossible unless it is clear which (single) referents or which parts belongs to an intended ensemble and which don't, and this condition automatically creates intended ensembles with a fixed quantity, as discussed in section 2.5.5 above. This problem is most pronounced with mass referents: a subdivision of a mass cannot be distinguished at all (no matter whether there is an intended reference or not) from others unless it is quantifiable via physical boundaries or measures.

Henceforth, referents whose quantity can in principle be determined will be called quantifiable and referents for which this does not hold will be said to be non-quantifiable. The biggest part of S/A detransitivisation can be explained if we assume that the transitive frame is used with quantifiable O and the detransitivised frame with non-quantifiable O. Let's reconsider the pair of examples given at the beginning of the preceding section from this angle:

(119) a. *Debi-ŋa seu kond-o-ko.*
Debi-ERG apple look.for-3[s]O-IND.NPST[.3sA]
'Debi is looking for the/an apple.'

b. *Debi seu kon-no.*
Debi apple look.for-IND.NPST[.3sA]
'Debi is looking for apples.' (elicitation PRAR 2010)

(120) a. *Abo sa tac-c-o.*
now meat bring-d-[SUBJ.NPST.1]d[iA.]3[s]O
'Now let's bring the meat.'

b. *Abo sa tac-ce.*
now meat bring-[SUBJ.NPST.1]d[iS]
'Now let's bring (some) meat.' (elicitation PRAR 2010)

In (119a), there is exactly one apple that Debi is looking for. Accordingly the A *Debi* carries ERG and the apple is indexed by *-u* [3O]. By contrast in (119b), there is no clear way to separate the

[16] It is not all pears, since it is possible to say *He likes pears, but only the green ones* but not *He likes all pears, but only the green ones.*

apples Debi is looking for from the ones that she doesn't want, so the detransitivised frame is used. What's more, the number of intended referents is indeterminate – Debi's search could equally well be said to have been successful if she found one, two, or twenty apples. *Debi found apples – twenty, to be precise* is natural, whereas *Debi found an apple – twenty, to be precise* sounds strange.

The examples in (120) are different in that *sa* 'meat' is per default homogeneous in the terms of Rijkhoff (2002), which means that one can subdivide a piece of meat and can still call the pieces "meat" (whereas one cannot call an arbitrary piece of an apple "apple"). Notwithstanding, tracking on the base of quantifiability works in parallel fashion. In (120a) there is a fixed amount of meat that is to be brought. My informant commented on this sentence that one would use it for instance at a wedding where there is a course of meat. This is why (120a) makes use of the transitive frame. By contrast, (120b) does not refer to a fixed amount – the sentence would be equally felicitous if the speaker group brought a whole bowl of meat or only a single piece. It is therefore impossible to separate one subdivision of meat from others and track it, and the detransitivised frame marks this.

The role of quantifiability becomes best visible with examples with overt quantifiers such as in (121) and (122), which are quantifiable by definition and almost always yield the transitive frame:

(121) *Etti~ti=kha kharayo-ce hicce u-tad-u-ŋs-u-c-e*
this.big~INTENS=NMLZ$_2$ hare-ns two 3[p]A-bring-3O-PRF-3O-ns-IND.PST
u-hik-ki-ce-ta.
3[p]A-keep-IND.NPST-3nsO-CONT
'He brought two hares as big as this and now he's keeping them.' (CLC:ctn_talk01.039)

(122) *Aseĩ a-mma Kathmandu khad-a-loĩs-a bela=ta*
last.time 1sPOR-mother Kathmandu go-PST-out-PST[.3sS] time=FOC
a-nicha-ce-ŋa bisauli sa u-c-o-hatt-e!
1sPOR-younger.sibling-ns-ERG 1.25kg meat 3[p]A-eat-3[s]O-AWAY.TR-IND.PST
'Last time my mother went to Kathmandu my brothers ate one and a half kilo of meat!'
(CLC:CLLDCh2R12S04.279)

The precise role of quantifiability will be discussed in detail in section 2.6.3, and its impact on S/A detransitivisation will be assessed in quantitative terms in section 2.7. However, before that we have to take a look at some other minor aspects that play a role for S/A detransitivisation.

2.6.2 Specificity and arbitrary reference

As we have seen above, quantifiability is a prerequisite for specificity and a central factor behind S/A detransitivisation. This section will show some cases where quantifiability alone does not provide an explanation and where one has to resort to a more general concept of specificity. Since quantifiability is a prerequisite for identification, there are no cases where a non-quantifiable referent is used with the transitive frame. However, there are some cases where a referent that is quantifiable cannot yet or need not be identified, and in these cases quantifiable referents may be used with the detransitivised frame.

In section 2.5.2 above we identified various steps in the identification of a referent. Since specificity equals identifiability on the part of the speaker only, most of these steps are irrelevant for Chintang. There are just two points left where typological variation is expected and where we therefore have to take a closer look. These are detailedness ("How much does a speaker have to know about a referent in order to consider it identifiable?") and mental spaces ("In which space does a referent have to be identifiable?").

The first of these questions is easily answered: any degree of detailedness makes a referent identifiable that suffices to set it apart from all referents that might be confused with it. This claim is illustrated by the mini-conversation below, translated freely from Lyons (1999:170):

(123) *Akka thitta iskule kakchya-be mai-khaŋ-yokt-u-ŋs-u-h-ẽ.*
1s one pupil class-LOC$_1$ NEG-see-PST.NEG-3[s]O-PRF-3[s]O-1sA-IND.PST
'I haven't seen one pupil in class.' (elicitation DKR 2011)

(124) *Sa-lo=kha?*
who-NOM=NMLZ$_2$
'Who is it?' (elicitation DKR 2011)

(125) a. *Huŋ=go Gita=kha, hana a-nis-o-ko hola.*
MED=NMLZ$_1$ Gita=NMLZ$_2$ 2s 2[s]A-know-3[s]O-IND.NPST maybe
'It's that Gita, maybe you know her.'

b. *Koni, krʌmsʌnkhya pʌndrʌ-jʌna u-ti-akt-a=kha, tʌrʌ etibela*
no.idea list fifteen-HUM.CLF 3[p]S-come-IPFV-PST=NMLZ$_2$ but now
somma coudhʌ-jʌna=le? u-yuŋ-no.
TERM fourteen-HUM.CLF=RESTR 3[p]S-be.there-IND.NPST
'I don't know, there were 15 people on the list, but so far there are only 14.' (elicitation DKR 2011)

(124) is felicitous with both (125a) and (125b) as its continuation, i.e. both if the speaker knows the missing pupil and if he doesn't. Lyons calls the NP *thitta iskule* in (125) referential in the first case and non-referential in the second case. This terminology is based on the ideal of real-world identifiability: in (125a), the speaker can identify the missing pupil in the real world, whereas in (125b) he doesn't know anything about him – not even his name if he hasn't checked yet who the 14 attending pupils correspond to. We have seen above that real-world identifiability is only a special case of identifiability in general, and this pair of examples shows that this type does not play any special role in Chintang. Even though in (125b) the teacher can't identify the missing pupil outside the classroom, it is easy to set him apart from all other pupils in the class by his not having attended yet. This makes him identifiable, so (125b) requires the transitive frame just as well as (125a).

Since minimal details are sufficient to track a referent, the predicate itself can also serve as a criterion to distinguish one referent from others. This makes utterances such as (126) possible. In the first sentence *ghãsa* 'grass' is non-identifiable because the speaker cut an indefinite amount of grass. However, this sentence creates a new referent – all grass cut by the speaker – which itself is identifiable and therefore used with the transitive frame in the second sentence:

(126) *Asinda akka ghãsa hekt-e-h-ẽ ni. Hana huĩ (ghãsa) hokko-i?*
yesterday 1s grass cut-PST-1sS-IND.PST ASS 2s MED grass which-LOC$_2$
a-khatt-o-ŋs-e?
2[s]A-take-3[s]O-PRF-IND.PST
'Yesterday I cut grass, right. Where did you take that grass?' (elicitation DKR 2011)

Mental spaces, on the other hand, do play a role for S/A detransitivisation. Consider the following pair of examples (again inspired by the English examples in Lyons (1999:167)):

(127) a. *Gita-ŋa bepari appi=go maʔmi num-ma=mo mitt-o-ko tara*
Gita-ERG merchant self=NMLZ$_1$ person make-INF=CIT think-3[s]O-IND.NPST[.3sA] but
huĩ-sa-ko u-bihor-a=ta ciʔ-no.
MED-OBL-GEN 3sPOR-behaviour-NTVZ=FOC be.bad-IND.NPST
'Gita would like to marry a merchant, but his manners are bad.'

b. *Gita bepari appi=go maʔmi num-ma=mo miʔ-no tara*
Gita merchant self=NMLZ$_1$ person make-INF=CIT think-IND.NPST[.3sS] but
huŋ=go miʔ-no likhi bepari
MED=NMLZ$_1$ think-IND.NPST[.3sS] EQU merchant
mai-chiʔ-yokt-a-ŋs-e.
NEG-find-PST.NEG-PST-PRF-IND.PST[.3sS]
'Gita would like to marry a merchant, but (so far) she hasn't found one that is like she imagines.' (elicitation DKR 2011)

Lyons distinguishes these sentences by saying that the counterfactual operator (here represented by *mitt-* + INF 'would like') has narrow scope in (127a) and wide scope in (127b) (i.e. it takes scope over the existential operator that creates the representation of the merchant). This can not fully explain what happens in the Chintang version of (127), though, because S/A detransitivisation can also be used in similar non-opaque contexts, for instance:

(128) a. *Akka asinda sum-bhaŋ u-tiy-a=go maʔmi-ce*
1s yesterday three-HUM.CLF 3[p]S-come-[SUBJ.]PST=NMLZ$_1$ person-ns
kond-u-ku-cu-ŋ-ta.
search-3O-IND.NPST-ns-1sA-CONT
'I'm searching for the three people who came (here) yesterday.'

b. *Akka sum-bhaŋ ka-pha-pa maʔmi koĩ-yã-ʔã-ta,*
1s three-HUM.CLF ACT.PTCP-help-REF person search-1sS-IND.NPST-CONT
jo=go nusayaŋ yaŋs-o.
whoever=NMLZ$_1$ CONCS be.good.for-[SUBJ.3sA.]3[s]O
'I'm searching for three helpers, anyone is okay.' (elicitation RBK 2012)

What brings (127b) and (128b) together is what I will call ARBITRARY REFERENCE. Although Gita wants to marry exactly one merchant and the speaker in (128b) is looking for exactly three persons, the speaker in both cases cannot set apart any corresponding referents from others and therefore cannot identify them. This is because the referents will only be fixed by the agents' efforts and cannot yet be tracked in the base space. They can, however, be tracked in the relevant derived space, that is, if one looks forward, for instance, into the time where Gita has found herself a husband, that husband is an easy to identify referent. It is therefore possible to say (129):

(129) *Gita bepari appi=go maʔmi num-ma=mo miʔ-no. Tara jibʌn bhʌri*
Gita merchant self=NMLZ$_1$ person make-INF=CIT think-IND.NPST[.3sS] but life long
maya mett-o-niŋ hola.
love do.to-3[s]O-NEG.[SUBJ.]NPST[.3sA] maybe
'Gita would like to marry a merchant. But she probably won't love him all her life.'
(elicitation JK 2012)

We must therefore now be more specific about the connection between identifiability and S/A detransitivisation. The transitive frame does not indicate specificity in general, but specificity in the mental space that presently gets most attention.

The focus of attention can sometimes change rather quickly. For instance, in (130) the relative clause refers to an event in a hypothetical space while the main clause refers to what happened in the base space. Both clauses contain an object referent *tauli* 'towel', but this referent is only identifiable in the hypothetical space, where the speaker has found and bought one towel she likes. It is therefore linked to O-AGR in the relative clause but not in the main clause:

(130) Akka khaŋ-ma les-u-ŋ=go tauli mai-chiʔ-yokt-a-ŋs-e-h-ẽ.
1s see-INF like-3[s]O-[SUBJ.]1sA=NMLZ$_1$ towel NEG-find-PST.NEG-PST-PRF-PST-1sS-IND.PST
'I haven't found a towel to my liking.' (elicitation DKR 2011)

The reverse case (detransitivised relative clause, transitive main clause) is illustrated by (131). Note that the relative clause is headless in this example.

(131) *U-cek-no=go=yaŋ u-toŋs-o-ko.*
3[p]S-say-IND.NPST=NMLZ$_1$=ADD 3[p]S-make.fit-3[s]O-IND.NPST
'They also coordinate what they say.' (CLC:chintang_now.882)

There are again two mental spaces involved here, one reflecting the general situation in which the A of *cekt-* say things, the other containing one concrete situation (or a set of such situations) in which all that has been said is coordinated (via modern media). Sentences such as (130) and (131)

raise interesting problems concerning the semantic makeup of relative clauses. While it is still true that in these sentences a referent is in some way shared between the relative and the main clause, that referent can apparently be represented in different mental spaces and can be viewed differently in terms of quantification.

Arbitrary reference of a simpler type is also frequently encountered in everyday conversation. For instance, in (132) the speaker uses S/A detransitivisation in order to express that he will fetch one stool (only one was needed in the context) but that it's not fixed yet which one:

(132) *Akka muda thap-ma-ʔã.*
1s stool fetch-1sS-IND.NPST
'I'll fetch a stool.' (field notes 2010)

Similarly, in (133) the speaker communicates that he will tell a story but isn't sure yet which one. This example is more striking because it contains an overt numeral.

(133) *Akka=yaŋ paĩ mi=kha thitta katha cek-ma=mo miʔ-ya-ʔã.*
1s=ADD today small=NMLZ$_2$ one story tell-INF=CIT think-1sS-IND.NPST
'I, too, want to tell a small story today.' (CLC:love_story.003)

An informant I asked when it would be appropriate to use the corresponding transitive form *mitt-u-ku-ŋ* [think-3[s]O-IND.NPST-1sA] said that this form would have been likely if the speaker had been prompted to tell a specific story.

The examples of arbitrary reference that we have seen so far are of a subtype that I will call OPEN REFERENCE because it requires that the link between a pointer and a referent is not fixed yet at event time. There is another subtype in which the arbitrarity of this link is retrospective rather than anticipatory and which I will call DISCARDABLE REFERENCE. Below is an example.

(134) *Lauri kekt-a-ŋs-e.*
stick hold-PST-PRF-IND.PST[.3sS]
'He has took hold of a stick.' (CLC:CLLDCh2R05S02.0849)

Here, the speaker uses S/A detransitivisation to indicate to all hearers that *lauri* is a discardable referent, that is, it does not have to be tracked. The reason why a speaker should wish to emphasise this is that the link seems arbitrary enough in order to feel that other links would have been equally possible. For instance, a lot of sticks lie around in Chintang, so picking up one is a relatively arbitrary decision.

Discardable reference is often used when a predicate and its object are felt to form a composite activity rather than two separate things. This mostly happens when a predicate-object combination acquires characteristics of its own. (135) and (136) show examples for this.

(135) *Cuwa a-thap-no?*
water 2[s]S-fetch-IND.NPST
'Do you fetch water?' (field notes 2010)

(136) *Hani a-sed-i-s-i-hẽ elo?*
2p 2S-kill-p-PRF-p-IND.PST or
'Have you killed (a pig)?' (field notes 2010)

Both (136) and (135) ignore that the object referents would be easily quantifiable and identifiable – the water was transported in a large metal vessel, and only one pig was killed which was visible at the time (135) was uttered. This is possible because both combinations are entrenched. Water supply in Chintang is incomplete, so people frequently have to go and fetch water from public wells, especially if they live in remote areas. This activity is different from fetching other things and doing other things to water because it is the most regular one and involves a specific path that normally doesn't change.

Killing pigs is a similar case. One pig is killed and its meat sold every Wednesday in Chintang. The whole process is highly standardised: the pig is always killed at the same place in the same

manner (by stabbing it and letting it bleed to death), its meat is always first sold above Devīthān and then packed in plastic bags and brought to the market at Pancakanyā, and the same people feature as helpers again and again. This makes this activity different from killing other things (for instance, chicks for rituals or chickens for private use) and doing other things to pigs (mainly feeding them, an activity that's in the responsibility of the keeper).

Since discardable reference does not indicate that a referent is not identifiable but that it probably won't be necessary to track it, it may be cancelled when the speaker changes his view on the subject. (137) shows an example for this.

(137) a. *Ram ko kina temma=kha luŋtak chiʔ-no.*
Ram walk.around[.SUBJ.NPST.3sS] SEQ nice=NMLZ$_2$ stone find-IND.NPST[.3sS]
'Ram walks around and finds a nice stone / nice stones.' (elicitation PRAR 2010)

b. *Huŋ=go luŋtak-be caĩ chikmakalok lukt-ad-a-s-e*
MED=NMLZ$_1$ stone-LOC$_1$ RETRV dirt stick-AWAY.ITR-PST-PRF-IND.PST[.3sS]
kina cuwa-ŋa wa-chid-o-ko.
SEQ water-ERG PVB-wash-3[s]O-IND.NPST[.3sA]
'That stone is dirty, so he washes it with water.' (elicitation PRAR 2010)

(137a) and (137b) were uttered in sequence, so they are part of a single paragraph. It is impossible to say whether (137a) taken alone refers to a single or to several stones – both interpretations are possible. It is only the following (137b) that forces a post-hoc singular interpretation (several stones would have to be referred to as *hun-ce* [MED-ns], and the verb would have to have 3nsO-AGR). The speaker of the paragraph presumably already had a single stone in mind when producing (137a). However, she still chose to present the referent as discardable for similar reasons as in (134) above. When the stone was unexpectedly referred to again in (137b), however, it was no problem to now use it with the transitive frame.

Discardable reference shows that the speaker has the final word on identifiability – even when a referent would be perfectly identifiable he is still free to present it as non-specific if he considers the link between the pointer and the actual referent to be particularly arbitrary.

Note, though, that importance in discourse is not a factor in identifiability, or in other words, whether a speaker thinks that a referent actually will get tracked or not is completely irrelevant to whether he considers it possible to track it. This is nicely illustrated by the sentences in (138).

(138) a. *Ʈhitta sɨŋtaŋ the=kkha yuw-a-kt-e=ta na, huŋ=go*
one tree big=NMLZ$_2$ be.there-PST-IPFV-IND.PST[.3sS]=FOC CTOP, MED=NMLZ$_1$
putt-o-ko.
pluck-3[s]O-IND.NPST[.3sA]
'There was a really big tree, and (now a man) plucks (one fruit).' (CLC:pear_1-1.011)

b. *Dhawa~dhawa pus-saŋa tis-o-ko, arko caĩ*
hurry~INTENS pluck-CVB.FGR put.in-3[s]O-IND.NPST[.3sA] other RETRV
u-ta-no, copt-and-u-ku-ce=le, ba=go
3[p]S-come-IND.NPST look.at-COMPL$_1$-3O-IND.NPST-[3sA.]3nsO=RESTR PROX=NMLZ$_1$
u-jhol-a-iʔ tis-o-ko.
3sPOR-bag-NTVZ-LOC$_2$ put.in-3[s]O-IND.NPST[.3sA]
'Hurriedly he plucks and puts it in(to a bag), others come (into view), he only looks at them, this one he puts into his bag.' (CLC:pear_1-1.012)

These are some of the first sentences of a Chintang Pear Story (cf. Chafe 1980). At the time of utterance, the speaker has not mentioned yet that there is a pear tree and a man plucking pears from it, and since the hearer doesn't know the story, she also doesn't know about these referents.[17]

What is remarkable here is that the speaker uses the transitive frame with all O referents. This is in perfect accordance with specificity and also shows once more the importance of quantifiability:

[17] This way of telling a story may seem completely ignorant of the needs of the hearer and almost brutal from the perspective of Western narrative traditions but is rather typical of Chintang and probably of other Kiranti languages, too – cf. section 2.3.1).

the first pear that can be seen in the movie is shown in isolation in a close-up, and the others referred to in (138) are arranged in neat individuated groups where the single pears are still easy to make out. The pears talked about here are thus easy to quantify and track in principle and therefore trigger the transitive frame, even though they never get mentioned again later.

This section has made the picture of the function of S/A detransitivisation more complete. We can now summarise the function with a couple of language-specific additions:

> S/A detransitivisation in Chintang marks specificity (= identifiability on part of the speaker). Transitive O are specific, detransitivised O are non-specific. The most important prerequisite for specificity is quantifiability, and S/A detransitivisation can be correctly predicted from this in most cases.
> The amount of information necessary for identifying the O referent is minimal in that it only has to be identifiable within the mental space that is in the focus of attention at speech time. If the link between the O pointer and a referent is not established yet in that space ("open reference"), the referent is viewed as non-identifiable even if it is quantifiable. The speaker may also present the referent as non-identifiable if the link is established but he views it as particularly arbitrary ("discardable reference").

2.6.3 Quantifiability in detail

2.6.3.1 The count/mass distinction and nominal number

We have already touched upon the connection between the count/mass distinction and identifiability: basically, count nouns are easy to identify and mass nouns aren't. We now need to make this statement more precise.

The familiar terms count noun and mass noun imply that the count/mass distinction is a lexical category. This is wrong, since "no noun fits absolutely into any one category", as already noted by Hewson (1972:46). "Count nouns" such as *cat* can be used like mass nouns in special contexts (*There was cat all over the street*), and "mass nouns" such as *cheese* can regularly be used like count nouns, for instance, in their type reading (*We sell various cheeses*). On the other hand, it cannot be denied that most nouns have a clear propensity for either of the two conceptualisations – cats are usually individuated and cheese is usually not.

An important property in this context is whether the combination of a noun with the numeral *one* evokes a clear mental image. For instance, the meaning of *one cat* can hardly be argued about. *One cheese* could again refer to a type of cheese, but if it was to refer to a token its shape would be more variable – although there still exists something like a prototype of a piece of cheese that has about the size of one sixth of a loaf of cheese. *One soil* brings us into a region where the combination with *one* starts to sound strange – *one soil* is certainly impossible with anything but a type reading, and even there it's unusual.

We will refer to this continuum as the individual-mass continuum. A noun that is more on the individual side will be called an INDIVIDUAL CONCEPT, and a noun that is more on the mass side will be called a MASS CONCEPT. If the combination with *one* yields a clear mental image, this will be called the BASE LEVEL of a noun.[18] Accumulations of referents belonging to an individual concept are easily perceived as constituted by their parts because it is possible to identify those parts with the base level. This is not possible with mass accumulations.

Note that the classification of concepts is language-specific. For instance, *ginger* in English does not seem to have a base level and is a quite clear mass concept. By contrast, the Chintang equivalent *phidaŋ* is more flexible: it does have a base level in the form of a single rhizome (*thitta phidaŋ* 'one ginger'), but a heap of ginger can be construed as consisting of several rhizomes (*phidaŋce* with *-ce* [ns]) or as non-quantifiable (*phidaŋ*). In yet another language ginger might even be a clear individual concept.

[18] Homogeneity, which is another popular criterion for separating individual and mass concepts (cf. e.g. Rijkhoff 2002), presupposes this notion. For instance, both *apple* and *ginger* can be used in English to construe quasi-homogeneous referents, only one needs to be pluralised (*they were selling apples*) and the other doesn't (*they were selling ginger*). It is only the base level of *apple* that is non-homogeneous in contrast to the base level of *ginger*.

The syntactic side of the individual-mass distinction is quantifiability. A count concept will usually be quantifiable, but it can be made into a non-quantifiable referent by various means to be discussed below, for instance, when it occurs in a large and hard to overlook group or when only parts of it are affected. Conversely, mass concepts tend to be non-quantifiable but can easily be made quantifiable by using containers and measures.

In Chintang, the distinction between individual and mass concepts is only weakly lexicalised, so in principle all nominal concepts can be marked as quantifiable or non-quantifiable by the same morphosyntactic means. For instance, *thitta makkai* 'one maize' is possible but does not refer to a single grain of maize but to a cob. But *makkai* can also be used to refer to heaps of maize grains where single cobs are no longer present.

Below are two more examples for this kind of flexibility.

(139) a. *I-bhuja c-o-hatt-o wa-ŋa.*
2sPOR-fried.rice eat-3[s]O-AWAY.TR-[SUBJ.3sA.]3[s]O
'The chicken will eat your fried rice.' (CLC:CLLDCh2R07S01.0269)

b. *Ʈhitta bhuja=yaŋ a-ham-c-o-ko=kha=lo naŋ?*
one fried.rice=ADD 2A-divide-d-3[s]O-IND.NPST=NMLZ$_2$=SURP but
'So you even divide a single grain of fried rice?' (CLC:CLLDCh2R07S01.1098)

(140) a. *Paĩ na wei? bhuŋ-na-da hou.*
today CTOP rain pile-LNK-come[.SUBJ.NPST.3sS] AFF
'Today there will be plenty of rain.' (CLC:CLLDCh3R02S06.457)

b. *Abo thitta wei?=yaŋ ma-ta-yokt-e.*
now one rain=ADD NEG-come-NEG-IND.PST[.3sS]
'Now it rained not even a single time.' (CLC:Chambak_int.0378)

Chintang also doesn't have a problem with pluralising what are mass concepts in English:

(141) *Ʒamma cuwa-ce khatt-u-c-a.*
all water-ns take-3O-ns-IMP[.2sA]
'Take all the water(*s).' (CLC:CLLDCh2R02S09.376)

While it is possible in English, too, to pluralise *water*, this process automatically selects quantifiable readings of the word, such as 'kind of water' (*We offer a fine selection of waters*).[19] Situations as in (141) require the use of a container word (e.g. *bottle*). Both the plural marker and determiners are associated with this word (*the bottles of water*, **bottles of the water*, **the bottle of waters*; *one bottle of water*, **bottle of one water*).

The same flexibility is seen in the use of S/A detransitivisation. For instance, mass concepts can be construed as quantifiable and accordingly be used with the transitive frame when the relevant referent is small and easy to overlook (142) or when it has physical boundaries (143):

(142) *Ghãsa na lab-o-ŋs-e.*
grass CTOP grab-3[s]O-PRF-IND.PST[.3sA]
'He has grabbed (a bunch of) grass.' (CLC:CLLDCh3R09S06.503a)

(143) *Rumpatti, hana ghãsa kekt-o-kh-o!*
Rumpatti 2s grass hold-3[s]O-CON-[IMP.2sA.]3[s]O
'Rumpatti, you hold the (bundle of) grass!' (CLC:CLLDCh4R13S04.371)

Similarly, concepts that tend toward the individual side can easily be construed as non-quantifiable. One possibility to do so is to disintegrate the base level marked by *one*. For instance, *kocuwa* 'dog' by default denotes a single dog, which would be a quantifiable referent. However, referring to a non-quantifiable subamount of dog becomes possible when the dog is acted upon in a way that ignores its unity. For example, when a dog dies and something eats it, its parts no longer serve different functions that together create *one dog*, but each part becomes just another part of the menu:

[19] In the case of *water* there are of course also lexicalised plural uses as in *These waters are dominated by the Americans.*

(144) *Ba kocuwa sa-lo ca-no=kha?*
PROX dog who-NOM eat-IND.NPST[.3sS]=NMLZ$_2$
'Who is eating (from) this dog?' (CLC:CLDLCH3R01S02.279)

Another frequent way to create a non-quantifiable referent from an individual concept is to multiply the base level as in (145):

(145) *A-nisa-ce sontoloŋ khali=ta u-toc-ce-ke.*
1sPOR-younger.sibling-ns tangerine always=FOC 3S-prong-d-IND.NPST
'My younger brothers prong at tangerines all the time.' (elicitation PRAR 2010)

Sontoloŋ by default refers to a single tangerine, but in this sentence the number of tangerines is indeterminate and the corresponding referent is not quantifiable.

It is interesting that English has to use the plural on *tangerines* whereas Chintang *sontoloŋ* is singular. On the one hand, this is another hint to the lack of grammaticalisation of the individual/mass distinction in Chintang: English uses the singular on non-quantifiable mass concepts (*He ate some porridge*) but the plural on non-quantifiable individual concepts (*He ate some tangerines*). On the other hand, it also shows a difference in the semantics of the English and the Chintang singular. In Chintang, the singular is the default number and the non-singular is only used when a speaker is sure that there is more than one individual referent, whereas the English plural already responds to the possibility of there being more than one such referent.

Another interesting property of the Chintang non-singular is that it implies quantifiability. It is therefore normally impossible to detransitivise an object marked by the non-singular:

(146) *Asinda akka paryatak maʔmi(*-ce) khag-e-h-ẽ.*
yesterday 1s tourist person-ns see-PST-1sS-IND.PST
'Yesterday I saw (some) tourists.' (elicitation PRAR 2010)

Exceptions can be found, though. One possibility are nested structures as in (147). *Khi* denotes a single yam root, *khi-ce* a small, quantifiable group of yam roots. The construction in (147) creates a group of such groups, which itself is non-quantifiable:

(147) *Kholakhi-ce tus-i-ki-ŋa.*
wild.yam-ns dig.out-1pS-IND.NPST-e
'We dig out wild yam roots.' (CLC:phidang_talk.045 + elicitation RBK 2012)

Another possibility are the circumstances discussed in section 2.6.2 where a referent may be quantifiable yet not identifiable. (148) shows an example where S/A detransitivisation is triggered by open reference.

(148) *Yo-ʔni bhai-ʔni dhami-ce kond-i-e-hẽ,*
DEM.ACROSS-DIR PROX-DIR shamen-ns search-1pS-e-IND.PST
kond-i-yakt-i-e-hẽ.
search-1pS-IPFV-1pS-e-IND.PST
'We searched for shamans, we were searching them (for some time).' (CLC:appa_katha_talk.045 + elicitation RBK 2012)

The special properties of the Chintang number system are bound to the object relation. With non-objects the non-singular behaves as the English plural, that is, non-quantifiable referents are marked:

(149) *Bhiya-ce=yaŋ u-ta-no-ta, maʔmi-ce=yaŋ u-si-no-ta…*
marriage-ns=ADD 3[p]S-come-IND.NPST-CONT person-ns=ADD 3[p]S-die-IND.PST-CONT
'Marriages are taking place, people are dying…' (CLC:Gen_talk.017-018)

This asymmetry can lead to one and the same referent triggering both S/A detransitivisation and ns-AGR when it is shared between two clauses. In (150), *maʔmi* represents a non-quantifiable, divisible referent ('people') which is shared between a relative clause, where it occupies A and is

linked to 3[p]A-AGR, and a main clause, where it occupies P and is not indexed. Also, because this relative clause is externally headed, *maʔmi* is assigned case and number by the main clause predicate and is therefore in the nominative singular.

(150) *Akka sʌhʌyog u-pi-ŋa-ʔa-ni-ŋ=go maʔmi koĩ-ya-ʔã.*
1s help 3A-give-1sO-IND.NPST-p-1sO=NMLZ$_1$ person search-1sS-IND.NPST
'I'm looking for people who can help me.' (elicitation PRAR 2010)

The precedence of quantifiability over divisibility in the number marking of objects and the resulting rarity of *-ce* [ns] on detransitivised objects seem to be the only aspect of nominal marking where detransitivised objects are restricted compared to transitive objects. As we have seen in section 2.4.3.3, they are full independent NPs in every other respect.

To summarise, there is no formal evidence for the existence of a distinction between individual and mass concepts in Chintang. The distinction is a useful construct to understand how reference is established but not a language-specific category. Thus, *sontoloŋ* could be translated as 'tangerine' just as well as 'tangerines' or 'piece of tangerine', *siŋ* means 'stick' as well as 'wood', and *maʔmi* means both 'person' and 'people'.

The lack of a grammaticalised individual/mass distinction does not mean, though, that Chintang is completely insusceptible to the difference between *cat* and *cheese*. In fact, most nouns have a clear preference for being construed as quantifiable or not, and this has consequences for how often they are used together with the transitive or the detransitivised frame. This is shown in Figure 2.4.[20]

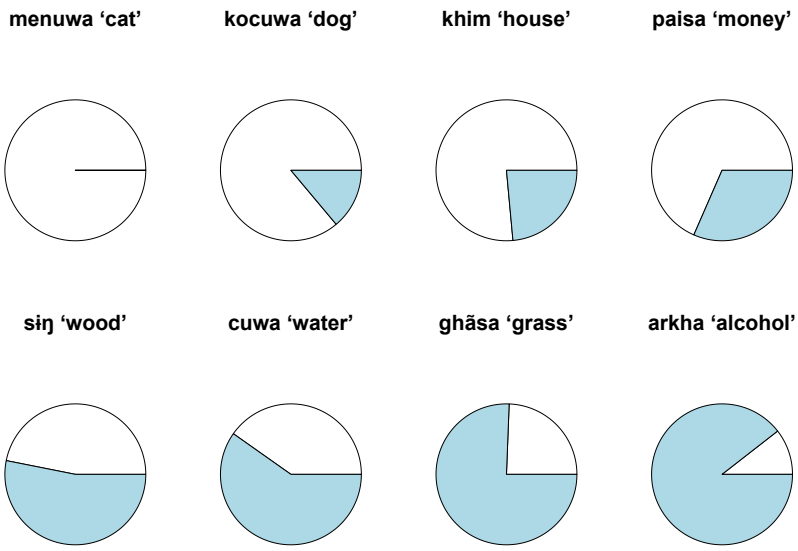

Figure 2.4: Framing for a couple of nouns (blue = proportion of S/A detransitivised clauses)

The factors determining whether a concept tends to be quantifiable are its intrinsic tendency to occur in certain quantities but also how it is perceived and handled by humans. For instance, the most frequent manifestation of 'cat' is a single cat. Neither are cats normally divided into equal parts (they are not part of the diet in Chintang), nor do they occur in groups (in contrast to dogs, which are notably less often quantifiable). Money usually comes in groups of similar individual referents (coins or notes), so its natural occurrence would point towards non-quantifiability. How-

[20] The counting procedure was as follows. I took a couple of nouns, extracted all clauses where they occupied the position of O and counted the number of detransitivised and transitive frames. All ambiguous cases were ignored, as they are – as far as I can see – independent of the semantics of S/A detransitivisation. With valency-manipulating vector verbs the O of the final verb was considered. Non-singular nouns in the nominative were counted like the same noun in the singular nominative. In order to get a more complete picture, not only the annotated part of the CLC but the complete corpus was considered.

ever, since money is being counted all the time and definite amounts of money play an important role in most societies as property and price, it is plausible that it should still be more frequently quantifiable.

Another interesting difference is that between water and alcohol. Both are maximally homogeneous and need external boundaries in order to become quantifiable. Both are stored in containers so that it is in principle possible to act on them as a whole. However, they are handled in different ways. The two main activities in Chintang with alcohol as their P are making (*heŋd-*) and consuming it (*thuŋ-*). When making alcohol, the quantity is not fixed beforehand – a woman stays at the cooking place for one or several days, and the output quantity varies according to her enthusiasm and her skills. Similarly, one drinks alcohol out of cups, but the quantity is rarely confined to one cup upon one occasion, so the affected referent is that in the next bigger container. The quantity that is drunk in the end varies depending on the generosity of the host and the endurance of the drinker and is rarely fixed beforehand. There are also no bars in Chintang where one could order one cup.

2.6.3.2 Construing quantifiability

This section explores under which circumstances referents are construed as quantifiable. It is again convenient to consider mass and individual concepts separately.

We have already seen above that mass concepts tend towards being construed as non-quantifiable. Here is one more example for this:

(151) *Akka yo?-ni* *sambok sop-ma-khaŋ-ŋa.*
1s DEM.ACROSS-DIR millet thresh-1sS-CON-[SUBJ.NPST.]1sS
'I'll try to thresh millet over there.' (CLC:CLLDCh3R02S02.009)

Sambok 'millet' is most often processed in a way that requires construing it as non-quantifiable. Harvesting millet is a time-consuming process because every ear has to be plucked separately. Therefore, usually only parts of a field are harvested at a time. After plucking, the millet is dried in the sun and then cooked (for making liquor) or ground (for making bread or millet porridge). All these processes only allow working on a small amount of millet at a time because plastic blankets of only about 2 to 3 m² are used for drying, medium-size pots for cooking, and grindstones driven by hand for grinding. So all conventional activities with millet as their object require acting on a subamount of variable size of the quantity that is available overall.

A slightly different case is presented by *ciya* 'tea':

(152) *Abo ciya thu-i* *o.*
now tea drink-[SUBJ.NPST.]1p[i]S okay
'Now let's have tea, okay?' (field notes 2010)

The word *ciya* can refer both to the plant and the beverage, but the plant and its leaves rarely occur in object position because the plant doesn't grow in Chintang and making the beverage is not described as 'cooking leaves' but as 'making tea'. The beverage tea, however, is frequently made and drunk. Because it loses its flavour quickly, it is not stored but always made freshly. Further, tea is always made for a certain quantity of persons (whether for customers in teashops, guests, or members of the own family) and is drunk from small cups. All this would point to a preference for constructing tea as quantifiable. However, the quantity of tea drunk is rarely restricted to one cup. To be sure, one cup is the conventional amount, but half a cup or two cups would be equally possible. For the same reason, making tea for five persons does not mean putting on five cups of water but a larger, approximate amount.

The easiest way to make a mass concept quantifiable is to refer to a complete accumulation such as a heap of firewood as in (153). Such accumulations have an outer boundary that constitutes an *ad-hoc* way of quantifying them via counting (one heap, two heaps…).

(153) *Kanchi, yo siŋ thapt-o-kh-o!*
youngest.daughter DEM.ACROSS wood bring.over-3[s]O-CON-[IMP.2sA.]3[s]O
'Kanchi, bring over that firewood!' (CLC:CLLDCh2R02S09.436)

Often it is more convenient to handle small amounts of masses at a time. Small accumulations in one place are created by adding physical boundaries:

(154) *Cuwa ek gilas thuŋ-c-o.*
water one glass drink-[1]d[iA]-[SUBJ.NPST.]3[s]O
'Let's have one glass of water.' (CLC:CLLDCh1R05S01.800)

Such physical boundaries do not have to be actually present; they can also be projected from another place or from the mind. The examples in (155) involve an amount of air that is bounded by an inflatable ball. Whereas in (155a) the air is in the ball and thus actually quantifiable, it is still outside of it in (155b) at the time it is affected by *tis-*.

(155) a. *Tott-e kina u-hawa loĩs-and-o-ŋs-e.*
poke-IND.PST[.3sS/A] SEQ 3sPOR-air let.out-COMPL$_1$-3[s]O-PRF-IND.PST[.3sA]
'He has poked (the ball) and let its air out.' (CLC:CLLDCh1R13S02.1379)
b. *Akka hawa tis-and-u-ŋ.*
1s air put.in-COMPL$_1$-3[s]O-[SUBJ.NPST.]1sA
'I'll let in (the) air.' (CLC:CLLDCh1R13S02.1378)

Another way of quantifying masses is to use units of measurements. Masses can be directly measured with dedicated units (156a) or indirectly with associated units (156b).

(156) *Sumci mana thukt-u-ku-m-ma kok.*
three mana cook-3[s]O-IND.NPST-1pA-e rice
'We cook three mana of rice.' (CLC:CLLDCh4R03S03.0224)

(157) *Ã, ghãsa akka paitis sai-ko khed-u-h-ẽ.*
yes grass 1s thirty.five hundred-GEN buy-3[s]O-1sA-IND.PST
'Yeah, I bought grass for thirty-five hundred.' (CLC:CLDLCh3R01S04.002)

Comparisons presuppose the possibility of measuring and accordingly make referents quantifiable, too:

(158) *Hana baddhe a-c-o-kko elo i-phuwa-ŋa?*
2s much 2[s]A-eat-3[s]O-IND.NPST or 2sPOR-elder.brother-ERG
'Do you eat more or your brother?' (CLC:CLLDCh1R02S01 0849)

Finally, boundaries may also be set in the domain of time:

(159) *Akka athomba redio-be sat baje-ko khabar-a khems-u-h-ẽ.*
1s before radio-LOC$_1$ seven o'clock-GEN news-NTVZ hear-3[s]O-1sA-IND.PST
'I just heard the 7 o'clock news on the radio.' (elicitation PRAR 2010)

A special case is presented by concepts which are (in the philosophical sense) rather accidents than substances. In English, such concepts are only quantifiable in connection with a referential carrier such as *thing*. Chintang is once more more flexible here in that both quantifiable and non-quantifiable referents can be construed from such concepts without the help of ancillary devices. (160) shows a pair of examples for this. *Halacoppa* 'red' is quantifiable in (160b), which in this case entails that there is exactly one red thing.

(160) a. *Akka halacoppa khag-e-h-ẽ.*
1s red see-PST-1sS-IND.PST
'I saw red.'

b. *Akka halacoppa khag-u-h-ẽ.*
1s red see-3[s]O-1sA-IND.PST
'I saw something red.' (elicitation RBK 2010)

We will now turn to quantifiability in individual concepts. We should once more remind the reader that the individual/mass distinction is not lexicalised in Chintang and that the syntactic concept of quantifiability is much more important for explaining S/A detransitivisation. The distinction does, however, make it easier to discuss quantifiability in a systematic way.

(161) shows the easiest case of a quantifiable individual concept, i.e. a single referent.

(161) *Ram-e patt-o-kh-o!*
Ram-NAME.NTVZ call-3[s]O-CON-[IMP.2sA.]3[s]O
'Call Ram!' (CLC:CLLDCh1R10S05.105)

(162) shows an apparent exception:

(162) *Ba sencak ci-a-ŋs-e.*
PROX mouse eat-PST-PRF-IND.PST[.3sS]
'A mouse has eaten from this (tangerine).' (field notes 2011)

Here we have a single individual concept ('tangerine') used with S/A detransitivisation even though it is quantifiable. This example can easily be explained, however, if one assumes that there are two referents involved here, viz. the tangerine and the small part nibbled away by a mouse. The referent that has been affected is the latter, and this referent is non-quantifiable because of course the mouse did not take out individual segments. One might still say that the eaten part is easy to distinguish from the rest and should therefore be quantifiable. However, this does not take into account that the action of the mouse did not affect the now visible hole (which itself is quantifiable) but a part of the tangerine that was still there at event time.

Predicate semantics play an important role in partial affectedness. Eating is a good example for an activity that affects its O gradually. Other activities have a more punctual effect and therefore do not allow for partial affectedness. For instance in (163), the pencil is clearly only partially affected, but not in a way that would justify using the verb *kɨpt-* 'cut, prune, shorten'. Put differently, the affected subamount may look different from the rest of the referent, but it is not *shortened* – cf. English *There was one rope and he shortened part of it*, which likewise sounds odd. What is shortened (or rather about to be shortened) is the whole pencil, so the transitive frame must be used:

(163) *Ram-e-ŋa chapmago kɨpt-o-ko tʌrʌ u-dhar-a*
Rame-NAME.NTVZ-ERG pencil cut-3[s]O-IND.NPST[.3sA] but 3sPOR-blade-NTVZ
manchi? kina na latt-and-o-ko.
be.not.there SEQ CTOP give.up-COMPL$_1$-3[s]O-IND.NPST[.3sA]
'Ram cuts off a piece from a pencil, but his knife is not sharp enough so he gives it up.'
(elicitation PRAR 2010)

The contrast between quantifiable referents and non-quantifiable subdivisions of theirs can also explain some discrepancies betwen English and Chintang, i.e. cases where English uses an article but Chintang uses S/A detransitivisation. The example in (164) is repeated from above:

(164) *Ca-saŋa=ta numd-a-kt-a-lok ek dini a-phe-ce*
eat-CVB.FGR=FOC do-PST-IPFV-[SUBJ.]PST[.3sS]-CVB.BGR one day 1sPOR-elder.brother-ns
bhai?-ni u-thab-a-ci-e.
PROX-DIR$_1$ 3[p]S-come.over-PST-COMPL$_2$-IND.PST
'While he was still taking the medicine, one day my brother's family came over for a visit.'
(CLC:appa_katha_talk.021-022)

Here, the medicine the speaker's father got from the hospital is quantifiable in the form of one or several containers – this and the fact that the medicine has been mentioned before yield the definite article in the English translation. However, what matters in Chintang is that the affected referent

is not the medicine as a whole but a non-quantifiable subamount of it.

Much variation is found in the treatment of larger groups of what is perceived as individual concepts. It has already been mentioned in section 2.6.1 that groups with an overt definite quantifier are always quantifiable. Here is another example for this:

(165) *Pãc-eda phultuŋ samet-a wad-u-ŋs-u-c-e aŋ.*
five-CLF underpants altogether-NTVZ put.on-3O-PRF-3O-ns-IND.PST[.3sA] QTAG
'Altogether he's put on five pairs of underpants, huh?' (CLC:CLLDCh3R11S07.267)

This is, of course, not true when quantification is only approximate:

(166) *Hardi athawa khair-a-ko sumce car-eda tukra=yaŋ tis-i-ki.*
turmeric or catechu-NTVZ-GEN three four-CLF piece=ADD put.in-1p[i]S-IND.NPST
'We also put in three or four pieces of turmeric or catechu.' (CLC:arkha_hengma.34)

Quantification can also be overridden in the case of nested referents. For instance in (167), the A does two things at a time. These two things together form a complex activity that is the relevant object referent for the predicate *numd-* 'do'. Since the beginning and the end of the activity are not fixed, it cannot be quantified in the relevant dimension of time:

(167) *Khoŋ-no?=yaŋ, ne-no?=yaŋ ni, maĩla na hicce=ta*
play-IND.NPST[.3sS]=ADD study-IND.NPST[.3sS]=ADD ASS second.son CTOP two=FOC
num-no.
do-IND.NPST[.3sS]
'He plays and he studies – Maila does two things (at the same time).'
(CLC:CLLDCh4R13S01.110)

The main factor determining the quantifiability of complex ensembles that are neither overtly quantified nor exhaustive is the ease of overlooking the ensemble. The smaller it is and the closer together its parts are, the higher is the probability that it will be construed as quantifiable. Consider the pair of examples in (168):

(168) a. *Sapphi sɨk u-tok-no ni.*
much louse 3[p]S-have-IND.NPST ASS
'They really have a lot of lice.'
b. *Sapphi sɨk u-tog-o-ko-ce.*
much louse 3[p]A-have-3O-IND.NPST-3nsO
'They really have a lot of lice.' (elicitation PRAR 2010)

The first sentence clearly is the default – when a person has lice there will usually be lots of them. The informant that gave me these sentences suggested that the second variant might be used after looking at a person's head. Thus, while in the first example the large number of lice and their lack of coherence motivates the non-quantifiable construction, (168b) uses the head as an anchor to (mentally) keep the lice together as a single group.

2.6.3.3 Quantifiability with indefinite quantifiers

The criterion of overlookability also plays a role in the way S/A detransitivisation interacts with indefinite quantifiers. In Chintang, such quantifiers generally do not distinguish between individual and mass concepts, which conforms with our claim that this distinction is not lexicalised in Chintang. For instance, *jamma* means both 'every' and 'all', *mi?muŋ* means 'few' and 'little', and *baddhe* means 'many' and 'much'.

Quantifiers such as *jamma* mark what I will call exhaustive reference. Exhaustive reference is always quantifiable because it does not tolerate deviations from a certain quantity. Thus, *all the apples* may refer to various numbers of apples when used in different contexts. However, in a specific context it can only refer to whichever number represents *all* apples, and if less apples than that are affected the use of *all* will be ungrammatical (*He ate all the apples, *but I kept two for*

you). (169a) illustrates exhaustive reference to a mass concept, (169b) exhaustive reference to an individual concept:

(169) a. *Abo=le jamma siŋ u-hutt-and-u-c-e.*
now=RESTR all firewood 3[p]A-burn-COMPL$_1$-3O-3nsO-IND.PST
'Only now did they burn all the firewood.' (CLC:martyr_story.325)
b. *Hun-ce-ko sipahi-ce sapai hun-ce-ŋa u-paŋs-u-ku-ce.*
MED-ns-GEN soldier-ns all MED-ns-ERG 3[p]A-send-3O-IND.NPST-3nsO
'They sent all their soldiers.' (CLC:rana_pilgrim.059)

This behaviour is also expected because exhaustive reference may be said to be an emphasised form of inclusive reference, which was described as an inherent characteristic of unique identifiability in section 2.5.5.

Apparent exceptions are once more possible when subamounts are involved:

(170) *Ek thaũ=ta jamma hon-na-dheĩ kina ca-no.*
one place=FOC all mix-LNK-COMPL$_1$[.SUBJ.NPST.3sS] SEQ eat-IND.NPST[.3sS]
'He mixes everything in one place and eats it.' (CLC:CLLDCh4R03S03.0203)

A Nepali meal generally consists of several independent dishes, typically rice, lentil soup, vegetable curry, and a small amount of pickles. These dishes are usually served separately on one plate. What the child occupying the role of A did here is to take something from every dish and then mix and eat it in one place. Thus, *jamma* relates to the total amount of food, whereas the actions coded by the verbs *hol-* and *ca-* only affect a subamount. The relevant referent is therefore not exhaustive.

A variant of exhaustive reference is the case where some referents are taken to represent a whole group. This can be due to various reasons. For instance, in (171) it is simply most animals that are affected by the tiger's tyranny, so it is a reasonable generalisation to say 'all'. By contrast, in (172) it is quite clear that the monkeys do not steal all or even most of the maize, but it is still enough to view the harvest as a whole as affected.

(171) *ɟaŋgal-a-be=ko-ce jamma=pho dukha pid-u-wakt-u-c-e.*
jungle-NTVZ-LOC$_1$=NMLZ$_1$-ns all=REP trouble give-3O-IPFV-3O-ns-IND.PST[.3sA]
'He (the tiger) gave trouble to all the jungle dwellers.' (CLC:story_tiger.024)

(172) *Ã, makkai u-c-o-kko, phidaŋ cahĩ u-c-o-kko-niŋ.*
yes maize 3[p]A-eat-3[s]O-IND.NPST ginger RETRV 3[p]A-eat-3[s]O-IND.NPST-NEG
'Yes, they (the monkeys) eat the maize, but they don't eat the ginger.'
(CLC:phidang_talk.483)

The range of tolerance becomes the greater the greater the ensemble is. In (173) no realistic speaker could wish that god conserve every single thing in the whole universe, so exhaustive reference is once more approximative:

(173) *ɟiu dan-a bar dan-a sab-ai kura a-yuŋs-u-m.*
life gift-NTVZ blessing gift-NTVZ all-FOC thing 2A-keep-3[s]O-[SUBJ.NPST.]2pA
'May you conserve the gift of life, the gift of the blessing, all things.'
(CLC:Budhohang_d.78)

Other indefinite quantifiers function quite differently from universal quantifiers. They can be combined both with the transitive and the detransitivised frame but have strong statistical associations. For small quantities, the transitive frame is the default:

(174) *Miʔyuŋ sag-a lett-u-ŋ=kha akka.*
a.few green.vegetables-NTVZ plant-3[s]O-1sA=NMLZ$_2$ 1s
'I've planted a few green vegetables here.' (CLC:CLLDCh1R07S02.516.)

This is expected because markers of small quantities presuppose an approximate benchmark in the mind of the speaker beyond which a quantity can no longer be said to be small. Since the quantity

must be greater than zero but stay below the benchmark it is easy to overlook, and as has been mentioned at the end of the last section, that is a factor favouring a quantifiable construal in itself.

Large quantities also involve a benchmark, but the category applies to everything *above* it. As a consequence, markers of large quantities have much more semantic leeway than those for small quantities: there are more cases to which they apply, and accordingly they are more appropriate in cases where one doesn't know the exact number of something and where the range of possible numbers is broad. This does, however, not mean that they are *only* appropriate in such cases, so what is expected for them is a lack of a clear preference for one of the two frames. This is what is found:

(175) *Ba=go jagga=yaŋ baddhe tok-no.*
PROX=NMLZ$_1$ land=ADD much have-IND.NPST[.3sS]
'He also owns much land.' (CLC:LH_Lal.0161)

(176) *Ghãsa=yaŋ Som-e-ŋa baddhe=ta hekt-o-ŋs-e.*
grass=ADD Som-NAME.NTVZ-ERG much=FOC cut-3[s]O-PRF-IND.PST[.3sA]
'Som has also cut a whole lot of grass.' (CLC:CLLDCh1R06S03.0309)

In (176a) the precise quantity of owned land is not important – the speaker just wants to say that the concerned person is rich. But it is important in (176b), where the context tells us that Som was expected to cut a certain amount of grass and the amount he managed to cut is compared to that benchmark (in fact, Som cut too little, so the utterance is ironic).

The default for small quantities can be overridden, for instance, when open reference is involved. Consider (177):

(177) *Ṭhi akka mi?muŋ thuŋ-ŋa-khaŋ-ŋa.*
beer 1s a.little drink-1sS-CON-[SUBJ.NPST.]1sS
'Let me try and have some beer.' (CLC:CLLDCh1R09S03.128)

This sentence is a polite request – the speaker might drink more or less, depending on how much is appropriate. On the other hand, the transitive variant *thu-u-ŋ-kha-ŋ* [drink-3[s]O-1sA-CON-[SUBJ.]1sA] sounds more demanding because the speaker has already fixed the amount in his head at the time of speaking.

2.6.3.4 Actual and virtual quantifiability

In all examples discussed so far, the relevant objects were quantifiable in reality. However, this is not always the case. Consider the following example:

(178) *Ã?, cuwa ni eŋs-o-ko.*
yeah water ASS divert-3[s]O-IND.NPST[.3sA]
'Yeah, he diverts the water.' (CLC:CLLDCh4R05S04 2069)

The background of this utterance is that water supply is incomplete in the village Chintang. There are a few major pipes from which people commonly divert water for their own use. The amount of diverted water is obviously not limited, as new water flows in all the time. The reason why *cuwa* is still construed as quantifiable here seems to be that the amount is measurable at least in theory. One could, for instance, measure the water at the end of each day, adding up the results, and would (as long as one can rely on one's counting abilities) never arrive at an indefinite quantity. This type of quantifiability may be called virtual.

Virtual quantifiability plays a great role for the explanation of S/A detransitivisation. Especially large amounts are rarely determinable in practice. For instance, separating maize grains from their cobs is a work that takes some time, since usually a whole basket of cobs is done at once. Even though nobody would count the cobs (let alone the grains) this is an easy enough job in theory, so the transitive frame is commonly used in sentences such as (179):

(179) *Makkai-ce na tak-ma a-hid-u-m-cu-mh-e?*
maize-ns CTOP break-INF 2A-finish-3O-2pA-3nsO-2pA-IND.PST
'Have you finished breaking off the maize?' (CLC:CLLDCh2R05S01.068)

Virtual boundaries also make examples such as the following possible:

(180) *Khakhutt-ad-a-s-e, huŋ=go-i? chʌ sat-jʌna*
become.dark-AWAY.ITR-PST-PRF-IND.PST[.3sS] MED=NMLZ$_1$-LOC$_2$ six seven-HUM.CLF
ma?mi-ce likhi khag-u-cu-h-ẽ, chutt-e num-ma=ta
person-ns EQU see-3O-ns-1sA-IND.PST separate-V.NTVZ do-INF=FOC
ma-hi-yakt-u-ŋs-u-ŋ-cu-h-ẽ.
NEG-be.able-PST.NEG-3O-PRF-3O-1sA-ns-1sA-IND.PST
'It had become dark and I saw about six or seven people there, but I couldn't keep them apart.' (elicitation PRAR 2010)

Here, even though the speaker herself admits that she doesn't know how many persons there are exactly, she uses the transitive frame because the group of people looks small enough for its quantity to be determinable in principle, and because it would be an easy thing to go and count them through.

A frequent form whose behaviour is best explained via virtual quantifiability is *asuk* 'how much, how many'. Even though a speaker who uses this form obviously does not know the quantity of something, *asuk* in O is almost always accompanied by the transitive frame. This is expected if we assume that *asuk* can only be used when the speaker presupposes that the quantity he is asking for can be determined. (181) shows an example.

(181) *Ba-sa-ŋa asuk khur-o-ko?*
PROX-OBL-ERG how.much carry-3[s]O-IND.NPST[.3sA]
'How much can she carry?' (CLC:CLLDCh1R05S02.177)

2.6.4 Interaction with other factors

2.6.4.1 Interaction with part of speech

A couple of parts of speech interact with S/A detransitivisation on the base of their specific reference. These are pronouns, demonstratives, and proper names (as a subclass of nouns).

One of the few unbreakable rules of S/A detransitivisation in Chintang says that it is ungrammatical with pronouns:

(182) **Akka u-cop-no.*
1s 3[p]S-look.at-IND.NPST
'They look at me.' (elicitation PRAR 2010)

Remember that the morphosyntactic class of pronouns in Chintang only comprises SAP pronouns (cf. section 2.2.1) – words referring to third persons via deixis fall into the class of demonstratives. Though the rule excluding pronouns from S/A detransitivisation is certainly well motivated, it does by no means fully follow from what we have so far said about specificity and quantifiability. Non-singular speech act participants may sometimes be hard to be quantified and may accordingly be on the edge of specificity, especially the first person inclusive plural, which in Chintang can function as a generic person similar to *you* in English. Compare the two sentences below, which virtually have the same meaning but where only the first can be detransitivised:

(183) a. *I-phak ma?mi nek-no?*
2sPOR-pig people bite-IND.NPST[.3sS]
'Does your pig bite people?'
b. *I-phak kha-nek-no?*
2sPOR-pig 1nsO-bite-IND.NPST
'Does your pig bite (us)?' (elicitation PRAR 2012)

Demonstratives are different from pronouns in allowing a minimal level of flexibility. In most cases, they are likewise incompatible with S/A detransitivisation:

(184) Akka (*ba=go) maʔmi koĩ-ya-ʔã.
1s PROX=NMLZ$_1$ person search-1sS-IND.NPST
'I am looking for (*this) people.' (elicitation PRAR 2010)

The reason for this is obvious: demonstrative NPs are inherently specific so that open and discardable reference are excluded. What's more, they restrict the scope of referential expressions to an area in space (or in one of its metaphorical extensions such as time and discourse), which makes it easy to overlook the referent and construe it as quantifiable.

Exceptions are possible with nested structures where the affected referent is a non-quantifiable subamount of the referent marked by the demonstrative:

(185) *To cuwa a-thuŋ-no=kha?*
DEM.UP water 2[s]S-drink-IND.NPST=NMLZ$_2$
'Do you drink (from) the water up there?' (CLC:CLLDCh1R03S06.221)

(186) *Akkai, ba=go ci-a=mo nusayaŋ...*
oh PROX=NMLZ$_1$ eat-IMP[.2sS]=CIT CONCS
'My, even if somebody told me to eat (from) this... (I couldn't.)' (CLC:CLDLCH3R01S02.281)

Note that sortal demonstratives are different from referential demonstratives. They do not point to referents directly but via their category and they also do not restrict the area of reference, so it's not surprising that they are well attested with S/A detransitivisation:

(187) *Ba-khiya waphuruk a-kha-i-s-i-hẽ?*
PROX-SORT cucumber 2S-see-p-PRF-p-IND.PST
'Have you seen such cucumbers before?' (CLC:CLLDCh3R04S01.137)

Proper names are another class where S/A detransitivisation is a 100% impossible:

(188) **Hari Lachman cop-no-ta.*
Hari Lachman look.at-IND.NPST[.3sS]-CONT
'Hari is looking at Lachman.' (elicitation SAR 2011)

This behaviour can be fully predicted from the general rules for S/A detransitivisation. Proper names are not only inherently specific, they also necessarily denote a singular and therefore quantifiable referent. Anything that is close enough to the world of human beings in order to be given a name is normally not acted upon as a mass of equal parts, so disintegration as a way of enabling non-quantifiable construals is excluded. Nor is adding an indefinite amount of other referents of the same name possible, since a name is not a category (i.e. several people called *Ram* cannot be referred to as *the Rams*, except in marginal contexts where *the Rams* will still refer to a small, quantifiable set).

2.6.4.2 Interaction with possession

Possessed O are similar to demonstrative O in that they are in most cases ungrammatical with S/A detransitivisation. The following example is from Bickel (2008b:4) (glosses adapted):

(189) a. *(A-)kam (a-)khim-be paŋs-u-h-ẽ.*
1sPOR-friend 1sPOR-house-LOC$_1$ send-3[s]O-1sA-IND.PST
'I sent (a/my) friend to (a/the/my) house.'
b. *(*A-)kam (*a-)khim-be paŋs-e-h-ẽ.*
1sPOR-friend 1sPOR-house-LOC$_1$ send-PST-1sS-IND.PST
'I sent friends home.' (in general)

As with demonstrative O, however, it again turns out that S/A detransitivisation is possible with possessed O and simply extremely rare due to functional reasons. Possession correlates with specificity because ownerhood is a good criterion for distinguishing one referent from others. It also correlates with quantifiability because one normally doesn't possess lots of things of one kind in different places and because possessed things are normally individual rather than mass concepts.

In (189), all readings that are compatible both with a possessed O and S/A detransitivisation are semantically strange. Since *kam* 'friend' is clearly individual, there are the following options:

- I sent home one of my friends with arbitrary reference. Human beings *per se* do not go together well with arbitrary reference, and that is all the more true of friends, which should be even easier to distinguish from each other than other people. It is hard to imagine a situation where somebody should send one of his friends home without having a good reason for choosing precisely that friend (such as that friend being tired or being the fastest runner).
- I sent home a non-quantifiable subamount of one of my friends. This is possible but sounds rather macabre and is therefore likely to be rejected.
- I sent home a non-quantifiable number of friends of mine. Although this sounds least weird out of all options, it is still difficult to find a matching situation. In a concrete situation it is not clear why one should send home two or three friends without caring about who exactly is in the set (and thereby determining quantity). In principle it would be possible to assume a general reading as suggested by Bickel (2008b) ('I used to send home one or the other friend'), but this would be expressed using *-yakt* [IPFV], not with the simple past tense.

(190) shows an example of a possessed O used together with S/A detransitivisation. The O referent is a non-quantifiable subamount of all the possessor has.

(190) *U-phuwa-go* *wa?-no=kha.*
3sPOR-elder.brother-GEN wear-IND.NPST[.3sS]=NMLZ$_2$
'He wears his brother's (clothes).' (CLC:CLLDCh1R08S05.0460)

A similar example is found in (191). A sells his aunt's tomatoes, but since it's not clear whether he will succeed in selling all of them and various people buy tomatoes from him over a longer period of time, S/A detransitivisation is possible:

(191) *Paĩ* *makku-ko* *golbheda in-no-ta.*
today younger.aunt-GEN tomato sell-IND.NPST[.3sS]-CONT
'Today he's selling his aunt's tomatoes.' (elicitation SAR 2010)

2.6.4.3 Interaction with aspect

It has long been known that there is a connection between aspect and the referential semantics of objects such that specific objects go together with perfective or telic events and non-specific objects with imperfective or atelic events (e.g. Verkuyl 1972, 1993; Dowty 1979). Some good examples for this connection are cited by Swart (2006:163): French *tricoter un chandail norvégien* 'knit a Norwegian sweater' is telic whereas *tricoter des chandails norvégiens* 'knit Norwegian sweaters' is not, and *réparer une bicyclette* 'repair a bicycle' is episodic whereas *réparer des bicyclettes* 'repair bicycles' is habitual.

This connection has led some scholars to propose that the two domains are not only associated with each other but are in fact functionally analogous. For instance, Rijkhoff (2002:59) speaks of "nominal aspect", and Leiss (2000:239) even goes so far to call articles and aspect "grammatische Synonyme" (grammatical synonyms). Kiparsky (1998) claims that VPs can derive what he calls their "unboundedness" either from an unbounded head (i.e. an imperfective state of affairs) or from an unbounded argument (i.e. a non-quantifiable object NP).

Whether this is a useful generalisation or one step to far is a difficult question in general but easy to answer when only looking at Chintang. In Chintang, there are no grammaticalised links between identifiability and aspect.

In order to show this, we first have to get an overview of the aspectual system of Chintang. This system is rich but asymmetrical in that the existing oppositions are not active in the whole language but tied up with tense. Whereas aspect is obligatorily marked in the subsystem found in the past tense, the nonpast subsystem is rather centered around one default form not expressing any aspect at all. Within and across both subsystems there are small niches occupied by highly specialised aspectual markers. We will ignore these and focus on the more frequent and abstract markers for our discussion:

- *-ŋs* [PRF] is only compatible with the past tense. It is similar to the English present perfect in marking events that took place prior to a reference time R but have consequences that are still to be perceived at R.
- *-yakt* [IPFV] is compatible with all tenses but has slightly different functions in each. In the past as well as in the imperative it is a true marker of imperfectivity, marking durative and habitual events and, as an addition, irreal consequences in combination with conditionals. In the non-past and tenseless nonfinite forms, however, it has a much narrower function – here it marks that an action is maintained with some effort, similar to English *keep doing*.
- *-ta* [CONT] only occurs in the non-past. It expresses that an event stretches without interruptions over a longer period of time including R. This excludes its use with habitual aspect and in many situations where in English the present continuous would be appropriate.
- *-dhend* [$COMPL_1$] and *-ca* [$COMPL_2$] mark completive aspect in all tenses and tenseless forms. They are complementarily distributed according to semantic verb classes: *-dhend* is the default marker, *-ca* is used with verbs of motion and in a few lexicalised cases such as with *ims-* 'sleep'. We will assume below that apart from this there is no functional difference between the two and will therefore ignore *-ca*.
- *-hat(t)* [AWAY]. The function of this marker is difficult to describe. In most cases it implies that after an action S or O is no longer where it was before (similarly to the English adverb *away* as in *go away*, *throw away*). Though this function does not correspond to any well-known typological canon, it usually goes together with the completion of an event and is therefore often exchangeable with *-dhend*.
  The form *-hat* is used with intransitive verbs, the form *-hatt* with transitive verbs. Since detransitivised verb forms are formally hybrid (intransitive inflection with transitive valency), it makes sense that both *-hat* and *-hatt* should be possible with them:

  (192) *Kapp-e aŋgreji=yaŋ nis-ad-a-ŋs-e/*
  Kalpana-NAME.NTVZ English=ADD know-AWAY.ITR-PST-PRF-IND.PST[.3sS]/
  nis-att-a-ŋs-e raicha.
  know-AWAY.TR-PST-PRF-IND.PST[.3sS] MIR
  'Kalpana has also learnt some English, too.' (elicitation DR 2010)

  I have not been able to identify any functional difference between these two variants. In the Chintang corpus, S/A detransitivisation with *-hat(t)* is extremely rare, anyway. For instance, for *ca-* 'eat', one of the verbs with the highest proportion of detransitivised objects, only two sentences are attested which contain this combination – one uses *-hat*, the other *-hatt*. It might thus be the case that speakers themselves are simply insecure about which aspect marker to choose, given the extreme rareness of the combination and the hybrid nature of the detransitivised frame. One speaker I consulted accepted both *ci-ad-e* [eat-AWAY.ITR-IND.PST[.3sS]] and *ci-att-e* [eat-AWAY.TR-IND.PST[.3sS]]; another one accepted *ciade* but found that *ciatte* sounded strange.

In addition to these markers there is the unmarked form. The unmarked form has a relatively clear function in the past, where it marks perfective, non-resultative events, but can only be described as default in the non-past and in the tenseless nonfinite forms. Table 2.8 shows a schematic overview of the aspectual system of Chintang.

	nonpast	nonfinite	imperative	past
Ø	(default)			perfective non-resultative
-ŋs	-			perfective resultative
-yakt	'keep doing'		imperfective	
-ta	continuative	-		
-dhend	completive			
-hat(t)	AWAY			

Table 2.8: The aspectual system of Chintang

The question now is whether each aspectual marker in each of its senses is compatible with S/A detransitivisation. The examples below illustrate that all theoretically possible combinations are attested. The imperative and the tenseless nonfinite forms have been ignored since the sense an aspectual marker assumes in these is always identical to either the sense in the nonpast or the past and since most nonfinite forms are indeterminate with respect to S/A detransitivisation, anyway. Although *-dhend* and *-hat(t)* are the only markers whose functions do not depend on tense, nonpast and past tense examples are given for the sake of interest.

(193) *-ŋs* [PRF]:

a. *Sa-ŋa sa-lo bug-o-ŋs-e?*
who-ERG who-NOM ask-3[s]O-PRF-IND.PST[.3sA]
'Who asked whom?' (CLC:CLDLCh3R01S02.209)

b. *Pheri biskut tad-a-ŋs-e.*
again biscuit bring-PST-PRF-IND.PST[.3sS]
'He has brought biscuits again.' (CLC:CLLDCh1R07S02 161)

(194) *-yakt* [IPFV]:

a. *Kani-dina khipt-u-yakt-u-ku-m.*
1piPOR-day count-3[s]O-IPFV-3[s]O-IND.NPST-1pA
'We keep counting our days.' (CLC:tangera_05.275)

b. *Khel-a u-num-ci-yak-ce-lok=ta khic-e*
game-NTVZ 3S-do-d-IPFV-[SUBJ.NPST.]d-CVB.BGR=FOC take.photo-V.NTVZ
numd-o-ko.
do-3[s]O-IND.NPST[.3sA]
'He takes a photo while they are playing games.' (CLC:CLLDCh4R07S05.0804)

c. *Asinda esbela bharkhari khatt-u-wakt-u-ŋ-ci-h-ẽ gor-ce.*
yesterday this.time just take.away-3O-IPFV-3O-1sA-ns-1sA-IND.PST ox-ns
'Yesterday I was just taking (back) the oxen at this time.' (CLC:CLLDCh1R04S06.0906)

d. *Anam=lo saĩla siŋ tad-a-kt-e, asinda?*
when=SURP third.son wood bring-PST-IPFV-IND.PST[.3sS] yesterday
'When was Saila bringing wood, yesterday?' (CLC:CLDLCh2R02S02.318)

(195) *-ta* [CONT]:

a. *To wamd-o-ko-ta u-taŋ.*
DEM.UP scratch-3[s]O-IND.NPST-CONT 3sPOR-head
'He is scratching his head up there.' (CLC:CLLDCh1R03S02.0737)

b. *Alu yakkheŋ a-ca-no-ta elo?*
potato curry 2[s]S-eat-IND.NPST-CONT or
'Are you eating potato curry or what?' (CLC:CLLDCh3R01S03.176)

(196) *-dhend* [$COMPL_1$]:

a. *Paisa=mo=go na=pho huĩ kampyutar-ŋa=ta*
money=CIT=$NMLZ_1$ CTOP=REP MED computer-ERG=FOC
loĩs-and-o-ko.
put.out-$COMPL_1$-3[s]O-IND.NPST[.3sA]
'As for the money, that computer produces (all of) it.' (CLC:CLDLCh3R01S03.124)

b. *Som-e u-hawa lon-na-dhen-no ni.*
Som-NAME.NTVZ 3sPOR-air let.out-LNK-$COMPL_1$-IND.NPST[.3sS] ASS
'Som lets out air (from the ball).' (CLC:CLLDCh1R13S02.1331)

c. *Ram-e-ŋa sed-and-o-ŋs-e.*
Ram-NAME.NTVZ-ERG kill-$COMPL_1$-3[s]O-PRF-IND.PST[.3sA]
'Ram has killed it.' (CLC:CLLDCh1R02S05.0028)

d. *Akka=yaŋ makkai koi thaŋ na tett-and-a-ŋs-e-h-ẽ=yaŋ.*
1s=ADD maize some place CTOP plant-$COMPL_1$-PST-PRF-PST-1sS-IND.PST=ADD
'I have also planted maize in some places.' (CLC:chintang_now.1377)

(197) *-hat(t)* [AWAY]:

a. *Bhewa-ce-ŋa u-c-o-hatt-o-ko=kha ni!*
insect-ns-ERG 3[p]A-eat-3[s]O-AWAY.TR-3[s]O-IND.NPST=$NMLZ_2$ ASS
'The insects eat it (up)!' (CLC:CLDLCh3R05S04.234)

b. *Kham lupt-ad-i-ki=ta? na.*
earth get.dirty.with-AWAY.ITR-1p[i]S-IND.NPST=FOC CTOP
'One gets all dirty with earth.' (CLC:CLLDCh1R10S03.185)

c. *Adha mil-att-o-ŋs-e.*
half swallow-AWAY.TR-3[s]O-PRF-IND.PST[.3sA]
'She swallowed (down) half of it.' (CLC:CLLDCh1R06S01.1629)

d. *Athomba anci ci-ad-a-c-e aŋ.*
before 1di eat-AWAY.ITR-PST-[1]d[iS]-IND.PST QTAG
'We've already eaten (up).' (CLC:CLLDCh4R02S02a.061)

While the combinability of S/A detransitivisation with the imperfective aspects is expected, the combinability with the perfective aspects PRF, AWAY, and especially COMPL is surprising from a European point of view. The relevant examples from above are discussed briefly below.

- Actions on non-quantifiable referents can produce results just as well as actions on quantifiable ones. For instance, bringing one or several biscuits as in (193b) results in cookies being there, just like bringing *the* cookie(s).
- *-dhend* can not only mark the completion of an action as a whole but also of single steps. An example of this is found in (196b). At the time of speaking it's not clear yet whether Som will let out all of the air from the ball – that would be exhaustive reference to a quantifiable amount and therefore require the transitive frame. But even though he only lets out more and more air, all the air that does go out is completely out and cannot be brought back into the ball.
- In (196d), the action is distributed over various places. The places and the maize planted in them represent non-quantifiable referents, so the sentence is detransitivised. *-dhend* marks that the overall act of planting maize has been completed.
- Example (197b) is rather similar to (196b). Although a non-quantifiable amount of dirt is transferred away from the soil to the speaker group, the transfer of each subamount is complete.
- Although an action like eating does not have an inherent *telos*, it can be construed as having one when it brings about a change of state. In (197d), eating however much results in the transition of the speaker from the state of not having had a meal to the state of having had a meal. It is this transition that is completed.

To summarise, S/A detransitivisation can be freely combined with all existing aspects, and semantic interaction can be fully explained from the individual semantics of the construction and the aspectual markers. Where S/A detransitivisation cannot be combined with an aspectual marker that can also be explained on the base of semantics. For instance, consider (198):

(198) *Menuwa sencak khoŋ-no(*-ta).*
cat mouse play.with-IND.NPST[.3sS]-CONT
'Cats play with mice (*right now).' (elicitation PRAR 2010)

The detransitivised sentence without *-ta* strongly tends towards a generic interpretation: it is a habit of all cats to play with mice. This reading is incompatible with *-ta*, which suggests an episodic action whose time span includes speech time. The only situation that would make (198) possible is one where there is a particular cat that spends some time playing with several mice, which is bizarre since cats normally play with only one mouse at a time.

2.6.4.4 Interaction with negation

There is no particular affinity between S/A detransitivisation and negation, contrary to what one might expect from prominent cases of interaction between negation and DOM in languages such as French or Russian. Negated verb forms can be detransitivised or not, just like all other verbs:

(199) a. *Phidaŋ u-c-o-kko-niŋ.*
ginger 3[p]A-eat-3[s]O-IND.NPST-NEG
'They don't eat the ginger.' (CLC:phidang_talk. 483)
b. *Mo=go kok ca-nik-niŋ.*
DEM.DOWN=NMLZ$_1$ rice eat-IND.NPST[.3sS]-NEG
'The one down there doesn't eat rice.' (CLC:CLLDCh2R09S01.080)

Both examples are indifferent with respect to the the scope of negation in relation to S/A detransitivisation. While the most natural interpretation of (199a) is 'they don't eat the ginger' (< 'there is a quantifiable amount of ginger that they don't eat'), 'they don't eat some ginger' (< 'there is an amount of ginger that they don't eat and it is quantifiable') is also possible. Likewise, (199b) would usually be taken to mean 'he doesn't eat (from that) rice' (< 'there is a non-quantifiable amount of rice that he doesn't eat'), but 'he doesn't eat any rice' (< 'there is an amount of rice that the doesn't eat and it is non-quantifiable') is also possible.

One interesting point is that *thitta* 'one' can be used in negated detransitivised sentences:

(200) *Ṭhitta riŋ khem-nik-niŋ.*
one word listen.to-IND.NPST[.3sS]-NEG
'He doesn't (even) listen to a single word.' (CLC:CLDLCh3R05S03.159)

In such sentences *thitta* is not in the scope of quantifiability; the meaning of (200) is not 'there is (exactly) one word he doesn't listen to' – this would require the transitive frame. This creates a strange situation: on the one hand there seems to be a non-quantifiable referent, *riŋ* (there is a non-quantifiable amount of words not listened to), on the other hand there is a quantifier with the meaning 'one' with unclear affiliation. It is, however, possible to reconcile these two components if one assumes that any negation is only justified as a contrast to an expected situation.

For instance, one will only say *It won't rain* in a situation where there is reason to expect rain. Similarly, *riŋ khemnikniŋ* will be uttered when there is reason to expect that somebody should listen to what one is saying (for instance, because that is what is expected of well-behaved children, at least in an idealised cognitive model as defined by Lakoff 1987). Thus, every negation creates a counterfactual mental space where what it negates is true. For *riŋ khemnikniŋ* there are two such spaces, depending on how one interprets the scope of negation. If negation is outside of quantifiability, the situation opposed to *riŋ khemnikniŋ* is one where some words are listened to. If negation is inside of quantifiability, it is one where at least one word is listened to. What *thitta* does in (200) is to select the second interpretation. Semantically it does not belong into the space

where words aren't listened to but into the pertaining counterfactual space where at least one word is listened to.

In some cases, negation has a direct influence on the used frame because non-action has a different effect than action. For instance in (201), people on the market buy some of the speaker's neighbour's ginger (thereby licensing detransitivisation) but none of his own. The expected frame would be the detransitivised frame in both cases because both the bought amount and the amount not bought are non-quantifiable. However, the transitive frame is used in the second case in order to express that all of the speaker's ginger is ignored by the customers:

(201) *Hath-a-be* *a-chimeki* *u-phidaŋ=le* *u-khe?-no-ta,*
market-NTVZ-LOC$_1$ 1sPOR-neighbour 3sPOR-ginger=RESTR 3[p]S-buy-IND.NPST-CONT
ak-ko *na* *u-khed-o-ko-niŋ.*
1s-GEN CTOP 3[p]A-buy-3[s]O-IND.NPST-NEG
'On the market they are only buying my neighbour's ginger, but they don't buy mine.'
(elicitation SAR 2011)

2.6.5 Conventionalisation

S/A detransitivisation is conventional in a couple of (partially frequent) contexts. Lexicalisation of S/A detransitivisation can be observed with certain verbs whose detransitivised objects are more frequently covert than those of other verbs and where there is the question of whether one should consider them as true S/A-ambitransitives (section 2.6.5.1). Another case are certain combinations of verbs with conventional object nouns where S/A detransitivisation is the default (section 2.6.5.2). Grammaticalisation of S/A detransitivisation is taking place with complex predicates (section 2.6.5.3), with pieces of information in O (section 2.6.5.4), and with certain adverbs seemingly replacing objects (section 2.6.5.5). Finally, intransitive infinitival complement clauses exhibit a peculiar system of assigning O-AGR that can be taken as a lexicalised and grammaticalised relative of S/A detransitivisation (section 2.6.5.6).

2.6.5.1 Non-specific or non-existent?

There are a couple of verbs that prefer the detransitivised frame over the transitive frame *and* drop their O more often than not in the detransitivised use. The detransitivised use with covert O can often be conveniently translated with an English intransitive verb. Here is a list:

- *cekt-* 'speak, say'
- *hand-* 'talk (about)'
- *hatt-* 'wait (for), watch out (for), look after'
- *haŋs-* 'be hot (for somebody; of food)'
- *khipt-* 'read, study, count' (Sambugaũ dialect)
- *khoŋs-* 'play (with)'
- *kupt-* 'perch, hatch'
- *nad-* 'refuse, do not eat'
- *pes-* 'vomit'
- *pokt-* 'leave'
- *ratt-* 'make noise, shout (at), scold'
- *rett-* 'laugh (at)'
- *yoŋs-* 'fast, abstain from'
- *ŋed-* 'read, study, count' (Mulgaũ dialect)

There are various reasons why these verbs display the mentioned behaviour. For most one can easily assume a detransitivised covert O from a language-internal perspective: *cekt-* 'speak' < 'say words', *hand-* 'talk' < 'talk about various matters', *haŋs-* 'be hot' < 'be hot for all kinds of people', *khipt-*, *ŋed-* 'study' < 'study various subjects', *khoŋs-* 'play' < 'play games', *nad-* 'do not eat' < 'refuse food', *pes-* 'vomit' < 'eject matter from one's stomach', *yoŋs-* 'fast' < 'abstain from

food'. These verbs do confirm with the semantics of S/A detransitivisation. The only thing that is special about them is that they drop their O more often than other verbs.

Some other verbs do not necessarily have a semantic object, whether specific or not: a fowl can simply perch (*kupt-*) without hatching anything, one can wait (*hatt-*) for Godot, make noise (*ratt-*) without addressing anybody, and laugh (*rett-*) without a good reason. *Pokt-* 'leave' must always have a person as its O in Chintang; if one wants to express that somebody left a place, the detransitivised form has to be used and O cannot be overt. For these verbs there also is a clear difference between the sense with an (overt or covert) non-specific O and no semantic O at all. 'Hatch chicks' does not mean the same as 'perch', 'wait' is not the same as 'wait for people', 'make noise' is not 'shout at people', 'laugh' is not 'laugh about things', and 'leave (a place)' is not 'leave people'. For these verbs it thus seems possible to distinguish a detransitivised variant from a truly intransitive variant, which is the one where O is not only non-specific but where it is simply not there.

That being said, a more detailed investigation shows that a clear-cut distinction between S/A ambitransitive and normal transitive verbs does not exist. For one thing, there are many cases where it's not clear whether based on the semantics of a predicate one should assume a detransitivised covert object or no object at all. For instance, cows are frequent "shouters" (*ratt-*) in the Chintang corpus. However, even if cows may not always moo at somebody, it is easy to conceive of them as if they would. Two cows standing by the road mooing could moo just for themselves, but they could also moo at passers-by. This construal is even more likely for other animals such as ducks or dogs, which "shout" at intruders into their territory.

Second, the verbs in this apparent class behave differently from each other. Being loud without addressing anybody may be okay, but, for instance, laughing without a reason is decidedly odd. Rather than saying that sometimes people do not laugh at something, it seems more correct to say sometimes nobody can understand what they are laughing at. Similarly, there may be situations where one is really waiting for nobody and nothing, but these are much rarer than one might think. Usually when one uses phrases like *they waited for time to pass by* or simply *they were waiting* there *is* an object-like referent (for instance, arrival time or any event of interest).

Another argument against a distinct class of S/A-ambitransitive verbs is that even verbs which clearly seem to necessitate an O semantically can be used as if there was no O under appropriate circumstances. For instance, there is no clear O for *khag-* 'see, watch' in the following example:

(202) *Pok-na-loĩ* *kina khaŋ-niʔ-niŋ.*
get.up-LNK-out[.SUBJ.NPST.3sS] SEQ see-IND.NPST[.3sS]-NEG
'After getting up he doesn't see (anything).' (CLC:CLLDCh1R07S01.067)

Thus, instead of posing a distinct class of S/A-ambitransitive verbs it is more useful to view verbs like the ones just discussed as special transitive verbs. All transitive verbs differ from each other with regard to how frequent they have a semantic object (most of them always have one). Very few verbs allow for no object at all under certain circumstances, but these verbs do not form a uniform class, and there are dubious cases where it's not clear whether an object is really there or not. Where the object is truly absent it is still not necessary to postulate a frame alternation different from that between the transitive and the detransitivised frame: since a non-specific referent is in a way similar to a non-existing referent, it makes sense that this variant should use the detransitivised frame.

This situation raises important theoretical questions. One is how to determine valency cross-linguistically. For Chintang, it is obviously impossible to draw a line between S/A ambitransitives and normal transitive verbs, so the easiest solution is to assume that an object is present in the valency of a verb whenever it can potentially be expressed overtly. But how to deal with the equivalents of verbs such as *hatt-* and *rett-* in other languages? Are English *wait* and *laugh* transitive and mark their P with the prepositions *for* and *at/about*? If they are transitive the number of transitive verbs in the lexicon of English will make a leap. If they are not, why are they not? A possible way out of this dilemma would be to say that transitivity in English does not pertain to lexical items at all but only to frames – which is actually common implicit practice in English dictionaries.

However, this does not solve the problem that there are languages like Chintang where there are no good criteria to distinguish between intransitive and transitive uses (except in a very limited sense, e.g. with respect to verbal morphology) and that such languages should be comparable with languages such as English as far as possible.

2.6.5.2 Frequent composite activities

In section 2.6.2 we introduced the concept of composite activities, i.e. combinations of a predicate and an object type that have some characteristics of their own and therefore tend to be viewed as a whole rather than as consisting of two components. This often happens when a predicate-object combination is frequent and always follows the same scheme. Such combinations tend to become lexicalised and are therefore to be treated under the heading of conventionalised S/A detransitivisation. (203) shows an example.

(203) a. *Ram-e-ŋa u-koŋcɨk wachid-o-ko.*
Ram-NAME.NTVZ-ERG 3sPOR-knee wash-3[s]O-IND.NPST[.3sA]
'Ram washes his knee.'
b. *Ram-e muk wachi-no.*
Ram-NAME.NTVZ hand wash-IND.NPST[.3sS]
'Ram washes (his) hands.' (elicitation PRAR 2010)

Washing a knee is a rare activity, whereas washing (both) one's hands is something one does all the time. Washing one's hands can thus be viewed as a composite activity, and indeed *muk (wa)chid-* is almost always used with the detransitivised frame in Chintang. The transitive frame is still possible under special circumstances:

(204) *Ram-e-ŋa thitta u-muk wachid-o-ko, phalto*
Ram-NAME.NTVZ-ERG one 3sPOR-hand wash-3[s]O-IND.NPST[.3sA] other
wachid-o-ko-nɨŋ.
wash-3[s]O-IND.NPST[.3sA]-NEG
'Ram washes one of his hands but not the other.' (elicitation PRAR 2010)

(205) *I-muk-ce chid-u-c-a temma.*
2sPOR-hand-ns wash-3O-ns-IMP[.2sA] well
'Wash (both) your hands well!' (CLC:CLLDCh2R02S09.609)

In (204), the two hands are treated in different ways so that it becomes necessary to keep them apart. In (205), the speaker emphasises that the hearer (a child) is to wash both his hands and not only one.

Another example is (206). *lus-* 'engage in a rhythmical activity' is conventionally used with the detransitivised frame with the objects *cham* 'song' (*cham lus-* 'sing') and *lak* 'dance' (*lak lus-* 'dance'). However, (206) contains a contrast between several songs that can be sung at the Wadhangmi festival and the only possible dance on that occasion, so *lak lus-* is exceptionally used with the transitive frame:

(206) a. *Wadhaŋmi na, ekdam akka cham lu-ma ni-ŋa-nɨŋ.*
wadhangmi CTOP very 1s song do-INF know.to-1sS-NEG.[SUBJ.]NPST
'On Wadhangmi I don't know at all how to sing (songs).' (CLC:chintang_sahid.223)
b. *Lak lu-ma na akka ekdam nis-u-ku-ŋ.*
dance do-INF CTOP 1s very know.to-3[s]O-IND.NPST-1sA
'(But) I know very well to dance (the dance).' (CLC:chintang_sahid.225)

Other composite activities with conventional S/A detransitivisation are *kok ca-* [rice eat] 'eat, have a meal', *maŋla khag-* [augury watch] 'inspect the augury' (at the Wadhangmi festival), *ŋalɨŋ tept-* [face wash] 'wash one's face', *topi wat-* [hat put.on] 'put on/wear a hat' (also with other clothes), *thal-a (wa)lekt-* [plate-NTVZ wash] 'do the dishes', *tei? wadhapt-* [clothes wash] 'wash clothes'.

2.6.5.3 Complex predicates

Complex predicates in Chintang can be defined as combinations of an abstract noun coding a state of affairs and called "N" below with one of the light verbs *lis-* 'be, become, happen', *numd-* 'do', or *mett-* 'do to, do with'. The light verb that is of interest in the present context is *numd-*.[21] If N is assigned a role by *numd-* it is P. Below is a first example of the complex predicate *kama numd-* 'work' with an intransitive verb form:

(207) *Milane kam-a numd-a-kt-e.*
Milane work-NTVZ do-PST-IPFV-IND.PST[.3sS]
'Milane was doing work/working.' (CLC:warisama_talk.417)

In most complex predicates N is a Nepali noun (e.g. *kam-a* < Nep. *kam*), but there are also a few combinations with Chintang (*maŋ* 'prayer' + *numd-* = 'pray, worship') and English nouns or verbs (*phon* 'telephone, phone call' + *numd-* = 'phone'). All these nouns semantically oscillate between a referential and a predicational reading. One can 'do a job' or 'work', 'perform a ritual' or 'worship', 'make a call' or 'phone'. However, formally they behave quite differently from each other in two important respects. One is whether N can be construed as an independent referent occupying P and can accordingly be used with the detransitivised and the transitive frame. This is possible, for instance, with *kama numd-*:

(208) a. *Lo, hani-ŋa ba-i etti kam-a numd-a-n-u-m-a!*
okay 2p-ERG PROX-LOC$_2$ this.much work-NTVZ do-IMP-2p-3[s]O-2A-IMP
'Okay now, do this much work here!' (CLC:story_tiger.061)

b. *Utti khei?yã=ta akka mi=kha themthemthem=kha kam-a*
then TMP.ABL=FOC 1s small=NMLZ$_2$ various=NMLZ$_2$ work-NTVZ
numd-a-k-e-h-ẽ.
do-PST-IPFV-PST-1sS-IND.PST
'After that I had various small jobs.' (CLC:lifestory_JK.17)

Other complex predicates such as *bola numd-* 'make an effort' do not allow the transitive frame:

(209) **Bol-a numd-o-ko.*
effort-NTVZ do-3[s]O-IND.NPST[.3sA]
'He makes an effort.' (elicitation SAR 2011)

The other point in which complex predicates differ is whether they allow an additional argument besides N. This is neither possible with *kama numd-* nor with *bola numd-*, although semantically an additional P would be conceivable here, e.g. 'work on' and 'struggle for'. An example for a complex predicate that can do this is *bihe numd-* 'marry', for which there are two possibilities. Marriage is considered an inherently reciprocal activity in Chintang, so a marriage partner can only be marked by the comitative, A getting S-AGR (210a). However, *bihe numd-* can also mean 'marry off', in which case T is marked by NOM and gets O-AGR and G (if overt) is marked by LOC (210b).

(210) a. *Kina akka the=kha~kha-niŋ=le biha num-ma-ʔã=kha.*
SEQ 1s big=NMLZ$_2$~INTENS-COM=RESTR marriage do-1sS-IND.NPST=NMLZ$_2$
'And I will only marry one that is really big.' (CLC:mouse_story.133)

b. *Ram-e u-ppa-ko car-jana u-chau-ce*
Ram-NAME.NTVZ 3sPOR-father-GEN four-HUM.CLF 3sPOR-child-ns
u-yuŋ-no. Abo hicci-baŋ bihe numd-o-s-u-c-e.
3[p]S-be.there-IND.NPST now two-HUM.CLF marriage do-3O-PRF-3O-ns-IND.PST[.3sA]
'Ram's father has four children. Now he has married off two.' (elicitation SAR 2011)

[21] *lis-* is intransitive and thus irrelevant to S/A detransitivisation. *mett-* is a double object ditransitive verb, so O-AGR is linked to G, the manipulated object. It is an interesting question whether *mett-* and frequent N merge to such an extent that N should no longer be considered T but part of the predicate (G thereby becoming P); however, this has nothing to do with S/A detransitivisation.

N	meaning with ***numd-***	O-AGR with N	additional argument
bihe	'marry'	0/37 (0%)	'marry P-COM' (no O-AGR) or 'marry T-NOM off to G-LOC' (O-AGR)
bola	'make an effort'	0/14 (0%)	no
gali	'insult'	1/15 (7%)	'insult P-NOM' (normally with *pid-* 'give')
kama	'work'	14/78 (18%)	no
khela	'play'	2/227 (1%)	'play with P-NOM'
maŋ	'pray, worship'	6/65 (9%)	no
pas	'pass (an exam)'	0/21 (0%)	'pass P-NOM'
phon	'phone'	14/16 (88%)	'phone P-NOM'
siya	'bow, greet'	2/11 (18%)	'bow to/greet P-NOM'

Table 2.9: Properties of complex predicates

Complex predicates that allow an additional argument often do not allow for an independent conceptualisation of N:

(211) Hun-ce(*-ŋa) ramma kai?-ma=go biha u-numd-e.
MED-ns-ERG joy come.up-INF=NMLZ$_1$ marriage 3[p]S/A-do-IND.PST
'They had a joyful marriage.' (elicitation RBK 2010)

Table 2.9 is a summary of the properties of some of the most frequent complex predicates in the CLC. The frame frequencies in brackets are relative to the number of all unambiguous frames (that is, unambiguous S/A detransitivisation or transitivity with either N or an additional argument).

For the majority of N in the table the ratio of S/A detransitivisation and the transitive frame (both as triggered by N) is reversed as compared to the normal situation, S/A detransitivisation being by far the most frequent frame. Several N (*bihe*, *bola*, *gali*, *pas*) do not allow O-AGR at all. The English N *pas* and *phon* behave exceptionally in being (almost) incompatible with S/A detransitivisation. Whereas for *pas* the reason for this is quite clear (*pas* has an additional P, the exam, in all 20 remaining cases), the behaviour of *phon* is unexpected. While it is the case that a phone call is easier to conceptualise as an independent referent than a "pass", the same is all the more true of N like *kama* 'work, job', which do not have similarly high proportions of O-AGR. What's more, although one almost always calls somebody when one makes a call, the callee only rarely triggers O-AGR (2 instances). Presently these facts cannot be explained.

Fortunately they do not influence the big picture, which is that complex predicates collocate with S/A detransitivisation. The question is whether this is because N and the light verb are fused to such a degree that they have to be viewed as a whole as an intransitive predicate or because N has properties that make it akin to normal detransitivised P. All in all the second solution seems to have more arguments on its side:

- It maintains parallelism between form and function – N is morphosyntactically and functionally independent, so it looks like an argument in every respect.
- It pays reference to the fact that N and normal detransitivised objects are functionally similar: the N in complex predicates is never quantifiable when the predicate is inflected intransitively.
- It explains why some N can trigger O-AGR under the same conditions as other nouns.

Neither solution is very good at dealing with additional arguments. A fused intransitive predicates should not allow such an argument at all. On the other hand, a predicate with N as its P does not have a slot for an additional argument, so one would have to assume an alternation between a monotransitive frame (N=P) and a ditransitive frame (N=T or G). However, N and additional arguments do not always map nicely to the ditransitive role set. While for the "ditransitive"

variant of *phon numd-* it is intuitive that the callee should be G and accordingly *phon* should be T, it is completely unclear which roles *khela* and the playee are mapped to since neither of them moves, whether physically or metaphorically. Moreover, N does not seem to have any referential properties in the presence of an additional argument.

I would therefore advocate a view where both variants of *khela numd-* and similar complex predicates are monotransitive. When there is no additional argument, N itself functions as P. When there is an additional argument, N loses its argument status and is semantically truly fused with the verb so that the additional argument can occupy P.

Beside complex predicates with a nominal component there are also combinations with Chintang reduplicated adverbials (212) and Nepali adjectives (213) that express a single meaning. Since these components are not interpretable as P, they are always used with the intransitive frame when there is no additional argument:

(212) *Cha-ce=yaŋ pheri carko=ta chululu-wa u-num-no?.*
child-ns=ADD again very=FOC fidgety-ADVZ 3[p]S-do-IND.NPST
'The children are also really being fidgety again.' (CLC:CLLDCh4R02S01.0885)

(213) *Utti bela caĩ mimoŋ khebak dhilo numd-a-ŋs-e=pho.*
that.much time RETRV a.little crab late do-PST-PRF-IND.PST[.3sS]=REP
'At that time, they say, the crab had become a bit late.' (CLC:khebak_tale.092)

2.6.5.4 Pieces of information as objects

There are two types of verbs coding the expression of information, *verba dicendi* and *verba cogitandi*. Since these are functionally alike and formally behave identically in Chintang, it is convenient to treat them as one class. All verbs coding the expression of information are transitive. Here is a list:

- *cekt-* 'say, speak, speak about, tell' (monotransitive)
- *dumd-* 'think about, ponder' (monotransitive)
- *lott-* 'speak (a language)' (monotransitive), 'tell, bring across' (double object ditransitive)
- *lud-* 'say to, tell' (double object ditransitive)
- *lus-* 'tell, recite, sing' (monotransitive), 'tell' (double object ditransitive)
- *lutt-* 'tell for, explain' (double object ditransitive)
- *mitt-* 'think, think of, remember' (monotransitive), 'consider, think of as' (double object ditransitive)
- *ŋis-* 'know' (monotransitive)
- *yok-mett-* 'tell, inform about' (double object ditransitive)

Here we are only interested in those verbs that have the piece of information as their O, that is, *cekt-*, *dumd-* and *ŋis-* and the monotransitive variants of *mitt-* and *lus-*. The examples below illustrate the transitive use of these verbs.

(214) *Thitta bhanai u-cekt-o-ko.*
one statement 3[p]S-say-3[s]O-IND.NPST
'They have their saying.' (CLC:exp_wadh_DK.256b)

(215) *Dumd-u-m kina cekt-u-m-kha-m-ne-na huŋ=go*
ponder-3[s]O-[SUBJ.]1pA SEQ say-3[s]O-1pA-CON-[SUBJ.]1pA-OPT-INSIST MED=NMLZ$_2$
tuŋdum.
matter
'Let's think about that matter and try to speak about it.' (CLC:chintang_now.1154)

(216) *A-kam-ce-niŋ khel-a numd-i-ŋa=go*
1sPOR-friend-ns-COM game-NTVZ do-1pS-[SUBJ.]e=NMLZ$_1$
mitt-u-ŋ-sed-u-h-ẽ kina.
think.of-3[s]O-1sA-DYSF.TR-3[s]O-1sA-IND.PST SEQ
'I thought of how I played with my friends.' (CLC:ctn_katha.009)

(217) *Paile=go katha-ce lus-u-ku-ce e?.*
earlier=NMLZ$_1$ story-ns tell-3O-IND.NPST-[3sA.]3nsO or
'He tells stories from the old times, huh?' (CLC:chintang_now.580)

(218) *Phidaŋ u-kott-a-kt-a=go ŋis-u-ku-ŋ=ta.*
ginger 3[p]S-carry-PST-IPFV-PST=NMLZ$_1$ know-3[s]O-IND.NPST-1sA=FOC
'I know that they were carrying around ginger.'
(CLC:phidang_talk.085 + elicitation RBK 2012)

As the sentences show, the piece of information in O can be coded via a noun (*bhanai* 'saying', *tuŋdum* 'matter', *katha* 'story') or via a complement clause (*khela numdiŋago* 'how/that we played', *phidaŋ ukottaktago* 'that they were carrying ginger'). Note that with complement clauses as the one in (218) S/A detransitivisation is impossible. This is because in order to be able to say that one knows a fact one has to be able to identify it.

Such examples are, however, not representative. Only *dumd-* and *lus-* are normally used as shown above. By contrast, the two most frequent verbs in this class, *cekt-* and *mitt-*, are much more frequently found in a different construction where the citation particle *=mo* is used and the verb is detransitivised:

(219) a. *Akka ko-si khai?-yã?=mo u-cek-no.*
1s walk.around-PURP go-[SUBJ.NPST.]1sS=CIT 3[p]S-say-IND.NPST
'They say "I'm going for a walk".' (CLC:tangera_05.058)
b. *Mek=mo cek-no elo.*
mek=CIT say-IND.NPST[.3sS] or
'It (the goat) says "mek", doesn't it?' (CLC:CLLDCh1R06S04.0812)

(220) *Kam-ma ekdamsita mai-pi-no=mo u-mi?-no cha-ce.*
friend-ERG a.lot 1piO-give-IND.NPST[.3A]=CIT 3[p]S-think-IND.NPST child-ns
'The children think "Friends give us a lot."' (CLC:CLLDCh1R03S01.0371)

Note that there is no difference in Chintang between direct and indirect speech, so (219a) could also be translated as 'They say they're going for a walk' and (220) as 'The children think that friends give them a lot'. *=mo* marks that the preceding elements cannot be interpreted in the present context, either because they have been uttered or thought in a different context than that of the speech act situation as in (219a) and (220), or because they only refer to themselves as *mek* in (219b). Thus, although the clause or word marked by *=mo* (the "citation") formally takes the place of P, its referent is no longer a piece of information but a linguistic expression potentially containing information.

This kind of referent resembles open referents (cf. section 2.6.2). In the case of open reference, the link between a pointer and a referent cannot be established yet at speech time in the mental space in focus. In the case of citations, the pointer can be linked to a linguistic expression, but what really would be of interest is the meaning of that expression (if there is one).

Similarly as with open referents, quantifiability does not matter for citations because of their special referential properties. For instance, in (221) it is completely clear from the context that the mouse only said a single sentence, yet that sentence triggers S/A detransitivisation in the matrix (NOM on the A *sencak*):

(221) *Them=yaŋ manche? naŋ ba-i? na=mo cekt-e=pho ni sencak.*
what=ADD be.not.there but PROX-LOC$_2$ CTOP=CIT say-IND.PST[.3sS]=REP ASS mouse
'The mouse said "But there is nothing here at all!"' (CLC:story_cat.136)

S/A detransitivisation with citations is grammaticalised to such a degree that the transitive frame has become ungrammatical after *mo* in most cases:

(222) *Paĩ=go bihe-be lak lu-no=mo u-cek-no/*
today=NMLZ$_1$ wedding-LOC$_1$ dance do-IND.NPST[.3sS]=CIT 3[p]S-say-IND.NPST
**u-cekt-o-ko.*
3[p]A-say-3[s]O-IND.NPST
'They say that there will be dancing at the wedding today.' (elicitation DKR 2010)

There are only two possible exceptions to this rule. One is where a citation is heard so often that it can be considered to have acquired a distinct referential identity:

(223) a. *Huŋ=go khali=ta namaste=mo cek-no.*
MED=NMLZ$_1$ always=FOC namaste=CIT say-IND.NPST[.3sS]
'He always says "Namaste".' (elicitation PRAR 2010)
b. *Huĩ-sa-ŋa khali=ta namaste=mo cekt-o-ko.*
MED-OBL-ERG always=FOC namaste=CIT say-3[s]O-IND.NPST
'He always says his namaste.' (elicitation PRAR 2010)

The other possibility is when something that has been said or thought is immediately referred back to by another predicate so that it becomes identifiable as a referent:

(224) *Akka numd-u-ku-ŋ-niŋ=mo cekt-a-ŋ=go/ cekt-u-ŋ=go hana*
1s do-3[s]O-IND.NPST-1sA-NEG=CIT say-PST-1sS=NMLZ$_1$ say-3[s]O-1sA=NMLZ$_1$ 2s
a-khems-e?
2[s]S/A-hear-IND.PST
'Did you hear that I said I won't do it?' (elicitation DKR 2010)

(225) *Temma lis-e=mo u-cek/ u-cekt-o*
nice become-IND.PST[.3sS]=CIT 3[p]S-say[.SUBJ.NPST] 3[p]A-say-[SUBJ.NPST.]3[s]O
nuseyaŋ akka huŋ=go mai-khem-ma num-ma-ʔã.
CONCS 1s MED=NMLZ$_1$ NEG-hear-INF do-1sS-IND.NPST
'Even if they say it's nice I don't listen to it.' (elicitation DKR 2010)

2.6.5.5 S/A detransitivisation with adverbials

The verb *numd-* 'do' is frequently used with modal adverbs derived from demonstrative roots by the suffix *-khiʔ* [MOD] and its derivates (*-khiʔniŋ* and *-khiʔni* [METHOD] < *-khiʔ-niŋ* [MOD-COM], *-khiʔni* [MOD-DIR]). The most common frame in this constellation is the detransitivised one:

(226) *Huĩ yo-khi num-no.*
MED DEM.ACROSS-MOD do-IND.NPST[.3sS]
'He does it like that.' (CLC:CLLDCh2R02S06.961)

Since the most natural translation of (226) into English involves *it*, these sentences can create the impression that the modal adverb itself occupies the role of P. However, a translation more faithful to the structure of Chintang would be 'He acts like that' or even better, 'He does things like that' with a non-quantifiable object *things*. That such an object can indeed be assumed is shown by the rare case of modal adverbs being used with the transitive frame, which requires a quantifiable object referent:

(227) *Ba-khiʔ numd-o-kh-o i-taŋ!*
PROX-MOD do-3[s]O-CON-[IMP.2sA.]3[s]O 2sPOR-hair
'Make your hair like this!' (CLC:CLLDCh3R06S05.544)

Nevertheless, it is possible that modal adverbs make it possible to drop non-quantifiable objects more easily in this construction than elsewhere because they make up for the informational gap left by the omitted object. This is also seen with other, non-demonstrative modal adverbs:

(228) *Hale, chito numd-i!*
let's.go quick do-[SUBJ.]1p[i]S
'Let's go, hurry up (let's do things quickly) now!'

(229) *Kani bekle mitt-i-ki.*
1pi different think-1p[i]S-IND.NPST
'We think differently (about things).' (CLC:tangkera_05.073)

One rather special word with mixed adverb/noun characteristics is *aŋ*. This word regularly has to be translated into English using *what*:

(230) a. *Aŋ lis-e?*
what happen-IND.NPST[.3sS/A]
'What happened?' (CLC:CLLDCh3R11S12.278)
b. *Lo? kina Monu esari aŋ num-no?*
okay SEQ Monu lately what do-IND.NPST[.3sS]
'Okay, and what is Monu doing these days?' (CLC:Tel_talk_01.022)

Nevertheless it is morphologically different from interrogative nouns like *sa-* 'who' and *them* 'what' in that it can neither carry the non-singular suffix *-ce* (vs *sa-ce* 'which people', *them-ce* 'which things') nor any case markers (vs e.g. *sa-ŋa* [who-ERG], *them-be* [what-LOC$_1$]). Thus from this viewpoint, it rather looks like another modal adverb.

However, differently from the modal adverbs above, *aŋ* can not be used with the transitive frame (231). In order to add an object that yields the transitive frame, *aŋ* has to be combined with double object ditransitive *mett-* 'do to, do with' so that *aŋ* occupies the role of T (232).

(231) **Aŋ a-numd-o-ko?*
what 2[s]S-do-3[s]O-IND.NPST
'What are you doing?' (elicitation RBK 2010)

(232) *Samjhana-ŋa dabai-ce aŋ mett-u-ŋs-u-c-e?*
Samjhana-ERG medicament-ns what do.to-3O-PRF-3O-ns-IND.PST[.3sA]
'What has Samjhana done with the medicaments?' (CLC:CLLDCh3R03S02a.754)

This is not because it is interrogative. Interrogative nouns are fully compatible with both frames, and so is the quasi-synonymous modal adverb *ho-khi?* [which-MOD]:

(233) a. *Maila them ca-no?*
second.son what eat-IND.NPST[.3sS]
'What is Maila eating?' (CLC:CLLDCh4R06S01.1395)
b. *I-them a-copt-o-ko ettikhera somma?*
2sPOR-what 2[s]A-look.at-3[s]O-IND.NPST this.time TERM
'What are you looking at (on you) that long?' (CLC:CLLDCh3R06S05.341)

(234) a. *Pacche na abo pakku cahi ho-khi a-numd-a-ŋs-e?*
later CTOP now younger.uncle RETRV what-MOD 2[s]S-do-PST-PRF-IND.PST
'And then later what did you do, uncle?' (CLC:chintang_now.1397)
b. *Ho-khi numd-o-ko?*
which-MOD do-3[s]O-IND.NPST[.3sA]
'How does he do it?' (CLC:CLLDCh4R06S03.0867)

The hybrid behaviour of *aŋ* can be easily accounted for if we take a closer look at its semantics. In contrast to *them*, *aŋ* cannot refer to quantifiable referents. This explains why *them* can be pluralised (*them-ce* 'what things') but *aŋ* can't and also why *aŋ* is incompatible with the transitive frame. Another, related property of *aŋ* is that it can only be used with highly abstract referents that would be hard to track, anyway. Whereas *them* can be combined with any verb, *aŋ* only collocates with a few frequent verbs having a NOM-marked argument position matching these characteristics, viz. *lis-* 'be, become, happen', *numd-* 'do', *mett-* 'do to, do with', *cekt-* 'say', and *lud-* 'tell'. We therefore do not have to assume a one-member part of speech for *aŋ* but can simply say that it is a noun

whose morphosyntactic characteristics are predictable from its semantics.

2.6.5.6 Complement clauses with intransitive infinitives

As we have seen in section 2.3.5.3 and section 2.3.5.4, there are a number of constructions in which a transitive matrix is combined with an intransitive embedded frame. Here we are only interested in those cases where the embedded clause can be viewed as the P of the matrix verb. All these constructions use the infinitive. The list below is repeated from section 2.3.5.3.

- {A-ERG P-[V.NONF] V-a(A).o(V.NONF)}
- {A-ERG/NOM P-[V.NONF] V-a(A).o(V.NONF)}
- {A-NOM P-[V.NONF] V-s(A)}

Below is one example for each complex frame.

(235) *Sa-ŋa im-ma tog-o-ko-niŋ?*
who-ERG sleep-INF get-3[s]O-IND.NPST[.3sA]-NEG
'Who doesn't get to sleep?' (elicitation SAR 2011)

(236) *I-khuwa(-ŋa) tuk-ma puŋs-o-ko?*
2sPOR-wound-ERG ache-INF start-3[s]O-IND.NPST[.3sA]
'Does your wound start aching?' (elicitation RMR 2011)

(237) *Ram-e wacak-ma ni-no.*
Ram-NAME.NTVZ swim-INF know-IND.NPST[.3sS]
'Ram can/knows how to swim.' (elicitation SAR 2011)

These constructions are of interest because there is the question to what extent the behaviour of infinitival P parallels that of nominal P. There is indeed an interesting pattern at work here: the choice of the complex frame can be partially predicted on the base of the telicity of the matrix verb. All telic matrix verbs have dummy 3sO-AGR:

- *chitt-* 'find the time to'
- *let-* 'stop doing'
- *mund-* 'forget to'
- *nad-* 'reject to'
- *pukt-/puŋs-/phind-* 'start to'
- *tok-* 'get to'

By contrast, all atelic matrix verbs only have S-AGR:

- *hid-* 'be able to'
- *lapt-* 'be about to'
- *mitt-* 'like to'
- *ŋis-* 'know to'

What telicity does not predict is the case of A – *chitt-*, *nad-* and *tok-* have A-ERG, the remaining verbs in the first group have A-ERG/NOM. Also note that there is one exception – *kond-* in the sense 'want, try' is atelic but has A-ERG and 3sO-AGR. Nevertheless, all infinitives of punctual matrix verbs behaves like transitive O with regard to O-AGR, and the majority of infinitives of atelic matrix verbs behave like detransitivised O. This makes sense insofar as punctual events have temporal boundaries on either side and do therefore loosely correspond to quantifiable referents, whereas atelic events do not have an inherent end point and are thus similar to non-quantifiable referents.

There is one major difference, though: if the infinitive behaved perfectly parallel to nominal P, it should be its own telicity and not that of the matrix which trigger S/A detransitivisation. However, atelic matrix verbs will even have S-AGR when the embedded verb has temporal boundaries (238), and telic matrix verbs will even have O-AGR when the embedded event does not have temporal boundaries (239).

(238) *I-chau ek minet ep-ma hi-no?*
2sPOR-child one minute stand-INF be.able-IND.NPST[.3sS]
'Can your child stand (upright) for a minute?' (elicitation RMR 2010)

(239) *Utti ghari yo-ʔni bha-iʔ-ni ko-ma*
that.much TMP.LOC DEM.ACROSS-DIR$_1$ PROX-LOC$_2$ wander-INF
led-and-u-ŋs-u-h-ẽ.
stop-COMPL$_1$-3[s]O-PRF-3[s]O-1sA-IND.PST
'At that time I had stopped wandering around.' (elicitation RMR 2010)

Thus, although there are some interesting parallels between S/A detransitivisation and the behaviour of matrix verbs with intransitive embedded frames, there are too many formal and functional differences to posit a synchronic link between the two constructions.

2.6.6 Some irrelevant variables

The preceding sections have shown that specificity is the central factor conditioning S/A detransitivisation and that this is in turn highly correlated with quantifiability. The quantitative data that will be presented in section 2.7 will further strengthen this picture. The other side of the centrality of specificity is that a lot of variables which are known to have an impact on O marking in other languages or which could be imagined to do so are not needed for the explanation of S/A detransitivisation in Chintang in that they don't add anything to what is already predicted by specificity.

These variables are briefly touched upon in this section. I included them because irrelevant variables (as absent things in general) are rarely talked about in linguistics and also because my initial elicitation work included the exploration of the relevance of a variety of variables, anyway. This section is, however, not meant as a comprehensive treatment of possible other factors in S/A detransitivisation, nor does it give any kind of proof that individual variables do not play any role in it – such a proof would not be trivial. Instead, a few examples are given for each variable which contrast very clearly and where one would thus expect a formal effect if the concerned relevant was functionally relevant.

2.6.6.1 Animacy and power of O

Animacy can be freely combined with S/A detransitivisation. The sentence below is a kit for six sentences (three animacy levels multiplied with two frames), all of which are grammatical.

(240) *Akka maʔmi/gohi/luŋtak khag-u-h-ẽ/khag-e-h-ẽ.*
1s person/crocodile/stone see-3[s]O-1sA-IND.PST/see-PST-1sS-IND.PST
'I saw a person/crocodile/stone.' / 'I saw people/crocodiles/stones.' (elicitation RBK 2010)

The related variable of power of O is likewise irrelevant. Both powerful and powerless animate beings can be freely combined with both frames:

(241) *Akka asinda jaŋgal-a-be bhalu/cikiyaŋ*
1s yesterday jungle-NTVZ-LOC$_1$ bear/ant
khag-u-h-ẽ/khag-e-h-ẽ.
see-3[s]O-1sA-IND.PST/see-PST-1sS-IND.PST
'Yesterday in the jungle I saw a bear/an ant.' / 'Yesterday in the jungle I saw bears/ants.'
(elicitation PRAR 2010)

2.6.6.2 Alienability of O

There is a strong yet predictable effect of possession on S/A detransitivisation (see section 2.6.4.2 for details). Alienability is hard to disentangle from possession, but one context where this is possible are conventionalised processes (section 2.6.5.2) like hand-washing, where possessed referents need not be marked as possessed. Here, alienable and inalienable referents alike occur with S/A detransitivisation:

(242) *Ram-e muk wachi-no.*
Ram-NAME.NTVZ hand wash-IND.NPST[.3sS]
'Ram washes (his) hands.' (elicitation PRAR 2010)

(243) *Ram-e thal-a wachi-no.*
Ram-NAME.NTVZ plate-NTVZ wash-IND.NPST[.3sS]
'Ram washes plates / does the dishes.' (elicitation PRAR 2010)

2.6.6.3 Kinship

It is hard to find examples of detransitivised clauses with a kinship term in O because the number of relatives of one kind is usually easy to overlook so that the transitive frame is the default. However, examples such as the following are possible:

(244) *Ak-ko a-yaŋme-ce u-bhuŋ kina kuneikunei*
1s-GEN 1sPOR-grandchild-ns 3[p]S-be.much[.SUBJ.NPST] SEQ some
mɨkseĩkhaŋ-ŋa-ʔã-nɨŋ.
recognise-1sS-IND.NPST-NEG
'My grandchildren are so many that I don't recognise some of them.'
(elicitation SAR 2011)

2.6.6.4 Social distance to and rank of O

Both of these related variables are irrelevant, as shown by the following pairs of examples. Friends, strangers, knowledgeable elders, and (poor) exchange workers in O can all be used with the detransitivised frame under appropriate conditions:

(245) *Akka kam/bidesi khag-u-h-ẽ/khag-e-h-ẽ.*
1s friend/stranger see-3[s]O-1sA-IND.PST/see-PST-1sS-IND.PST
'I saw a friend/stranger.' / 'I saw friends/strangers.' (elicitation SAR 2011)

(246) a. *Weiʔnakma ghari pujari-ŋa budha-ce katt-u-c-e.*
rain.ritual TMP.LOC priest-ERG elder-ns bring.up-3O-3nsO-IND.PST[.3sA]
'At the time of the rain ritual the priest invited the elders (to come up to the mountain where the ritual is performed).'
b. *Weiʔnakmak ghari pujari budha katt-a-ŋs-e.*
rain.ritual TMP.LOC priest elder bring.up-PST-PRF-IND.PST[.3sS]
'At the time of the rain ritual the priest invited elders.' (elicitation SAR 2011)

(247) *Akka boniwala patt-u-s-u-h-ẽ/patt-a-s-e-h-ẽ.*
1s exchange.worker call-3[s]O-PRF-3[s]O-1sA-IND.PST/call-PST-PRF-PST-1sS-IND.PST
'I called an exchange worker.' / 'I called exchange workers.' (elicitation SAR 2011)

2.6.6.5 Discourse topicality of O

Discourse topicality is likewise irrelevant to S/A detransitivisation. An O referent that is mentioned for the first time can be used with the transitive frame when it is trackable, as the crab in the last clause in (248), where the frame is indicated by the ERG on the postposed A.

(248) *Huŋ=go wacak-ma hid-e kina bahira lond-e kina*
MED=NMLZ take.bath-INF finish-IND.PST[.3sS] SEQ outside go.out-IND.PST[.3sS] REP
pho luŋghek-ko kap-a-be thitta khebak copt-e pho
stone-GEN crack-NTVZ-LOC$_1$ one crab see-IND.PST[.3sA] REP MED-OBL-ERG
huĩ-sa-ŋa.

'After he finished swimming and came out (from the river), he saw a crab in a crack on a stone.' (CLC:khebak_tale.009-010)

Conversely, a referent that has been mentioned many times before can be used with the detransitivised frame as long as a non-trackable subamount of it is affected. In the group of examples in (249), rice and religious practices connected to it have been the subject of the conversation for quite a few sentences, but since not exactly the same rice is affected every time, the frame keeps oscillating between transitive and detransitivised.[22]

(249) a. *Ma, kok na huŋ=go-iʔ u-thuk-nik-niŋ=kha naŋ.*
Q rice CTOP MED=NMLZ-LOC$_2$ 3pS-cook-IND.NPST-NEG=NMLZ$_2$ but
'But they don't cook rice there, do they?' (CLC:phidang_talk.381)

b. *Hokko-iʔ-ya u-tad-o-koʔ*
which-LOC$_2$-ERG 3pA-bring-3[s]O-IND.NPST
'Where do they bring it from?' (CLC:phidang_talk.381)

c. *Kok u-bhokt-o-ko… huŋ=go-iʔ u-thuk-no ni.*
rice 3pA-stick.on-3[s]O-IND.NPST MED=NMLZ-LOC$_2$ 3pS-cook-IND.NPST ASS
'They stick rice (on that stone), and… they cook it right there.'
(CLC:phidang_talk.382-383)

2.6.6.6 Contrastive focus on O

Contrastive constituent focus is possible with both transitive and detransitivised O, as shown by (250a) and (250b):

(250) a. *Akka maʔmi khag-u-h-ẽ, piʔ caĩ mahaʔ.*
1s person see-3[s]O-1sA-IND.PST cow RETRV be.not
'I saw a human, not a cow.'

b. *Huŋ=go murali muʔ-no, bãsuri caĩ mahaʔ.*
MED=NMLZ$_1$ a.type.of.flute blow-IND.NPST[.3sS] a.type.of.flute RETRV be.not
'He plays the *murali* flute, not the *bãsuri*.' (elicitation RBK 2010)

2.6.6.7 Tense

Tense is not only well known to be frequently relevant to DAM but is also closely related to aspect, which was shown to interact (if in predictable ways) with S/A detransitivisation in section 2.6.4.3. Nevertheless, no direct effect on S/A detransitivisation was found. Both nonpast (251a) and past (251b) can be freely combined with the transitive and the detransitivised frame.

(251) a. *Akka maʔmi kha-u-ku-ŋ/khaŋ-ŋa-ʔã.*
1s person see-3[s]O-IND.NPST-1sA/see-1sS-IND.NPST
'I see somebody/people.'

b. *Akka maʔmi khag-u-h-ẽ/ khag-e-h-ẽ.*
1s person see-3[s]O-1sA-IND.PST see-PST-1sS-IND.PST
'I saw somebody/people.' (elicitation PRAR 2010)

2.6.6.8 Typicality of A

Typicality of A is an important factor in Nepali DAM, so initially the possibility was considered that DAM in Chintang could be independent of DAGR and depend on similar factors as in Nepali. This is, however, not the case, as shown by (252). Although mum is the prototypical rice-cooker in a Nepalese family, she is marked by the nominative in (252a) just like the third son in (252b). The reason is, of course, the non-specificity of O.

[22] One can of course argue in this case that there is no single referent but several referents that overlap only partially. If reference is taken in this strict sense, the detransitivised frame indeed never occurs with highly topical referents, because whenever a referent is being tracked and the criteria for identifying instances of it with each other are clear it will always get O-AGR. But this is not because the referent is highly topical but because prolonged tracking of a strict referent presupposes specificity.

(252) a. *A-mma kok thuk-no.*
1sPOR-mother rice cook-IND.NPST[.3sS]
'Mum cooks rice.'
b. *Maila kok thuk-no.*
third.son rice cook-IND.NPST[.3sS]
'Maila cooks rice.' (elicitation PRAR 2010)

2.6.6.9 Volitionality

Volitionality is another factor that is rather connected to A than to O marking. Since S/A detransitivisation includes DAM, this factor was also tested. (253) and (254) each show one volitional and one non-volitional action. The sentences in (253) are transitive, those in (254) are detransitivised, so the two factors can be freely combined:

(253) a. *Asinda Ram-e-ŋa akka heŋd-u-ŋs-u-ŋ=go*
yesterday Ram-NAME.NTVZ-ERG 1s make-3[s]O-PRF-3[s]O-[SUBJ.]1sA=NMLZ$_1$
arkha jamma thu-o-ŋs-e.
alcohol all drink-3[s]O-PRF-IND.PST[.3sA]
'Yesterday Ram drunk all the alcohol I had made.'
b. *Kina arkha-ŋa sed-e kina c-o=go jamma*
SEQ alcohol-ERG kill-IND.PST[.3sS/A] SEQ eat-[SUBJ.3sA.]3[s]O=NMLZ$_2$ all
pes-o-ŋs-e.
throw.up-3[s]O-PRF-IND.PST[.3sA]
'Then he got drunk and threw up all that he had had.' (elicitation SAR 2011)

(254) a. *Asinda Ram-e sapphi arkha thu-a-ŋs-e.*
yesterday Ram-NAME.NTVZ a.lot alcohol drink-PST-PRF-IND.PST[.3sS]
'Yesterday Ram drank a lot of alcohol.'
b. *Kina arkha-ŋa sed-e kina guwakguwak*
SEQ alcohol-ERG kill-IND.PST[.3s>3s] abundantly throw.up-PST-PRF-IND.PST[.3sS]
pes-a-ŋs-e.

'Then he got drunk and he threw up all over the place.' (elicitation SAR 2011)

2.6.6.10 S/O detransitivisation

Verbs which are known to participate in S/O detransitivisation are not more or less prone to S/A detransitivisation. For example, *hutt-* 'burn' can do S/A detransitivisation under exactly the same conditions as any other transitive verb:

(255) a. *Ana-chimeki-ce-ŋa kailekaile phohor-a u-hutt-o-ko.*
1pePOR-neighbour-ns-ERG sometimes garbage-NTVZ 3[p]A-burn-3[s]O-IND.NPST
'Sometimes our neighbours burn (a certain amount of) garbage.'
b. *Ana-chimeki-ce kailekaile phohor-a u-huʔ-no.*
1pePOR-neighbour-ns sometimes garbage-NTVZ 3[p]S-burn-IND.NPST
'Sometimes our neighbours burn garbage.' (elicitation RBK 2010)

(256) shows an S/O detransitivised example from the corpus for comparison:

(256) *Asindaʔ u-taŋ hutt-ad-a-ŋs-e.*
yesterday 3sPOR-head burn-COMPL.ITR-PST-PRF-IND.PST[.3sS]
'Yesterday his head got burnt.' (CLC:CLLDCh1R06S02.0318)

2.7 Quantitative analysis based on corpus data

2.7.1 Introduction

For the quantitative analysis, parts of the Chintang Language Corpus (see section 0.4) were annotated for the central variable of quantifiability and various related information. The annotation was done by a German student of linguistics, who was writing her Master's thesis at the relevant time, and by myself. The guidelines the annotation was based on will be summarised below. The full guidelines can be found in the appendix (Appendix A). Altogether 6606 sentences containing 28,345 words were annotated. The annotations were checked for consistency and extracted from the corpus using Perl scripts (attached in the appendix, sections C.2, C.1). The final analysis was done with R (R Development Core Team 2012) based on a CSV file output by the Perl scripts (see appendix C.3).

Counting proportions of frames in all sentences is not completely trivial. First, there is the question as to whether all possible frames should be considered (including frames resulting from S/O detransitivisation, reflexivisation, causativisation and the like) or whether it is legitimate to compare only the transitive frame and the S/A detransitivised frame. There are two answers to this, both pointing into the same direction. First, the two frames of interest together make up about 84% of all annotated frames, so the remaining frames are negligible in terms of numbers. Second, the factors conditioning the other alternations are completely different from those conditioning S/A detransitivisation, so the subset "transitive and S/A detransitivised frames" is meaningful.

Another problem is the large number of cases where morphosyntax is ambiguous with respect to S/A detransitivisation. The central marker *-u* [3O] is dropped before vocalic suffixes such as *-a* [IMP] and *-e* [IND.PST], so surface verb forms are often indeterminate as to their transitivity. When in addition A is covert as usual, it becomes completely impossible to tell the two frames apart. Such cases make up about 34% of all clauses with lexically transitive verbs or about 37% of all clauses that are either transitive or S/A detransitivised. The indeterminate clauses have to be ignored because their occurrence is due to factors other than those relevant to S/A detransitivisation, viz. simple morphological coincidence.

This leaves us with 1368 observations of the transitive and 544 observations of the S/A detransitivised frame. The transitive frame thus covers about 72% of all relevant forms and is clearly the default choice. Note that this does not contradict the claim made in section 2.4.1 that neither of the two frames is derived from the other. As we have seen in section 2.6.3.1, the proportion of the transitive frame varies greatly across different types of referents and is sometimes smaller than that of the S/A detransitivised frame.

2.7.2 Syntactic annotation and primary variables

The annotation for Chintang started at an early stage in the analysis when it was already clear that S/A detransitivisation was functionally comparatively simple but other things such as the role of arbitrary reference, the relation between quantifiability and identifiability, and the full range of parameters relevant to identification processes had not been discovered yet. For this reason, quantifiability and identifiability were annotated side by side rather than as facets of one and the same phenomenon. The applied definition of identifiability was somewhat more conservative than the radical view presented in section 2.5, and arbitrary reference was not annotated at all. In addition to quantifiability and identifiability, various syntactic information was annotated that was not of direct interest to this but to other ongoing research projects on Chintang.

A minimal amount of syntactic structure is represented by the variable domain, which marks elements that are syntactically associated by identical numeric IDs. For instance, all constituents (arguments and predicate) of the first sentence in a text get the domain ID `1`, those of the next sentence get `2`, and so forth. Nested structures can be indicated by slashes; for instance, `2/1` is the first clause embedded into sentence 2, and `2/1/3` is recursively embedded into sentence 2 and the third element on level `2/1`. Domains are not directly relevant for S/A detransitivisation but have a couple of indirect uses. First, they make it possible to locate objects within files easily. Second,

they establish a link between an object and its predicate. Third, they make the understanding of annotated texts easier and allow to do some general statistics.

Another basic variable is role. The central roles were S, A, P, T, G with definitions based on Dowty (1991) and Bickel (2011) with some simplifications. In the definition of ditransitives, actual or metaphorical movement was taken as the central criterion distinguishing T and G. For copular clauses, the special labels CT (copular theme) and CR (copular rheme) were used instead of roles. Three other pseudo-roles were N.EXP (the experiencer noun featuring in the experiential frames), CSR (for the causor in causatives), and BEN (for an additional benefactor in benefactives).

The other, more directly relevant variables are shown with their values below. All variables had an additional value x that was to be used in cases of insecurity and that was ignored in the statistical evaluation. More complete definitions can be found in the appended annotation guidelines (Appendix A).

- **verb class** – the lexical class of the verb as defined by its characteristic frame (cf. section 2.3.3):
  - **itr** – intransitive
  - **tr** – monotransitive
  - **dido** – direct object ditransitive
  - **dipo** – primary object ditransitive
  - **dioo** – double object ditransitive
  - **exptr** – transitive experiential
  - **expitr** – intransitive experiential
  - **uninf** – uninflected verboid
  - **aux** – auxiliary
  - **other** – any other minor class
- **alternation** – various syntactic alternations modifying the base frame. Where no alternation was present this variable stayed empty.
  - **sad** – S/A detransitivisation
  - **idt** – indeterminate as to S/A detransitivisation
  - **sod** – S/O detransitivisation
  - **refl** – reflexive
  - **recp** – reciprocal
  - **ambrec** – ambitransitive reciprocal
  - **pass** – passive
  - **caus** – causative
  - **ben** – benefactive
  - **cop** – copulative frame
  - **poss** – possessive frame
  - **dumA** – dummy A-AGR in the transitive experiential frame
  - **OtoS** – S-AGR with embedded O in infinitival constructions
- **quantifiability** – as defined in section 2.6.1:
  - **qnt** – quantifiable
  - **nonq** – non-quantifiable
- **identifiability** – a less sophisticated version of the definition in section 2.5:
  - **def** – identifiable for both speaker and hearer
  - **spec** – identifiable for the speaker only
  - **idf** – identifiable for neither speaker nor hearer

In the initial phase of the annotation, two files were worked through by both annotators independently and Cohen's Kappa was calculated. Cohen's Kappa (Cohen 1960) measures the proportion of interannotator agreement that is not due to chance. Table 2.10 shows the results for the central variables role, quantifiability, and identifiability.

	observed agreement	expected chance agreement	Cohen's Kappa
story_rabbit			
role	86%	17%	0.83
quantifiability	93%	78%	0.67
identifiability	95%	78%	0.79
kamce_talk			
role	77%	21%	0.71
quantifiability	89%	78%	0.49
identifiability	89%	71%	0.61

Table 2.10: Interannotator agreement for Chintang

As is well known in the literature (e.g. Carletta 1996, Sim and Wright 2005), there is no cutoff value for Cohen's Kappa that is meaningful for all applications and thus generally accepted. One of the first proposals for evaluating Kappa is found in Landis and Koch (1977:165), according to which most of the values in Table 2.10 indicate "substantial" agreement (Kappa between 0.61-0.80). The only case where only a "moderate" level of agreement was reached (Kappa between 0.41-0.60) was quantifiability in the session kamce_talk. Note, however, that Cohen's Kappa is not only influenced by inter-annotator agreement. As noted by Sim and Wright (2005:261), the measure penalises high probabilities for chance agreement so that the higher this probability the lower Kappa. Both quantifiability and identifiability have high probabilities for chance agreement between 70 and 80% because some of their values are much more frequent than others (`qnt` 76%, `def` 69%). I therefore accepted the low value in question as sufficient.

2.7.3 The centrality of quantifiability

The results of the annotation confirm the central role of quantifiability for S/A detransitivisation. 95% of all non-quantifiable O referents co-occur with the S/A detransitivised frame, and 97% of all quantifiable O referents co-occur with the transitive frame. Unsurprisingly, a Fisher's exact test on these numbers indicates an extremely high level of significance ($p < 0.01$) for the interaction between the two variables. Figure 2.5 visualises the proportions.

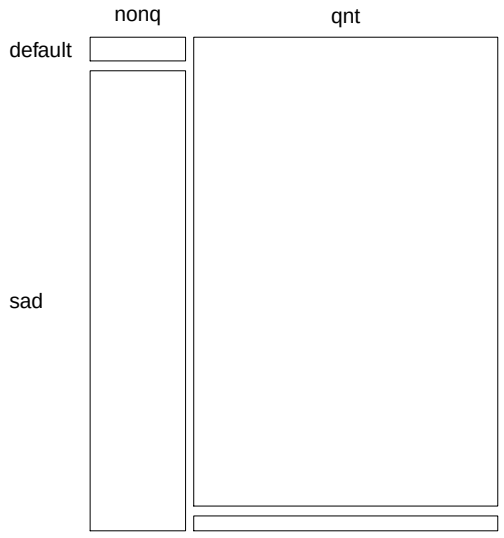

Figure 2.5: Quantifiability and S/A detransitivisation

The strength of association between two categorial variables can be measured by the coefficient

Cramer's V, which ranges between 0 (no association) and 1 (perfect association). Cramer's V for quantifiability and S/A detransitivisation is 0.90. This value is far above anything that is reached for single variables in Nepali DOM (cf. section 3.6.4). S/A detransitivisation can thus be said to be linked much more tightly to a single variable and to be functionally less complex than DOM.

For the other variable under investigation, identifiability, the numbers are less easy to read. While there are clear associations which also do reach significance ($p_{\chi^2} < 0.01$), the numbers fall back behind those for quantifiability: 73% of all indefinite O have S/A detransitivisation, and 78/92% of all specific/definite O have the transitive frame. Figure 2.6 visualises this.

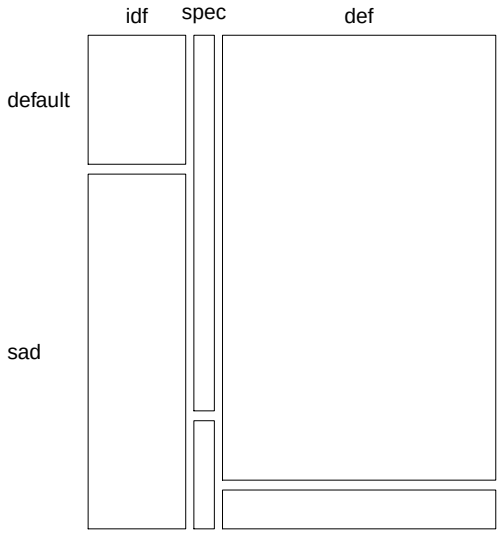

Figure 2.6: Identifiability and S/A detransitivisation

Since definiteness was defined in section 2.5 as entailing specificity, it is possible to fuse the values `spec` and `def` to a category with the meaning 'at least specific'. This category gets the transitive frame in 91% of all cases and therefore still stays slightly behind quantifiability.

Cramer's V for identifiability and S/A detransitivisation is 0.64, which is still high compared to the values for DOM but low compared to the 0.90 reached by quantifiability. Cramer's V for identifiability with `def` and `spec` fused also rounds to 0.64.

These results are unexpected given the discussion in section 2.5 and section 2.6, where quantifiability was viewed as a precondition for specificity, complemented by arbitrary reference. If this truly was the case, the fusion of `def` and `spec` should have produced equally good or better predictive results than just quantifiability. However, as mentioned before, the definition of identifiability used for the annotation reflects an earlier stage of the analysis. Therefore, the aberrant behaviour of identifiability is rather an artifact of the annotation than a reflex of what is really going on. In some cases, mismatches between quantifiability and identifiability defined in a somewhat more conservative fashion point out some interesting differences between various conceptions of identifiability.

There are two kinds of such mismatches. One are cases where one referent was annotated as non-quantifiable but also as definite or specific. The majority of these cases can be traced back to the role of subamounts for quantifiability. As discussed in section 2.6.3.1, it is important whether a whole referent is affected or just some non-quantifiable subamount. While this distinction should in principle also be applied to identifiability, this would be somewhat less intuitive and was not done in the annotation. Consider, for instance, the example in (257):

(257) *Pache ciya kha-pid-e* *kinana ciya thu-i-hẽ.*
then tea 1nsO-give-IND.PST[.3sA] SEQ tea drink-1p[i]S-IND.PST
'Then she gave us tea and we drank the tea.' (CLC:Lok_yatra.189-190)

From the perspective of quantifiability it's immediately clear that not all the tea is affected at once in the process of drinking but only a subamount, so the referent behind *ciya* was considered as non-quantifiable. For identifiability one could also have said that the precise affected amount is indefinite, but this seemed a bit awkward given the fact that all of the tea, which has just been mentioned, is much more relevant as a referent than the affected subamount. Therefore, referents like *ciya* in last were usually tagged as `def`.

Similarly, the other kind of mismatch – quantifiable indefinite referents – is to the biggest part due to examples like those shown in (258) and (259). Here, the O referents are clearly quantifiable because they are single referents. However, it's not so clear whether they are also specific or definite. They are if one takes into account that sometimes very little information may be sufficient to identify a referent within a certain mental space, but in a more conservative view they are not:

(258) *Sel-a ekthopa ta-ma u-pi-c-o-niŋ.*
jackal-NTVZ at.all come-INF 3A-allow-d-3[s]O-NEG
'They wouldn't allow a jackal to come near at all.' (CLC:ctn_talk01.153)

(259) *Ani-yiŋ=le u-nis-o-ko, aru u-nis-o-ko-niŋ.*
1piPOR-language=RESTR 3pA-know-3[s]O-IND.NPST other 3pA-know-3[s]O-IND.NPST-NEG
'They only know our language, they don't know (any) other.' (CLC:Durga_Exp.55-56)

To summarise, quantifiability is of central importance to S/A detransitivisation. Identifiability would have been expected to yield the same predictive results under the strict definition presented in section 2.5, but since its definition for the annotation reflects an earlier stage of the analysis, the results of the annotation are not as relevant for the discussion of the function of S/A detransitivisation as they are for the meta-question which definition of identifiability works best.

2.7.4 The role of exceptions

As stated above, quantifiability is the central variable for S/A detransitivisation. However, there are some cases which it does not explain. If we look at the counts from the perspective of prediction, quantifiability correctly predicts 97% of all frames (within the binary choice we are looking at here), or 98% of all transitive frames and 91% of all S/A detransitivised frames.

In a way this is what is expected. In section 2.6.2 it was stressed that quantifiability can be overridden by open (unknown or arbitrary) reference. In section 2.6.5 several cases were discussed where S/A detransitivisation is conventional. An inspection of the 38 annotated cases where a quantifiable O referent was used with S/A detransitivisation shows that almost all of them fall into one of the categories discussed in the two mentioned sections. 18 (47%) have arbitrary reference (partially with grammaticalisation in the case of frequent composite activities), 13 (34%) have pieces of information in O, and another 5 (13%) fall into other minor categories. The only two cases that cannot be explained at all (5%, 0.005% of all instances of the S/A detransitivised frame) are shown in (260) and (261).

(260) *Hun-ce u-nis-a=kha raicha!*
MED-ns 3pS-know-PST=NMLZ$_2$ MIR
'They had known it!' (CLC:phidang_talk.447)

(261) *Pa u-chau pokt-e kina huŋ=go khad-a-kt-e=ta.*
father 3sPOR-child leave-IND.PST[.3sS] SEQ MED=NMLZ$_1$ go-PST-IPFV$_1$-IND.PST[.3sS]=FOC
'After the father$_i$ had left his child$_j$, he$_j$ himself was going away.' (CLF:sadstory_RM.122)

The cases where a non-quantifiable referent was used with the transitive frame were fewer (19). The subtype that is easiest to understand here is one where the speaker chose to represent a non-quantifiable referent by a non-singular NP (where S/A detransitivisation is extremely rare, as discussed in section 2.6.3.1) or by a non-singular agreement prefix (*kha-* [1nsO], *mai-* [1nsiO]). Two examples are shown below.

(262) *Warisa-ce cahĩ mai-mek-no=mo kina na hun-ce dukkha pi-ma-ce*
young.woman-ns RETRV 1nsiO-like-IND.NPST=CIT SEQ CTOP MED-ns trouble give-INF-3nsO
maha?=kha gonei.
be.not.good=NMLZ$_2$ ATTN
'When we think that the girls like us we shouldn't give trouble to them.'
(CLC:khinci_talk.081-082)

(263) *Koikoi lik-ma kha-pi-ni?-niŋ=pho.*
some enter-INF 1nsO-allow-IND.NPST-NEG=REP
'(I heard that) they don't allow some of us to enter.' (CLC:chintang_sahid.263)

This subtype constitutes 7 (37%) of the cases in question. Another, less transparent subtype seems to occur when O is filled by a dummy referent corresponding to the general situation. This type covers another 6 cases (32%). (264) is an example.

(264) *Anaŋ mett-u-m, huŋ-khi=ta nahaŋ!*
what do.with-3[s]O-[SUBJ.NPST.]1pA MED-MOD=FOC but
'What shall we do (with this whole thing), that's just the way it is!' (CLC:tang_talk.146)

Finally, in 5 cases (26%) the speaker seems to have picked out an exemplar in order to refer to a category:

(265) *Masala kiya mai-ta-yokt-a-kt-e ba=go teĩ-be, abo na*
spice oil NEG-come-NEG-PST-IPFV$_1$-IND.PST[.3sS] PROX=NMLZ$_1$ village-LOC$_1$ now CTOP
gududu-wa u-tad-u-l-o-ŋs-e!
steady-ADVZ 3pA-bring-3[s]O-back-3[s]O-PRF-IND.PST
'Spices and oil were not coming to this village, but now they bring the stuff non-stop!'
(CLC:khim_ring.106-107)

There is only a single case (5%, or 0.0008% of all instances of the transitive frame) where I do not have the slightest idea what could have conditioned the frame. This sentence is shown in (266).

(266) *Tara hana=yaŋ sapphi a-numd-o-ko onei.*
but 2s=ADD much 2[s]A-do-3[s]O-IND.NPST ATTN
'But you also do a lot.' (CLC:tang_talk.218)

To summarise, the majority of exceptions can be attributed to arbitrary reference, grammaticalisation, and varying construals. Thus, S/A detransitivisation is a phenomenon with a rigid core conditioned by quantifiability and with flexible fringes. It would be an interesting question to ask whether there was more or less flexibility in earlier times, but since there are no written records prior to the arrival of CPDP, this is impossible to answer.

2.8 Summary

The preceding section has discussed the formal and functional properties of S/A detransitivisation in Chintang in all detail. It is now time to sum up.

S/A detransitivisation was defined as a kind of differential framing. This term is analogous to differential marking and differential indexing and can in principle cover all grammatical mechanisms where roles can be coded in different ways and where several morphosyntactic features are tied together. In Chintang these are A case, A-AGR, and O-AGR. S/A detransitivisation forms an non-directional link between an abstract transitive frame of the form {A-ERG O-NOM V-a(A).o(O)} and a detransitivised frame {A-NOM O-NOM V-s(A)}. Constructions that make direct reference to the transitivity of a frame treat the detransitivised frame as a hybrid, i.e. they can react to its intransitive or to its transitive characteristics. S/A detransitivisation resembles various other phenomena but is different from all of them: differential case and agreement are isolated phenomena, S/A ambitransitivity is lexically restricted, O is downgraded in antipassives, and O loses its status

as argument and NP in noun incorporation.

The grammatical relation O (defined as a NOM-marked argument linked to O-AGR in the transitive frame) is central for S/A detransitivisation since it is O whose referential properties determine the frame: specific O require the transitive frame, non-specific O the detransitivised frame. With regard to O-AGR, the most prominent manifestation of S/A detransitivisation in the majority of cases, one could thus say that O that are trackable get tracked and O that are not don't. This rule does not only cover simple clauses but also various more complex environments: non-quantifiable O can neither be raised to O-AGR nor to S-AGR, and they cannot be indexed on infinitives or via nominal possessive prefixes on the purposive.

The function of S/A detransitivisation can be determined as specificity. Specificity was described on a par with definiteness as being based on identifiability, and quantifiability was shown to be a central prerequisite for this. The use of quantifiability in description highlights some aspects of identifiability that are not often taken into account otherwise, for instance, the role of the count/mass distinction, of pluralisation, and of partial affectedbess. Quantifiability is in turn connected to various factors such as overlookability, overt quantification, and exhaustive reference. All kinds of nouns can be easily construed as quantifiable or non-quantifiable in Chintang, and there are no signs of a lexicalised count/mass distinction. On the other hand, not all quantifiable referents are also identifiable. The concept of arbitrary (= open or discardable) reference was introduced to account for those cases where quantifiability is of no use.

In a couple of areas S/A detransitivisation is more or less conventional. This applies to verbs where it is not clear whether a semantic object should still be assumed or not, to composite activities and complex predicates, where O tends to merge with the predicate, to propositional verbs that take a piece of information as their O, and to adverbials "replacing" O. Whilst S/A detransitivisation is clearly motivated in all of these constructions, the extent to which detransitivisation can be predicted from their presence varies.

The last section discussed S/A detransitivisation from a quantitative perspective. Corpus annotations confirm the central role of quantifiability, which correctly predicts 97% of all transitive or S/A detransitivised frames. Almost all exceptions can be explained with the help of arbitrary reference or via conventionalisation.

2.9 S/A detransitivisation in other Kiranti languages

2.9.1 Overview

S/A detransitivisation in exactly the same form as in Chintang is also found in a number of other Kiranti languages. The languages for which syntactically informed reference grammars are available are surveyed in this section.

One interesting generalisation is that S/A detransitivisation appears in a large coherent region in the southeast corner of the Kiranti area. This region is constituted by the languages (from east to west) Limbu, Yakkha, Athpare, Belhare, Chiling, Chintang, Bantawa, and Puma, which are discussed in detail further down. This statement is relativised, though, by the problem that one tradition of writing Kiranti grammars has it that these languages have such a complex morphology that their syntax is negligible. This is obviously wrong, as evidenced by the rich literature that is now available on the syntax of, for instance, Chintang, Belhare, and Puma (see e.g. Bickel 2004b, Bickel 2004a, 2007b, 2007a, 2010, Paudyāl et al. 2010, Gaenszle 2011, Stoll and Bickel in press, 2012, Schikowski et al. forthcoming).

Still, many Kiranti grammars do barely or not at all provide information on syntax, so just because S/A detransitivisation is not mentioned for a language that does not necessarily mean that it does not exist there. One striking example for this is Limbu. S/A detransitivisation has been clearly described for this language as "middle conjugation" by Weidert and Subba (1985) but is nevertheless completely ignored by Driem (1987).

Languages whose status is insecure due to this problem are (again from east to west) Yamphu, Kulung, Dumi, Wambule, Jero, and Sunwar. These are briefly discussed below.

Rutgers (1998) in his Yamphu grammar does not mention anything like S/A detransitivisation. However, the appended dictionary contains rudimentary information about syntactic verb classes in the form of the labels "v.intr." and "v.tr.". Interestingly, the majority of verbs which are labelled as transitive at the same time carry the label "v.intr.". Verbs with both labels seem to be either S/A or S/O ambitransitive. The two classes can be distinguished by the placement of labels. Only for S/O ambitransitives are the labels repeated before each sense (apparently following the intuition that intransitive 'break' and transitive 'break' (with S=O) are semantically more different than intransitive and transitive 'cut' (with S=A)). Compared to the ubiquitous S/A ambitransitives, S/O ambitransitives seem to be somewhat rarer. Among the S/A ambitransitives are not only classical cases like *uŋma* 'drink' or *cama* 'eat' but also less typical verbs like *aŋma* 'chop up', *hamma* 'ration' or *khupma* 'scratch'. I therefore consider Yamphu as a likely candidate for another language having S/A detransitivisation, which would extend the S/A detransitivisation area further to the north.

For Kulung, Tolsma (2006) does also not talk about syntactic verb classes or valency. Differently from Rutger's Yamphu grammar, his dictionary appendix also does not contain ambitransitives. On p.137 there is one example that looks as if the verb was detransitivised (glosses adapted here and in the following citations from grammars; translation unchanged):

(267) *C^{h}am-ci so ləi-ya-ke.*
song-ns also sing-1p[iS].NPST-ASS
'Let's also sing songs.' (Tolsma 2006:137)

However, *c^{h}am ləima* 'sing a song' might have special properties as in Chintang where it belongs to the class of conventional V/N combinations (section 2.6.5.2). In fact, several other examples with apparently non-specific objects have bipersonal agreement:

(268) *Samk^{h}e so let-a-m.*
potato also plant-3O-[NPST.]1p[i]A
'We plant potatoes.' (Tolsma 2006:151)

All in all the evidence thus is against S/A detransitivisation.

Van Driem states in his grammar of Dumi (Driem 1993:228) somewhat enigmatically that "the concept of transitivity is versatile" but then discusses only a few verbs with diverse syntactic properties none of which comes close to S/A detransitivisation. His chapter on transitivity is long but unsystematic, and since he also overlooked S/A detransitivisation in Limbu (Driem 1987) it is probably a bit early to conclude that Dumi does not have S/A detransitivisation.

By contrast, it seems relatively safe to say that Wambule does not have S/A detransitivisation. Opgenort (2004:151) mentions differential agent marking conditioned by the markedness of transitive scenarios and thus seems to be aware of differential marking in general. In his discussion of verb classes he mentions what he calls "middle verbs" (p.250), but these have the frame {A-ERG P-NOM V-s(A)}. They are also functionally quite different from S/A detransitivisation as it is found in Chintang and other eastern Kiranti languages in that they mark reflexivity.

Another completely unclear candidate is Jero. Allen (1975:42) mentions that a few verbs can be used both transitively and intransitively, but only one of his examples is clearly S/A ambitransitive (*hut-* 'fly' or 'fly to somebody'). Opgenort (2005) also does not provide more detailed information.

Finally, Sunwar as described by Borchers (2008) is very unlikely to have S/A detransitivisation. Sunwar occupies a special position among the Kiranti languages because the older bipersonal agreement has broken down and made way for monopersonal agreement with S or A. Borchers mentions that a few verbs such as *cīcā* 'wash, bathe' and *mecā* 'vomit' can be used with both intransitive and transitive inflection (that is, either with the verbal affixes normally indexing S or with those normally indexing A) and says that the use depends on the presence of an object. However, she also emphasises that these verbs are few compared to the rest, which are always used with either intransitive or transitive inflection, so Sunwar seems to have lexicalised S/A ambitransitivity but nothing more. Genetti (1988) describes an earlier stage of the language where bipersonal agreement was still in use but focusses on morphology and morphophonology and does not give any information about alternations.

Apart from these problematic languages, there are a few more western languages which are very unlikely to have S/A detransitivisation. Nothing in the direction is mentioned for Camling by Ebert (1997a), who did recognise and describe S/A detransitivisation for Athpare in Ebert (1997b). Thulung and Koyu are described in Lahaussois (2002) and Lahaussois (2009). Thulung has split DAM depending on lexical factors (nouns and 3s/3d/3p/2p pronouns vs other pronouns), whereas Koyu has a fluid DAM system where the ergative is apparently used to mark the agent in scenarios where otherwise O could thought to be A. Both languages also have a borrowed case marker *-lai* (< Nep. *-lai* [DAT]) which they use for DOM in a similar fashion as Nepali (however, without any effects on O-AGR). These independently existing patterns make it unlikely that S/A detransitivisation should exist in these languages.

For the remaining Kiranti languages either no data are available at all or only articles with a focus on a topic other than syntax exist. These are Bahing, Chiling, Dungmali, Hayu, Khaling, Khambu, Lohorung, Mewahang, Mugali, Nachiring, Sam, Sampang, and Tilung.

2.9.2 Limbu

S/A detransitivisation in Limbu is described early by Weidert and Subba (1985) and later in a dedicated article by Angdembe (1998). It is ignored by Driem (1987).

Weidert takes a somewhat eccentric view on the phenomenon. He speaks of an "anti-passive transformation", which fulfills the formal definitional criteria for S/A detransitivisation, and says that "presumably most" verbs are open to this transformation. Below are examples given by him.

(269) a. *Am-baˑ-rɛ pit-nuˑ-n thuŋ-u-rɔ yak.*
1sPOR-father-ERG cow-milk-DEF drink-[3sA.]3[s]O-CONJ.PTCP stay[.3sS]
'My father is drinking milk.'

b. *Am-ba pit-nu thuŋ-lɔ yak.*
1sPOR-father cow-milk drink[.3sS]-CONJ.PTCP stay[.3sS]
'My father drinks milk.' (Weidert and Subba 1985:108)

(270) a. *Aŋgaʔ sɔksɔkk-in niːr-u-ŋ-lɔ yakk-aʔ.*
1s book-DEF read-3[s]O-1sA-CONJ.PTCP stay-NPST.1sS
'I am reading the/a book.'

b. *Aŋgaʔ sɔksɔk niːt-aʔ-rɔ yakk-aʔ.*
1s book read-NPST.1sS-CONJ.PTCP stay-NPST.1sS
'I read books; I am a reader of books.' (Weidert and Subba 1985:109)

What is special about Weidert's view is that he does not consider S/A detransitivisation in isolation. Instead, he views all predicate frames containing a nominative argument and a verb with monopersonal agreement as related. This does not only include S/A detransitivised transitive verbs but also normal intransitive verbs and reflexives. He calls the bipersonal paradigm the "active" and the monopersonal paradigm the "middle conjugation" and accordingly analyses S/A detransitivisation as active verbs in middle conjugation. He presents Table 2.11 to summarise the contrasts between the two conjugation types in various dimensions.
A few comments are in place here to explain Weidert's terms:

- "Directionality" refers to the existence of a "cause-effect relationship" and is similar to Hopper and Thompson's (1980) concept of affectedness. "Corporeal" is supposed to mean that in the active conjugation this relationship becomes visible in the form of a (often physical) effect of the action initiated by the agent on the goal (i.e. the patient). In the middle conjugation there are "no determinable causal consequences" for the goal.
- The terms "centrifugal" and "centripetal" are not further explained. They are introduced in the discussion of "directionality" and seem to signify whether the action performed by the agent (Weidert's cover term for S and A) is directed towards a goal outside its origin (centrifugal) or not (centripetal).
- "Possessivity" refers to whether the goal is inalienably possessed by the agent or not. This criterion is not meant to be relevant for all cases.

	active/'transitive verbs [sic]	**middle/'intransitive' verbs**
directionality	corporeal	Ø (for one-argument predicates) diffuse (in anti-passive construction type)
volitionality	strong	reduced
attentional focus	ergative agent	absolutive agent
attention vector	centrifugal	centripetal
actor animacy	animate, preferentially human	unconstrained
ergativity	necessary	Ø
possessivity	alienable	inalienable

Table 2.11: Limbu active and middle verbs (Weidert and Subba 1985:122)

Treating intransitive verbs and detransitivised transitive verbs as one category as Weidert does is problematic. If S/A detransitivisation in Limbu is indeed possible for most verbs, it is likely that is not lexically conditioned, as in Chintang. Being an intransitive verb, by contrast, is a lexically fixed property. Weidert thus tries to bring together two constructions with very different degrees of freedom. That this does not work very well can be seen in Table 2.11. Most of the dimensions of contrast listed are not confirmed by convincing elicited examples, and none are corroborated by corpus data. Some of the concepts such as directionality, attentional focus, and attentional vector are ill-defined, and for all but the formal criterion of ergativity, counterexamples can be easily found. For instance, in (269) it's not clear why there should be a difference in directionality, volitionality, attentional focus, or attention vector between the two examples, nor why human A should be preferred in the first sentence of each pair but not in the second.

Apart from his attempt to give a generalised characterisation of all monopersonal predicate frames, Weidert also talks about the Limbu antipassive as a construction in its own right in some places. He mentions the known formal features and also that the antipassive cannot be used with personal pronouns, demonstrative pronouns, numeral expressions above 1, and the plural suffix *-haʔ* on the goal noun (Weidert and Subba 1985:108). As concerns function, Weidert says that the antipassive removes the argument position for the goal (P) and that as a consequence the noun becomes a part of the verb, that is, incorporated (although Weidert does not use the word). This is meant both as a formal and a functional property. Weidert does, however, not expand on this; for him, the main functional characteristic of the antipassive is its being "middle".

One interesting difference between the Limbu antipassive and Chintang S/A detransitivisation is that the antipassive is unproblematic with possessed objects. Weidert even claims that the antipassive is obligatory with objects which are inalienably possessed by A with two verbs. One is shown in (271):

(271) a. *Aŋgaʔ a-bik kɔmm-aʔ.*
1s 1sPOR-cow herd-NPST.1sS
'I'll herd my cow.'
b. *Aŋgaʔ ku-bitt-in kɔm-u-ŋ.*
1s 3sPOR-cow-DEF herd-3[s]O-[NPST.]1sA
'I'll herd his cow.' (Weidert and Subba 1985:117)

Weidert's work is hard to appreciate because of his special terminology and eccentric ideas. Still, he presents a solid analysis of the formal properties of the construction to which the dedicated paper by Angdembe (1998), does not add much. What's especially interesting is that even the central claim of Angdembe's paper that the phenomenon has to be analysed as noun incorporation has in principle been anticipated by Weidert. Although he does not use the term, he compares antipassivised objects to the nominal component of lexicalised noun-verb combinations that can no longer be used separately from the verb and form a morphological unity with it.

Angdembe mentions a few more formal characteristics *en passant* but does not present data: the object cannot be dropped, it cannot be modified by adjectives, numerals (without Weidert's

restriction "greater than 1"), and the definite article, and it cannot be a proper noun or (contradicting Weidert) an inalienably possessed noun. He also touches on ditransitive verbs (where G triggers O-AGR) and says that the verb is only detransitivised when both direct and indirect object (i.e., T and G) are "incorporated" but does not present examples for the crucial case of G being incorporated and T not.

Angdembe explicitly rejects an analysis of the Limbu antipassive in terms of definiteness. His argument, however, is not very convincing: "the situation in (3) is the same as the situation in (2)" (where (3) and (2) are the default and detransitivised versions of a sentence translated as 'The friend killed the buffalo'; Angdembe 1998:21). If the "situation" was indeed the same in both clauses any functional explanation including noun incorporation would have to fail.

2.9.3 Yakkha

Schackow (In preparation) mentions S/A detransitivisation as "detransitivization" in her Yakkha grammar and says that "any verb in Yakkha can basically be inflected intransitively" (Schackow In preparation:61). She does not talk explicitly about the case of A and presents examples with dropped A:

(272) a. *Cog-uks-u=na.*
do[PST.3sA]-TEL-3O=NMLZs
'He did it.'
b. *ekdam cog-a-nuŋ cog-a-nuŋ*
very do[3sS]-PST-while do[3sS]-PST-while
'while he worked hard/while he did a lot' (Schackow In preparation:61)

She goes on to say that the default frame is used when the object is "definite or specific", whereas detransitivisation is used when it is "unspecific or generic" or when "the matter is rather about the structure and manner of the event".

2.9.4 Athpare

S/A detransitivisation in Athpare is described by Ebert (1997b) as "undergoer demotion". Ebert does not mention the connection of this phenomenon to case marking explicitly but shows with an example that the pattern is just as expected, A being zero-marked in the detransitivised frame:

(273) a. *Un-na laribo choŋs-u-na.*
he-ERG banana sell-[3sA.]3[s]O-NMLZ
(no translation provided)
b. *Un laribo choŋ-na.*
he banana sell[.3sS]-NMLZ
'He sells bananas.' (Ebert 1997b:122)

Ebert only touches on the function of undergoer demotion. She mentions that it is used when "the undergoer noun does not denote a specific entity" and that the noun is "quasi-incorporated" (p. 122), all of which sounds very similar to Chintang.

As a special case she mentions "inherent objects" without defining the term but listing the combinations 'speak a language', 'sing a song', and 'cook food' as examples. What is intuitively inherent about these objects is that they correspond more or less to the type of referent required by the verb. For instance, 'song' comes close to covering all possible objects of 'sing'. There is some variation, though: 'language' is obviously an important but not the only type of object licensed by 'speak', and food is not the only type of thing that can be cooked. Such cases correspond to what has been called composite activities here (see section 2.6.5.2).

2.9.5 Belhare

Bickel uses the terms "object downgrading" (Bickel 2003a) and "detransitivisation" (Bickel 2004a, Bickel et al. 2010) for S/A detransitivisation in Belhare. He adds to the defining criteria that the concerned object cannot be pluralised, possessed, or specified by a demonstrative or any other attribute, and that it cannot be moved to the right of the verb. Detransitivisation is also impossible with "inherently specific" objects (Bickel 2004a:167).[23] He presents various arguments against a formal analysis as noun incorporation: even though the object cannot stand to the right of the verb, other elements can intervene between the two, for instance, focussed agents. The object can be dropped if "the context is clear enough" (Bickel 2004a:169), and it is accessible to information structuring processes such as topicalisation, focalisation, and questioning. He concludes that detransitivisation leaves the argument status and role of the object untouched. Here are his examples:

(274) a. *(I-na) wa khuʔ-yu.*
DIST-DEM chicken steal-NPST[.3sS]
'This [guy] steals chicken.'
b. *(I-na-ŋa) wa khuiʔ-t-u.*
DIST-DEM-ERG chicken steal-NPST-[3sA.]3[s]O
'This [guy] will steal a/the chicken.' (Bickel 2003a:557)

The function of the construction is described in familiar terms: "The nominal does not refer to a specific referent but to a *kind* of referent" (Bickel 2004b:167, emphasis by Bickel). This points to the token : type distinction being important for Belhare. However, Bickel does not go deeper into this and also mentions less specifically in his other paper that detransitivisation "partially fulfills an antipassive function" (Bickel 2003a:556). In Bickel et al. (2010:388) it is claimed that detransitivisation marks non-specificity and may imply the notion of a "general activity".

An interesting detail has to do with the Belhare perfect. There is a perfect marker with suppletive allomorphs, *-sa* after intransitive stems and *-ŋa* after transitive stems. In detransitivised frames the transitive allomorph is used. This is different from Chintang, where both variants of the marker *-hat(t)* [AWAY], which has similar formal behaviour, are allowed with S/A detransitivisation.

2.9.6 Chiling

Chiling (Chiliŋ) has so far not been linguistically documented at all. However, since it is spoken in Ākhisallā, a VDC neighbouring Chintang, I had the chance to do some preliminary elicitation work in 2012 and 2013. The data below clearly show that Chiling is part of the Eastern Kiranti area where S/A detransitivisation is common. The intransitive verb in (275a) carries the same agreement marker as the transitive verb in (275b). (275c) shows a contrasting transitive form of the same verb. The motivation for S/A detransitivisation seems to be the count/mass distinction (cf. section 2.6.3.1 on the same factor in Chintang).

(275) a. *Mu-bak yuŋ-yu-wa.*
DEM.DOWN-LOC be.there-IND.NPST-1sS
'I'm down.'
b. *Cama ca-yu-wa.*
rice eat-IND.NPST-1sS
'I eat rice.'
c. *Sontorok cay-u-ku-ŋ.*
orange eat-3[s]O-IND.NPST-1sA
'I eat an orange.' (elicitation RKU 2013)

How similar this pattern is to S/A detransitivisation in the neighbouring languages can only be shown by further research.

[23] Note, however, that his example for this is the possessed noun *ucha* 'his child', so it's not clear whether this is really an independent factor.

2.9.7 Bantawa

The first grammar of Bantawa is Rāī (1984), who does, however, not talk about S/A detransitivisation. The relevant reference work is therefore Doornenbal (2009). Similarly to Weidert and Subba (1985), Doornenbal views intransitive and transitive inflection as conjugation classes and says that "many" verbs can be inflected both transitively and intransitively. He is apparently not aware of the ambiguity of this statement with respect to S/A and S/O detransitivisation but makes it clear in the following that he is talking about S/A detransitivisation by using the term "antipassive".

He distinguishes two antipassives, an "implicit" (i.e. unmarked) and an "explicit" one that makes use of the "dummy object marker" *kha*. The explicit antipassive is syntactically different from S/A detransitivisation in that it requires ERG on A. This particle is likely to be cognate to Chintang *kha-*, which codes [1nsO] in the dialect of Sambugaũ and is used as a detransitiviser on a handful of verb roots in both dialects (e.g. *copt-* 'look at' vs *khacopt-* 'look around, stare'). There is also a corresponding prefix in Puma (see below).

The implicit antipassive, on the other hand, fulfills the formal criteria for S/A detransitivisation. Here are Doornenbal's examples:

(276) a. *Ŋa laʔ-u-ŋ.*
fish catch-3[s]O-1sA
'I caught a fish (the fish).'

b. *Ŋa laʔ-a-ci-ʔa.*
fish catch-PST-[1]d-e[S]
'We (dual, excl) went fishing.' (Doornenbal 2009:223)

Concerning the function of the implicit antipassive, Doornenbal first says that the antipassive is the default for "verbs where the object is less specific or less obviously affected" (p. 223). After that he gives a couple of other conditions that are very different from what is found in Chintang: the antipassive is preferred where "time reference is less relevant".[24] For transitive inflection it is also important that the action be completed. The antipassive may even interact with phase semantics, as in *kɨkt-a* [hold-PST[.3sS]] 'he held it (for a long time)' vs *kɨkt-u* [hold-[3sA.]3[s]O] 'he grabbed it' (p. 224).

It is not only the function that leaves room for investigation here. Doornenbal also makes some confusing statements about A. He says that A is most naturally omitted in antipassives and that if A is overt A-ERG is "doubtful" (p. 224). This is in contrast to his own earlier claim (p. 222) that ERG is in fact necessary on antipassivised A. His examples do not make the situation any clearer: most of them have A-NOM, but there is indeed one with A-ERG.

Another formal property that distinguishes the Bantawa antipassive from S/A detransitivisation in Chintang is that it cannot be used with all verbs. Doornenbal mentions that it is impossible for "more transitive" verbs like 'kill', 'take', or 'kick' (p. 225).

2.9.8 Puma

Puma is another language with a dedicated paper on detransitivisation (Bickel et al. 2007b). Like Bantawa it has two S/A detransitivising constructions, one of which (the "ø-detransitive") corresponds to Chintang S/A detransitivisation and Doornenbal's (2009) "implicit antipassive". The other corresponds to Doornenbal's (2009) "explicit antipassive" and is marked by the cognate prefix *kha-*. Differently from Bantawa this latter construction has the same syntactic consequences as the ø-detransitive, including the marking of A by NOM, and requires a human object referent. The functional analysis presented in Bickel et al. (2007b) is the most sophisticated available description of S/A detransitivisation in another Kiranti language.

First, here are two representative examples for the construction in question:

[24] This contradicts his own earlier statement that "if one has been peeling already, it is acceptable to say 'I have peeled' in an intransitive form".

(277) a. *Doromen lam-u-ŋ.*
something search-3[s]O-1sA
'I looked for something.'
b. *Doromen lam-oŋ.*
something search-PST.1sS
'I looked for stuff.' (Bickel et al. 2007b:6)

The function of the ø-detransitive is, according to Bickel et al., to delete "any entailment to the cardinality of referents", so that it is "generally used for non-denumerable or generic reference" (p. 11). The unusual term "cardinality" is not defined but seems to be borrowed from mathematics, where it designates the number of elements in a set. Bickel et al. use the term in a slightly different way, though, because for them referents do not *have* a cardinality but *are* cardinal (if they have a countable number) or non-cardinal.

This paper also seems to be the only work on S/A detransitivisation apart from the present one that is so consequent as to assign a function not only to the detransitivised frame but also to the default frame. This function depends on the number indicated by O-AGR: singular indicates that there is at most one referent, dual that there are at most two referents,[25] and plural that there are more than two referents.

There are some problems with these functions. For instance, the meaning of a simple negated sentence such as 'Nobody saw it' (p. 11) with 3sO-AGR is neither 'There is at most one who didn't see it' nor 'There is not at most one who saw it'. Further, the meaning of plural O-AGR is not characteristic of the default frame but may also be found in the detransitivised frame, for instance, when there is a quantifier like 'many' – although Bickel et al. do not mention whether such quantifiers are compatible with the ø-detransitive, this seems likely given its semantics.

Bickel et al.'s concept of cardinality has been the main inspiration for them term quantifiability as used in the present work. There are, however, some important differences:

- Quantifiability can also be applied to the description of mass nouns, where cardinality does not make much sense.
- Bickel et al. view cardinality as an alternative to specificity and not as a closely related concept. In Chintang, quantifiability is only relevant as a prerequisite to specificity.
- Bickel et al. associate cardinality with knowledge (p. 13: "a detransitive form signals that the cardinality of the set of object referents is unknown"), whereas for quantifiability it is more relevant whether the quantity of a referent *could* be known.

An interesting detail is that Puma seems to have borrowed the Nepali dative marker *-lai* along with DOM. According to Bickel et al. (2007b:7), this marker is optional on indexed P but banned from the ø-detransitive.

[25] Note that as in Chintang and many other Kiranti languages, the distinction between 3d and 3p is neutralised in Puma 3O-AGR, so this seems to be a mistake.

	Limbu	Yakkha	Athpare	Belhare	Chiling	Chintang	Bantawa	Puma
term	antipassive	detransitivisation	undergoer demotion	detransitivisation	-	S/A detransitivisation	implicit antipassive	ø-detransitive
agreement	V-s(A)	V-s(A)	V-s(A)	V-s(A)	V-s(A)	V-s(A)	V-s(A)	V-s(A)
A case	NOM	?	NOM	NOM	?	NOM	NOM	NOM
applicability	most verbs	all verbs	?	?	?	all verbs	many verbs	?
O droppable	no	?	?	yes	?	yes	?	no
banned POS	pronoun, proper noun	?	?	?	?	pronoun	?	?
O + ADJ	no	?	?	no	?	yes	?	yes
O + REL	?	?	?	yes	?	yes	?	yes
O + POSS	insecure	?	?	no	?	yes (rare)	?	?
O + NUM	no	?	?	no	?	yes (rare)	?	?
O + PL	no	?	?	no	?	yes (rare)	?	?
argumenthood	incorporated	?	quasi-incorporated	independent	?	independent	?	independent
semantics of O/V	incorporated	unspecific or generic, focus on event	non-specific	kind/generic	?non-specific	non-specific	less specific/affected; time less relevant	non-cardinal

Table 2.12: S/A detransitivisation in Eastern Kiranti

2.9.9 Summary

Table 2.12 shows a summary of the properties of S/A detransitivisation in those Kiranti languages where it is known to exist. Inspite of the common coding base other properties are diverse. Note, though, that several phenomena that were initially thought to be impossible in Chintang, too, could be shown to be marginally possible by careful elicitation and a large corpus. The heterogeneity in the table might thus be reduced if similar in-depth studies were conducted for the other languages.

Chapter 3

Nepali: Differential A and O marking

3.1 Language background

Nepali ([n̪eˈpaːli], IAST *Nepālī*, Devanagari नेपाली) is an Indo-Aryan language spoken by more than 15 million speakers in Nepal and elsewhere.

As in the case of Chintang, the name of the language is derived from a toponym. However, Nepali being a language with a long history, it has come along a more winding path. The ethnic group who originally introduced the language into Nepal were the Khaśa (Whelpton 2011:8), so Nepali is also known as *Khaśa Bhāṣā*. A later term is *Parvatī*, which is linked to the term *Parvatīya* 'hill people' (cf. Nep. *parvat*/*pʌrbʌt* 'hill'), a cover term for all Nepali-speaking castes (Whelpton 2011:264). The word *Nepālī* only came into use much later. *Nepāla* originally was the name of the Kathmandu valley and is likely to be etymologically related to the ethnonym *Newar* (Whelpton 2011:14). The Newars are the indigenous Tibeto-Burman inhabitants of the valley and still constitute one of the largest minorities of Nepal.

The term started to be used for a larger area after Prithivī Nārayaṇa Śāha, head of the kingdom of Gorkha, first conquered the valley and then large parts of the territory of present-day Nepal in the second half of the 18th century (Whelpton 2011:35). During this process the Parvatī language spoken by the conquerors (also known as *Gorkhalī* by that time) gained influence, and, as the political center of the kingdom shifted to the Kathmandu valley, came to be known as *Nepālī*. The ambiguity of the term *Nepāla* continues even today – for instance, older people in Chintang still refer to the Kathmandu valley as *Nepala*.

Nepali is the biggest language of Nepal and is nowadays spoken all over the country, but not everywhere in equal proportion. As a rule of thumb, the stance of Tibeto-Burman languages becomes stronger the farther north or east one goes. In the flatlands, closely related Indo-Aryan languages such as Maithili, Bhojpuri, and Awadhi form large minorities, and in the Kathmandu valley itself Newari is spoken by several hundred thousand people. Outside of Nepal, Nepali is spoken in a number of neighbouring areas, especially in the Indian states Sikkim, Assam, and West Bengal (Darjeeling district), in Bhutan, and in Myanmar (Schmidt 1993:x, Lamsāla 2062 V.S.:1). Due to excessive labour migration in recent years, substantial communities do now also exist in the United States, in the Gulf states, and in various Asian countries such as Korea and Singapur.

Nepali is by no measures an endangered language. On the contrary, it is thriving both as the national language of Nepal and the monolingual Parvatī population and as the *lingua franca* used among speakers of numerous other first languages. In the younger generation of Nepal, virtually everybody speaks Nepali at least as a second language. In almost all cases of language endangerment in Nepal, Nepali is the primary attacker. Its safe stance is due to its being the largest and the only officially supported language.

The number of speakers of Nepali is hard to estimate due to two factors. One is the low reliability of the official data in the last National Census of Nepal (Central Bureau of Statistics 2001), the other the incongruence of the language area with political structures. Genetti (1994:5) cites the

number of 9,300,000 from the 1991 census and says that it is not clear whether this number includes L2 speakers or not. Yadava (2003:141) gives the number of 11,000,000 native speakers from the 2001 census. The number of 15,000,000 given above is a rought estimation based on assumptions about population growth, language shift, and Nepali speakers outside Nepal.

Nepali has a number of dialects which differ on all levels of linguistic analysis. An example for differences on the syntactic level is the marking of A. As many other Indo-Aryan languages, Nepali has split A marking, A being marked by the nominative (= zero) or the ergative *-le* depending on tense, aspect, and a couple of other minor factors (Abadie 1974, Li 2007b). Impressionistically, ERG tends to be more frequent the farther one moves to the East, possibly due to the influence of Tibeto-Burman substrates.

By contrast, object marking seems to be homogeneous across dialects, at least as far as can be told at the moment. Nepali dialects are poorly documented, but several linguistically informed native speakers I consulted indicated that the use of nominative and dative was similar among speakers from all regions. The Nepali National Corpus does not allow for the investigation of dialects because unfortunately, the birthplace and dialect of speakers haven't been documented and the written texts are standardised, anyway.

Differently from Chintang, a lot of linguistic work exists on Nepali. However, much of it lacks typological informedness and is thus of little use for the present work. Works influenced by the prescriptive traditions of Sanskrit and Latin grammar writing or intended for language learners will only be cited in the central sections on DOM, where every available piece of information should be considered. The overview sections are based on a couple of papers and on my own research.

3.2 Overview of relevant morphology

3.2.1 Parts of speech

Traditional grammars of Nepali suffer from well-known problems when it comes to parts of speech, most importantly inconsistency in applying formal and functional criteria. They will thus be ignored here. There is one recent classification which has been incorporated into the Nepali National Corpus (NNC), the written parts of which have been POS-tagged automatically. The tag set is laid out in Hardie (2005) and is typical for computational linguistic approaches in lumping together lexical and syntactic criteria. For instance, the masculine and feminine forms of adjectives or the various tenses of verbs each receive different tags. Such a classification is likewise not in the spirit of general linguistics. I therefore give a proposal of my own in Table 3.1 below. For a survey of treatments of parts of speech by Nepalese scholars see Prasain (2011:7ff.).

As in Chintang, SAP and NSAP deixis are expressed by different parts of speech, but the situation is a bit more complicated. One class (pronouns in the technical sense) is formed by all SAP forms excluding *tʌpaĩ* [2HH] (which morphosyntactically is a noun) but including *aphu* [REFL]. Two other, closely related classes (nominal and versatile demonstratives) are used for NSAP deixis. When it comes to DOM, *tʌpaĩ* patterns with the pronouns in necessitating DAT (see section 3.5.8).

It is useful to assume a superclass of nominals characterised by their ability to carry case markers. This class comprises nouns, pronouns, both types of demonstratives, adjectives, numerals, and determiners. Apart from numerals, all of these can also be marked for number. Differently from Chintang, there are no differences in the distribution of case markers after the various nominal subclasses – all nominals can be combined with all case markers.

3.2.2 Nominal morphology

The two central categories of Nepali nominals are number and case. Number is simple: there are two numbers, a zero-marked singular and a plural marked by *-hʌru* [PL]. As in Chintang, the plural marker also has an associative use (*Ram-hʌru* 'Ram and the others').

The case system is a bit more complicated. Some grammarians deem it important to distinguish between case markers and adpositions (in the case of Indo-Aryan languages: postpositions). I do

	dependency	inflection	syntactic use
verb	no	TMA, polarity, person/number of one argument	predicate
noun	no	number, case	referent
pronoun	no	number, case, special GEN	deictic referent
nominal demonstrative	no	number, case, special obl. case	deictic referent
versatile demonstrative	no	number, case, special PL and obl. case	deictic referent or modification
adjective	no	gender, number, case	qualification of referent
numeral	no	classifier, case	quantification of referent
determiner	no	number, case	other modification of referent
adverb	no	no	modification of verb
interjection	no	no	equivalent to clause
conjunction	clause	no	clause chaining
particle	any other word	no	grammatical
affix	specific p.o.s.	no	grammatical

Table 3.1: Nepali parts of speech

not consider this a useful distinction, since case is a functional category (cf. section 2.2.2) and the distinction between adpositions and affixes is a formal one. If one wants to make a distinction for some reason it should be between morphologically dependent and independent case markers. By the criterion used for affixhood in section 3.2.1 (morphological dependency and requiring a host belonging to a specific part of speech), all of the "postpositions" found in Nepali are suffixes, even if they are only used once in conjoined NPs such as *Gita rʌ Ram-lai* [Gita and Ram-DAT] 'to Gita and Ram'. Table 3.2 shows the most important case markers. Since there are a lot more than in Chintang and there is no strong paradigmaticisation, many are not glossed with a grammatical term but with their English translation.

All markers in the table are of relatively recent origin. Traces of old Indo-European cases have survived in present day literary Nepali in nouns which have a rectus form in *-o* or *-u* (e.g. *baʈo* 'way') and an oblique form used with case and number markers (e.g. *baʈa* in *baʈa-ma* [way-LOC] and *baʈa-hʌru* [way-PL]). This distinction is, however, only maintained in the written language, whereas the spoken language uses one of the two forms in all environments, depending on the dialect. A similar distinction that is fully intact in all registers is that between the singular rectus and singular oblique of demonstratives, e.g. *yo* [PROX] vs *es-ma* (older *yʌs-ma*) [PROX-LOC]. Forms like *baʈa* and *es* will be treated as stem allomorphs here.

All essive locative cases can also be used as allatives, so for instance *-ma* is not only 'at, in, on' but also 'to, into, onto'. They can be combined with specialised allatives and with ablatives to yield more specific meanings, e.g. *-ʌgadi-sʌmmʌ* [ANTE-TERM] 'up to (the place) in front of', *-mathi-baʈʌ* [SUPER-ABL$_1$] 'from above'. A noun marked by the genitive *-ko* can without further marking be interpreted as referring to a possessum, e.g. *Ram-ko* [Ram-GEN] 'Ram's (thing)'. As a consequence, the genitive can be followed by all other case markers, e.g. *Ram-ko-sʌŋgʌ* [Ram-GEN-COM$_4$] 'with Ram's (thing)'. The genitive can express relations between referents and predicates but mostly marks relations between referents.

For the present study, the most important cases are the zero-marked nominative and the dative marked by *-lai*. The nominative is the default case and marks the majority of S, many P/T/G, and many A. The dative marks all G in one verb class, P/T/G with certain referential properties in another, and most S of experiencer verbs. Other cases marking argument roles are the ergative *-le* (A and instrument-like T), the locative *-ma* (many G and all kinds of essive and allative relations),

Ø	NOM	nominative	*-lai*	DAT	dative
-bahek	'except'		*-le*	ERG	ergative
-bairʌ	EXTRA	extraessive	*-mathi*	SUPER	superessive
-baṭʌ	ABL_1	ablative I	*-ma*	LOC	locative
-bhitrʌ	INTRA	intraessive	*-muni*	SUB	subessive
-bhʌnda	COMP	comparative	*-mʌddhe*	'amidst, among'	
-bic	'between'		*-nirʌ*	'near'	
-bina	ABESS	abessive	*-pari*	TRANS	translative
-biruddhʌ	'against'		*-pʌchi*	TMP.POST	temporal postessive
-dekhi	ABL_2	ablative II	*-sath*	COM_1	comitative I
-dwara	'by, by means of'		*-sitʌ*	COM_2	comitative II
-jʌsto	EQU_1	equative I	*-sʌmbʌndhi*	'concerning'	
-jhʌĩ	EQU_2	equative II	*-sʌmet*	COM_3	comitative III
-ko	GEN	genitive	*-sʌmmʌ*	TERM	terminative
-(ko)barema	'about'		*-sʌŋgʌ*	COM_4	comitative IV
-(ko)lagi	FIN_1	final I	*-tirʌ*	DIR	directional
-(ko)nimti	FIN_2	final II	*-tʌrphʌ*	'on behalf of'	
-(ko)pʌchaḍi	POST	postessive	*-ʌghi*	TMP.ANTE	temporal antessive
-(ko)ʌgaḍi	ANTE	antessive	*-ʌnusar*	'according to'	

Table 3.2: Nepali case markers

and the genitive *-ko* (possessors, some S/A in subordinate clauses). See section 3.3.2 for some more details on the use of core cases.

3.2.3 Verbal morphology

Verbal morphology is much more complex than nominal morphology in that more categories can be expressed by a single word form and the interaction between the relevant markers is less regular. Verbs index one argument (usually S/A) and are marked for one composite TMA category. Non-finite forms can be combined with auxiliaries (partially tending towards univerbation) to express an even greater range of TMA functions. In addition, polarity is also marked morphologically. See section D.2 in the appendix for paradigm tables. Genetti (1994) and Prasain (2011) offer more detailed accounts of verbal morphology.

The referential properties that may be indexed are person, number, gender (masculine/feminine), and honorificity (no/mid/high). There are complex interactions between these properties: singular and plural are only distinguished in the first person and in the non-honorific third person, masculine and feminine gender are not distinguished in the second person and with high honorificity, and dedicated non-honorific forms only exist in the singular. In addition, there are interactions with TMA: the tense that allows most distinctions is the simple nonpast with 11 agreement suffixes, whereas, for instance, the positive simple past has maximally 9 different agreement suffixes. The whole system varies according to register and dialect. The category that is most vulnerable is gender: for instance, in the spoken Nepali of Kathmandu, all gender distinctions tend to get neutralised to the effect that masculine forms are used in all contexts. While honorificity is stable in the second person (except for some L2 speakers, who tend to make simplifications), the threefold distinction in the third person is mainly maintained in the written language, whereas the spoken language uses either non-honorific or high honorific forms. See Genetti (1999) for more details on sociolinguistic variation in agreement.

The TMA system is no less complex. Tense, aspect, and mood cannot be functionally separated. One distinct TMA cluster will be referred to as a screeve here, a term borrowed from the Kartvelian tradition of grammar writing. The seven simple screeves (i.e. screeves that can be expressed without an auxiliary) are the simple nonpast, the future/habitual nonpast (only distinct from the simple

nonpast for the copula), the simple past, the habitual past, the optative, the imperative, and the simple future. If one combines the past participle *-eko*, the nonpast participle *-ne*, and the continuative form *-dʌi* with all available forms of the copula one gets as much as 21 additional tenses. The more well-known of these are the present perfect (past participle + *ch-*), the present progressive (progressive + *ch-*), the composite future (nonpast participle + *ch-*), the past perfect (past participle + *thi-*), the past progressive (progressive + *thi-*), and the future in the past (nonpast participle + *thi-*).

There is a wealth of non-finite forms, among them also the ones already mentioned. Altogether there are two infinitives (*-nu*, *-nʌ*), two participles (*-eko*, *-ne*), the continuative *-dʌi*, six converbs (*-erʌ*, *-i*, *-ikʌnʌ*, *-da*, *-dakheri*, *-unjel*) and the conditional/nominaliser *-e*.

Negation is marked by a prefix *nʌ-* in optative, imperative, probable future, and all non-finite forms. In the other forms (including the composite tenses) it is fused with the agreement and TMA suffixes.

Apart from inflection there is one highly productive derivational process that also plays a role for DOM, which is the passive (see section 3.3.3.2).

3.3 Overview of relevant syntax

3.3.1 Word order

Word order in Nepali is governed by very similar principles as in Chintang (see section 2.3.2). The default word orders are SV, APV, and AGTV. Highly topical elements tend to be placed farther to the left of the verb than other elements, and the postverbal position is typically used for NPs that were originally planned as covert. Figure 3.1 on the next page shows the frequencies of various word orders in fully expanded frames in the annotated part of the Nepali National Corpus.

The word orders APV, PAV, and PVA are illustrated by the examples below. (1) shows the monotransitive default word order APV. (2) has PAV because various potatoes have been the topic of the talk for a couple of paragraphs. The sort represented by *telai* is therefore also a contrastive topic. Finally, (3) has PVA because the question already makes it clear that the speaker is not talking about his own motorbike so that it wouldn't have been necessary to mention A overtly.

(1) *Mʌi-le ʌlikʌti kuro bujh-in-ʌ.*
1s-ERG a.bit talk understand-NEG.PST-1s
'I didn't quite understand what you said.' (NNC:A001011002.374)

(2) *Te-lai ni manche-le ruca-ch-ʌn.*
MED-DAT ASS person-ERG prefer.PRF-NPST-3p
'People really like this (kind of potato).' (NNC:A001011002.740)

(3) *Tel~sel kʌti kha-nch-ʌ tim-ro-le?*
oil~and.stuff how.much eat-NPST-3s 2s-GEN-ERG
'How much oil does your (motorbike) consume?' (A001017003.143)

3.3.2 Frames and classes

This section lists the most important verb classes of Nepali together with examples. Differently from Chintang, one class rarely employs several completely different frames, so it makes more sense to base the description on classes instead of frames. In particular, Nepali has neither S/O detransitivisation nor a single S/O ambitransitive verb. What is quite frequent, though, are alternative cases on one or more roles within a frame that otherwise remains unchanged. No valency dictionary of Nepali is available, so the sizes of the classes cannot be given.

3.3.2.1 Intransitive verbs {S-NOM V-s(S)}

This is the simplest class. There is a single argument which is always marked by the nominative and linked to verbal agreement. Examples are *sut-* 'sleep', *sunni-* 'swell', *jʌl-* 'burn', or *ʌlmʌli-* 'be

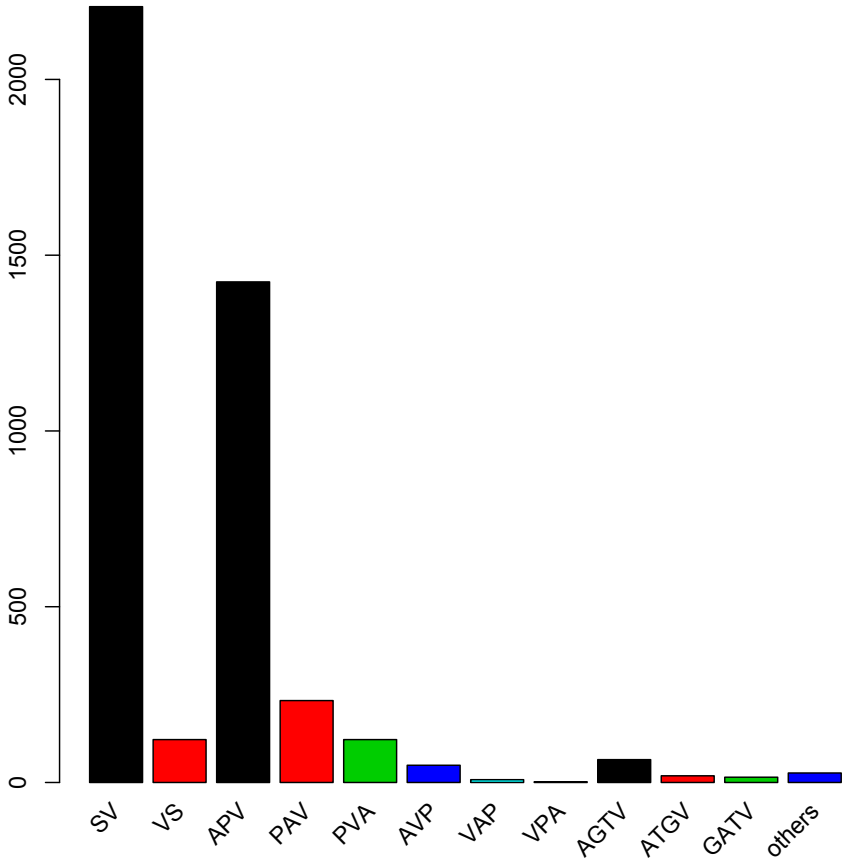

Figure 3.1: Attested word orders in fully expanded frames

confused':

(4) *Laṭo sathi ʌlmʌli-y-o.*
stupid fellow be.confused-PST-3s
'The stupid fellow was confused.' (NNC:kantipur-art-2061-12-20.16)

3.3.2.2 **Unergative verbs** {S-NOM/ERG V-s(S)}

The few verbs in this class work like intransitive verbs except that their S can be marked by the ergative in the past (or more precisely, under the same conditions as transitive A – see section 3.3.3.1). All verbs in this class involve activities where air or matter is exhaled/excreted from the human body and which cannot be completely controlled. An example is *khok-* 'cough' as shown in (5); others given by Adhikārī (2052 V.S.:95ff) are *ḍʌkar-* 'belch', *mut-* 'piss', *ghur-* 'snore'.

Below is one example each for S-NOM and S-ERG with *khok-* 'cough'.

(5) a. *U khok-ch-ʌ.*
DIST cough-NPST-3s
'He coughs.' (elicitation SAR 2011)

b. *Us-le bistarʌi khok-y-o.*
DIST-ERG slow cough-PST-3s
'He coughed slowly.' (NNC:s27.101)

3.3.2.3 Bivalent motion verbs {A-NOM P-NOM/LOC V-s(A)}

This is another relatively small class with mixed characteristics. Its members are bivalent and thus have the same role set as mono- and ditransitive verbs (see below); however, their A can never be marked by ERG and their P shows an altogether different DOM pattern (LOC vs NOM), which is exemplified in (6) below. All verbs in this class express self-inducted motion towards a goal, e.g. *ja-* 'go', *au-* 'come', *phʌrk-* 'return', *lag-* 'set forth for'.

(6) a. *Mʌ koʈha-ma gʌ-ẽ.*
1s room-LOC go-PST.1s
'I went to my room.' (NNC:book-fiction-yojangandha-2063.2259)

b. *Mʌ Kaʈhmaɖʌũ gʌ-ẽ.*
1s Kathmandu go-PST.1s
'I went to Kathmandu.' (NNC:book-biography-vyakti-ek-drishti-anek-2058.1144)

3.3.2.4 Monotransitive verbs {A-ERG/NOM P-NOM/DAT V-s(A)}

Monotransitive verbs are the largest bivalent class. They are characterised by the presence of an A and a P with differential case marking on each (cf. sections 3.3.3, 3.4.1), A being linked to verbal agreement. Below is an example for *ʈhʌg-* 'deceive, cheat'. Other verbs in this class are *sun-* 'hear, listen', *pol-* 'burn', *khol-* 'open'.

(7) *Belaet-le hami-lai ʈhʌg-y-o.*
UK-ERG 1p-DAT cheat-PST-3s
'The UK have cheated us.' (NNC:janaastha-news-2061-12-03.81)

There are a few bivalent verbs which superficially look like normal monotransitive verbs but which always mark P by DAT, even in sentences which are otherwise semantically completely parallel to sentences with P-NOM. (8a) shows another example with *ʈhʌg-*, where P is marked by NOM because it is non-specific. (8b) shows a contrasting example with *sʌgau-* 'help', where even a non-specific P must be marked by DAT.

(8) a. *Ɗhãʈuwa-le manche thʌg-ch-ʌ.*
liar-ERG people deceive-NPST-3s
'A liar deceives people.' (elicitation SAR 2011)

b. *Debta-le manche*(-lai) sʌgaũ-ch-ʌ.*
god-ERG people-DAT help-NPST-3s
'God helps people.' (elicitation SAR 2011)

The verbs displaying this behaviour are semantically diverse – other examples include *phon gʌr-* 'phone', *sod-* 'ask', *ʌsʌr par-* 'affect'. They must therefore be viewed as constituting a lexical subclass.

3.3.2.5 Transfer ditransitives {A-ERG/NOM T-NOM/DAT G-DAT/LOC/NOM V-s(A)}

This is the most complex class in terms of case marking. There are three arguments, all of which may be differentially marked. The constitutive element is a T that is open to DOM. Beside that, there is an A with the DAM pattern found in most transitive classes and a particularly versatile G. The marking of G makes it possible to distinguish several subclasses:

- **Ia**: one case on G (DAT), e.g. *di-* 'give', *sikau-* 'teach', *bec-* 'sell', *khuwau-* 'feed'
- **Ib**: one case on G (LOC), e.g. *bʌsal-* 'seat, put on', *osar-* 'transport'
- **IIa**: two cases on G (DAT/LOC), e.g. *rakh-* 'put', *bãɖ-* 'distribute (to)', *hal-* 'put in', *phal-* 'throw'
- **IIb**: two cases on G (LOC/NOM), e.g. *la-* 'take (away)', *lyau-* 'fetch', *phʌrkau-* 'return', *sar-* 'move'
- **III**: three cases on G (DAT/LOC/NOM), e.g. *pʌʈhau-* 'send', *puryau-* 'deliver'

So far it is not clear to which degree this division is motivated or arbitrary, with one exception: class Ia only contains verb whose semantics require an animate G, and all verbs that license an animate G are in one of the classes which allow G-DAT.[1] For the present work G is irrelevant, anyway; what is important is that the case alternation on T follows the same scheme as DOM in other verb classes. Below are examples for each subtype.

Ia – *di-* 'give':

(9) *Joseph-le daju-hʌru-lai ʌnnʌ di-y-o.*
Joseph-ERG elder.brother-PL-DAT grain give-PST-3s
'Joseph gave grain to his brothers.' (NNC:book-criticism-paschimka-kehi-sahityakar-2062.5034)

Ib – *bʌsal-* 'put on':

(10) *Us-le ciya culo-ma bʌsal-y-o.*
3s-ERG tea stove-LOC put.on-PST-3s
'He put tea on the stove.' (elicitation SAR 2011)

IIa – *rakh-* 'put':

(11) a. *Yo skul-ma pʌni kʌmpyuʈʌr rakh-nu pʌr-y-o.*
PROX school-LOC also computer put-INF_1 fall-PST-3s
'In this school, too, a computer should be installed.'
b. *Yo bʌcca-lai nam rakh-nu pʌr-y-o.*
PROX child-DAT name put-INF_1 fall-PST-3s
'This child should be given a name.' (elicitation SAR 2011)

IIb – *la-* 'take (away)':

(12) a. *Beatrise aph-n-a premi-lai nʌwʌũ sorgʌ-ma la-nch-in.*
Beatrice REFL-GEN-OBL lover-DAT ninth heaven-LOC take-NPST-3fMH
'Beatrice takes her lover to seventh heaven.' (NNC:book-criticism-paschimka-kehi-sahityakar-2062.672)
b. *Ram-le Sita-lai ghʌr la-nch-ʌ.*
Ram-ERG Sita-DAT house take-NPST-3s
'Ram takes Sita home.' (elicitation SAR 2011)

III – *pʌʈhau-* 'send':

(13) a. *Aŋgrej-le sʌine dʌl dherʌi yuddhʌ chetrʌ-hʌru-ma pʌʈha-y-o.*
English-ERG military group many fight area-PL-LOC send-PST-3s
'The English sent troups into many fighting areas.' (NNC:himalkhabarpatrika-2059-03-16.1490)
b. *Mʌntralʌe-le Renbo ephem-lai goppe pʌtrʌ pʌʈha-y-o.*
ministery-ERG Rainbow F.M.-DAT secret message send-PST-3s
'The ministery sent a secret message to Radio Rainbow F.M.'
(NNC:book-essay-radio-patrakarita-2062.1895)

c. *Uni-hʌru-k-a bauama-le keʈi-hʌru-lai skul pʌʈha-ek-a rʌhech-ʌn.*
DIST-PL-GEN-OBL parents-ERG girl-PL-DAT school send-PST.PTCP-PL MIR-3p
'Their parents sent the girls to school.' (NNC:book-academic-jana-siksha-2058.1458)

[1] This does not mean that all animate G are marked by DAT, though. DAT is certainly the default here, but there are cases where animate G are marked by LOC.

3.3.2.6 **Instrumental ditransitives** {A-ERG/NOM T-ERG G-NOM/DAT V-s(A)}

This is a smaller class of ditransitive verbs where T takes on an instrument-like function. A takes part in DAM, and G takes part in DOM. Examples are *han-* 'hit', *lip-* 'smear', *kaʈ-* 'cut', *chop-* 'cover':

(14) *U dubʌi hʌtkela-le mukh chop-ch-ʌ.*
DIST both palm-ERG face cover-NPST-3s
'He covers his face with both palms.' (NNC:book-fiction-ek-paluwa-anekaun-yam-2026.7347)

3.3.2.7 **Equational ditransitive frame** A-ERG/NOM T-NOM G-DAT V-s(A)

This is not a class but an isolated frame – all verbs that use it also employ other frames. It is nevertheless important because both T and G have fixed cases in it and are therefore not open to DOM. It is called equational here because all verb senses that use it equate an existing G with a T, either mentally or physically. All senses that are known so far to use this frame are *bʌnau-* 'make G a T', *bʌtau-* 'announce G to be a T', *bhʌn-* 'call G a T', *man-* 'consider G a T', *tulyau-* 'make G a T', *ʄhan-* 'consider/deem G a T', *sʌmjh-* 'consider G a T'. (15) shows an example with *bʌnau-*.

(15) *Es-ʌi ʌnubhʌb-le un-lai mʌhakʌbi bʌna-y-o.*
PROX.OBL-FOC experience-ERG 3sMH-DAT great.poet make-PST-3s
'It was that experience that turned him into a great poet.'
(NNC:book-criticism-paschimka-kehi-sahityakar-2062.2404)

3.3.2.8 ***hu-*** {S-NOM V-s(S)}, {S-DAT N.EXP V-s(N.EXP)}, {CT-NOM CR-NOM V-s(CT)}, {A-GEN P-NOM V-s(P)}

The copula *hu-* forms a single-member class. It is nevertheless important because of its high token frequency. The copula is both formally and functionally more complex than any other verb, its most important characteristic being the use of several suppletive stems depending on tense, aspect, aktionsart, and the ontological type of the predicate nominal (beside *hu-* these are *h-*, *ch-*, *bhʌ-*, and *thi-*). See section D.2.6 in the appendix for detailed paradigms.

The various frames of *hu-* depend on its function. Below is one example for each.

S-NOM V-s(S) – existence:

(16) *Tʌrkari rakh-ne thaw-ʌĩ ch-ʌin-ʌ.*
vegetables put-IPFV.PTCP space-FOC be.there-NEG.NPST-3s
'There is no place to put the vegetables.' (NNC:V001005002.40)

S-DAT N.EXP V-s(N.EXP) – experience:

(17) *Us-ko pʌral-ko chapro pʌni ʄin-ko chapro-ko rup-ma bʌdʌli-nʌ*
DIST-GEN straw-GEN hut also tin-GEN hut-GEN form-LOC change-INF_2
sʌk-la bhʌnne us-lai biswas ch-ʌ.
be.able-PROB.FUT.3s CIT.ADN DIST-DAT belief be.there.NPST-3s
'He believes that his straw hut could change into a tin hut.'
(NNC:book-fiction-ek-paluwa-anekaun-yam.4648)

CT-NOM CR-NOM V-s(CT) – identity, attribution:

(18) *Timi kun kisim-ko syal h-ʌu?*
2s which kind-GEN jackal be[NPST]-2MH
'What kind of jackal are you?' (NNC:book-criticism-samakalin-nepali-natak-2057.4945)

A-GEN P-NOM V-s(P) – possession:

(19) *Me-ro kam ch-ʌ.*
1s-GEN work be.there.NPST-3s
'I have work.' (NNC:book-criticism-samakalin-nepali-natak-2057.1448)

3.3.2.9 *lag-* {S-NOM V-s(S)}, {S-DAT N.EXP V-s(N.EXP)}, {A-DAT P-ABL N.EXP V-s(N.EXP)}

Another unique but highly frequent verb is *lag-* 'be on, be attached to, be there'. While morphologically simple, this verb uses rather different frames in different context. Most experiences are expressed using this verb and a noun that refers to an emotion and is linked to verbal agreement.

S-NOM V-s(S) – state:

(20) *Gham lag-eko hu-nch-ʌ.*
sun be.on-PRFV.PTCP be-NPST[.HAB]-3s
'The sun is always shining.' (NNC:book-fiction-pagalbasti-2059.2464)

S-DAT N.EXP V-s(N.EXP) – experience:

(21) *Mʌ-lai bhok lag-y-o.*
1s-DAT hunger be.on-PST-3s
'I'm hungry.' (NNC:book-essay-hindai-garda-2061.935)

A-DAT P-ABL N.EXP V-s(N.EXP) – bivalent experience:

(22) *Tʌpaĩ-lai keṭi-dekhi ḍʌr lag-ch-ʌ?*
2/3H-DAT girl-ABL$_2$ fear be.on-NPST-3s
'Are you afraid of girls?' (NNC:book-fiction-dosro-prahar-2062.1309)

3.3.3 Differential marking and valency manipulation

Beside a few patterns pervading the whole verbal lexicon, Nepali is characterised by a large number of small-scale alternations that only apply to some classes or subclasses (described in great detail in Adhikārī 2052 V.S.). This section describes one pervasive pattern, differential agent marking (section 3.3.3.1), because it is highly relevant for Nepali syntax in general, and the passive (section 3.3.3.2), which interacts with DOM. Also see section 3.4.7 for a special kind of ambitransitivity that is only found with complex predicates. DOM itself is another pervasive pattern and is discussed in detail in sections section 3.4, section 3.5, and section 3.6.

3.3.3.1 Differential agent marking

In Nepali, all A that can be marked by ERG can also be marked by NOM. The reverse statement ("all A-NOM can be exchanged with A-ERG") is true with one exception: the A of bivalent motion verbs such as *ja-* 'go' never get ERG. Li (2007b) makes some important observations concerning the conditions behind DAM:

- ERG is obligatory on all inanimate A.
- ERG is obligatory with animate A in the "perfective domain", that is, when the predicate is marked by one of the following tenses:
  - simple past
  - present perfect
  - past perfect
- ERG is optional with animate A in the "imperfective domain":

- simple present (= simple nonpast in this work)
- present progressive (= nonpast progressive)
- past progressive
- past habitual
- simple future

One important corollary of the statements above is that A-ERG is always possible whereas A-NOM is much more restricted. This is the reversal of the historical situation, where A-NOM was the standard and A-ERG was an innovation (Wallace 1981, Hutt 1988), and makes Nepali different from all other modern Indo-Aryan languages, where it is rather A-ERG that is marked in terms of functional conditions. The original function of *-le* [ERG], which is still preserverd, was to mark instruments.

Li does not treat the remaining mood-like screeves (optative, imperative), the non-finite forms, and the majority of composite tenses formed from their base, nor does he mention which factors condition the choice between NOM and ERG for animate A in the imperfective domain.

An earlier paper, Abadie (1974), is all in all not as detailed as Li but gives a useful hint to one of these factors. Abadie mentions (p. 163) that A-ERG is also obligatory in "certain constructions in the semantic area of permission/obligation", for instance, in the deontic *-nu pʌr-* [-INF$_1$ fall] construction. Here is an example from the NNC:

(23) *ʄo-sʌŋgʌ dherʌi ch-ʌ us-le dherʌi tir-nu pʌr-ch-ʌ.*
who.REL-COM$_4$ much be.there.NPST-3s DIST-ERG much pay-INF$_1$ fall-NPST-3s
'Whoever has much has to pay much.'
(NNC:book-academic-swasthya-samaj-ra-rajniti-2062.2161)

Obligations are usually more characteristic of one or a few referents than other predications. This is because specific duties tend to get attributed to specific responsible persons. I believe that the characteristicness of an action for a referent is actually the major factor conditioning the use of the ergative on animate A in the imperfective domain. A similar idea is put forward by Butt and Poudel (2007), who claim that the main factor for A-ERG in imperfective tenses is the distinction between stage- and individual-level predicates. Consider the following example, where both clauses contain an ERG-marked, animate A in an imperfective tense and both describe actions that are characteristic of these A:

(24) *Biralo-le musa-lai khela-i khela-i mar-e-jhʌĩ, manche-le pʌni aphu-bhʌnda nirdha-lai sidhʌi mar-i-hal-dʌin-ʌ.*
cat-ERG mouse-DAT play-CVB$_2$ play-CVB$_2$ kill-NMLZ-EQU$_2$ person-ERG also REFL-COMP weak-DAT directly kill-LNK-COMPL-NPST.NEG-3s
'Just like cats kill mice after playing a while with them, people also don't kill those weaker than them directly.' (NNC:freenepal-fiction-2061-12-11.261)

Note that there are two ways in which a predicate can be characteristic of an A: either an action is particularly frequent among all actions carried out by an A over the course of time (for instance, cows eat grass all the time), or one A is particularly frequent among all A that are observed to induce an action (for instance, sowing is usually done by farmers). These two cases may coincide as in (24) but do not have to – grass is also eaten by other animals, and farmers have other things to do than to sow.

3.3.3.2 Passive

Nepali has a passive that is marked by a verbal suffix *-i* [PASS]. The verb stays fully inflected, but agreement shifts to O. A usually stays covert, but if it is to be mentioned overtly, it must carry the case marker *-dwara* 'by' or *-baṭʌ* [ABL$_1$] instead of NOM or ERG. Since both roles remain expressible, we will assume that the passive does not change valency. Below is an example.

(25) *Rin-ma pʌr-erʌ timi malik-dwara bãdh-i-y-ʌu.*
debt-LOC fall-CVB_1 2s master-by bind-PASS-PST-2s
'You had debts, and so you were bound by the master.'
(NNC:book-drama-prempinda-2058.4908)

Intransitive verbs and bivalent motion verbs, whose A never gets marked by ERG, can also be passivised to express an impersonal reading:

(26) *Ajʌ ʌphis ja-nu pʌr-dʌin-ʌ, ajʌ din-bhʌri sut-i-nch-ʌ.*
today office go-INF_1 fall-NEG.NPST-3s today day-full.of sleep-PST-NPST-3s
'Today we don't have to go to the office, today we will sleep the whole day.'
(NNC:madhuparka-humor-2061-11.121-122)

(27) *Seks-ko nʌjik esʌri ja-ũ, jʌsʌri mʌndir-ma gʌ-i-nch-ʌ.*
sex-GEN near PROX.METHOD go-[OPT.]1p REL.METHOD temple-LOC go-PASS-NPST-3s
'Let's approach (the topic) sex as when going to a temple.'
(NNC:deshantar-misc-2061-11-02.67)

The most interesting point about the Nepali passive in the present context is that it preserves DOM, that is, O can still be marked by NOM or DAT after passivisation. See section 3.4.6 for details.

3.4 Formal properties of DOM

3.4.1 DOM as an isolated pattern

Nepali DOM is much less inter-connected than Chintang S/A detransitivisation. It concerns only a single argument (the object O), which can be marked by either NOM or DAT. The case of the other arguments and especially the case of the agent are not touched by this process, nor is verbal agreement (with one minor exception, see section 3.4.6). The examples below show that all combinations of A and P case are grammatical.[2]

A-ERG P-NOM:

(28) *Gaibʌstu kʌs-le her-ch-ʌ?*
cattle who-ERG look-NPST-3s
'Who will look after the cattle?' (NNC:book-academic-swasthya-samaj-ra-rajniti-2062.1710)

A-ERG P-DAT:

(29) *Ibsʌn-le jʌnsadhʌrɔ̃ɖ-ko ãkha-le naʈʌk-hʌru-lai her-e.*
Ibsen-ERG ordinary.people-GEN eye-ERG drama-PL-DAT look-PST.3MH
'Ibsen looked at plays with the eyes of the ordinary people.'
(NNC:book-criticism-paschimka-kehi-sahityakar-2062.3118)

A-NOM P-NOM:

(30) *U bhitteghʌɖi her-ch-ʌ.*
DIST wall.clock look-NPST-3s
'He looks at the wall clock.' (NNC:book-fiction-kaiphiyat-prativedan-2062.1206)

[2] A consequence of this is that Nepali and other Indo-Aryan languages do not have a uniform alignment pattern even within verb classes (cf. Li 2007b), nor do they have a simple split-system (e.g. ergative perfective vs accusative imperfective). Instead, DAM and DOM represent two orthogonal splits. All claims to the contrary, which consider Nepali to be ergative or accusative as a whole, rely on some notion of basicness (e.g. nominative is basic compared to dative and ergative for some reason).

A-NOM P-DAT:

(31) *Ma her-ch-u mrtyu-lai.*
1s look.at-NPST-1s dead-DAT
'I look at the dead.' (NNC:book-criticism-samakalin-nepali-natak-2057.7036)

There are relatively few formal links between DOM and other areas of the language. This is a direct consequence of its isolation. One link to the rest of the language system would be A: A case marking is conditioned by factors lying outside the referential properties of A itself, such as TMA and typicality in connection with the predicate (section 3.3.3.1), and all Nepali clause chaining mechanisms that have a coreferentiality constraint have an S/A pivot. However, this is irrelevant because DAM is independent of DOM. Similarly, agreement would link DOM to even more areas of grammar but also does normally not interact with it.

3.4.2 Arguments selected by DOM

Most transitive verb classes have exactly one argument that is open to DOM. There are no classes with more than one O argument, but bivalent motion verbs, equational ditransitives, and transitive experiencer predicates have no O at all. The role mapping for the relevant classes is as follows:

- P for monotransitive verbs,
- T for transfer ditransitives (all subclasses), and
- G for instrumental ditransitives.

The relevant argument can also be determined looking at the cases of its co-arguments. O then covers:

- all P where A case alternates between NOM and ERG,
- all T except instrumental T-ERG, with a high threshold for T-DAT G-DAT, and
- G with T-ERG.

Below is one example for DOM in each class.

Monotransitive:

(32) a. *Us-le sʌrkari jagir roj-y-o.*
DIST-ERG governmental job choose-PST-3s
'He chose a job in the government.' (NNC:book-fiction-atripta-aakash-2061.3970)
b. *Me-ro bibek-le prʌtisod-lai roj-y-o.*
1s-GEN reason-ERG revenge-DAT choose-PST-3s
'My reason chose revenge.' (NNC:book-criticism-samakalin-nepali-natak-2057.4270)

Transfer ditransitive (here IIb):

(33) a. *Byapar-hʌru bãs aph-n-a chetrʌ-ma lʌija-nch-ʌn.*
trader-PL bamboo REFL-GEN-OBL region-LOC take-NPST-3p
'The traders take the bamboo to their own region.'
(NNC:himalkhabarpatrika-2059-01-01.430)
b. *Paʈhʌk-lai es-le soj-ʌi pʌscim Nepal-ko yʌuʈa Tharu bʌsti-ma*
reader-DAT PROX-ERG direct-FOC western Nepal-GEN one.CLF Tharu settlement-LOC
lʌija-nch-ʌ.
take-NPST-3s
'(The book) takes the reader straight away to a Tharu settlement in western Nepal.'
(NNC:himalkhabarpatrika-2059-11-16.228)

Instrumental ditransitive:

(34) a. *Murti-hʌru mathi-baʈʌ pʌrda-le ɖhak-ch-ʌ.*
statue-PL above-ABL$_1$ veil-ERG cover-NPST-3s
'He covers the statues with a veil from above.' (NNC:himalkhabarpatrika-2059-06-16.1913)

b. *Mʌila-ko biyogpʌrũɖa jiwʌn-lai kʌpʌɖa rʌ gʌhna-le dherʌi ɖhak-dʌin-ʌ.*
woman-GEN secluded.suffering life-DAT clothes and jewellery-ERG much cover-NEG.NPST-3s
'He doesn't cover women's life of secluded suffering with a lot of clothes and jewellery.'
(NNC:kantipur-misc-2061-11-24.77)

Transfer ditransitives of subclass III have an alternation between NOM and DAT on G, too. However, this alternation is different from DOM in that it also includes LOC and is conditioned by different functional factors. Roughly, DAT is used with animate G into whose possession T changes, whereas LOC/NOM are used with inanimate G. For the distinction between LOC and NOM, it is useful to assume two prototypes of movement (possibly among others). One, which may be called *routing*, takes place along a specific course with a definite end point. The other, which may be called *searching*, does not have a definite end point, and accordingly the direction and course of movement cannot be construed immediately. Routed G tend to take NOM, whereas searched G tend to take LOC. Below is one typical example for each G case.

(35) *Bʌidesik rojgari-le sʌmaj-lai phaida puryaũ-dʌin-ʌ.*
foreign employment-ERG society-DAT advantage bring-NEG.NPST-3s
'Foreign employment doesn't bring any advantage to society.' (NNC:book-belleletter-nepalma-garibiko-bahas-2061.4012)

(36) *Mʌi-le Kʌncʌnjʌŋgha-lai bimanstʌl purya-eko thi-ẽ.*
1s-ERG Kancanjangha-DAT airport bring-PRFV.PTCP be.there-PST.1s
'I had brought Kancanjangha to the airport.' (NNC:madhuparka-memoir-2060-12.4)

(37) *Ghumʌnte kʌrmʌ-le us-lai espʌʈʌk nepal-k-a dui sundʌr sʌhʌr-ma purya-y-o.*
roaming fate-ERG DIST-DAT this.time Nepal-GEN-OBL two beautiful city-LOC bring-PST-3s
'This time his fate of roaming brought him to two beautiful cities of Nepal.'
(NNC:himalkhabarpatrika-2060-08-16.1084)

These examples should make it clear enough that the G of transfer ditransitives of type III are not open to DOM but display a separate alternation, which is a peculiarity of this class. The additional examples below show that this alternation is independent of DOM, which is found on T in this class. There is one example for each combination of T and G cases for the verb *pʌʈhau-* 'send'.

T-NOM G-NOM:

(38) *Yʌs bheg-le bemʌusʌm-k-a bela pʌni samanne pãc ʈrʌk-ko harahari-ma golbheɖa dʌinik rajdhani pʌʈhaũ-ch-ʌ.*
PROX.OBL area-ERG bad.weather-GEN-OBL time also generally five truck-GEN average-LOC tomato daily capital send-NPST-3s
'Even during bad weather, this area normally sends an average of five trucks of tomatoes to the capital.' (NNC:book-academic-swasthya-samaj-ra-rajniti-2062.4200)

T-NOM G-LOC:

(39) *Purus-le dut-hʌru gaũ-ma paʈha-y-o.*
man-ERG messenger-PL village-LOC send-PST-3s
'The man sent messengers (in)to the village.' (NNC:book-fiction-utsarga-prem-2058.460)

T-NOM G-DAT:

(40) *Mʌhendrʌpulisklʌblelʌdalʌtlaijʌwaphpʌʈhayo.*
Mahendra police club-ERG court-DAT answer send-PST-3s
'The Mahendra Police Club sent an answer to the court.' (NNC:himalkhabarpatrika-2060-08-16.575)

T-DAT G-NOM:

(41) *Timi pʌni tʌ manche-lai pʌrʌmdham pʌʈhaũ-ch-ʌu ni.*
2s also CTOP person-DAT heaven send-NPST-2s ASS
'You also send people to heaven.' (NNC:book-drama-prempinda-2058.5017)

T-DAT G-LOC:

(42) *ʄʌs-lai hami ramro swastekʌrmi-ko rup-ma gaũ-ma pʌʈhaũ-dʌi*
who.REL-DAT 1p good health.worker-GEN form-LOC village-LOC send-PROG
thi-y-ʌũ…
be.there-PST-1p
'Whoever we were sending to the village as a good health worker…'
(NNC:book-academic-swasthya-samaj-ra-rajniti-2062.2663)

T-DAT G-DAT:

(43) *Ram-le aph-no nokʌr-lai Sita-lai pʌʈhay-o.*
Ram-ERG REFL-GEN servant-DAT Sita-DAT send-PST-3s
'Ram sent his own servant to Sita.' (elicitation SAR 2011)

Even though the dative found on the G of transfer ditransitive verbs is functionally different from the dative found on O, this use is – out of all the numerous uses of the dative (cf. section 3.5.1) – probably the one that is most similar to DOM. As mentioned above, G-DAT is mainly found with animate G; and as mentioned earlier in section 3.3.2.5, all verbs that semantically require an animate G are in the transfer ditransitive class. What's more, the G of subclass Ia, which are classical recipients without exception and where DAT is obligatory, are usually not only highly animate but also specific and topical, properties that place them very closely to O-DAT (cf. section 3.5). Thus, very similar factors motivate the alternation DOM and the lexicalised G-DAT of this subclass.

3.4.3 Position of the marker

Like all case markers in Nepali, *-lai* attaches to NPs as a whole, not to single nouns. When the relevant NP is complex, *-lai* marks the rightmost element:

(44) *Ajʌratimʌilesʌpnamasano(*-lai)bʌcca(-lai)dekh-ẽ.*
today at.night 1s-ERG dream-LOC small-DAT child-DAT see-PST.1s
'Last night I saw a small child in a dream.' (elicitation GP 2010)

This is also true of more complex NPs containing several simple NPs joined by conjunctions like *rʌ* 'and' or *ki* 'or':

(45) *Tini-hʌru-le Ram(*-lai)rʌ us-ko kukur(-lai)dekh-e.*
MED.MH-PL-ERG Ram-DAT and DIST-GEN dog-DAT see-PST.3p
'They saw Ram and his dog.' (elicitation KP 2012)

Interestingly, although the position of the DAT marker is a purely formal constraint, it seems to be the reason why some speakers give greater weight to the last element of a complex NP when it comes to determining case. This can be seen in groups where the referential properties of the joined NPs point into different directions, e.g. in terms of animacy (cf. section 3.5.3):

(46) a. *Tini-hʌru-le Ram rʌ yʌuʈa gaɖi(?-lai) dekh-e.*
MED.MH-PL-ERG Ram and one.CLF car-DAT see-PST.3p
'They saw Ram and a car.'
b. *Tini-hʌru-le yʌuʈa gaɖi rʌ Ram*(-lai) dekh-e.*
MED.MH-PL-ERG one.CLF car and Ram-DAT see-PST.3p
'They saw a car and Ram.' (elicitation KP 2012)

3.4.4 Double datives

Nepali has a strong preference against more than one dative in a clause. This is evidenced by three phenomena. The first is rather trivial: the O of predicates whose A is marked by DAT can carry various cases but never another DAT. They are thus excluded from DOM. This mostly concerns experiencer predicates:

(47) *Us-lai timi-sʌŋgʌ/*-lai ris uʈh-y-o.*
DIST-DAT 2s-COM/-DAT anger rise-PST-3s
'He is angry with you.' (elicitation NP 2012)

(48) *Mʌ-lai un-ko/*-lai yad a-y-o.*
1s-DAT 3sMH-GEN/-DAT remembrance come-PST-3s
'I remembered her.' (elicitation NP 2012)

Another frequent case is the verb *cahi-* 'need', a lexicalised passive of *caha-* 'want'. 'I need it' is literally expressed as 'it is wanted to me', with A marked by DAT and P marked by NOM. DAT on O is impossible, independently of whether A-DAT is covert or not:

(49) a. *Tehi ʌusʌdi(*-lai) mʌ-lai cahi-nthy-o.*
MED.FOC medicament-DAT 1s-DAT need-PST.HAB-3s
'I would have needed precisely that medicament.'
b. *Tehi ʌusʌdi(*-lai) cahi-nthy-o.*
MED.FOC medicament-DAT need-PST.HAB-3s
'Precisely that medicament would have been needed.' (elicitation KP 2012)

Second, when DAT is used to mark the A of a deontic expression (cf. section 3.5.1 below), DAT becomes doubtful on the P (50). When there is a G that is obligatorily marked by DAT, deontic A-DAT is not possible at all (51).

(50) *Ram-lai tyo manche(?-lai) bheʈ-nu pʌr-ch-ʌ.*
Ram-DAT MED person-DAT meet-INF_1 fall-NPST-3s
'Ram has to meet that person.' (elicitation NP 2012)

(51) *Ram-le/*-lai ciʈʈhi Sita-lai di-nu pʌr-ch-ʌ.*
Ram-ERG/-DAT letter Sita-DAT give-INF_1 fall-NPST-3s
'Ram has to give the letter to Sita.' (elicitation NP 2012)

Third, although in principle all T can be marked by DAT, the threshold for this is much higher when there is a fixed G-DAT in the same clause. Although the claim made by Gupta and Karmacharya (1981) as well as by Li (2007b) that DAT is impossible on "indirect objects" (that is, T of transfer ditransitives of class Ia) is not true, T-DAT G-DAT frames are extremely rare. Here is an elicited example for *sikau-* 'teach':

(52) *Yo tʌrika goppe h-o. Es-lai kʌs-ʌi-lai pʌni sik-au-nu*
PROX technique secret be[NPST]-3s PROX-DAT who-FOC-DAT also learn-CAUS-INF_1
hũ-dʌin-ʌ.
be.okay-NEG.NPST-3s
'This technique is secret. It must not be taught to anyone.' (elicitation SAR 2011)

T-DAT G-DAT is also attested in natural discourse, but only in highly complex sentences with a lot of intervening material between T and G. In (53), both the G (*sʌrbʌsadhʌrʌn* 'common people') and the postverbal T (*yo kura* 'this matter') of *bujhau-* 'explain' are marked by DAT.

(53) *Kʌsʌri bujh-au-ne sʌrbʌsadhʌrʌn-lai tyo muluki ʌin-ko*
Q.METHOD understand-CAUS-IPFV.PTCP common.people-DAT MED national law-GEN
eghar-ʌũ sʌmsodhʌn-le di-eko sʌmpʌti-mathi-ko kehi ʌdhikar rʌ ʌbo
eleven-ORD amendment-ERG give-PRFV.PTCP property-SUPER-GEN some right and now
yo kura-lai?
PROX matter-DAT
'Now how to explain to the common people the couple of rights concerning property that the eleventh amendment to the national law has given us and (how to explain) this (other) matter?' (NNC:A001017001.73)

The usual solution to avoid a double dative is to assign DAT only to G and mark T by NOM. Only T which rank very high with respect to the referential factors relevant to DOM can overcome this restriction and get DAT even in the presence of another DAT. The questions of which functional factors favour DAT are dealt with in detail in section 3.5.

Beside this there is another solution. The ban is against two overt datives in one clause, so T-DAT becomes much less marked as soon as G is covert, even if G would have been marked by DAT if it had been overt. This is illustrated in (54) with the verb *bec-* 'sell':

(54) a. *Hijo hami-le ham-ro gai(-lai) bec-y-ʌũ.*
yesterday 1p-ERG 1p-GEN cow-DAT sell-PST-1p
'Yesterday we sold our cow.'
b. *Hijo hami-le ham-ro gai(*-lai) Ram-lai bec-y-ʌũ.*
yesterday 1p-ERG 1p-GEN cow-DAT Ram-DAT sell-PST-1p
'Yesterday we sold our cow to Ram.' (elicitation BP/KP 2012)

Occasionally a dislike for double datives can be observed across clause boundaries, too. Consider:

(55) *Mʌi-le tʌpaĩ-lai bhʌn-eko manche bheṭ-nubhʌyo?*
1s-ERG 2HH-DAT tell-PRFV.PTCP person meet-PST.2/3HH
'Have you met the person I told you about?' (elicitation GP/SAR 2010)

I had expected *manche* as highly animate and definite to be marked by DAT. When I asked the speaker why she didn't say *manche-lai* she corrected herself and said that DAT was actually better. Since this sounded like a trace of prescriptive school grammar to me, I tried the same sentence on another speaker and again got NOM as the first suggestion. When I asked this speaker why she hadn't said *manche-lai*, she said that it sounded odd to have so many *-lai* in one sentence.

The NNC also contains sentences where a tendency against more than one dative within a certain spread of words seems to be the only explanation for NOM. For instance, in (56) the DAT on *pani* is more or less well motivated by contrastive focus (see section 3.5.11). However, if focus is at work here, there is no obvious reason why the other NP involved in the contrast, *nun rʌ cini*, should be marked by NOM, except that this avoids having too many datives in one sentence:

(56) *Pani-lai tehã nʌi nap-i ʌbhilekh gʌr-i-eko*
water-DAT MED.LOC FOC measure-CVB_2 documentation do-PASS-PRFV.PTCP
thi-y-o bhʌne nun rʌ cini caĩ poko par-i prʌyogsala-ma
be.there\be.there-PST-3s although salt and sugar RETRV bundle make-CVB_2 laboratory-LOC
lya-erʌ mapʌn rʌ bislesʌ̃ḍ gʌr-i-eko thi-y-o.
take-CVB_1 measurement and analysis do-PASS-PRFV.PTCP be.there\PST-PST-3s
'Whereas water had been measured and documented in the same place, salt and sugar had been packaged and measured and analysed after taking them to a laboratory.'
(NNC:book-academic-swasthya-samaj-ra-rajniti-2062.2272)

3.4.5 The question of incorporation

In the literature on DOM in Hindi, certain O-NOM are frequently referred to as "incorporated" since Mohanan (1994). There is the question whether such objects exist in Nepali, too, and if so, how this construction is related to DOM.

The term "incorporation" as used in the literature on Hindi is unfortunate for two reasons. First, this construction is very different from well-known examples of incorporation in polysynthetic language families such as Algonquian or Eskimo. Most importantly, there is no evidence for univerbation on the syntactic level. In Hindi, an incorporated object still triggers agreement in the perfective tenses, and the agent is still marked by *-ne* [ERG]. In Nepali, all finite verb forms ever only agree with S/A except in the passive, but the syntactic transitivity of clauses with "incorporated" objects is still visible in the A-ERG. (57) shows an example from Mohanan and its translation into Nepali.

(57) a. *Anil-ne kitaab-ẽ bec-ĩĩ.*
Anil(m)-ERG book(f)-PL sell-PRFV.PTCP.pf
'Anil sold books.' / 'Anil did book-selling.' (Mohanan 1994:106)

b. *Anil-le kitab bec-yo.*
Anil-ERG book sell-PST.3s
'Anil sold books.' / 'Anil did book-selling.' (elicitation NP 2012)

A full list of the features Mohanan claims to be characteristic of incorporation in Hindi is shown below.

- the object noun (O) has "generic" reference[3]
- no modifiers are allowed on O
- no material can intervene between O and the verb (V)
- O and V share a single intonation contour
- O cannot be gapped
- O cannot be conjoined with another noun
- V cannot be passivised

Looking at this list, another problem about incorporation becomes apparent: Mohanan mixes formal (phonological and morphological) and functional criteria, and the latter remain vague and apodictic. At least for Nepali, these two kinds of criteria do not always go together. For instance, the examples in (58) below have type reference but violate the constraints on modification and intervening material. Intonation seems to be yet another independent factor, because whereas O and V may (but need not) share a single contour in (58a), they may not do so in (58b).

(58) a. *Us-le purano kitab bec-ch-ʌ.*
DIST-ERG old book sell-NPST-3s
'He does old-book-selling.'

b. *Us-le kitab nʌyã bʌjar-neri bec-ch-ʌ.*
DIST-ERG book new market-near sell-NPST-3s
'He does book-selling near the new market.' (elicitation NP 2012)

On the other hand, gapping and conjoining seem to preclude a type reading in Nepali, too, as shown in (59). In neither of the two sentences may O and V lie under a single intonation contour, so in this case all factors point into a single direction:

(59) a. *Us-le kitab bec-ch-ʌ ʌni Ram-le kin-ch-ʌ.*
DIST-ERG book sell-NPST-3s and Ram-ERG buy-NPST-3s
'He sells books and Ram buys them', but *'He does book-selling and Ram does buying.'
(elicitation NP 2012)

[3]Mohanan uses this term in a rather narrow sense. Generic reference for her seems to be identical to type reference.

b. *Us-le kitab rʌ philim bec-ch-ʌ.*
DIST-ERG book and film sell-NPST-3s
'He sells books and films', but *'He does book-and-film-selling.' (elicitation NP 2012)

However, when the NP in O consists of several nouns but no conjunction is used between them, a single intonation contour becomes possible again, as shown in (60):

(60) *Ram-le bhai bʌini kuʈ-ch-ʌ.*
Ram-ERG younger.brother younger.sister beat-NPST-3s
'Ram does brother-sister-beating (his own or others).' (elicitation NP 2012)

These examples show that in Nepali there is no simple feature cluster of the type claimed for Hindi by Mohanan (1994). To be sure, there are O that display all the defining features. The problem rather is that none of these features is fully dependent on the others, so one category to cover them all is not of much descriptive use. Clearly more research is needed in this area.

The question of incorporation is not highly relevant for DOM. So far it does not seem like incorporation exists in Nepali, but even if it does, it can be viewed as a subtype of O-NOM. The only question of interest would be whether the (assumed) incorporated objects can be interpreted as the next stage beyond ordinary O-NOM. It is conceivable, for instance, that specific O should get DAT, non-specific O get NOM but retain their stress, and yet less specific or "generic" O (possibly with additional formal characteristics) get NOM and lose their stress. This question was not investigated in the present work because of a methodological dilemma: intonation contours are easy to hear in oral elicitation, but degrees of specificity and related semantic distinctions are extremely difficult to elicit. On the other hand, degrees of specificity could have been inferred from corpus annotations, but the NNC only contains written resources.

The remainder of this section shows a few examples that illustrate the (trivial) fact that O-NOM do not have any inherent properties that make them similar to incorporated O in a narrow sense, i.e. they do not form a syntactic unit with the verb form that governs them. First of all, O-NOM do not have to be adjacent to a verb form. Compare (61a) and (61b):

(61) a. *Saikʌl cal-au-ne baʈo ch-ʌin-ʌ.*
bike move-CAUS-IPFV.PTCP road be.there-NEG.NPST-3s
'There is no road for riding bikes.' (NNC:nispaksha-interview-2061-11-04.34)
b. *Yo saikʌl mʌ cʌl-aũ-ch-u.*
PROX bike 1s run-CAUS-NPST-1s
'I ride this bike.' (NNC:madhuparka-fiction-2060-10.210)

O-NOM can also be modified, e.g. by the demonstrative *yo* [PROX] in (61b). They can thus head complex NPs. Other modificators are also possible, for instance, adjectives (62a), numerals (62b), or relative clauses (62c).

(62) a. *Hijo rati mʌi-le sʌpna-ma sano bʌcca dekh-ẽ.*
yesterday at.night 1s-ERG dream-LOC small child see-PST.1s
'Last night I dreamt of a small child.'
b. *Hijo rati mʌi-le ek hʌjar-wʌʈa hatti dekh-ẽ.*
Yesterday at.night 1s-ERG one thousand-CLF elephant see-PST.1s
'Last night I dreamt of one thousand elephants.'
c. *Ɉun pʌni nam sun-ne pustʌk pʌɖ-ne gʌr-ch-ʌ.*
whichever also name hear-IPFV.PTCP book read-IPFV.PTCP do-NPST-3s
'He's in the habit of reading any book he hears about.' (elicitation GP 2010)

A third standard question – whether O-NOM can be covert or not – is irrelevant for Nepali because it is impossible to determine the case of a covert referent. What may be said, though, is that both highly specific and non-specific referents can be covert, as shown in (63).

(63) a. *Goʈhalo-baɖʌ phʌrkʌ̃-dakheri saph pani-le cuʈ-e-ch-ʌ.*
cowhearding-ABL$_1$ return-CVB$_5$ frank rain-ERG thrash-PRF-NPST-3s
'When he returns from cowhearding, the hard rain has thrashed him.'
(NNC:A001011002.296-298)

b. *ʈelibhijʌn-ma gau-nʌ-kolagi matrʌi gaek bʌn-ni*
television-LOC sing-INF$_2$-FIN$_1$ only singer become-IPFV.PTCP
'to become a singer just for singing (songs) on TV' (NNC:A001013001.905)

3.4.6 DOM and agreement

Many Indo-Aryan languages have a rule stating that only NOM-marked arguments can trigger agreement. This is, for instance, the case in Hindi (Gair and Wali 1989, Butt 1993, Mohanan 1994), where the transitive agreement pattern in imperfective tenses is very different from that found in perfective tenses: the former have the frame {A-NOM O-NOM/DAT V-s(A)}, whereas in the latter A is marked by ERG so that AGR gets either relinked to O-NOM, yielding {A-ERG O-NOM V-s(O)}, or to a dummy 3s when both arguments are non-NOM, yielding {A-ERG O-DAT V-s(3s)}.

In Nepali, only NOM- and ERG-marked S/A may and must trigger agreement (Bickel and Yadava 2000:348)[4] so that DOM does not play a big role for agreement. There are two minor contexts, however, where O does get linked to AGR and where accordingly DOM can have an effect on it.

The first of these is the passive (section 3.3.3.2), where the agreement system works similar to what has just been described for Hindi. When the O of a passive is marked by NOM it gets linked to AGR. This is not possible when O is marked by DAT, and since A is also non-NOM if it is expressed overtly at all (the possible markers are *-dwara* 'by' and *-baʈʌ* [ABL$_1$]), AGR can only be set to a dummy 3s.

The examples below illustrate O-NOM (64a) and O-DAT (64b) in natural discourse. (65) shows the connection between case marking and agreement.

(64) a. *Mʌ ʌhile pʌkr-i-ẽ bhʌne santi-sitʌ kʌile bheʈ gʌr-ne?*
1s now arrest-PASS-PST.1s COND peace-COM$_2$ when meeting do-IPFV.PTCP
'If they arrested me now, how should I meet with peace?'
(NNC:book-fiction-shanti-2058.2469)

b. *Akhir, bises ʌdalʌt-ma mʌ-lai lʌg-i-y-o.*
finally special court-LOC 1s-DAT take-PASS-PST-3s
'Finally I was taken to a special court.' (NNC:himalkhabarpatrika-2060-08-01.1173)

(65) a. *Iskul-ma mʌ*(-lai) kuʈ-ch-ʌn.*
school-LOC 1s-DAT beat-NPST-3p
'In school they beat me.'

b. *Iskul-ma mʌ(*-lai) kuʈ-in-ch-u.*
school-LOC 1s beat-PASS-NPST-1s
'I am beaten in school.'

c. *Iskul-ma mʌ*(-lai) kuʈ-in-ch-ʌ.*
school-LOC 1s-DAT beat-PASS-NPST-3s
'I am beaten in school.' (elicitation GP 2010)

The function of DAT in the passive is identical to that in the active, but the threshold for using it seems to be much higher. The passive also overrides the few hard rules that require DAT, for instance the rule that personal pronouns in O must be marked by DAT (cf. section 3.5.8 below). The examples in (65) show that even a first person singular pronoun can be marked by NOM in the passive.

Kärkkäinen (1994) notes that the passive is used with inanimates in about 88.9% of all cases. Since inanimates appear almost always in the third person and plural agreement with third persons

[4] This is of course only true if one assumes that O becomes S in the passive. In a purely semantic role system such as the one used here, O stays O in the passive because valency doesn't change – the A can still be expressed. The rule would then have to be rephrased as "Only NOM- and ERG-marked arguments can trigger agreement in Nepali".

is optional in Nepali, that means that it is in most cases impossible to decide whether a passive verb agrees with O or whether there is dummy 3s-AGR. Only 2.3% of the passive verb forms in Kärkkäinen's sample unambiguously agree with O.

Another, even more marginal construction where there is a link between DOM and agreement is described by Wallace (1985). This is optional raising of embedded S/A to matrix O with verbs of perception. As long as the embedded S/A gets its case marking from the predicate of the embedded clause, it may optionally trigger AGR as in (66a) and (67a). However, as soon as it raises to the matrix and is marked by DAT it loses this property, as in (66b) and (67b):

(66) a. *Ram(-le) Sita bahirʌ-baʈʌ a-i-rʌh-ek-i sun-ch-ʌ.*
Ram-ERG Sita outside-ABL$_1$ come-LNK-CONT-PST.PTCP-F hear-NPST-3s
'Ram hears how Sita comes from outside.' (Wallace 1985:89)

b. *Ram(-le) Sita-lai bahirʌ-baʈʌ a-i-rʌh-eko sun-ch-ʌ.*
Ram-ERG Sita-DAT outside-ABL$_1$ come-LNK-CONT-PRFV.PTCP hear-NPST-3s
'Ram hears Sita coming from outside.' (Wallace 1985:90)

(67) a. *Timi(-le) uni-hʌru-le phuʈbʌl khel-i-rʌh-ek-a dekh-ʌu-la.*
2MH-ERG DIST-PL-ERG football play-LNL-CONT-PST.PTCP-PL see-2MH-PROB.FUT
'You will see how they play football.' (Wallace 1985:89)

b. *Timi(-le) unihʌru-lai phuʈbʌl khel-i-rʌh-eko dekh-ʌu-la.*
2MH-ERG DIST-PL-DAT football play-LNL-CONT-PRFV.PTCP see-2MH-PROB.FUT
'You will see them playing football.' (Wallace 1985:90)

3.4.7 DOM in complex predicates

Complex predicates in Nepali can be defined parallel to Chintang (cf. section 2.6.5.3): they consist of a noun ("N") coding a state of affairs combined with a light verb (mostly *gʌr-* 'do'), which together can be viewed as a single semantic predicate. Most abstract notions in Nepali can only be expressed as complex predicates, so this construction is very frequent, especially in the written language. The inventory of complex predicates is likely to be several hundred times larger than in Chintang. Below is a first example.

(68) *Murkhʌ manche-lai kʌs-ʌi-le pʌni adʌr gʌr-dʌin-ʌn.*
idiotic person-DAT who-FOC-ERG also respect do-NEG.NPST-3p
'Nobody respects an idiot.' (NNC:book-popularlore-balsukti-2061:864)

Morphosyntactically *adʌr* 'respect' clearly is a noun, as shown in (69), where it is modified by an adjective and marked by ERG:

(69) *Mʌ-lai bhʌemisrit adʌr-le her-ne gʌr-th-e.*
1s-DAT mixed.with.fear respect-ERG look-IPFV.PTCP do-PST.HAB-3p
'They used to look at me with respect mixed with fear.' (NNC:book-autobiography-mero-aviral-jivangit-2060.2495)

Still, *adʌr* does not behave like an argument in (68). Instead, *adʌr gʌr-* looks like a single predicate with two arguments, an A (*kʌsʌile*) and a P (*manchelai*).

The questions regarding the relation between complex predicates and DOM are similar as for S/A detransitivisation. The possibility of O-AGR corresponds to the possibility of marking N by DAT. The question of whether an additional argument besides N is allowed is the same. An additional question is whether A gets marked by ERG or not. In Chintang this factor is tied up with O-AGR, but in Nepali A and O marking are independent of each other (section 3.4.1), so this point is worth looking at. Further, since the question is relevant for Nepali whether some O may be considered as incorporated (see section 3.4.5), the same may also be asked for complex predicates.

So far I have not come across any DAT-marked N in natural spoken Nepali. Corpus searches over the NNC for about a dozen highly frequent N with the V *gʌr-* 'do' also did not yield any matches. N-DAT is sometimes marginally possible in elicitation but never preferred over N-NOM.

This makes it very different from O-AGR with N in Chintang, which likewise was not the default but well possible with many complex predicates, and justifies treating complex predicates as a formal rather than a functional factor in Nepali DOM. One of the rare cases where an N-DAT could be elicited is shown in (70), where DAT is marginally acceptable on *yatra* 'journey':

(70) *Us-le bis bʌrsʌ-ʌgaḍi bihe gʌr-e-pʌchi gʌr-eko yatra(?-lai) pheri gʌr-y-o.*
DIST-ERG twenty year-ANTE marriage do-NMLZ-TMP.POST do-PRFV.PTCP journey-DAT again do-PST-3s
'He made the (same) journey again that he had made after marrying.'
(elicitation BP/KP 2012)

The possibility of DAT seems to depend on whether N can be construed as an independent referent of which several instances can be easily separated and identified with each other. But this alone is not sufficient for DAT. (71) shows an example of a referential N which has properties that are typical of O-DAT (highly specific, marked by one of the focal demonstratives that are otherwise frequently used with DOM – cf. 3.5.11) but which still can only be marked by NOM.

(71) *Us-le tehi kam(*-lai) gʌr-y-o.*
DIST-ERG MED.FOC work-DAT do-PST-3s
'He did that very same work.' (elicitation BP/KP 2012)

Also note that the reason why *kam* cannot be marked in (71) is not that it codes a process. The very same noun can be marked by DAT in sentences such as (72), confirming once more that the disfavouring of N-DAT in complex predicates is a formal factor:

(72) *Hʌusʌla, prerãḍa rʌ ramro kam-lai prʌsãsa gʌr-ne gʌr-nu pʌr-ch-ʌ.*
encouragment motivation and good work-DAT praise do-IPFV.PTCP do-INF_1 be.necessar-NPST-3s
'One should encourage and motivate them and praise good work.'
(NNC:sadhana-psychology-2061-10.105)

Interestingly, in spite of the near-ungrammaticalness of N-DAT, the A of a complex predicate must always be marked by ERG in perfective tenses, no matter whether there is an O-like argument besides N or whether N itself is P. This is even the case with N that are minimally referential and have no chance of ever becoming the head of an expanded NP, such as *hʌttar* 'hurry' in (73):

(73) *Us*(-le) hʌttar gʌr-y-o.*
DIST-ERG hurry do-PST-3s
'He hurried.' (elicitation BP/KP 2012)

This shows that A and O case marking react to very different criteria. The fact that complex predicates like *hʌttar gʌr-* require A-ERG is a strong argument for analysing N as a special kind of argument (with role = P) even when it is marked by NOM, at least as long as there is no additional object-like argument. Where such an argument is present as in (74), N-DAT is never even marginally grammatical, although there are no restrictions other than the usual ones on the case of the additional P:

(74) Us-le tyo misin(-lai) nas(*-lai) gʌr-y-o.
DIST-ERG MED machine-DAT destruction-DAT do-PST-3s
'He destroyed that machine.' (elicitation NP 2012)

Both complex predicates with an additional P and where N itself is P do not show characteristics of noun incorporation. Words can intervene between N and V (75a), the order of N and V can be reversed (75b), and V can be gapped when it has just been mentioned (75c). Examples where N is the head of a complex NP were already shown above, e.g. in (70).

(75) a. *Chʌlphʌlmatrʌigʌrchʌn.*
discussion only do-NPST.3p
'They only discuss.'
b. *Kaʈhmanḍu-ma gʌrchʌ, kam.*
Kathmandu-LOC do-NPST-3s work
'He works in Kathmandu.'
c. *Tʌrʌʌḍkʌlrʌnapgʌrne kam phʌrʌkchʌ.*
but estimation and measure do-IPFV.PTCP work different be-NPST-3s
'But estimating and measuring are different.' (elicitation SAR 2011)

A special subgroup of complex predicates is constituted by etymologically related N-V combinations. In these predicates N is truly semantically empty because exactly the same meaning is also coded by V. Nevertheless, the behaviour of predicates with *figura etymologica* is identical to that of other complex predicates: DAT is mostly ungrammatical as in (76) but rarely possible with highly referential N as in (77). ERG is obligatory on A, as shown in (76).

(76) *Hami*(-le) tehi khel(*-lai) khel-thy-ʌũ.*
1p-ERG MED.FOC game-DAT play-PST.HAB-1p
'We used to play that very same game.' (elicitation BP/KP 2012)

(77) *Prithbi Naraĕḍ Sa-ko jibʌn(?-lai) ʌrko manche-le jiu-nʌ sʌk-dʌinʌ-thy-o.*
Pṛthvī Nārāyaṇ Śāha-GEN life-DAT other person-ERG live-INF_2 be.able-NEG-PST.HAB-3s
'Another person couldn't have lived the life of Pṛthvī Nārāyaṇ Śāha.' (elicitation BP/KP 2012)

Many N take cannot only take *gʌr-* 'do' as their V but also the copula *hu-*. The effect is a passive: A is removed completely so that it cannot be re-introduced by *-dwara* 'by', and verbal agreement is re-linked to O. N does not change its shape in this process. Below is an example for *nikasi* 'export' (*nikasi gʌr-* 'export', *nikasi hu-* 'be exported').

(78) a. *Sʌhʌkarileʲʌḍibuʈi khʌrid gʌri Bharʌt nikasi gʌrchʌ.*
cooperative-ERG herbs purchase do-CVB_1 India export do-NPST-3s
'The cooperative purchases herbs and exports them to India.'
(NNC:kantipur-business-2061-12-20.27)
b. *Nepali gai, bhʌĩsi-ko posilo dudh Bharʌt nikasi bhʌyo.*
Nepalese cow buffalo-GEN nutritious milk India export happen-PST-3s
'The nutritious milk of Nepalese cows and buffalos was exported to India.'
(NNC:himalkhabarpatrika-2061-01-01.696)

Differently from the morphological passive discussed in section 3.4.6, the light verb passive cancels the possibility of DOM. This can be attributed to the fact that while the morphological passive keeps A in the valency, the light verb passive completely removes it so that O truly becomes S. (79) shows contrasting examples for a sentence in the active and in the two passives with the complex predicate *khʌtʌm gʌr-* 'ruin' / *khʌtʌm hu-* 'be ruined':

(79) a. *Rasʈrʌkhʌtʌm gʌrchʌn.*
state-DAT end do-NPST-3p
'They ruin the state.'
b. *Rasʈrʌkhʌtʌm gʌrinchʌ.*
state-DAT end do-PASS-NPST-3s
'The state is (being) ruined.'
c. *Rasʈrʌkhʌtʌm hunchʌ.*
state-DAT end become-NPST-3s
'The state gets ruined.' (elicitation SAR 2011)

Another peculiarity of complex predicates is that when there is an O besides N, this O may frequently not only be marked by NOM or DAT but also by GEN (i.e. as the possessor of N). An

example of NOM alternating with GEN is shown in (80) below. Since O-GEN are never possible with simplex predicates, they will be ignored in the remainder of this work.

(80) *Nepal-ma ciya(-ko) utpadʌn dherʌi gʌr-i-nch-ʌ.*
Nepal-LOC tea-GEN production much do-PASS-NPST-3s
'A lot of tea is produced in Nepal.' (elicitation SAR 2011)

In summary, complex predicates in Nepali work similarly to those in Chintang (see section 2.6.5.3). In the case of Nepali, the main argument for treating N as P in the absence of other non-S/A arguments is the obligatoriness of ERG on A. Independently of that, the threshold for marking N with DAT is much higher than the threshold for linking N to O-AGR, to the extent that N-DAT is so far unattested in corpus data. Complex predicates with *gʌr-* 'do' as their V can frequently also use *hu-* 'be' instead. An O besides N becomes S in this process and accordingly cannot be marked by DAT any longer.

3.5 Functional properties of DOM

3.5.1 Uses of the dative

The marker *-lai* appears in a number of diverse functions, which are summarised below. *-lai* marks:

- O in DOM, that is, P of monotransitive verbs, T of transfer ditransitives, and G of instrumental ditransitives
- G of transfer ditransitives, with differences between the subclasses:
  - I: obligatory
  - IIa: alternating with LOC
  - III: alternating with LOC/NOM
- S and A of experiencer expressions
- beneficiaries
- S and A in deontic constructions
- S and A of various idiosyncratic verbs

The use of *-lai* that is central for this work is of course the one in DOM. This is illustrated once more in (81), where the personal pronoun *hami* [1p] features as P and is marked by *-lai*:

(81) *Tʌpaĩ-k-ʌi prʌhʌri-le hami-lai sʌtaũ-ch-ʌ.*
2/3HH-GEN-FOC police-ERG 1p-DAT trouble-NPST-3s
'Your policemen trouble us.' (NNC:himalkhabarpatrika-2059-10-01.1079)

Another frequent use of the dative is on G. This use also participates in alternations, which are, however, independent of DOM (cf. sections 3.3.2.5, 3.4.2). Below is an example for a transfer ditransitive of subclass Ia, where only DAT is allowed on G:

(82) *Belaet-le dui-ʈa miliʈeri helikʌpʈʌr Nepal-lai bec-y-o.*
U.K.-ERG two-CLF military helicoptar Nepal-DAT sell-PST-3s
'The U.K. sold two military helicopters to Nepal.'
(NNC:book-belleletter-nepalma-garibiko-bahas-2061.2450)

Most experiencers are also marked by the dative. Differently from DOM, there is a surprising amount of dedicated literature on dative experiencers in Nepali; see Gupta and Tuladhar (1979), Ichihashi-Nakayama (1994), Ghimire (2002), Bickel (2004b). The most frequent verbs in experiencer expressions are *hu-* 'be, be there, become' and *lag-* 'be on, be attached to, be there', as in (83):

(83) *Mʌ-lai khusi lag-y-o.*
1s-DAT happiness be.on-PST-3s
'I am happy.' (NNC:nepal-story-2062-11-30.xml.143)

Less frequently, the dative can mark beneficiaries which are not part of the frame of the verb. More usual ways to express such beneficiaries are to mark them by the final case *-(ko)lagi* or to increase valency by using the benefactive vector verb *-di*, which introduces a dative-marked beneficiary that can be mapped to G. Accordingly, sentences such as the following have so far not been observed in the NNC and are not accepted by all speakers:

(84) *Bikrʌm-le chora-lai ghʌr bʌn-a-y-o.*
Bikram-ERG son-DAT house be.built-CAUS-PST-3s
'Bikram built a house for his son.' (Paudyal 2009:15)

A use of the dative that is akin to benefactives but more common is to mark affected possessors:

(85) *Bimla-lai ākha-ma dhulo pʌr-y-o.*
Bimla-DAT eye-LOC dust fall-PST-3s
'Dust got into Bimla's eye.' (Adhikārī 2052 V.S.:97)

An argument class that has particularly variable case marking is the S/A of deontic expressions. Intransitive S and A of bivalent motion verbs can be marked by NOM or DAT, and transitive A can be marked by ERG or DAT. (86) and (87) show examples for all possible cases including S-DAT and A-DAT.

(86) a. *ʌbʌ mʌ phʌrkʌ-nu pʌr-ch-ʌ.*
now 1s return-INF$_1$ fall-NPST-3s
'Now I have to return.' (NNC:madhuparka-fiction-2060-08.166)

b. *Pʌti-lai jel ja-nu pʌr-y-o rʌ duniyã-ko-samu un-lai ʌpʌmanit hu-nu*
husband-DAT jail go-INF fall-PST-3s and world-GEN-before 3MH-DAT insulted COP-INF
pʌr-y-o.
fall-PST-3s
'Her husband had to go to jail and she had to tolerate being insulted before the world.'
(NNC:book-fiction-sanghu-tarepachhi-2062.2302)

(87) a. *Swasthe sewa sʌb-ʌi nagʌrik-le pau-nu pʌr-ch-ʌ.*
health service all-FOC citizen-ERG get-INF$_1$ fall-NPST-3s
'All citizens have to get health care.'
(NNC:book-academic-swasthya-samaj-ra-rajniti-2062.4814)

b. *Sãguro ghʌr-ma rani-lai aph-no sʌrir khumcy-au-nu pʌr-ch-ʌ.*
narrow house-INF queen-DAT REFL-GEN body be.bent-CAUS-INF$_1$ fall-NPST-3s
'In the narrow house the queen has to bend her body.'
(NNC:book-belleletter-mauripalan-2062.338)

Finally, there are various scattered verbs and multi-word expressions that require the dative on S or A. These are partially highly frequent but do not fit into any of the existing classes. Below is an example for a verb (*cahi-* 'be needed, be necessary', lexicalised passive of *caha-* 'wish, want, need'), for a combination of verb and noun (*tha hu-* [knowledge be.there] 'know'), and for a combination of verb and adjective (*sʌnco hu-* [well be] 'be well, be fine'):

(88) *Hami-lai uni-hʌru cahi-nch-ʌn rʌ uni-hʌru-lai hami-hʌru cahi-nch-ʌũ.*
1p-DAT DIST-PL be.needed-NPST-3p and DIST-PL-DAT 1p-PL be.needed-NPST-1p
'We need them and they need us.' (NNC:book-fiction-ek-chihan-2056.2440)

(89) *Yo kura sʌbʌi grahʌk-hʌru-lai tha ch-ʌ.*
PROX thing all customer-PL-DAT knowledge be.there.NPST-3s
'All customers know this.' (NNC:himalkhabarpatrika-2059-11-01.3152)

(90) *Mʌ-lai sʌnco hu-nthy-o.*
1s-DAT well COP-PST.HAB-3s
'I used to be fine.' (NNC:madhuparka-prose-2060-07.15)

Apart from NPs, the dative can also used on verb forms and adverbs. DAT on either of the two infinitives has a final reading:

(91) a. *Bhʌkti gʌr-nu-lai atma suddhʌ hu-nu pʌr-ch-ʌ.*
devotion do-INF$_1$-DAT soul pure be-INF$_1$ fall-NPST-3s
'In order to practice devotion, one's soul has to be pure.' (NNC:d04.18)

b. *Pʌɖ-nʌ-lai kunʌi niscit sʌmʌe wa sthan ch-ʌ?*
study-INF$_2$-DAT any fixed time or place be.there.NPST-3s
'Is there any fixed time and place for studying?' (NNC:nepal-misc-2061-11-23.290)

With adverbs of time DAT emphasises that a state of affairs is restricted to a certain time span:

(92) *ʌbʌ ajʌ-lai tʌ upae nʌi ch-ʌin-ʌ.*
now today-DAT CTOP means EMPH be.there-NEG.NPST-3s
'For today, there is no way to do it.'
(NNC:book-fiction-upasamhar-arthat-chautho-anta-2058.4500)

Looking at the numerous uses of the Nepali dative, there is of course the question whether all these can be brought together under a common function. This question will have to wait until we have considered the functional factors behind DOM in detail. It is taken up again in section 3.6.6.

3.5.2 Literature review

Although so far there is no dedicated study dealing with Nepali DOM, there is a number of articles, grammars, and teaching materials that touch on the topic. This literature is reviewed in this section. The various publications mostly do not refer to each other and thus do not form a continuous tradition. They will therefore not be presented in chronological order below but in an order that facilitates understanding. While the zero-marked case is invariably called nominative, the case marked by *-lai* is either called dative, accusative, or objective case. I will keep using the term dative to prevent confusion.

First of all, there is a surprising number of grammars and overviews which have missed or at least ignore the fact that objects in Nepali can be marked in two different ways. The interested reader is therefore *not* encouraged to consult any of Turnbull (1923 [1992]), Meerendonk (1949), Clark (1963), Verma (1992), Genetti (1994), Pokharela (2054 V.S.), Khaḍkā (2055 V.S.), Riccardi (2003), Lamsāla (2062 V.S.).

The most rudimentary treatments of DOM are found in coursebooks. They reduce the distinction between nominative and dative to a simple binary opposition. For instance, Sommer (1993:28) assumes that DAT marks all O except "wenn das Object ein unbelebtes Wesen ist", i.e. inanimate O get NOM and animate O get DAT. Gupta and Karmacharya (1981:84) draws the line between human and non-human referents, including animals among the kind of referents that must be marked by NOM. Matthews (1984) uses the same distinction and explicitly includes proper nouns and pronouns on the human, DAT-marked side.

The next level of analytical depth is constituted by some grammars and articles. While remaining simplistic, these works are superior to the binary approaches in assuming several factors behind DOM. For instance, Hughes (1947) claims that DAT is obligatory on human objects and impossible on inanimate objects. With non-human animate objects, DAT may be used "for emphasizing the true object of the sentence" (Hughes 1947:52). Although Hughes does not make any clearer what this means, it could be taken as a hint to the role of disambiguation in DOM (because the true object only needs to be emphasised when there is something else that also looks like an object, i.e. when there is ambiguity).

Abadie (1974:160) first mentions that DAT is obligatory on pronouns in P. She then goes on to talk about animacy and says that animate nouns "mostly" take DAT in P. Importantly, she adds that DAT on animate nouns "carries with it an implication of definiteness", thereby introducing information structure as a second functional dimension.

Li (2007a) makes a more precise statement by claiming that DAT can only be used with animate,

specific referents (p. 1471) and that within this class it is only obligatory with proper nouns and pronouns (p. 1472). Proper nouns and pronouns are among the word classes which are definite by default, so from Li's claims it is only a small step to a more elaborate system where one would say that DAT is possible on all animate, specific referents and obligatory on all animate, definite referents. In an earlier paper (Li 2007b), Li gives less detailed information but mentions *en passant* that DAT may also be used on nouns which are not animate and specific when they are "socially important", without further elaborating this idea.

Another work on this level is Korolev (1965). Although this grammar has been published earlier than both Abadie's and Li's papers, it is more detailed and explicit than them in some respects. Korolev (p. 133) mentions three factors that play a role for DOM: part of speech, animacy, and degree of semantic generality ("семантическая обобщенность"). He does not talk about part of speech in detail but focusses on the latter two factors. Human referents (independently of word class) are always marked by DAT, animals at least often so. For inanimate referents NOM is the default case, but DAT may be used for picking out or specifying referents ("выделить или уточнить"). Infinitives never carry DAT.

The most complex rule system so far is posited by Wallace (1985:25), who states the following four rules for the marking of "direct objects":

- "All referential direct object NPs (names, pronouns whose antecedents are persons etc.) must be marked by *-lai*.
- All nonreferential but human direct object NPs may be marked by *-lai* or Ø.
- All direct object NPs may be marked by *-lai* to indicate emphasis or definiteness.
- All inanimate direct object NPs are otherwise unmarked."

Neat as this system looks at first sight, it does have its internal weaknesses: Wallace does not define what exactly he means by "referential" and "emphasis". Further, rule 1 was probably intended to apply to all *human* referential objects – this is what is suggested by the note in parentheses, and otherwise it wouldn't be correct that inanimate objects are per default marked by NOM. It's also not clear why it is necessary to emphasise that non-referential human objects can be marked by *-lai* if all objects can be marked by *-lai* for "emphasis", anyway.

One last approach that should be mentioned is Acharya (1991). Acharya is the only grammarian who includes verb class as a factor in DOM, albeit in a rather unsystematic way. He distinguishes the following four classes of transitive verbs (p. 160):

- verbs with direct object (= monotransitive verbs and instrumental ditransitives in this work)
- verbs with direct and indirect object (= transfer ditransitives Ia)
- verbs with direct object and "object complement" (= equational ditransitive frame; not a class)
- verbs with direct object and "locative complement" (= transfer ditransitives Ib, IIa, IIb, III)

Acharya indirectly claims that DOM is only present in the first and in the fourth class, where animates are marked by DAT and inanimates by NOM. In the other two classes the "direct object" is invariably marked by NOM. Differently from the other multifactorial approaches, Acharya does not talk about definiteness.

To summarise, two general approaches to Nepali DOM can be distinguished. Monofactorial approaches try to reduce the DAT/NOM alternation to a single functional opposition, whereas multifactorial approaches take into account several factors. The most important variable seems to be animacy, which is recognised by all works dealing with DOM and which is given a vague priority by the multifactorial approaches. Other variables that are mentioned are definiteness and part of speech (both of the object and of the verb).

The account presented in the next section will deviate from this base in several respects. First, it will be shown that there are more factors behind DOM than have been assumed so far. Second, none of the relevant factors is sufficient or necessary in all instances – there are always examples where some value does not yield the expected case or where the expected case is given but with an unexpected combination of values. Last but not least, I will compare the benefits of a rule system and a probabilistic system for modelling the interaction of the relevant factors and will argue that the latter does a slightly better job and is theoretically sounder.

3.5.3 Animacy

It is no coincidence that animacy is the only variable that features in all descriptions of Nepali DOM. It is certainly the variable whose influence is most easily visible, and also one of the weightiest (cf. section 3.6.4.18). Human or animate referents are associated with DAT, inanimate referents with NOM. This is shown in (93).

(93) a. *Gai-lai lʌura-le piʈ-nʌ thal-e.*
cow-DAT stick-ERG beat-INF_2 start-PST.3p
'They started beating the cow with a stick.' (NNC:book-fiction-ek-chihan-2056.3732)

b. *Un-le ʌbʌ phʌlam-ko ɖʌɳɖa, khukuri rʌ kei ɖhuŋga pʌni ochyan-muni*
3MH-ERG now iron-GEN rod knife and some stone also bed-SUB
luka-erʌ rakh-ek-i ch-ʌn.
hide-CVB_1 put-PST.PTCP-F be.there-3MH
'She hid the iron rod, the knife and also some stones under the bed.' (NNC:s02.91)

However, contrary to what's assumed in works like Gupta and Karmacharya (1981), Matthews (1984), Sommer (1993) (see section 3.5.2 above), animacy is neither sufficient nor necessary for DAT. (94a) and (94b) show NOM-marked animate referents, (95c) a DAT-marked inanimate referent.

(94) a. *Guruŋ-hʌru bhʌīsi, gai pʌni pal-ch-ʌn.*
Gurung-PL buffalo cow also keep-NPST-3p
'The Gurung also keep buffalos and cows.' (NNC:book-anthropology-sabai-jatko-fulbari-2055.989)

b. *Mʌ yʌuʈa manche khoj-i-ra-ch-u.*
1s one.CLF person search-LNK-CONT-NPST-1s
'I've been looking for someone.' (NNC:book-academic-rupantaran-2062.868)

c. *Carkune ɖhuŋga-lai latti-le tin baji han-in.*
rectangular stone-DAT kick-ERG three time hit-PST.3fMH
'She kicked at the rectangular stone three times.' (NNC:book-fiction-bircharitra-2060.193)

The same fluidity in the marking of animate objects also becomes visible in elicitation. When speakers are asked which of the two relevant cases is correct in a given sentence, they are in many cases able to pick out one but also often say that both are correct. This fluidity can be measured by allowing informants to grade grammaticality judgements on a scale of subjectively judged commonness. Since we are dealing with a binary variable, the commonness of one value determines that of the other, so I was able to use the following scale:

- grammatical/ungrammatical: only one variant is possible, the other variant is never found
- normal/odd: only one variant is normally used, the other is decidedly odd or not even acceptable for all speakers
- common/uncommon: one variant is possible but notably less common than the other one
- more/less common: one variant is somewhat less common than the other one
- equal: both variants are equally common

For instance, various O in the framing sentence

(95) *Ajʌ rati mʌi-le sʌpna-ma … dekh-ẽ.*
today at.night 1s-ERG dream-LOC … see-PST.1s
'Yesterday night I saw … in a dream.'

were judged by the speakers GP (elicitation 2010) and SAR (elicitation 2011) as shown in Table 3.3:

While commonness judgements vary depending on the speaker and on the sentence containing the object, there is one very clear and stable tendency: DAT is more common with animates, NOM is more common with inanimates. In fact, if one assumes an animacy hierarchy instead of a binary

O noun	commonness of DAT for GP	commonness of DAT for SAR
manche 'person'	normal	more common
kukur 'dog'	less common	less common
putʌli 'butterfly'	ungrammatical	uncommon
ɖhuŋga 'stone'	ungrammatical	uncommon
pani 'water'	ungrammatical	odd

Table 3.3: Commonness of *-lai* [DAT] on various nouns

opposition animate/inanimate, it even becomes possible to say that DAT becomes the less common the further one moves down the ladder: DAT is overall more common with humans than with animals and with mammals compared to non-mammals. Interestingly, though, usually no difference is made between low animals such as reptiles and insects and plants or inanimate individual concepts such as *ɖhuŋga* 'stone'. Note that dead referents also usually count as low:

(96) *Maowadi bhʌn-ek-a dui-jʌna-k-a aphʌnta-le tʌ uni-hʌru-ko las*
Maoist say-PST.PTCP-OBL two-HUM.CLF-GEN-OBL relatives-ERG CTOP DIST-PL-GEN corpse
dekh-nʌ pa-en-ʌn.
see-INF_2 get-NEG.PST-3p
'The relatives of the two alleged Maoists didn't get the change to see their dead bodies.'
(NNC:himalkhabarpatrika-2056-10-16.130)

Inanimate mass concepts such as *pani* 'water' regularly feature lowest on the hierarchy. It is virtually impossible to get DAT on these concepts in elicitation. This is expected insofar as mass concepts are even less similar to humans and high animals than inanimate individual concepts: individual concepts can be easily moved around and identified at different times in different places, whereas mass concepts are hard to move without an appropriate container and are hard to identify once the container gets removed (one can say *This is the same cup of water as before*, but *This is the same water as before* sounds odd).

It should be remembered that the individual/mass distinction corresponds to quantifiability on the syntactic side (cf. section 2.6.3.1). This factor can also be looked at independently of animacy in Nepali (section 3.5.5).

There is another interesting difference within inanimates. DAT is more common with static concepts such as *problem*, which can be defined independently of time, than with procedural concepts such as *education*. Consider the following two sentences from the same text, where the pronoun in (97a) refers to a finished song (a static concept) but the one in (97b) refers to its rearrangement (a procedural concept):

(97) a. *ʌnurodh pʌni gʌr-ya th-ẽ ki, el-lai nikal-ne ki*
request also do-PRFV.PTCP be.there-PST.1s or PROX-DAT bring.out-IPFV.PTCP or
bhʌnerʌ.
CIT
'I had already made a request to them asking whether we should bring it out.'
(NNC:A001013001.219)

b. *Rʌ yo gʌr-i-sʌk-e-pʌchi pheri nikal-ni bhʌnni kura*
and PROX do-LNK-COMPL-NMLZ-TMP.POST again bring.out-IPFV.PTCP CIT.ADN talk
hũ-dakheri…
$COP\text{-}CVB_5$
'And after I had done that and the idea of bringing it out came up again…'
(NNC:A001013001.243)

This difference can be integrated into the animacy hierarchy, too: all animates are beings that can be defined independently of time, so procedural concepts are even less similar to animates than are static concepts.

If we take together everything that has been said above we get the following, somewhat odd hierarchy:

human > mammal > lower static individual concept ("thing") > mass/process

This hierarchy captures nicely some tendencies, especially those found in elicitation. However, it is not that useful for quantitative analysis because the distances between the steps on it cannot safely be said to be equal. Animacy was therefore split into several variables for the quantitative analysis in section 3.6 (see esp. section 3.6.4.2, section 3.6.5).

3.5.4 Specificity

The second-most frequent factor used to explain Nepali DOM is definiteness. Another concept mentioned in many descriptions of other Indo-Aryan languages (cf. section 3.9) is specificity. These two were brought together in section 2.5 on the base of unique identifiability, where specificity was defined as identifiability on the part of the speaker and definiteness as identifiability on the part of both hearer and speaker.

Inspite of the literature for Nepali and of what is described for many other Indo-Aryan languages, definiteness is irrelevant for Nepali DOM. Although there are many instances of definite O-DAT and indefinite O-NOM, I have so far not been able to find a single example where the case of O can *only* be explained with reference to definiteness. Moreover, in all cases where definiteness seems to be at play at first glance, closer inspection reveals that it is really specificity that makes the difference. An example is shown in (98).

(98) a. *mʌnoj bida hũ-da manche bheʈ-nʌ ja-nch-ʌ.*
Manoj free.time COP-CVB$_4$ person-DAT meet-INF$_2$ go-NPST-3s
'In his free time, Manoj goes to meet people.' (elicitation SAR 2011)

b. *mʌnoj bida hũ-da manche-lai bheʈ-nʌ ja-nch-ʌ.*
Manoj free.time COP-CVB$_4$ person-DAT meet-INF$_2$ go-NPST-3s
'In his free time, Manoj goes to meet someone.' (elicitation SAR 2011)

c. *mʌnoj bida hũ-da tyo manche-lai bheʈ-nʌ ja-nch-ʌ.*
Manoj free.time COP-CVB$_4$ MED person-DAT meet-INF$_2$ go-NPST-3s
'In his free time, Manoj goes to meet that person.' (elicitation SAR 2011)

Manche referring to a non-specific referent as in (98a) gets NOM by default. As soon as the referent becomes specific as in (98b), DAT becomes the default. With a definite O as in (98c), DAT is equally common as with O that are only specific.

In some cases, differences in specificity can even lead to strict grammaticality judgements, as in (99):

(99) a. *mʌ tin-jʌna sʌgau-ne manche(*-lai) khoj-dʌi ch-u, jo*
1s three-HUM.CLF help-IPFV.PTCP person-DAT search-PROG be.NPST-1s who.REL
bhʌ-e pʌni hu-nch-ʌ.
COP-COND also be.okay-NPST-3s
'I'm looking for three helpers, anyone is okay.'

b. *mʌ tin-jʌna hijo pʌni yã a-eko manche*(-lai)*
1s three-HUM.CLF yesterday also PROX.LOC come-PRFV.PTCP person-DAT
khoj-dʌi ch-u.
search-PROG be.NPST-1s
'I'm looking for the three people who also came here yesterday.' (elicitation SAR 2011)

Note, though, that such judgements may vary from speaker to speaker – for instance, another speaker, GP, differed from SAR by allowing ?DAT on the non-specific referent in (99a) and equating NOM and DAT on the specific referent in (99b). The observed tendency is, however, the same: DAT is less common with non-specific than with specific object referents.

An interesting twist to specificity in Nepali is that it may sometimes matter how much is known about a referent. For instance, the sentence in (100) below was elicited with two different backgrounds. In both cases the speaker was told that there had been a series of spectacular thefts, which the police suspected to have been committed by the same person. In the first case that person was well known to the police but could never be arrested because of lack of evidence, whereas in the second case the police was completely unclear about the identity of the thief until they caught him in the act. Although the thief is identifiable to the police in both cases (once in the real world, once via his deeds), DAT was preferred when the thief was a known person and NOM when he was not.

(100) *Pulis-le cor(-lai) pʌkr-y-o.*
police-ERG thief(-DAT) arrest-PST-3s
'The poliece arrested the thief.' (elicitation NP 2012)

NOM is also licensed by reference to types (as opposed to tokens). This may seem like a separate phenomenon but is actually motivated by similar reasons as (100). Since a type is an abstraction over many tokens, it has necessarily fewer specific properties than a token. Knowledge about types is therefore always more restricted than knowledge about tokens. An example is shown in (101), where O is marked by NOM even though it is coded by a demonstrative (*yo* [PROX]), definite, and has been mentioned a couple of times. The reason is that the speaker does not want to see the same shoe in black but the same model in black:

(101) *Kalo-ma her-um yo.*
black-LOC see-[OPT]1p PROX
'Let's see this (shoe) in black.' (NNC:V001001002.17)

Another fact pointing into a similar direction is that speakers often indicate in elicitation that the use of DAT is related to how present a referent is on one's mind. For instance, DAT is the default in (102) (a variant of the pair of examples presented above), where the relevant object *manche* is both human and definite. However, NOM is marginally possible if the speaker just saw the three people but did not talk to them and therefore doesn't remember them well now:

(102) *Hijo pʌni yʌhã a-ek-a tin-jʌna manche?(-lai) mʌ khoj-dʌi ch-u.*
yesterday also PROX.LOC come-PST.PTCP-PL three-HUM.CLF person-DAT 1s search-PROG be.there.NPST-1s
'I'm looking for the three people who also came here yesterday.' (elicitation NP 2012)

A similar situation is given in (103), where the speaker forgot to give a farewell present to one person:

(103) *Bastʌbma mʌi-le sʌbʌi-lai bidai-ko upʌhar di-nʌ cahʌ-nch-u, tʌrʌ me-ro bicar-ma mʌi-le ek-jʌna(?-lai) bhul-ẽ.*
really 1s-ERG all-DAT farewell-GEN present give-INF_2 want-NPST-1s but 1s-GEN thought-LOC 1s-ERG one-HUM.CLF-DAT forget-PST.1s
'Actually I wanted to give a farewell present to everybody, but I think I forgot one person.' (elicitation KP 2012)

This sentence was situationed in three different contexts:

1. The speaker added *...because there is one present left, but I have no idea who it is.*
2. The speaker added *...uhm, right, I forgot Peter.*
3. Another speaker asked *Whom?* and the first speaker answers *Peter, of course.*

Although the speaker theoretically knows much more about the referent in variant 2 than in variant 1, NOM is preferred in both variants because the referent is not sufficiently present on the speaker's mind and the knowledge about him thus cannot be easily accessed. By contrast, DAT was preferred in variant 3, where the speaker doesn't have to think to produce the referent's name.

Thus, in order for DAT to be preferred, it seems like the referent does not only have to be identifiable but should be readily identifiable.

As for arbitrary reference (cf. section 2.6.2), both open reference (104) and discardable reference (105) usually go together with NOM.

(104) *Chan-nu pʌr-da bibhinnʌ bekti-le ʌruʌru nʌi lekhʌk*
choose-INF$_1$ be.necessary-CVB$_4$ various individual-ERG other FOC writer
chan-la-n.
choose-PROB.FUT-3p
'If they had to choose various persons would choose some other author.'
(NNC:book-criticism-paschimka-kehi-sahityakar-2062.71)

(105) *Dosro bissoyuddhʌ-le pãc kʌroɖ manche mar-y-o.*
second world.war-ERG five ten.million person kill-PST-3s
'The Second World War killed 50 mio. people.' (NNC:himalkhabarpatrika-2059-02-16.2065)

Discardable reference can become conventionalised in composite activities (cf. section 2.6.5.2). An example is *bʌcca pau-* 'get/bear a child', where the child is almost always marked by the nominative in spite of being human and specific:

(106) *Pʌcpʌnnʌ bʌrsʌ-ko-le ʌsti bhʌkkhʌr ei hʌspiʈʌl-ma bʌcca pa-erʌ*
fifty.five year-GEN-ERG recently just PROX.FOC hospital-LOC child get-CVB$_1$
gʌ-y-o.
go-PST-3s
'Just recently a (woman) of 55 years got a child in this very hospital and went (home again).' (NNC:V001014002.63)

Just like animacy, specificity cannot explain all instances of O-DAT. There are cases like (107) where specific objects are marked by NOM and cases like (108) where DAT is present although the object is non-specific.

(107) *Aph-n-ʌi ghʌr-ʌgadi-ko sʌɖʌk-ma rʌh-eko bʌm khelaũ-da... balʌk-hʌru*
REFL-GEN-FOC house-ANTE-GEN street-LOC be-PRFV.PTCP bomb play.with-CVB$_4$ child-PL
mar-i-ek-a hu-n.
kill-PASS-PRFV.PTCP-PL be[NPST]-3p
'The children were killed while playing with the bomb, which lay on the street before their own house.' (NNC:bbc-news-2061-12-14.13)

(108) *Manche-lai ʧyakk-ʌi kyac gʌr-nʌ sʌk-ne chemʌta ham-ro lokgit-ma*
person-DAT right-FOC catch do-INF$_2$ be.able-IPFV.PTCP power 1p-GEN folk.song-LOC
euɖa ʌjib-ko gũɖ ch-a.
one.CLF peculiar-GEN characteristic be.there.NPST-3s
'The power to capture people right away is characteristic of our folk songs.'
(NNC:A001013001.1261)

Also note that there is no rigid interaction between specificity and negation. Although there is a tendency to interpret a negated predicate in combination with O-DAT as having narrow scope and as having wide scope with O-NOM, both cases in principle allow both scope interpretations. Put more simply, NOM and DAT are no more rigidly associated with specificity in negated than in non-negated clauses. This is illustrated by (109).

(109) *Mʌi-le manche(-lai) bheʈ-in-ʌ̃.*
1s-ERG person-DAT meet-NEG.PST-1s
both 'I didn't meet anybody.' or 'I didn't meet somebody.' (elicitation BP 2012)

3.5.5 Quantifiability

Quantifiability does not play the same big role in Nepali as in Chintang (cf. section 2.6.1), for the simple reason that specificity in Nepali is not as important as in Chintang. That being said, there are still many cases that are easiest to explain via quantifiability. For instance, demonstratives in object position are usually marked by DAT (see section 3.5.8 below), but this can be cancelled when the referent in question is non-quantifiable. In (110), the speaker has repeated a statement an indefinite number of times, thereby making the corresponding referent non-quantifiable:

(110) *MʌI-le yo dorya-ko h-ũ.*
1s-ERG PROX repeat-PST.PTCP be[NPST]-1s
'I have been repeating this.' (NNC:V001004001.7)

A more complex example is found in the following sequence of sentences, which is about sugarcane cultivation. In (111a), an indefinite amount of sugarcane is planted. Since sugarcane is also inanimate, it's natural that it should get the nominative. Note that the dative on *tellai* in (111a) is not due to DOM but because *bhʌn-* has a fixed G-DAT in the sense employed here. One year after planting, all planted sugarcane has to be weeded (111b). Exhaustive reference entails quantifiability because it does not admit (too great) deviations from a certain number (cf. section 2.6.3.3). This together with the use of the demonstrative *tyo* [MED] motivate DAT in this sentence. Finally, (111c) shows that *tyo* alone is not enough to trigger DAT: since over two years indefinite amounts of sugarcane are harvested and then removed at a time, NOM is used again.

(111) a. *Ũkhu caini ek bʌrsʌ rop-e-pʌchi dui bʌrsʌ caini, tei- phedi*
sugarcane RETRV one year plant-NMLZ-TMP.POST two year RETRV PROX.FOC phedi
bhʌn-th-e tel-lai.
call-PST.HAB-3p PROX-DAT
'After planting sugarcane for one year, two years, uhm, *phedi*, that's what they used to call it.' (NNC:A001011002.508)

b. *Tei, tei jʌra-baɖʌ caini pheri pʌla-erʌ, pheri pʌla-erʌ pheri*
MED.FOC MED.FOC root-ABL RETRV again sprout-CVB$_1$ again sprout-CVB$_1$ again
tel-lai goɖmel gʌr-y-o.
MED-DAT weeding do-PASS.PST.3s
'And that, after it grew for some time, after it grew it was weeded again.'
(NNC:A001011002.508)

c. *Tyo dui bʌrsʌ-sʌmmʌn caĩ la-ko bʌrsʌ, tyo pʌchi ugal-ne*
MED two year-TERM RETRV take-PRFV.PTCP year MED later remove-IPFV.PTCP
rʌ ʌni mʌkʌi chʌr-ni tyo thaũ-ma.
and and maize sow-IPFV.PTCP MED place-LOC
'After harvesting it for up to two years, it had to be removed and maize had to be sown in the same place.' (NNC:A001011002.508)

The role of quantifiability for DOM becomes especially clear with mass concepts (cf. section 2.6.3.1) in object position. Mass concepts can be either looked at from the perspective of animacy (section 3.5.3) or from that of quantifiability. Whereas animacy can help explain why mass concepts are not marked by DAT in general, only quantifiability can explain this *and* why DAT sometimes is possible. Mass concepts do not have natural boundaries, so their default construal is non-quantifiable as in (112a), where DAT is impossible. However, once boundaries are added to them they can get DAT as in (112b). All attested examples of DAT on mass concepts are of this type.

(112) a. *Us-le pani(*-lai) matrʌi paũ-ch-ʌ.*
DIST-ERG water-DAT only get-NPST-3s
'He only gets water.' (elicitation GP 2010)

b. *ʌbʌ kitli-ko pani-lai thal-ma selau-nʌ thal-ch-ʌ.*
now kettle-GEN water-DAT plate-LOC cool-INF_2 start-NPST-3s
'Now she starts to cool the water from the kettle on a plate.'
(NNC:book-fiction-ek-paluwa-anekaun-yam-2026.6352)

Mass concepts, too, can be made quantifiable via exhaustive reference:

(113) *Bãki nʌbbe prʌtisʌt watawʌrãɖ-ma phʌili-erʌ hawa, maʈo rʌ pani-lai*
remaining ninety percent environment-LOC spread-CVB_1 air soil and water-DAT
prʌdusit gʌr-nʌ pug-ch-ʌ.
polluted make-INF_2 be.enough-NPST-3s
'The remaining ninety percent (of poison) are enough to pollute (all the) air, soil, and water.'
(NNC:book-academic-swasthya-samaj-ra-rajniti-2062.4145)

On the other hand, again similarly to Chintang, individual concepts can yield non-quantifiable referents when it is indefinite subamounts of theirs that are affected by an event. In (114) only some of the food available at a feast is taken and given to a dog, so NOM is used in spite of the referring expression being a demonstrative:

(114) *Tyo ʌliʌli jhik-erʌ lya-erʌ cʌini kukur-lai dī-dakheri kukur ʧhʌhʌrʌi mʌr-y-o.*
MED a.bit take-CVB_1 bring-CVB_1 RETRV dog-DAT give-CVB_5 dog at.once die-PST-3s
'He took some of it and brought it, and when he gave it to the dog it died right away.'
(NNC:A001011002.241-243)

Generic but non-exhaustive reference is one type of non-quantifiable reference and therefore generally goes together with NOM. However, when generic reference is achieved via the construal of a type representing a whole species, that type may be marked by DAT, as in (115).

(115) *Sukkha roʈi-lai ʧhulo matra-ma utpadʌn gʌr-i-ne pauroʈi-le*
dry bread-DAT big scale-LOC production do-PASS-IPFV.PTCP toast-ERG
lʌghar-eko ch-ʌ.
push.out-PRFV.PTCP be.there.NPST-3s
'Toast produced on a big scale has replaced dry bread.' (NNC:f23.11)

This is especially common when types are compared. For instance, in (116a) generic reference to a kind of potato is achieved via a non-quantifiable amount of tokens and the potato is marked by NOM. In (116b) a type is construed via the derivational suffix *-e* and marked by DAT.

(116) a. *Rato ɖʌllo alu lau-thy-ʌũ.*
red round potato apply-PST.HAB-1p
'We used to plant red, round potatoes.' (NNC:A001011002.725)
b. *Swad-ma caini bheʈ-tʌin-ʌ tyo ɖʌlle alu-lai.*
taste-LOC RETRV meet-NEG.NPST-3s MED round.type potato-DAT
'It (another kind of potato) doesn't match the round potato in taste.'
(NNC:A001011002.731)

3.5.6 Interplay of animacy and specificity

So far we have seen that animacy and specificity (backed up by quantifiability) play an important role in determining object case, but also that none of their values is sufficient or necessary in isolation for either NOM or DAT. Before we go on to examine less prominent factors involved in DOM, the question should be asked what their combined impact looks like.

Interestingly, although both factors are frequently mentioned in the literature on DOM in Indo-Aryan, it is rarely made clear how they work together. The work which is most explicit in relating animacy and identifiability is Mohanan (1994) on Hindi. According to Mohanan, a high value in either category (human or definite) is enough to yield DAT (and only DAT). The same is true for two

high values (human and definite). NOM becomes the only possibility when intermediate or low values are combined (e.g. specific and inanimate, "incorporated" and animate). See section 3.9.6 for more details.

For Nepali the picture is not so simple. As we have already seen, there is no simple value – however high it may be – that has enough weight to always yield DAT. Some works on Nepali have made more sophisticated claims, but those do not stand up to scrutiny, too. Abadie (1974) states that animate referents marked by DAT have to be definite, but this is contradicted by examples such as (117), where O is not even specific.

(117) *Gai-lai mar-nʌ pa-ĩ-dʌin-ʌ, dharmik sotʌntrʌta cah-ĩ-dʌin-ʌ!*
cow-DAT kill-INF$_2$ get-PASS-NEG.NPST-3s religious freedom need-PASS-NEG.NPST-3s
'Killing cows is unacceptable, we don't need religious freedom!' (field notes 2011)

Li (2007a) claims that DAT is only found on animate specific referents, which is contradicted by numerous examples such as (119).

(118) *ʌbʌ kʌsto~kʌsto caĩ ʌdhikar ham-le pa-k-a ch-ʌm bhʌnne*
now what.kind.of~DISTR RETRV right 1p-ERG get-PST.PTCP-PL be-1p CIT.ADN
kura-lai *ham-le ʌlikʌti khojiniti pʌni gʌr-nu pʌr-y-o.*
matter-DAT 1p-ERG a.bit inquiry also do-INF$_1$ fall-PST-3s
'Now we also have to ask the question which rights precisely we have achieved.'
(NNC:A001017001.513-515)

What can be said, however, is that the combination of the highest values (human and specific) or the lowest values (mass/process and non-specific) yield DAT and NOM, respectively, in virtually all cases. The only counterexample so far features a human specific referent marked by NOM and is shown in (119). I am grateful to Balarām Prasāiṃ for drawing my attention to this.

(119) *Mʌ tim-r-i ama bheṭ-nʌ a-eko.*
1s 2s-GEN-F mother meet-INF$_2$ come-PRFV.PTCP
'I've come to meet your mother.' (NNC:book-fiction-alikhit-2058.635)

Why *ama* did not get a DAT here is not fully clear. One possibility is that *ama* is a rather unexpected referent here. Just before this scene, the speaker (Ṛṣirāma) gave a kiss to the hearer (Matiyā). Because Matiyā is not married to Ṛṣirāma but to somebody else, that was a bold thing to do and Matiyā broke into tears. Ṛṣirāma left without a word but came back after thinking for half an hour about why Matiyā cried and what could be the consequences if another villager heard about this.

When he sees Matiyā cooking rice as if nothing had happened, he says the sentence in (119) just in order to excuse himself, but in such a clumsy manner that he makes a fool of himself. The nominative might reflect the fact that Ṛṣirāma had not been thinking about Matiyā's mother before but just inserted her spontaneously into the object slot. If that is correct it would be another argument for the claim made in section 3.5.4 above that a referent must not only be identifiable but readily identifiable to the hearer in order to license DAT. *Ama* in (119) would then not be specific in a narrow sense and would no longer present a counterexample.

I also tried to find examples for specific human referents marked by NOM in O via elicitation but failed. One interesting example is the following, where NOM on an seemingly definite referent becomes marginally possible if one adds an expression specifying that the speaker is actually not sure about whether he can really identify the referent:

(120) a. *Mʌi-le hijo pʌni tyo manche*(-lai) dekh-ẽ.*
1s-ERG yesterday also MED person-DAT see-PST.1s
'Yesterday I saw that guy, too.'
b. *Mʌi-le hijo pʌni tyo manche?(-lai) dekh-ẽ jʌsto lag-ch-ʌ.*
1s-ERG yesterday also MED person-DAT see-PST.1s like seem-NPST-3s
'I think I saw that guy yesterday, too.' (elicitation KP 2012)

Similarly, it was impossible to find DAT-marked non-specific masses/processes (and, in fact, any DAT-marked non-specific objects below mammals), both in the NNC and via elicitation. This has also been stated by Korolev (1965), who says that inanimate referents marked by DAT have to be specific.

To summarise, there seems to be an island of regularity within the huge amount of variation in DOM: "double-high" and "double-low" referents must be marked by DAT and NOM, respectively. The variation takes place in heterogeneous constellations where the variables involved point into different directions.

3.5.7 Topicality

Topicality is a factor that has so far been neglected by the literature on DOM in Indo-Aryan languages. This seems strange in the light of the fact that according to Iemmolo (2011), topicality is one of the most prominent factors behind DOM world-wide. One possible explanation is that topicality is very hard to investigate via elicitation, another that it is a notoriously fuzzy concept. I will assume the following definition:

> Topicality is the presence of a referent on the mind of a person from whose perspective an utterance is made (this is by default the speaker). A referent becomes the more topical the more often that person thinks of it and the less topical the more time passes since its last mention and the more other referents it is surrounded by.

Of course it is so far impossible to measure this kind of topicality, and in many cases – especially in corpus texts where one doesn't know the context – it is even difficult to estimate it. For more practical purposes I will therefore approximate mental presence as frequency in a text (written or spoken), assuming that this is one of the most important factors mirroring (and contributing to) mental presence. This approximation produces good results, as will be shown below.

Note that both the definition above and its approximation make reference to discourse. In a strict sense we are therefore talking about discourse topics (in the sense of Lambrecht 1994), excluding clause and sentence topics. However, since this is the only sense that is relevant for DOM in Nepali, I will simply keep speaking of topics.

Here is a first simple example for the relevance of frequency from elicitation. I repeatedly asked a speaker to determine the best case for *tyo* [MED] while increasing the frequency of the corresponding referent (a chili pod) in a preceding mini-discourse. In the first case, *tyo* was cataphoric. A child had left a single chili pod on his plate, and his mother asked:

(121) *Tyo* *khã̃-dʌin-ʌs?*
MED-DAT eat-NEG.NPST-2sLH
'Don't you eat that one?' (elicitation BP 2012)

As can be seen, the best case was NOM. This didn't change in the second case, where there was some more background: there had originally been three chili pods and the child had eaten two and given the third to this brother, ordering him to eat it. When the mother asks why he left that pod, *tyo* is anaphoric with one mention so far:

(122) *Tyo* *kinʌ* *choɖ-ch-ʌs?*
MED why leave-NPST-2sLH
'Why do you leave that one?' (elicitation BP 2012)

In the next situation, the child was about to eat the third chili pod when it discovers that one side of it is black and announces that it won't eat it (first mention). The mother starts scolding him, saying that just because it (second mention) is ugly that doesn't mean it (third mention or covert) doesn't taste as good as the bright red pods, and orders him to eat it (third or fourth mention). This finally produced DAT as the preferred case:

(123) *Tes-lai pʌni kha!*
MED-DAT also eat[IMP.LH]
'Eat that one, too!' (elicitation BP 2012)

Note, though, that the same game did not work with the mass concept *ʌcar* 'pickles, chutney' – here, NOM was still used in (123). Thus, as in virtually all instances of DOM, topicality is not the only factor that is at work here – quantifiability as a precondition for specificity matters, too.

A more complex but telling example for topicality effects is the short story *Ṭēbalamāthiko tyasa ākāśavāṇī* ("That telegram on the table") by Paraśu Pradhāna (Pradhāna 1997). The biggest part of the text describes the thoughts of its protagonist, a man called Kṛṣṇa. His thoughts keep revolving around a telegram lying on a table in his room, wander away from it and are attracted to it again, until he finally accepts the sad truth contained in it that his wife has died. The telegram is thus the most important topic on Kṛṣṇa's mind, and this status is also reflected in text frequencies.

The frequency counts for the ten most frequent referents in the story are shown in Table 3.4. For absolute frequencies, I counted all overt mentions and all covert mentions in argument roles. Relative frequencies were calculated against the number of all referent pointers (i.e. again overt mentions or covert mentions in argument roles). In addition to standard relative frequencies, I calculated normalised relative frequencies by mapping the absolute frequencies isomorphically to values between 1 and 0 (column "ranked" in the table).

rank	identity	frequency		
		absolute	relative	ranked
1	Kṛṣṇa	94	0.30	1.00
2	telegram	23	0.07	0.25
3	friend of Kṛṣṇa's	8	0.03	0.09
4	mountains	6	0.02	0.06
5	Kṛṣṇa's wife	5	0.02	0.05
6	Kṛṣṇa's room	5	0.02	0.05
7	sympathetic words	4	0.01	0.04
8	Kṛṣṇa's hometown	4	0.01	0.04
9	Kṛṣṇa's eyes	3	0.01	0.03
10	table	3	0.01	0.03

Table 3.4: Top ten of referents in *Ṭēbalamāthiko tyasa ākāśavāṇī*

While Kṛṣṇa as the protagonist occupies rank 1, the telegram is the second-most frequent referent and is clearly set apart from the next lower referent, both in terms of absolute numbers and by the steep drop in frequency between them which does not get repeated anywhere below rank 3. The exceptional status of the telegram is reflected by its case marking: it appears six times in overt object position and is marked by DAT five times. (124) and (125) show the first and the second to last sentence of the story, which contain the first and last mention of the telegram.

(124) *Ʈebʌl-mathi-ko tes akaswãɖi-lai pheri pʌɖ-y-o.*
table-SUPER-GEN MED.OBL telegram-DAT again read-PST-3s
'Again he read that telegram on the table.' (Pradhāna 1997:75)

(125) *Us-le tes akaswãɖi-lai dhujadhuja par-i cyat-i-di-y-o rʌ ɖãko*
DIST-ERG MED.OBL telegram-DAT to.pieces make-CVB_2 tear-LNK-BEN-PST-3s and cry
chaɖ-erʌ ru-nʌ lag-y-o.
let.out-CVB_1 weep-INF_2 start-PST-3s
'He tore the telegram to pieces, let out a cry and started to weep.' (Pradhāna 1997:79)

With the wide definition of topicality in mind that was presented above it makes sense that the telegram should be DAT-marked from the very first sentence – that sentence opens a window into Kṛṣṇa's mind, where the telegram has been thought about again and again. Although the

reader hears of the telegram for the first time here (apart from the title of the story itself), this sentence immediately lets him know that it is specific (marked by *tes* [MED.OBL]) for Kṛṣṇa, whose perspective the text takes, and that it has been an important topic on his mind (marked by *-lai* [DAT] and enhanced by *pheri*). From the narrower perspective of frequency, (124) shows that (at least in this case) overall frequency in a text is a better measure of mental presence than frequency up to a certain point: the latter would have been 0 at this point so that DAT would have looked rather unexpected, whereas DAT is easily motivatable from the overall high frequency of the telegram.

The only instance where the telegram is not marked by DAT in object position is (126). Note that the first instance (*tei akaswāḍi*) is governed by a complex predicate with the light verb passive (see section 3.4.7), so DAT is impossible there for formal reasons.

(126) *Us-ko lʌcche-k-a nimittʌ saed tei akaswāḍi praptʌ hu-nu awʌsek*
DIST-GEN aim-GEN-OBL reason probably MED.FOC telegram obtained be-INF_1 necessary
thi-y-o rʌ ***tyo*** *pa-erʌ saed, u khusi ch-ʌ.*
be.there-PST-3s and MED get-CVB_1 probably DIST happy be.there-NPST-3s
'For his aim it was probably necessary that that telegram was obtained, and having received it he is probably happy.'
(Pradhāna 1997:76]

This sentence looks rather mysterious at first sight: the overall high frequency of the telegram is unchanged, and the frequency up to this point is also already comparatively high (6 mentions). The nominative becomes understandable when we take on again the wider definition of topicality. The event expressed by *tyo paerʌ* is a past event. At the time when Kṛṣṇa received the telegram it was completely new to him. It only acquired a special status in his mind after he opened and read it and kept thinking about it for a considerable time. The NOM in (126) reflects the state of Kṛṣṇa's mind before receiving the telegram. This sentence demonstrates that frequency can only approximate mental presence for practical purposes but not replace it.

Mental presence and frequency can also be incongruent the other way round, that is, referents with low text frequency can get the dative based on high mental presence. An example for this comes from the film *Yatīko khojīmā* by Santośa Ḍhakāla. Four friends go on a trip to find the legendary Yeti. One of them (Devon) has the secret aim of killing the Yeti and uses another one (Lucy) to distract Mohit, the clever leader of the group. Mohit falls in love with Lucy and doesn't realise Devon's intentions until it's almost too late. When he discovers that Lucy's affair with him was part of Devon's devilish plan he curses and leaves her. Shortly after that Lucy is hit by a bullet in a dramatic series of events. When Mohit finds her dying and talks to her for the last time it turns out that she had really been in love with him and had changed sides without anybody knowing. At this point, Mohit says:

(127) *Mʌi-le tim-ro maya-lai buj-nʌ sʌk-in-ʌ.*
1s-ERG 2s.MH-GEN love-DAT understand-INF_2 be.able-NEG.PST-1s
'I couldn't understand your love.' (Ḍhakāl 2008)

Lucy's love has not been mentioned before and is not mentioned later, as Lucy dies right after this sentence. In terms of frequency it is thus minimally topical. Nevertheless *maya* gets the dative because it has been an important topic on Mohit's mind for well over half of the film's length.

DAT as a marker of topicality in a more abstract, fuzzy sense also seems to be at play in (128), another sentence from *Ṭēbalamāthiko tyasa ākāśavāṇī*:

(128) *Nyuyark-k-a akas chu-ne ghʌr-hʌru-lai u sʌdʌī sʌpna-ma dekh-ch-ʌ.*
New.York-GEN-PL sky touch-IPFV.PTCP house-PL-DAT DIST always dream-LOC see-NPST-3s
'He always sees the skyscrapers of New York in his dreams.' (Pradhāna 1997:76)

The skyscrapers marked by DAT have again not been mentioned before, nor are they ever mentioned again later. Differently from (127), however, the reader also does not have any reason whatsoever to construe the skyscrapers as topical based on indirect and/or non-verbal evidence – this sentence hits him out of the blue. It is part of a description of Kṛṣṇa's dreams, and the following

sentences make it clear that those dreams revolve around America. The reader can infer from the use of the dative that skyscrapers feature prominently in Kṛṣṇa's dreams and for him are a symbol for a different life full of promises.

Topicality is even more opaque in (129):

(129) *etro lamo ʌntʌral-ma akas-baʈʌ bʌrsa-ko pani-lai kur-nu*
PROX.EXT long interval-LOC heaven-ABL$_1$ rainy.season-GEN water-DAT wait-INF$_1$
pʌr-ne
be.necessary-IPFV.PTCP
'having to wait for such a long time for the rains from heaven'
(NNC:book-belleletter-nepalma-garibiko-bahas-2061.1852)

The text in which this sentence is embedded is about poverty in Nepal. The present section deals with agriculture and how it depends on environmental on political factors. Rain has only been mentioned once before, and that was several paragraphs earlier. Although the text is generally written in an objective style, (129) can be viewed as a short lapse into the subjective perspective of poor farmers where the annual rainfalls are a recurrent important topic.

3.5.8 DOM with demonstratives and pronouns

At first sight, part of speech looks like a purely formal factor. However, there are always reasons why a speaker chooses a pronoun or a demonstrative to refer to a referent rather than another nominal form or a zero, and those reasons are tangled up with specificity and topicality. Only specific referents can be referred to by a pronoun or a demonstrative. The use of anaphoric demonstratives and of reflexive pronouns is limited to contexts where the last mention of the referent is not too far away, which is also an important factor in topicality. The referents of SAP pronouns can be considered as inherently highly topical because every utterance in a conversation is produced by or addressed to them so that they have a high mental presence even if they do not feature themselves as referents in the discourse.

That being said, the functional factors that motivate the use of pronouns and demonstratives cannot fully explain the interaction of DOM with these parts of speech. The following three strict rules apply (in active sentences – NOM is always possible in the passive, see section 3.4.6):

- Pronouns in object position must always be marked by DAT. Pronouns as defined in section 3.2.1 are *mʌ* [1s], *hami* [1p], *tɅ̃* [2sLH], *timi* [2MH], and *aphu* [REFL].
- The noun *tʌpaĩ* [2HH] must always be marked by DAT.
- Demonstratives must be marked by DAT when they refer to a human being. Nominal demonstratives as defined in section 3.2.1 are *u* [DIST], *ini* [PROX.MH], *tini* [MED.MH], and *uni* [DIST(.MH)], but also *ko* 'who'; versatile demonstratives are *yo* [PROX] and *tyo* [MED].

When the functional factors correlating with the choice of pronouns and demonstratives are given but a different part of speech is used, DAT is still likely but not mandatory (cf. section 3.5.4, section 3.5.7 above). We are thus dealing with an island of grammaticalisation island within DOM. The mentioned rules are illustrated by the examples below.

(130) shows the obligatoriness of DAT on pronouns and on the noun *tʌpaĩ*. (130a) also shows that social status is irrelevant for this rule.

(130) a. *TɅ̃*(-lai) dekh-ch-ʌ.*
2sLH-DAT see-NPST-3s
'He sees you.'
b. *Timi*(-lai) dekh-ch-ʌ.*
2[s]MH-DAT see-NPST-3s
'He sees you.'
c. *Tʌpaĩ*(-lai) dekh-ch-ʌ.*
2[s]HH-DAT see-NPST-3s
'He sees you.' (elicitation GP 2010)

The reflexive pronoun *aphu* also requires DAT. This is independent of the animacy of O:

(131) a. *Biralo-le aphu*(-lai) caʈ-dʌi ch-ʌ.*
cat-ERG REFL-DAT lick-PROG be.there.NPST-3s
'The cat is licking itself.' (elicitation GP 2010)

b. *Gʌɖgʌɖaũ-do Hyaŋsi khola-le Kʌrnali-ma aphu*(-lai) bilin gʌr-ch-ʌ.*
thunder-CHAR.PTCP Hyaṅsī river-ERG Karṇālī-LOC REFL-DAT merger do-NPST-3s
'The thundering Hyaṅsī river merges (itself) with the Karṇālī.'
(NNC:book-travelogue-humla-bolchha-2062.966 + elicitation NP 2012)

Aphu can also be used as a polite reference to second person. Strangely enough, it does not require DAT in that sense:

(132) a. *Yo sʌhʌr aphu*(-lai) bigar-ch-ʌ.*
PROX city REFL-DAT destroy-NPST-3s
'This city destroys itself.'

b. *Yo sʌhʌr aphu(-lai) bigar-ch-ʌ.*
PROX city 2s-DAT destroy-NPST-3s
'This city destroys you.' (elicitation NP 2012)

(133) shows the obligatoriness of DAT with versatile demonstratives with human reference.

(133) a. *Es*(-lai) dekh-ch-ʌ.*
PROX-DAT see-NPST-3s
intended: 'He sees him/her.'

b. *Es(-lai) dekh-ch-ʌ.*
PROX-DAT see-NPST-3s
intended: 'He sees it.' (elicitation GP 2010)

The nominal demonstratives *u* [DIST] and *ko* 'who' have inherently human reference and must therefore always be marked by DAT:

(134) *Ajʌ kʌs*(-lai) bheʈ-eko?*
today who-DAT meet-PRFV.PTCP
'Whom have you met today?' (elicitation GP 2010)

U has another, probably cognate sense in which it functions as a filler replacing a word that doesn't come to the mind of the speaker. This variant is pronounced slightly longer and may be followed by a short pause. It allows only inanimate reference and is almost always marked by NOM:

(135) *Mʌi-le u(*-lai) tʌyar gʌr-ẽ.*
1s-ERG FILLER-DAT ready make-PST.1s
'I prepared that, uhm, thing…' (NNC:A001013001.241 + elicitation KP 2012)

In the rare case that the filler *u* is marked by the dative, the stem does not change to the oblique form *us* as with *u* [DIST]. This shows that the filler has started to split away from the demonstrative not only semantically but also morphologically:

(136) *rʌ tyo ʌsʌman u-lai hʌtau-nʌ-lai…*
and MED unequal FILLER-DAT get.rid.of-INF_2-DAT
'and in order to get rid of that unequal thing…' (NNC:A001017001.565)

The other nominal demonstratives *ini* [PROX.MH], *tini* [MED.MH], and *uni* [DIST.MH] are inherently human in the singular but become flexible when combined with *-haru* [PL].

The versatile demonstratives *yo* [PROX] and *tyo* [MED] have flexible reference in all contexts. Li (2007a:1472) claims that DAT is ungrammatical with *tyo* when its referent is inanimate and that it is obligatory when its referent is animate (but not necessarily human). Both claims are wrong, as shown by (137) and (138).

(137) *Golbheɖa-ko abʌsekta pʌr-eko sʌmʌe-ma es-lai prʌyog gʌr-nʌ*
tomato-GEN necessity fall-PRFV.PTCP time-LOC PROX-DAT use do-INF_2
sʌk-i-nch-ʌ.
be.able-PASS-NPST-3s
'This (sugo) can be used in times when one needs tomatoes.'
(NNC:saptahik-art-2061-12-05.504)

(138) *Tyo thulo kukur tel(-lai) her-dʌi ch-ʌ.*
MED big dog MED-DAT watch-PROG be.there-NPST-3s
'That big dog is watching it.' (elicitation NP 2012)

A word which is technically a demonstrative but does not pattern with this group in terms of case marking is the relative pronoun *jo* 'whoever, whatever'. This form allows NOM even when it refers to humans:

(139) *Apʌt-ma jʌs(-lai) bheʈ-ch-ʌu tes-lai pʌtter gʌr-nu pʌr-ch-ʌ.*
misfortune-LOC who.REL-DAT meet-NPST-2MH MED-DAT belief do-INF_1 fall-NPST-3s
'In misfortune you have to trust whomever you meet.' (elicitation SAR 2011)

3.5.9 DOM with proper nouns

Besides pronouns and demonstratives, there is one more word class where DOM has become grammaticalised – proper names require the dative in object position (in active sentences):

(140) *Sita*(-lai) maya gʌr-ch-ʌ.*
Sita-DAT love do-NPST-3s
'He is in love with Sita.' (elicitation GP 2010)

However, a couple of comments are in place here. First, proper names are not morphosyntactically distinct from common nouns in Nepali. Proper nouns can be pluralised (141) and form the head of complex NPs (142) as any other noun.

(141) *Krisnʌ-hʌru tʌ ch-ʌin-ʌn.*
Krishna-PL CTOP be.there-NEG.NPST-3p
'Krishna and the others are not there.' (NNC:book-fiction-radha-2062.2764)

(142) a. *Ram-ko bichoɖ-ma bicʌr-i Sita ro-i-rʌh-ek-i ho-l-in.*
Ram-GEN leaving-LOC poor-F Sita cry-LNK-CONT-PST.PTCP-F COP-PROB.FUT-3fMH
'Poor Sita was probably crying after Ram's leaving.'
(NNC:book-fiction-ekadeshki-maharani-2059.1859)

b. *Mʌʈan-ma uklʌ-nʌ lag-ne Gita rʌ orlʌ-nʌ*
veranda-LOC climb.down-INF_2 be.about-NPST.PTCP Gita and climb.up-INF_2
lag-ne Kesʌri-ko bheʈ hu-nch-ʌ.
be.about-NPST.PTCP Kesari-GEN meeting happen-NPST-3s
'Gita, who is about to climb down onto the veranda, and Kesari, who is about to climb up, meet.' (NNC:book-drama-prempinda-2058.3867)

Further, the DAT constraint only applies to proper names referring to human beings. Place names, for instance, can have both NOM (143a) or DAT (143b):

(143) a. *Pokhʌra choɖ-da gham ʈupi-mathi thi-y-o.*
Pokhara leave-CVB_4 sun peak-SUPER be.there-PST-3s
'When (I) left Pokhara, the sun was above the peak.'
(NNC:book-travelogue-anam-pahadma-2062.3768)

b. *Machapuchre, ʌnnʌpurnʌ, Phewa-hʌru-lai jhik-i-di-ne*
Machapuchre Annapurna Phewa-PL-DAT take.away-LNK-BEN-IPFV.PTCP
h-o bhʌne, tɑ̃ Pokhʌra-lai dekh-nʌ sʌk-ch-ʌs?
be[NPST]-3s 2sLH Pokhara-DAT see-INF_2 be.able-NPST-2sLH
'If somebody took away Mt. Machapuchre, Mt. Annapurna, and Lake Phewa, would you still be able to recognise Pokhara?' (NNC:book-essay-paila-agatma-tekera-2055.107)

The DAT constraint is also not bound to certain nouns but to the way they are used in sentences. For instance, the word *Netrʌ* can (as a name) refer to a person or (like all words) to itself. While the former interpretation is the default and requires DAT (144a), the latter is also possible and allows NOM (144b):

(144) a. *Tʌpaĩ Netrʌ-ji-lai cin-nuhunchʌ?*
2HH Netra-HON-DAT know-NPST.2/3HH
'Do you know Netra?'
b. *Tʌpaĩ Netrʌ-ji cin-nuhunchʌ?*
2HH Netra-HON know-NPST.2/3HH
'Does "Netra-ji" ring a bell with you?' (elicitation BP/NP 2012)

DOM with human proper names is thus different from DOM with pronouns and demonstratives in that the class it defines cannot be pinned down using morphosyntactic criteria.

3.5.10 Modification

Modification would be expected to be relevant for DOM because nominal modifiers can often change an NP's referential profile substantially. Versatile demonstratives can mark a referent as specific or definite, and numerals and other quantifiers can make it quantifiable. Possession in a narrow sense (= ownedness) can bring referents closer to the human sphere, and adjectives and relative clauses can single out referents.

Inspite of all this, the effect of modification on DOM is minimal. In most cases adding modifiers to an NP does not change anything at all, as shown in the examples for modifying demonstratives, numerals, and adjectives below.

(145) a. *Yo kʌlʌm(-lai) dekh-ẽ.*
PROX pen-DAT see-PST.1s
'I saw this pen.'
b. *Mʌi-le ek hʌjar-wʌʈa hatti-hʌru(-lai) dekh-ẽ.*
1s-ERG one thousand-CLF elephant-PL-DAT see-PST.1s
'I saw one thousand elephants.'
c. *Hijo rati mʌi-le sano bʌcca(-lai) dekh-ẽ.*
yesterday at.night 1s-ERG small child-DAT see-PST.1s
'Last night I saw (a/the) small child(ren).' (elicitation GP 2010)

The only type of modifier that sporadically seems to have an impact is possessors. Possessors that are high with respect to DOM-relevant factors (highly animate, specific, highly topical etc.) may "rub off" on their low possessums, causing them to be marked by DAT where it otherwise wouldn't be expected:

(146) *Uni-hʌru-le me-ro parthʌkke-lai buj-nʌ sʌk-en-ʌn.*
DIST-PL-ERG 1s-GEN secession-DAT understand-INF_2 be.able-NEG.PST-3p
'They couldn't understand my secession.'
(NNC:book-criticism-samakalin-samalochanako-swarup-2061.3505)

This phenomenon is, however, rather a functional one and has nothing to do with the syntactic status of the possessor. In (147), the possessor is not in the genitive but expressed by a relative clause but still has the same effect:

(147) *Mʌ-sə̃ bhʌ-ek-a kwaliʈi-hʌru-lai bistari bistari rekʌɖ-ma lyau-nu pʌr-ch-ʌ.*
1s-COM_4 COP-PST.PTCP-PL quality-PL-DAT slowly slowly record-LOC take-INF_1 fall-NPST-3s
'The qualities I have should be taken into a record slowly.' (NNC:A001013001.823)

3.5.11 Unexpectedness

Besides topicality, a range of other information-structural phenomena are relevant to DOM in Nepali. I will summarise these under the heading of unexpectedness because they share the trait of having an unexpected referent in O. Unexpectedness is, of course, closely related to focus. I have chosen not to use the latter for two reasons. First, focus is a dangerously ambiguous term. Gundel (1994) distinguishes between psychological focus (i.e. the current center of attention), semantic focus (new information), and contrastive focus ("prominence" with the purpose of contrasting or emphasising). König (1991) spots even more types – see König (1991:32) for a list and proponents of the various views. Second, focus is in some terminological traditions a complementary term to topic. In this perspective, saying that both topic and focus influence DOM would be tantamount to saying that everything influences DOM. What's more, unexpected topics (a possible combination) would have to be translated with the paradoxical expression "focused topics".

The term "contrastive" would seem to offer a way out of this dilemma ("contrastive focus" and "contrastive topic" are both usual), but then on the other hand, contrastiveness in a classical sense is only one of several motivations for unexpectedness. Below are two first examples where DAT seems to be motivated by O being under contrastive focus (148) and a contrastive topic (149), respectively. These are long sentences, so the relevant objects are marked bold.

(148) *Bhuʈan-k-a Nʌres Ɉiŋme Siŋge Waɲcuk-le bhʌn-nubhʌeko ch-ʌ sʌbʌi*
Bhutan-GEN-OBL Nareś Jiṁme Siṅge Wāñcuk-ERG say-2/3HH.PRFV.PTCP be.there-3s all
nagʌrik-hʌru-le so sʌmbidhan-ko mʌsyʌuda hosiyaripurbʌk pʌɖ-i
citizen-PL-ERG that.same constitution-GEN draft carefully read-CVB_2
aph-n-a sujhab-hʌru sʌmʌe-mʌĩ pʌʈha-e-ma ***upʌyogi***
REFL-GEN-PL suggestion-PL time-LOC.FOC send-NMLZ-LOC helpful
ʈhan-i-ek-a sujhab-hʌru-lai *sʌmabes gʌr-nʌ sʌk-i-ne*
deem-PASS-PRFV.PTCP-OBL suggestion-PL-DAT inclusion do-INF_2 be.able-PASS-IPFV.PTCP
ch-ʌ.
be.there-3s
'Nareś Jiṁme Siṅge Wāñcuk of Bhutan has said that if all citizens read the present constitution draft carefully and send their own suggestions in time, it will be possible to include suggestions that are deemed helpful.' (NNC:bbc-news-2061-12-14.18)

(149) ***Pʌriwʌrtit phlor mulle-lai*** *li-erʌ gʌlʌĩca nikasikʌrta-hʌru-bic mʌtʌikke kaem*
changed floor price-DAT take-CVB_1 carpet exporter-PL-between consensus settled
hu-nʌ sʌk-eko ch-ʌin-ʌ.
be-INF_2 finish-PRFV.PTCP be.there-NEG.NPST-3s
'So far no compromise has been reached between the carpet exporters with respect to accepting a changed floor value.' (NNC:a01.16)

In these examples, *sujabhʌru* 'suggestions' and *mulle* 'price' are both nouns, and their referents are inanimate and new in discourse. DAT is therefore not due to any of the factors that have been discussed so far. In (148), the citizens' suggestions are mentioned for the first time in the text with *pʌʈhau-* 'send' and then again with *sʌmabes gʌr-* 'include'. The second, DAT-marked instance is unexpected because although everybody can give sugggestions, only the "helpful" ones will be taken into consideration. In this case the DAT-marked reference is also in contrastive focus.

In (149), the floor price of woolen carpets has been mentioned a couple of times before and can therefore be considered topical. The last floor price was lower than the present one and exports went down after it was marked up, so now some exporters make demands to lower it again. The new price that is not agreed upon yet is in contrast with the accepted present price and therefore unexpected. Since the floor price is topical, this can be considered as a case of contrastive topic.

We will now look at cases where contrastiveness alone has no effect or where unexpectedness is due to other factors. Let us first consider the former. DAT is impossible in (150) and (151) in spite of both O being under contrastive focus.

(150) *Mʌi-le jhyal hʌinʌ ɖhoka(*-lai) khol-nʌ bhʌn-ẽ.*
1s-ERG window NEG door-DAT open-INF_2 say-PST.1s
'I said open the door, not the window!' (elicitation KP 2012)

(151) *Us-le nilo hʌinʌ hʌriyo kitab(*-lai) pʌɖ-ch-ʌ.*
DIST-ERG blue NEG green book-DAT read-NPST-3s
'He reads the green book, not the blue one.' (elicitation KP 2012)

(151) is more interesting than (150) because the contrasting alternatives belong to the same category (books). This puts this example very close to (148), yet DAT is impossible here. It is unexpectedness that distinguishes the two. In isolated examples like (150) and (151), there is no reason why one referent should be less expected than the other – the only motivation is given by the *hʌinʌ* [NEG] in the sentences themselves, but that doesn't seem to be enough to cross the threshold for DAT. By contrast, the unexpectedness of the "useful suggestions" in (148) has a broader base: after hearing that everybody is allowed to make suggestions, one might think that all suggestions will be considered. But such expectations are cancelled in the following clause, and the first marker of that cancellation is the dative at its head.

Once one assumes that contrastive focus has to be combined with unexpectedness in order to work, it becomes easy to produce focused O-DAT in elicitation, too. Below are two more examples. The O in (152) is unexpected because the situation of men is more often looked at than the situation of women. (153) is even more telling: throwing away stuff is nothing unusual, especially out of context, so the first O is not unexpected and therefore can't have DAT, but keeping just what one can sell (instead of what one likes most) is decidedly odd, so the second O is unexpected and therefore can have DAT.

(152) *Purus-ko hʌinʌ mʌila-ko ʌbʌstha(-lai) her-nu pʌr-ch-ʌ.*
man-GEN NEG woman-GEN situation-DAT look.at-INF_1 fall-NPST-3s
'One should look at the situation of the women, not of the men.' (elicitation BP/KP 2012)

(153) *Us-le dherʌi kura(*-lai) phal-i-sʌk-y-o tʌrʌ us-le pʌchi bec-nʌ*
DIST-ERG much thing-DAT throw-LNK-COMPL-PST-3s but DIST-ERG later sell-INF_2
sʌk-ne(-lai) rakh-eko ch-ʌ.
be.able-IPFV.PTCP-DAT keep-PRFV.PTCP be.there.NPST-3s
'He has thrown away a lot of things but he has put aside what he can sell later.'
(elicitation BP/KP 2012)

To be fair it should be noted, though, that unexpectedness does not explain everything. There are sentences like (155) where DAT is possible under simple contrastive focus and in spite of both alternatives being about equally usual:

(154) *Tini-hʌru-le Mao-ko hʌinʌ Marks-ko siddantʌ(-lai) pʌchya-e.*
MED.MH-PL-ERG Mao-GEN NEG Marx-GEN theory-DAT follow-PST.3p
'They embraced Marx' theory, not Mao's.' (elicitation BP/KP/NP 2012)

One might speculate that (154) is possible because Maoism is much more popular than Marxism in Nepal and even a default for the whole political left, so following Marx is deviant. However, DAT stays grammatical when the statement is reversed ('They embraced Mao's theory, not Marx'.'). A weak factor favouring DAT here may be the human possessor of *siddantʌ* 'theory' (cf. section 3.5.10 above).

We will now turn to various cases that can be explained via unexpectedness even though contrastiveness is not involved at all. In the first example in (155), the DAT-marked object refers to the fact that a certain Nepali song became a world hit. This was unexpected for the hearer (who composed the song) as well as for many other people:

(155) *Es-lai cʌ tʌpai-le kʌsʌri her-nubhʌko ch-ʌ?*
PROX-DAT RETRV 2/3HH-ERG Q.METHOD look.at-PST.PTCP.2/3HH be.there-3s
'How have you experienced this?' (NNC:A001013001.36)

Another case is presented in (156). Here the object does not refer to an unlikely referent but is unlikely as an object, the reason being that it is so abstract and broad that an event effecting the whole of it is a very rare case:

(156) *Tini-hʌru-ko bicar thi-y-o, tehã test-a kei* ***bʌstu-hʌru***
MED.MH-PL-GEN thought be.there-PST-3s MED.LOC MED.SORT-PL some thing-PL
bheʈ-i-ne ***sʌmbhabʌna*** *ch-ʌ jun bheʈ-i-y-o* ***bhʌne***
find-PASS-IPFV.PTCP possibility be.there-3s which.REL find-PASS-PST-3s COND
sə̃sar-ʌi-le ***manʌbiyʌ sʌbheta-ko itihas wa bikas-lai*** *ʌrko kisim-le*
world-FOC-ERG human civilisation-GEN history or development-DAT other type-ERG
byakkhe gʌr-nu pʌr-ch-ʌ.
description do-INF_1 fall-NPST-3s
'They believed that there was the possibility of discovering some things, and if these were discovered the world would have to rewrite the history or development of human civilisation in another form.' (NNC:book-fiction-alikhit-2058.156)

The sentence in (157) is taken from a newspaper article. The object is unexpected both for the hearer of the original utterance and for the readers of the article. The object is, however, not unlikely by itself; rather, it is simply an aspect that has (unjustifiedly) not been looked at so far, so it derives its unexpectedness from discourse:

(157) *ʌddhecche Sresʈhʌ-k-a ʌnusar* ***utpadʌk tʌtha nikasikʌrta-hʌru-ko khas***
chairman Srestha-GEN-OBL according.to producer and exporter-PL-GEN special
lagʌt-lai *dhyan-ma rakh-erʌ mulle nirdharə̃ɖ gʌr-i-nu awʌsek ch-ʌ.*
cost-DAT mind-LOC put-CVB_1 price assessment do-PASS-INF_1 necessity be.there.NPST-3s
'According to chairman Srestha, it is necessary to consider the special costs of producers and exporters before fixing the price.' (NNC:a01.18)

In (158), unexpectedness results again from the discourse, but the unexpected object is not one of several of a kind as in (157). Instead, its unexpectedness is a result of its association with a certain context. The last couple of sentences before (158) were about the problem of large numbers of Kurdish and Shiite refugees crossing the Iranian border. The pressure on Saddam Hussein has not been mentioned at all, nor has the text said anything about whether he is considering other points, too. Yet the pressure is an unexpected referent in this context for the reader for at least two reasons. First, it belongs to a different stage with different actants (Iraq instead of Iran, Saddam Hussein instead of Iranian and American officials). Second, the refugee problem suggests a development to the worse, whereas the growing pressure and Hussein's moves as described in the rest of the sentence (loosening the ban on travelling abroad, dissolving the Ba'ath militia) point into the opposite direction.

(158) *ʌrkotirʌ iraki rastrʌpʌti Sʌddʌm Husen-le* ***aphu-mathi bʌɖ-do***
on.the.other.hand Iraqi president Saddam Hussein-ERG REFL-SUPER increase-PTCP.CHAR
dʌbab-lai *dristigʌt gʌr-dʌi…*
pressure-DAT review do-PROG
'On the other hand, Iraq's president Saddam Hussein, considering the growing pressure on him…' (NNC:a01.75)

There is no clear-cut border between this kind of unexpectedness and topic shifting. Consider, for instance, (159). Here, the A arrived at the Queen's Pond just a few sentences ago. He looked at the door to the park around the pond and thought about sticking a letter that he got before to it. Then his attention shifts to the water in the pond, which is marked by DAT:

(159) *Raniphokhʌri-ko carʌitirʌ-baʈʌ bar-i-rakh-eko phʌlame bar-baʈʌ us-le*
Queen's.Pond-GEN four.sides-ABL fence-LNK-put-PRFV.PTCP iron fence-ABL DIST-ERG
raniphokhʌri-ko pani-lai *nihal-y-o.*
Queen's.Pond-GEN water-DAT scrutinise-PST-3s
'He scrutinised the water in the Queen's Pond.' (NNC:book-fiction-prem-ra-mrityu-2057.4188)

(160) presents a rather complicated example. Kābhre and Palāncoka are two areas that together constitute the district of Kābhrepalāncoka. When the districts were created, in principle any other two areas could have been merged so that a district called Kābhrepalāncoka would never have come into existence. At that point the choice of Kābhre and Palāncoka was not particularly surprising, and today the existence of Kābhrepalāncoka is taken as given. Unexpectedness only comes in through the comparison of these two perspectives: if the creation of Kābhrepalāncoka is viewed as a historical *telos* (from the present-day perspective), it was a great coincidence that that *telos* was reached.

(160) *Bhʌn-ʌũ duiʈ-ʌi-lai sʌmeʈ-erʌ Kabhrepʌlancok bhʌ-eko*
say-[OPT.]1p two.CLF-FOC-DAT keep.together-CVB$_1$ Kabhrepalancok COP-PST.PTCP
h-o.
be[NPST]-3s
'Say Kābhrepalāncoka has come into being by keeping precisely these two (areas) together.' (NNC:A001011002.106-112)

O-DAT is also possible when it is not a single referent that is unexpected but a range of several referents that is wider than expected. This is called exhaustive focus by Krifka (2007) and is illustrated by (161) and (162).

(161) *ʈoʈʈʌl tyo renc-lai nʌi blenɖ gʌr-eko ch-ʌ.*
total MED range-DAT FOC blend do-PRFV.PTCP be.there-3s
'It (Nepali music) has blended that whole range (of influences from Pakistan to Bangladesh).' (NNC:A001013001.1291)

(162) *Yo bibhinnʌ bhasa-hʌru-ko git sʌŋgit-lai tʌpaĩ-le yʌuʈ-ʌi yalbʌm-ma sʌmabes*
PROX various language-PL-GEN song music-DAT 2HH-ERG one-FOC album-LOC include
gʌr-ne kunʌi bicar gʌr-nubhʌko ch-ʌ?
do-IPFV.PTCP any thought do-PRFV.PTCP.2/3HH be.there-3s
'Have you ever thought about including these songs and this music from various languages on a single album?' (NNC:A001013001.614)

Another subtype of unexpectedness that is especially far away from contrastiveness can be seen in (163) and (164).

(163) *ʌbʌ teskarʌn caĩ pa-ko euɖa ʌdikar-lai hami-le prʌyog gʌr-nʌ sʌk-nʌ*
now therefore RETRV get-PST.PTCP one.CLF right-DAT 1p-ERG use do-INF$_2$ be.able-INF$_2$
pʌr-y-o.
fall-PST-3s
'Therefore we have to be able to make use of the one right we've got.' (NNC:A001017001.511)

(164) *Us-le sʌ̃sar-ko ʌntim syau*(-lai) kha-i-hal-y-o.*
DIST-ERG world-GEN last apple-DAT eat-LNK-COMPL-PST-3s
'He ate up the last apple in the world.' (elicitation SAR 2011)

The elicited example in (164) is rather extreme since it only allows DAT in spite of the object referent being inanimate and non-topical. What both examples have in common is that O is without alternatives: there are no other rights women could make use of in (163), and all other apples have gone in (164). This subtype shows therefore most clearly that unexpectedness cannot be equalled with contrastiveness, at least not in its usual sense.

The mental setting that is relevant here is rather complex. The first requirement is a world that is perceived as the norm and where there are many tokens of a certain type. In (163) that norm is an ideal world where women should have more than one right, in (164) the norm is the real world where there are lots of apples. There must then be a world that contrasts with the norm in that there are very few or only a single token of the relevant type. In such a world, an O of that type is unexpected because an agent will rarely find his way to it. Using a right or eating an apple is an ordinary activity in a world full of rights of apples, but where there is only a single token of each it becomes rather noteworthy.

This kind of contrast often makes DAT possible where simple unexpectedness would not suffice. For instance, consider the two sentences in (165). DAT on *pʌisa* 'money' is impossible in (165a) even though it is unexpected that a thief should just steal a passport and leave the money in its place. By contrast, DAT is possible in (165b). The difference between the two sentences is that in the first case all the money in a wallet or bag contrasts with a single document, whereas in the second case the money is a small thing compared to all the other things that were stolen. So in the first case it is not surprising that an A found his way to the money but only what he did with it, whereas in the second case both types of unexpectedness are involved because there was relatively little money (compared to all other things that could be stolen).

(165) a. *Tyo cor-le tʌ radani(*-lai) matrʌ lʌg-e-ch-ʌ, tʌrʌ sʌbʌi pʌisa(*-lai)*
MED thief-ERG CTOP passport-DAT only take-PRF-NPST-3s but all money-DAT
choɖ-e-ch-ʌ.
leave-PRF-NPST-3s
'That thief only took my passport but left all the money.' (elicitation BP/KP/NP 2012)

b. *Cor-le sʌbʌi(*-lai) lʌg-e-ch-ʌ, tʌrʌ pʌisa(-lai) choɖ-e-ch-ʌ.*
Thief-ERG all take-PRF-NPST-3s but money-DAT leave-PRF-NPST-3s
'The thief took everything, but he left the money.' (elicitation NP 2012)

A last subtype of unexpectedness can only be established across several clauses. (166) shows an example.

(166) *Ehã aphʌi-le niem bʌn-au-ne rʌ* ***tei niem-lai***
PROX.LOC REFL-ERG rule be.created-CAUS-IPFV.PTCP and PROX.FOC rule-DAT
kaʈ-ne thupr-ʌi mʌhanubhab-hʌru ch-ʌn.
cut-IPFV.PTCP many-FOC squire-PL be.there.NPST-3p
'There are many squires here who create a rule and then break it themselves.'
(NNC:freenepal-fiction-2061-12-8.283)

Niem 'rule' in the second clause is not an unexpected referent by itself, and the breaking of rules is (in isolation) not a particularly surprising event. The rule also does not contrast with another rule that was not broken, nor is it a type of rule that is rarely broken. It only becomes unexpected in connection with the first clause – it is unexpected that one should break a rule one has created oneself, at least in an ideal world.

To summarise, unexpectedness is an important factor in Nepali DOM. It is hard to spot because first all other possible factors must be excluded – the referent in question must at least be inanimate or non-specific and in addition weakly to non-topical. However, where it does become visible it is all the more conspicuous. An object can be unexpected for various reasons, for instance, because it represents itself an unlikely event, because its combination with a certain predicate is unlikely, because its affectedness in general or its repeated affectedness is unlikely, or because the agent's finding it is unlikely; because it is in contrast to something in the discourse or because it is more comprehensive than one might expect. Unexpectedness may be combined with contrastiveness (as in contrastive focus and contrastive topic), but contrastiveness alone does not necessarily yield it.

The big problem with unexpectedness is that although it nicely explains many otherwise exceptional cases, it can hardly be used to predict anything. It is virtually impossible to assess the relevant kind of unexpectedness independently of form (i.e., case marking). If one tries to, DAT will be massively overpredicted because there are many more objects which are in some general sense

unexpected or which have the potential to be unexpected than objects which are unexpected from the subjective perspective of a speaker/writer. Unexpectedness is therefore descriptively useful but not suitable for building predictive models.

3.5.12 Disambiguation

Disambiguation is a prominent candidate when it comes to determining *the* function of DOM – cf. the discussion of "distinguishing approaches" to DOM in Iemmolo (2011:25ff). In Nepali, disambiguation is clearly if marginally relevant.

The default word order in monotransitive clauses is AOV (section 3.3.1), as illustrated by the sentences in (167). If one changes this to OAV as in (168), the preferred case by default does not change. This is especially remarkable in the case of O-NOM, since in OAV the first argument looks like an A in terms of position and its case is ambiguous.

(167) a. *mʌ tyo kitab bhʌre bheʈ-ch-u.*
1s MED book later find-NPST-1s
'I'll find that book later.'
b. *mʌ tyo manche-lai bhʌre bheʈ-ch-u.*
1s MED person-DAT later find-NPST-1s
'I'll find that person later.' (elicitation NP 2012)

(168) a. *Tyo kitab mʌ bhʌre bheʈ-ch-u.*
MED book 1s later find-NPST-1s
'I'll find that book later.'
b. *Tyo manche-lai mʌ bhʌre bheʈ-ch-u.*
MED person-DAT 1s later find-NPST-1s
'I'll find that person later.' (elicitation NP 2012)

Note that in (168a), even though position and case are ambiguous, the role of the first argument is still indicated by its semantics: a book is an unlikely agent, especially when its co-argument is a second person (although of course occasionally one may say things like *So that book finally found you*). Position starts to interact with case as soon as one takes away this indication. For instance in (169), one speaker saw a cat that was standing still and looking alarmed. He asked another speaker, who had a better view on the scene, what had happened, and got the answer that the cat was being watched by a big dog (implying that it knew it was). When the cat occupies the default position for O between A and V as in (169a), DAT is possible. When the cat is fronted to OAV as in (169b), however, DAT becomes obligatory:

(169) a. *Tyo thulo kukur tel(-lai) her-dʌi ch-ʌ.*
MED big dog MED-DAT watch-PROG be.there-NPST-3s
'That big dog is watching it.'
b. *Tel*(-lai) tyo thulo kukur her-dʌi ch-ʌ.*
MED-DAT MED big dog watch-PROG be.there.NPST-3s
'It is being watched by that big dog.' (elicitation NP 2012)

This effect is only possible because in the argument set {cat dog} there is no clear default for role distribution. This effect is not inhibited by the case marking of A but is just the same when A is marked by ERG:

(170) a. *Tyo ʈhulo kukur-le tel(-lai) dekh-y-o.*
MED big dog-ERG MED-DAT see-PST-3s
'That big dog saw it.'
b. *Tel*(-lai) tyo thulo kukur-le dekh-y-o.*
MED-DAT MED big dog-ERG see-PST-3s
'It was seen by that big dog.' (elicitation NP 2012)

This is expected on an incremental processing background: even though (170b) as a whole is un-

ambiguous (the only meaningful interpretation for *-le* here is as an A marker, so the other, fronted argument must be P), the crucial marker comes relatively late in the sentence. Since speakers tend to integrate words as soon as possible into the syntactic structure they have have built so far (Gompel and Pickering 2007:289), even a temporary ambiguity is not desirable. In this case it holds long enough to make the DAT on the fronted P obligatory.

In the following group of examples, A and P marking can be observed independently of each other. The sentences become more grammatical when A is marked by ERG because killing bears is something conceived as typical of a hunter (see section 3.3.3.1 for functions of DAM). AP is better than PA, and fronted P only becomes fully acceptable when it is at the same time marked by DAT. Note that specificity is irrelevant here, so all sentences could mean 'The hunter kills the bear', 'A hunter kills a bear', or 'Hunters kill bears'. Grammaticality is marked at the beginning of each sentence, with "!" marking the default.

(171) a. ?*Sikari bhalu mar-ch-ʌ.*
hunter bear kill-NPST-3s
'Hunter kills bear.'
b. *Sikari-le bhalu mar-ch-ʌ.*
hunter-ERG bear kill-NPST-3s
'Hunter kills bear.'
c. ?*Sikari bhalu-lai mar-ch-ʌ.*
hunter bear-DAT kill-NPST-3s
'Hunter kills bear.'
d. *!Sikari-le bhalu-lai mar-ch-ʌ.*
hunter-ERG bear-DAT kill-NPST-3s
'Hunter kills bear.' (elicitation KP 2012)

(172) a. **Bhalu sikari mar-ch-ʌ.*
bear hunter kill-NPST-3s
'Hunter kills bear.'
b. ?*Bhalu sikari-le mar-ch-ʌ.*
bear hunter-ERG kill-NPST-3s
'Hunter kills bear.'
c. ?*Bhalu-lai sikari mar-ch-ʌ.*
bear-DAT hunter kill-NPST-3s
'Hunter kills bear.'
d. *Bhalu-lai sikari-le mar-ch-ʌ.*
bear-DAT hunter-ERG kill-NPST-3s
'Hunter kills bear.' (elicitation KP 2012)

Apart from sentences where the word order of A and O is reversed, DAT for disambiguation is also found in sentences where O is far away from the predicate. (173) shows an example for this, where the relevant O *sʌmpʌti* 'property' is separated from the associated predicate *bec-* 'sell' by a converbial clause. Leaving DAT away is marginally possible here but according to an informant makes the sentence harder to understand. By contrast, if the O NP is moved next to *bec-*, both DAT and NOM become equally possible.

(173) *ʌbʌ* ***aphu-le ʌŋsʌ pa-eko sʌmpʌti?(-lai)*** *ʌbʌ chorachori-sʌŋgʌ mʌnjuri*
now REFL-ERG share get-PRFV.PTCP property-DAT now children-COM permission
nʌ-li-i bec-nʌ pa-e bhʌnerʌ tes-lai caĩ ʌbʌ durupʌyog gʌr-nʌ
NEG-take-CVB$_2$ sell-INF$_2$ get-COND CIT MED.OBL-DAT RETRV now abuse do-INF$_2$
bhʌ-en-ʌ, hʌinʌ.
be-PST.NEG-3s QTAG
'Now just because one gets the chance to sell property one holds a share of without taking one's children's permission that doesn't mean one will abuse this right.'
(NNC:A001017001.519 + elicitation SAR 2011)

This phenomenon can be explained in the same terms as AO inversion. (173) allows two interpretations for the location of the A of *bec-*. Either it is marked overtly by *aphule*, or *aphule* syntactically pertains to the relative clause predicate *pau-* 'get' and the A of *bec-* is covert. In the first case the order of the relevant A and O is regular, in the second case A does not have a position at all, so AO inversion is not given in either case.

However, a NOM-marked *sʌmpʌti* would still be uncomfortably unambiguous here for two reasons. First, in a long elaborate sentence such as this one, it can't be taken as granted that a NOM-marked argument that is not A is O. Thus, even though *sʌmpʌti* is semantically unlikely as an agent, it could, for instnce, belong to another subordinate clause as S. Second, the crucial factor for disambiguating the affiliation and role of *sʌmpʌti* here is the predicate *bec-* itself, which requires a T (= O) referent. Thus, the greater the distance between T and the predicate, the longer the hearer has to wait until he can fix the role of *sʌmpʌti*. Using DAT facilitates an early role assignment. Of course DAT is by no means an unambigous marker of O either, but O is in this sentence the most likely function marked by it.

A similar case is found in (174). There are again several linked predicates sharing an A, which occupies the first position in the clause. The T (= O) of the second predicate *gʌwau-* 'make sing' directly follows the A but is separated from *gʌwau-* by the first predicate together with its own ornate T and is therefore marked by DAT:

(174) *Mʌi-le* ***Nepali git-lai*** *ʌru pʌni bidesi kʌlakar-hʌru, ramr-a ramr-a utkristʃʌ*
1s-ERG Nepalese song-DAT other also foreign artist-PL good-PL good-PL excellent
kʌlakar-hʌru lya-erʌ mʌi-le gʌw-a-ẽ.
artist-PL bring-CVB_1 1s-ERG sing-CAUS-PST.1s
'I also brought other foreign artists, really good, excellent artists, and had them sing Nepalese songs.' (NNC:A001013001.1397-1401)

In this sentence it is even more apparent that AO inversion is not the reason for case marking – both A and O occupy their regular positions in relation to each other. That the distance to *gʌwau-* is really the factor conditioning DAT here is shown by the fact that DAT becomes impossible as soon as one moves *git* to the left of *gʌwaẽ* (elicitation KP 2012). As in (173), fronted O-NOM is again marginally possible here, but only with a small iconic pause after *git* indicating that it does not belong to the following nonfinite clause.

In (175), neither AO inversion nor distance to the predicate can explain the dative on *prʌstab-hʌru* 'proposals'. Disambiguation may still be involved, though, because there is a great distance between A and O. The dative towards the end of the sentence fits together with the ergative from its beginning and thus makes it easier for the reader to associate both of them with the matrix predicate *ʌnumodʌn gʌr-* 'approve':

(175) *Rasʃtrʌsʌŋghiyʌ surʌcche pʌrisʌd-le Irak-k-a rasayʌnik hathʌtiyar-hʌru-ko*
United.Nations security council-ERG Iraq-GEN-PL chemical weapon-PL-GEN
niricchɑ̃ɖ gʌr-nʌ ja-ne bisesʌgge-hʌru-ko mukti-kalagi ***Irak-dwara prʌstut***
inspection do-INF_2 go-IPFV.PTCP specialist-PL-GEN freedom-FIN_1 Iraq-by presented
prʌstab-hʌru-lai *ʌnumodʌn gʌr-eko ch-ʌ.*
proposal-PL-DAT approval do-PRFV.PTCP be.there.NPST-3s
'The UN Security Council has approved proposals made by Iraq for (improving the) freedom of specialists going (there) to inspect Iraq's chemical weapons.' (NNC:a02.74)

3.5.13 Affectedness

One last factor that plays a role for Nepali DOM is the degree of affectedness of the object. The more strongly a referent is affected, the more likely it is to be marked by DAT. This explains, for instance, why DAT is ungrammatical in (176a) but at least marginally possible in (176b) in spite of the object being inanimate and non-specific:

(176) a. *ʌru-ko saman(*-lai) cor-nu hũ-dʌin-ʌ.*
other-GEN thing-DAT steal-INF$_1$ be.good-NEG.NPST-3s
'One shouldn't steal others' things.'
b. *ʄun-sukʌi saman(?-lai) bigar-ch-ʌ.*
which-ever thing-DAT destroy-NPST-3s
'He destroys all kinds of things.' (elicitation BP/KP 2012)

Affectedness can also explain differences such as those in (177). Unexpectedness combined with a "weak" verb as in (177a) does not yield any results, but unexpectedness in combination with a "strong" verb as in (177b) makes DAT acceptable:

(177) a. *Us-le aiphon hʌinʌ samsuŋ(*-lai) roj-y-o.*
DIST-ERG iPhone NEG Samsung-DAT choose-PST-3s
'He didn't choose an iPhone but a Samsung.' (elicitation KP 2012)
b. *Tini-hʌru-le purano ghʌr bhʌtk-a-en-ʌn tʌrʌ nʌyã ghʌr(-lai)*
MED.MH-PL-ERG old house fall.down-CAUS-PST.NEG-3p but new house-DAT
bhʌtk-a-e.
fall.down-CAUS-PST.3p
'They didn't pull down the old house but the new one.' (elicitation KP 2012)

The only verb known so far that can have both NOM and DAT based exclusively on affectedness is *sun-* 'hear'. When something is heard directly, DAT may be used if the remaining factors allow it. However, when something is heard *of* (i.e. indirectly), DAT is never used:

(178) a. *Tel-lai sun-ch-ʌu?*
MED-DAT hear-NPST-2s
'Do you hear that?' (field notes 2011)
b. *Tyo tʌ sun-ya ch-ʌin-ʌ.*
MED CTOP hear-PRFV.PTCP be.there-NEG.NPST-1s
'I haven't heard of it.' (field notes 2011)

If the activities that affect referents most severely are those that make them cease to exist (such as *bigar-* 'destroy' and *bhʌtkau-* 'pull down' in the examples above), the other extreme would be activities that make a referent come into existence, i.e. activities with effectuated objects. Adhikārī (2052 V.S.) notes that such objects cannot be marked by the dative:

(179) *Bʌtti(*-lai) kat-y-o.*
light-DAT light-PST-3s
'He lit a light.' (Adhikārī 2052 V.S.:77)

Other effectuating verbs are *bun-* 'weave', *bʌnau-* 'build', *khic-* 'shoot (a photo)' (Adhikārī 2052 V.S.:77).

All in all affectedness is the least important factor in DOM. In most cases it only licenses DAT or excludes it as in (179), but it cannot be *marked* by case in the sense that DAT alone could mark the degree of affectedness or NOM could mark effectuation – the house in (178b) is not necessarily less affected when it has NOM than when it has DAT. Affectedness is, however, rather interesting for theoretical reasons: since it is to a large degree determined by the lexical semantics of the verb, it is the only factor which is not located on the object referent itself.

3.5.14 Some irrelevant variables

After the long list of variables presented above it may appear to the reader that simply everything is relevant to DOM in Nepali that could be imagined to be relevant. Although this comes close to the truth, there are a few notorious variables that do not seem to have any influence on DOM. These are briefly listed below with a few illustrating examples. As with the variables irrelevant for Chintang S/A detransitivisation (section 2.6.6), this section is not meant to provide an in-depth discussion but simply lists some data for the sake of interest. Also note that the irrelevance of some

other variables has already been noted *en passant* above (social status in section 3.5.8, modification in section 3.5.10, contrastiveness in section 3.5.11).

3.5.14.1 Aspect

The most famous case where aspect is known to be relevant for DOM is Finnish (see e.g. Kiparsky (1998) and references therein). DOM in Finnish is formally rather exceptional in that it does not exhibit a privative but an equipollent opposition of partitive vs genitive. In Nepali, where the dative contrasts with zero, aspect does not seem to play any role at all, as illustrated by the examples in (180).

(180) a. *Us-le khana pʌka-y-o.*
DIST-ERG food prepare-PST-3s
'He prepared food.'
b. *U khana pʌkaũ-dʌi thi-y-o.*
DIST food prepare-PROG be.there-PST-3s
'He was preparing food.'
c. *Us-le khana pʌkaũ-thy-o.*
DIST-ERG food prepare-PST.HAB-3s
'He used to prepare food.' (elicitation NP 2012)

(181) a. *Us-le sathi-lai sod-y-o.*
DIST-ERG friend-DAT ask-PST-3s
'He asked a friend.'
b. *U sathi-lai sod-dʌi thi-y-o.*
DIST friend-DAT ask-PROG be.there-PST-3s
'He was asking a friend.'
c. *Us-le sathi-lai sod-thy-o.*
DIST-ERG friend-DAT ask-PST.HAB-3s
'He used to ask a friend.' (elicitation NP 2012)

3.5.14.2 Polarity

Polarity may also sometimes interact with O marking, as in the case of French, where mass concepts in O require the marker *du* (< *de le* [GEN DEF.M.SG]) with positive predicates but only *de* [GEN] with negative predicates. No such distinction is observed in Nepali, where both NOM and DAT can be freely combined with negation:

(182) a. *Us-le bhat kha-y-o.*
DIST-ERG rice eat-PST-3s
'He ate rice.'
b. *Us-le bhat kha-en-ʌ.*
DIST-ERG rice eat-NEG.PST-3s
'He didn't eat some/any rice.' (elicitation NP 2012)

(183) a. *Us-le sathi-lai bheʈ-y-o.*
DIST-ERG friend-DAT meet-PST-3s
'He met a friend.'
b. *Us-le sathi-lai bheʈ-en-ʌ.*
DIST-ERG friend-DAT meet-NEG.PST-3s
'He didn't meet a/any friend.' (elicitation NP 2012)

Also note that as mentioned in section 3.5.4, the case of O does not entail consequences for the scope of negation – hence the alternative translations in (182b) and (183b).

3.5.15 One form, one function?

There is a certain ideal of marking in linguistics that says that there should be a one-to-one correspondence between forms and functions. This ideal seems to be grounded in the assumption that it cannot be chance that one and the same marker is used in various situations: there has to be a common functional element in all usages, and if that element is to be characteristic of the form in question it must not to be linked to other forms, too.

Nepali *-lai* is a far cry from this ideal. On the one hand, it is a case marker because it forms a paradigm with other, less problematic case markers and because it frequently serves to disambiguate between roles. The dispreference of double datives described in section 3.4.4 can also be interpreted as a tendency against usages of *-lai* where it does not disambiguate. On the other hand, the disambiguating power of *-lai* only unfolds as soon as one knows the predicate class – before one sees the predicate, an argument marked by *-lai* could be in any role: S or A with experiencer predicates, P with monotransitives, T and G with various classes of transfer ditransitives. What's more, none of these roles requires DAT across the lexicon. The roles summarised as O do not even do so within single predicate classes because of DOM.

Is there a common denominator in all uses of DAT apart from role? One candidate is a mismatch between referential status and control: one could say that DAT is used whenever a high referent occupies a role where it is not in control, i.e. a role covered by O or G (where the high referent would be expected to be in A) or an experiencer or deontic S/A (where the high referent can not control as much as he would be expected to). However, this view is problematic for several reasons. First, it is not clear at all what a high referent is – as we have seen, there are various criteria for determining referential status which do not always move in the same direction, and DAT may be triggered by any of them. We have also seen numerous cases where DAT marks referents that wouldn't be viewed as high under any current definition – cf. e.g. section 3.5.11 on unexpectedness and section 3.5.13 on affectedness. Moreover, status-control mismatches are only marked by DAT within a restricted area. For instance, A lacking control can (and must) only be marked by DAT within experiencer and deontic predicates but get NOM or ERG like all other A with other predicates.

If there is no common function for *-lai*, why does it exist at all? Should it be treated as a simple case of homophony? Since there are clear functional links between the various usages of *-lai*, this would mean going unnecessarily far into the opposite direction. One important link is the mentioned concept of status-control mismatch, the other one the complex of functional correlations between the many factors associated with *-lai* (animacy, specificity, topicality etc.). One way to reconcile the existence of such links with the lack of a single function for *-lai* is to assume that *-lai* historically started out with a single function, which was subsequently extended based on precisely these links.

As will be shown below in the section on the history of DOM (3.7), the first attested function of *-lai* is to mark recipients (animate G), which is later extended to experiencer S/A and O. There is a metaphorical bridge between G and experiencer S/A in that the latter can be construed as the locations of experiences (i.e. the place where an emotion or a sensation manifests itself). The constant use of the same marker with G of a certain type (high recipients) and experiencer S/A is likely to create an association between *-lai* and high referential status independently of role, which opens up a pathway for the extension of *-lai* to high O referents.

Since most O referents are not high in any of the relevant dimensions (cf. section 3.6 below), *-lai* can acquire an additional function from this point as a marker of O with unusual properties. This explains why *-lai* can mark unexpectedness and also why it is used in cases of ambiguity – position is yet another aspect that may make an O unusual.

Finally, the use of *-lai* with highly affected O as well as with deontic S/A can be thought of as an extension of the idea of status-control mismatches to cases where the status of an O referent is not particularly high but the effect exerted by A is particularly big, or where A has a high status but lacks control and therefore becomes similar to high O, respectively.

In this view, DOM in Nepali is not a single homogeneous phenomenon with one function but rather a result of the diachronic correlation of multiple grammaticalisation paths. The existence

of a single marker is not due to the existence of a single function but to continuous functional extension to domains where no marker had been present before. Synchronically *-lai* is conditioned by a plethora of intertwined functions, most notably so within DOM.

3.6 Quantitative analysis based on corpus data

3.6.1 Introduction

For the quantitative analysis parts of the Nepali National Corpus (cf. section 0.4) were annotated for various types of information. The annotation was carried out by two native speakers who had a Master's degree in linguistics and myself and was based on guidelines that will be discussed below. Altogether 9558 sentences containing 35,776 words were annotated. Annotations were examined for consistency and extracted from the corpus using Perl scripts (see appendices C.5, C.4). A CSV file containing one observation (= NOM/DAT marked object) per line was output and used as the base for statistical analysis with R (R Development Core Team 2012). I used the additional packages ltm (Rizopoulos 2011) and rms (Harrell 2011). The analysis scripts are appended in sections C.6, C.4.

The advantage of getting the help of native speakers for the annotation was that they could parse Nepali sentences much more quickly than I could, so they were most efficient in annotating syntactic structure, roles, and referential identity. However, when it came to some of the central variables bound to referents (for instance, animacy and specificity), it turned out that they were not familiar with a lot of basic typological concepts and had a steep learning curve before them. For this reason we resorted to a twofold strategy: while syntactic structure, roles, and referential identity were annotated without my help, I supervised the annotation of all other variables, paying special attention to fuzzy concepts such as specificity. I double-checked all of their annotations and discussed cases where I would have chosen a different value, bringing together my linguistic with their native speaker intuitions, until we found a solution that was agreeable to both of us. Nevertheless, this procedure introduced a considerable degree of uncertainty into the data.

In order to collect as much data as possible, only eligible objects were annotated for the full set of variables. Eligible objects were defined as overt O (in the technical sense, see section 3.4.2) marked by NOM or DAT. It was considered irrelevant whether the other case would have been possible given the full set of annotated variables. For instance, pronouns in object position must be marked by DAT as described in section 3.5.8 but were nevertheless annotated for the full set of variables. The aim of this was to make as few prejudgements as possible. The drawback of ignoring non-eligible objects is that statements about P/T/G in general may be biased – only about 46% of all P/T/G are eligible.

Figure 3.2 shows counts for some argument types in the annotated subcorpus. The most interesting fact emerging from these data is that the dative is to some degree characteristic of P/T/G and O and vice versa. Whilst about 8% of all overt arguments are marked by DAT, this proportion is notably higher with P/T/G and O (13% and 12%, respectively). Similarly, DAT more often indicates P/T/G than other argument roles – 74% of all DAT-marked arguments are P/T/G. Note, however, that only 40% of all DAT-marked arguments have DAT due to DOM (bar "O-DAT"); the remaining datives are fixed by the predicate frame (experiencer S/A-DAT, animate G-DAT etc.). The datives that are due to DOM constitute 54% of all DAT-marked P/T/G. This illustrates nicely that DOM is only one out of many factors determining DAT and that speakers are exposed to roughly equally many DAT that are due to DOM as are due to other factors.

When looking at the proportions of DAT- and NOM-marked arguments, NOM clearly emerges as the default O case with 2788 instances (88% percent of O). This justifies formulating rules with DAT as the marked case as we have usually done so far (e.g. "highly animate referents are marked by DAT" instead of "lowly animate referents are marked by NOM").

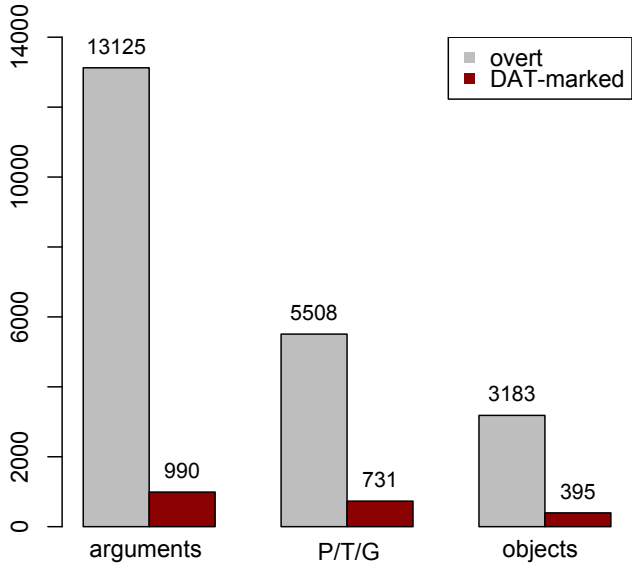

Figure 3.2: Counts of overt and DAT-marked arguments

3.6.2 Syntactic annotation and primary variables

Altogether eleven primary variables were annotated. Nine of these were directly relevant for the modelling of DOM, whilst two (domain and identity) were only relevant for the calculation of derived ("secondary") variables. Below is a short summary of all variables with their values. See Appendix B for the full annotation guidelines.

As for Chintang, a syntactic skeleton is provided by domains in the form of numeric IDs. Arguments and predicates that syntactically belong together get the same ID, and subordinate clauses get IDs under the matrix ID (e.g. `2/1/1`).

Another indirectly relevant variable is referential identity. This variable is crucial for calculating topicality-related secondary variables (see section 3.6.3 below). A unique alphanumeric ID was given to each referent, and when that referent came around later the same ID was used again. All NPs and all zero arguments were assumed to represent referents.

Two steps were taken in order to prevent incidental ID splits (one referent referred to by several IDs) and ID mergers (one ID referring to several referents). First, after finishing annotating a file, a Perl script (appendix C.4) was used to extract a list of all referents sorted by frequency. Annotators were instructed to inspect the referent list and look out for referents that they thought were important in the text but featured low on the list (an indicator of a potential ID split). ID mergers mostly occur where several similar unimportant referents occur with long stretches of text between them. For instance, although *problem* is a frequent word in newspaper articles, most problems are not identical to each other. In order to avoid a merger caused by giving the ID `problem` to all of them, a system was introduced whereby the IDs for referents that would presumably not come up again could be optionally composed from an ordinary ID and the domain in which the referent occurred (e.g. `problem37`, `problem90`).

A first variable that is directly relevant to DOM is role. The available roles were S, A, P, T, G. The definition was based on Dowty (1991) and Bickel (2011) with some simplifications, especially in the definition of ditransitives, where actual or metaphorical movement was taken as the central criterion distinguishing T and G. For copular clauses, the special labels CT (copular theme) and CR (copular rheme) were used instead of roles.

Role alone is not enough to determine eligibility for DOM. Since neither a syntactic parser nor an electronic valency dictionary were available for Nepali, eligibility was annotated by hand. For this, a variable called DOM was tagged on all P/T/G. If an argument was eligible, DOM would

take one of the values NOM, DAT, or GEN. GEN-marked O were filtered out later. Non-eligible P/T/G (including all zeros) got the value NA (= non-applicable).

The remaining variables are those which are most relevant to DOM. They were only tagged on eligible O. All variables have a value x for cases where the annotator didn't feel sure. These cases were ignored in the statistical evaluation. Below is an overview of all variables and values. Details can again be found in the full guidelines (Appendix B).

- **animacy** – animacy, partly amalgamated with closely related variables:
  - **human** – an ordinary human referent
  - **human.fam** – relative or in-law
  - **human.prop** – human proper name
  - **human.group** – a group of human beings designated by a singular noun such as *army*
  - **high.anim** – non-human mammals and birds
  - **mid.anim** – all other animals
  - **low.anim** – plants, mushrooms, bacteria
  - **thing** – non-animate, touchable object
  - **state** – non-touchable object that can be defined independently of time
  - **process** – non-touchable object that can only defined with reference to time
- **quantifiability** – as defined in section 2.6:
  - **qnt** – quantifiable
  - **nonq** – non-quantifiable
- **situation** – the fixedness of the time and place of the event associated with an object:
  - **concrete** – the event has a clear place and time
  - **exemplary** – the event is singular but could take place at any time
  - **general** – the event has a place but no clear time
  - **abstract** – the event has neither place nor time
- **ctag** – the variable representing part of speech. Parts of the NNC had been automatically tagged for parts of speech using this attribute. The name was kept, but for the statistical evaluation the elaborate NNC system documented in Hardie (2005) was mapped to the same simple system we used for untagged texts. The allowed values were:
  - **n** – noun
  - **adj** – adjective
  - **pro** – a pronoun as defined in section 3.2.1 or *tʌpaĩ* [2HH]
  - **dem** – any demonstrative
  - **other** – all other parts of speech
- **modification** – various kinds of modifying elements:
  - **none** – bare NP without any modifiers
  - **adj** – adjective
  - **relclause** – relative clause
  - **humposs** – a human possessor in the form of a possessive pronoun or a genitive NP
  - **latposs** – a human possessor coded in another way
  - **poss** – non-human possessive pronoun or genitive NP
  - **num** – numeral
  - **dem** – demonstrative
  - **interrog** – interrogative
  - **sortal** – a sortal modifier such as *esto* [PROX.SORT]
  - **sortal.q** – the sortal interrogative *kʌsto* [Q.SORT]
  - **other** – any other modifier

    - **several** – several modifiers at a time

- **focus** – various types of focus:

    - **nofoc** – no focus
    - **contrast** – contrastive focus
    - **fragile** – the kind of focus that comes about when the illocutionary force of an utterance is felt to crucially depend on an object that is rare, singular, unlikely, or hard to achieve

- **diathesis** – the diathesis of the verb:

    - **Ø** – verb forms without passive marking were not annotated for diathesis at all. Their value was later automatically converted to `active`
    - **passive** – verb forms with the suffix *-i* [PASS] and a transitive role set. Spontaneous passives without an A such as *pani rok-i-y-o* [rain stop-PASS-PST-3s] 'the rain stopped' were annotated as having a single argument, as were light verb passives (section 3.4.7).

As is evident from the list above, the variables used for the annotation are not fully congruent with the variables discussed in section 3.5. Some variables have more fine-grained values (animacy, modification, focus), some are missing (specificity, topicality, disambiguation, affectedness), and yet others have not been discussed at all before (situation). I will briefly summarise the reasons for these divergences before proceeding.

- Animacy and modification were annotated in great detail because they belong to the least controversial variables, so broadening the spectrum of values does not decrease reliability. With animacy there was the question whether an extended, intuitively arranged hierarchy would be confirmed by annotation data. The individual/mass distinction is covered by quantifiability and was therefore not integrated into animacy. The distinction between states and processes is also useful for capturing the effect of complex predicates, where N mostly codes a process.

- Modification did not have a strong effect in elicitation, so here the question was whether individual specific types of modifiers would have a clearer impact.

- Focus/unexpectedness has not been discussed in the previous literature, nor did my initial elicitation work provide any hints as to its relevance. It was only after I started working with the NNC that more and more examples came up where inanimate, lowly topical referents were marked by DAT and which could best be explained via some kind of focus. "Classical" contrastive focus was introduced on suspicion, and "fragile" focus was the initial, fuzzy label for the remaining known examples. When it became clear that contrastive focus was as good as irrelevant and fragile focus was all about unexpectedness, so many texts had already been tagged that is was no longer possible to change the system. This is hopefully excusable given that focus/unexpectedness is still the variable that is hardest to grasp. The approximate congruence between fragile focus and unexpectedness may give some hints to its relevance, but further theoretical work is needed here, anyway.

- In the beginning, identifiability was directly annotated with the values `definite`, `specific`, and `non-specific`. However, it soon turned out that it was very hard to reach a satisfactory level of agreement on this variable. My own interpretations and those of the native annotators would rarely coincide, and this maybe was no wonder given the difficulty of defining identifiability and my own unavoidable bias as the native speaker of a language with articles. When it turned out at an early stage that definiteness was irrelevant, anyway, I took this as an additional reason and abandoned identifiability. It was replaced by quantifiability, which was already known to be relevant from Chintang, and by the experimental variable situation – the idea here was that specific referents would most often be found in a situation with a clear place and time. Interestingly, although quantifiability remained a frequent reason

for disagreement among annotators, it turned out to be much easier to annotate than specificity. This is remarkable because the two refer to almost the same thing from two different perspectives.

- Topicality and disambiguation were not annotated manually but were calculated from other variables. See section 3.6.3 below.
- The role of affectedness was discovered when it was already too late to include it in the annotation. Since it is a minor variable by any standards, this should not have done great harm to the evaluation.

3.6.3 Calculation of secondary variables

From the primary variables discussed in the last section, several secondary variables were calculated. All are described in the list below. Note that although there are several variables which approximate topicality, I refrained from calculating a single topicality value by feeding all variables into a heuristic formula. I did so because I wanted to keep the effects of the individual contributors visible.

Some extralinguistic variables (genre, sex and age of speaker) were not included in the qualitative discussion in section 3.5. These variables were taken in here because they could be extracted from the corpus texts without much additional effort and because it is an interesting question in general how linguistic and extralinguistic factors work together in shaping grammar. However, the discussion of these variables will not be as deep as that of the linguistic ones, and the results should be considered exploratory rather than conclusive.

- **givenness** – referents whose ID was used in a text for the first time got the value `new` while referents which had been mentioned before got `given`. Given referents are more likely to be topical.
- **ranked frequency** – the frequency of a referent divided by the highest referent frequency. This value can be calculated based on the number of referents up to the point where the referent in question is mentioned ("ranked frequency so far") or on the number of all referents in the text ("ranked frequency total"). Absolute frequencies were ignored because they are not comparable across texts. Conventional relative frequency (the frequency of a referent divided by the number of all referent presentations, overt or zero) is very similar to ranked frequency, but ranked frequency has the additional advantage that it has a fixed range (1 as the maximum in every file and lower values approximating 0). This makes ranked frequency a more robust predictor for the probability of NOM/DAT. The closer the value gets to 1, the more topical a referent is on average.
- **distance to last mention** – for given referents, the number of words between the present and the last mention. Zero arguments do not increase the distance to the last mention because their position is not defined. For instance, if object referent X was mentioned twice in two adjacent clauses and A was zero in both, including the zeros in the count could yield distances between the two X ranging from 0 to 2, depending on where the zeros were placed in the annotation (`X-0-0-X`, `0-X-X-0` etc.). For new referents the distance to their last mention is `NA` (non-applicable). The greater the distance to the last mention, the less likely a referent is to be topical. Since this variable can take on very large values in exceptional cases, it was logged to the base of 10.
- **competitors** – the number of other potentially topical referents in the neighbourhood of a referent. Referents were taken as competitors of an object referent O when their absolute frequency was higher than that of O at the time O was mentioned and when their last mention was fewer than 50 words away. These values are of course arbitrary but did yield interesting results (see section 3.6.4 below). Further research would be needed to determine whether other definitions of competition would fare even better. In particular, it would be interesting

to see whether the topicality of competitors is best determined relatively to O (as in the present definition) or relative to all referents (e.g. making use of ranked frequency). The more competitors there are, the lower topicality becomes on average.

- **relative position** – this variable reflects one aspect of disambiguation. It takes on the following values:
  - **AO** – A precedes O
  - **OA** – O precedes A
  - **zero A/O** – one or both of the two arguments are zero so that their relative positions cannot be determined
- **distance from predicate** – another variable related to disambiguation. It contains the number of words between an object referent and its predicate. The distances for referents preceding their predicate are negative, those for referents following it positive.
- **co-argument case** – the case of the G accompanying a T. The values are:
  - **T with G-DAT** – T with dative G
  - **T with other G** – T with G marked by any case other than DAT
  - **T with zero G** – T with zero G (thus indeterminate with respect to case marking)
  - **P/G** – objects other than T
- **genre** – the genre of the text. For most subgenres there wouldn't have been enough files, so only two supergenres were recognised:
  - **spoken** – transcriptions of spoken language
  - **written** – all other texts
- **identity**, **sex** and **age** of speaker. Interesting as these extralinguistic variables are, they are of limited use for two reasons. One is that there are many gaps in the metadata provided for the spoken part of the NNC, the other that there are no metadata at all for the written part. Sex and age of the speaker are therefore not known for about 53% of all O, and taking them together with the other variables would make it necessary to ignore a great many O for which information is otherwise complete. What's more, identity is not a useful predictor variable because there are too many different speakers in the world. For these reasons, identity, sex and age are only discussed in isolation below.

A special case for topicality-related variables occurs when a single NP represents several referents that are kept separate in other sentences. In that case the frequency of the composite NP is defined as the highest of all contained frequencies, and all further measures (givenness, distance to last mention etc.) are based on that frequency.

3.6.4 Impact of individual variables

This section summarises the frequency distributions of the variables introduced above and discusses their impact on DOM based on various types of evidence. For the categorial variables the following was done:

- Contingency tables were created. For variables with more than two values, a χ^2 test was done to test the significance of their interaction with DOM. For variables with only two values, the contingency table had 2×2 cells, so significance could be tested using Fisher's exact test. Note that both tests make the assumption that the sample they are based on is a random sample and that all members of the underlying population have equal probabilities for getting into the sample. Since a text, where every word interacts with its surroundings, is obviously not a sample of this kind, the p values produced by these tests can only be viewed as approximations. They are given as p_{χ^2} and p_{FT} below each table.

- In addition, Cramer's V was calculated as a measure for the strength of the association between each predictor variable and DOM. Compared to Pearson's contingency coefficient C, Cramer's V has the advantage that its values do not vary depending on the number of rows and columns in the contingency tables, so it is easier to compare across the different variables influencing DOM. It is given as CV below each table.

- In a second step, the significance of the interaction of single values with DOM was tested by collapsing all other values into a single one and doing a Fisher's exact test on the resulting 2×2 table. The results of this test row are shown in an additional row in the contingency tables.

- Finally, each variable was tested as the single predictor in a logistic regression model. Since there is no single variable that can predict a satisfying amount of object cases – let alone all –, a realistic model has to include several predictors. Such a model will be built in section 3.6.5. Since the impact of a variable may change considerably when it is looked at in connection with other variables, the values from this test should also be viewed as an approximation. The statistic that is given here is R^2. R^2 is defined in the `rms` package on the base of Nagelkerke's (1991) improvement of the ideas formulated in Cox and Snell (1968). It is therefore a measure of how much better the fitted model is in comparison to the null model (i.e. a model that does not use any predictors and always predicts the default outcome (NOM) instead). R^2 ranges from 0 (no improvement) to 1 (all errors implied by the null model are successfully removed). The p value for a variable being a significant predictor is given as p_{LR} in the rare case that it expresses a level of significance that is different from p_{χ^2}/p_{FT}.

- In order to visualise the proportions of NOM and DAT in every value, a mosaic plot was drawn for every variable. The mosaic plot does not only reflect the frequencies of NOM and DAT but also those of the predictor values, so the size of each block is proportional to N.

Continuous variables require a different treatment:

- As above, each variables was tested as a predictor in a logistic regression model. The R^2 value is again given as the most important indicator of the goodness of the predictor. The significance value for the interaction with DOM is also based on logistic regression and is thus given as p_{LR}.

- An appropriate measure of the strength of the association between a random variable which is at least interval-scaled and a dichotomous dependent variable (NOM/DAT) is the point-biserial correlation coefficient r_{pbi}. Differently from Cramer's V, r_{bpi} ranges from 1 (perfect positive association) to -1 (perfect negative association). Thus, Cramer's V can be compared to the absolute value of r_{bpi}.

- For visualisation, plots were drawn that show the values of the random variable on the x axis and their frequencies in general (black line), with NOM (green line) and with DAT (red line) on the y axis. In order to level outliers (e.g. relative frequency values recurring more often than expected due to the dominance of one longer corpus text), the relation between values and frequencies was linearly interpolated using the R function `approx()`.

3.6.4.1 Role

The interaction of role and DOM is highly significant, that is, DAT is more or less likely depending on the role O is mapped to. The numbers for this are shown in Table 3.5, and the corresponding proportions are visualised by Figure 3.3. Note that there are too few attestations of G-DAT (expected value: 3.48), so Yate's correction was applied.

Most notably, DAT is much less frequent than usual with T, where it is found only in about 8% of all instances (vs 13% on average). This effect is, however, an artifact – the factor that is really relevant here is co-argument case (see section 3.6.4.15 below). With accompanying G-DAT, T-DAT

	NOM	DAT	significance
P	2292	353	yes ($p_{FT} < 0.01$)
T	482	40	yes ($p_{FT} < 0.01$)
G	14	2	no ($p_{FT} = 1$)

Table 3.5: Role and DOM ($p_{\chi^2} < 0.01$, CV = 0.06, $R^2 = 0.01$)

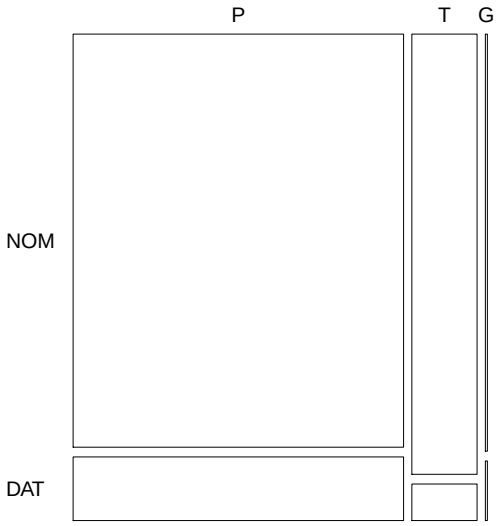

Figure 3.3: Proportions of role and DOM

is only attested a single time, in accordance with what was said in section 3.4.4. When T is not accompanied by an overt G-DAT, the proportion of DAT equals the average proportion.

The proportion of DAT in P is only slightly higher than the average (13.35% vs. 12.58%), but the difference turns out to be highly significant. By contrast, the proportion of DAT in G is almost exactly equivalent to the average proportion, and there is no significant difference. Note that the data for G are problematic because G which are eligible for DOM are extremely rare – most G have fixed DAT or participate in other alternations. The only G that can participate in DOM at all are G of the instrumental ditransitive class. This class is very rare in the annotated part of the corpus, and nothing can be said for sure about its behaviour from the few attestations.

3.6.4.2 Animacy

Some values of animacy turned out to be very rare in the annotated part of the NNC and were therefore fused with other values. `human.fam` (13 instances) was fused with `human`, and `high.anim` (10 instances) and `mid.anim` (2 instances) were fused with `low.anim` to a common category `anim`. Further, cross-tabulation of animacy and DOM showed that `human.group` does not behave too differently from `human`: `human` got DAT in 67% of all instances and `human.group` in 78%. These two were therefore fused, too. Finally, `human.prop` did behave differently from the other human categories in that it had DAT in 100% of all cases, as described in section 3.5.9. Nevertheless, it had few instances overall (27) and was therefore also fused with `human`.

Table 3.6 shows the contingency table of the frequencies of the resulting values with NOM/DAT. P_{χ^2} is highly significant, and animacy has the highest Cramer's V of all variables. All values are significant in isolation, too, although `state` and `anim` reach notably lower p values than the other three values. The proportions of NOM and DAT in each value are shown in the mosaic plot in Figure 3.4.

The mosaic plot reveals some interesting facts. First of all, the proportions do not point to a

	NOM	DAT	significance
process	1011	18	yes ($p_{FT} < 0.01$)
state	898	148	yes ($p_{FT} = 0.04$)
thing	714	35	yes ($p_{FT} < 0.01$)
anim	77	4	yes ($p_{FT} = 0.04$)
human	88	190	yes ($p_{FT} < 0.01$)

Table 3.6: Animacy and DOM ($p_{\chi^2} < 0.01$, CV = 0.55, $R^2 = 0.48$)

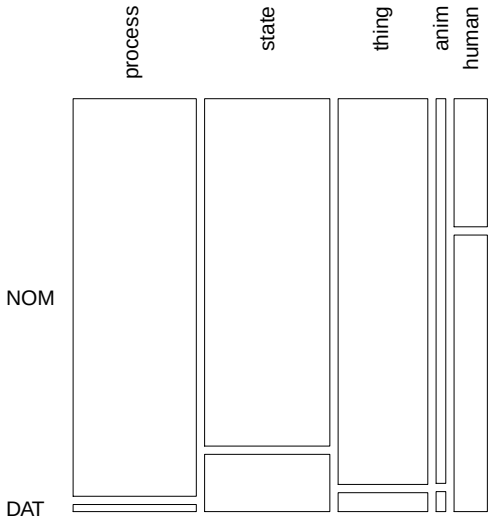

Figure 3.4: Proportions of animacy and DOM

hierarchy as the one proposed in section 3.5.3. Things and non-human animates get approximately equal proportions of DAT (5%), whereas states, which should rank lower than both of these, get 14%. The only clear candidates are at the two ends of the spectrum: processes barely ever get DAT (2%), and human O get more DAT than any other referent class (68%).

A possible explanation for the presence of the hierarchy in elicitation vs its absence from corpus data is that a hierarchy is only construed when the speaker takes a relatively conscious approach to his language (similar to that of linguists, who also tend to see hierarchies everywhere). However, this is an explanatory construction rather than an extraction of patterns that are actually to be found in language data.

A second point to note is that `human`, the value that most frequently yields DAT, is at the same time very rare: only 9% of all O have it. Only `anim` is rarer (3%), whereas `thing` (24%), `state` (33%), and `process` (32%) are all much more frequent. This confirms the hypothesis put forward in section 3.5.15 that *-lai* is (among other functions) a marker of properties that are unusual for O.

3.6.4.3 Quantifiability

Quantifiability only has two values, so no values had to be (or could have been) fused. Table 3.7 shows the contingency table and Figure 3.5 summarises the proportions. As with animacy, the interaction between quantifiability and DOM is highly significant, though the degree of association as measured by CV is much smaller. As expected, DAT is more frequent with quantifiable O (23%) than with non-quantifiable O (3%). However, this time the value associated with DAT is overall more frequent (60% of all O are quantifiable).

	NOM	DAT	significance
qnt	1559	356	yes (= p_{TOTAL})
nonq	1229	39	yes (= p_{TOTAL})

Table 3.7: Quantifiability and DOM ($p_{FT} < 0.01$, CV = 0.23, $R^2 = 0.12$)

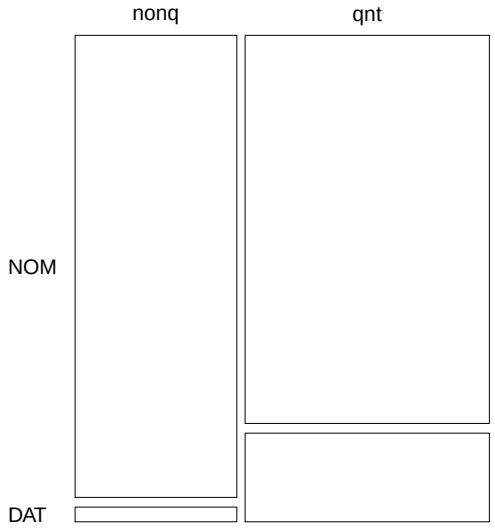

Figure 3.5: Proportions of quantifiability and DOM

3.6.4.4 Situation

Of the four levels of this variable, two had low attestation numbers and were therefore abandoned: `abstract` (18 instances) and `exemplary` (16 instances) were fused with `general` to a composite category `non-concrete`. The resulting 2×2 contingency table is shown in Table 3.8. The corresponding proportions are illustrated by Figure 3.6.

	NOM	DAT	significance
non-concrete	1329	152	yes (= p_{TOTAL})
concrete	1459	243	yes (= p_{TOTAL})

Table 3.8: Situation and DOM ($p_{FT} < 0.01$, CV = 0.06, $R^2 < 0.01$)

As expected, the proportion of DAT is slightly higher in concrete situations. That this is significant is confirmed by p_{FT}. However, the low values for p, CV, and R^2, which trail behind the values for the variables we have seen so far, make situation a less interesting predictor.

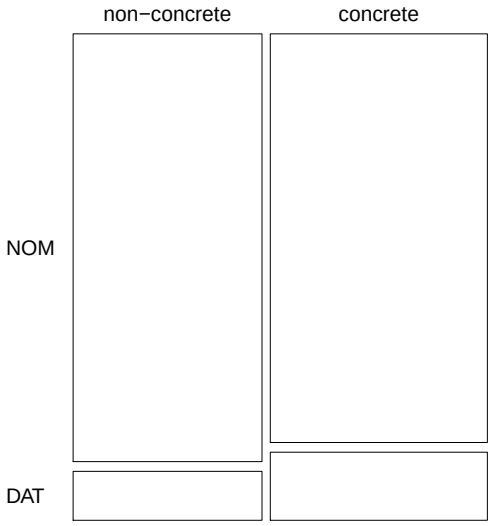

Figure 3.6: Proportions of situation and DOM

3.6.4.5 Ctag (part of speech)

Table 3.9 shows the contingency table for DOM and part of speech. This is the second-most important variable after animacy based on its high values for Cramer's V and R^2. Nouns, demonstratives, and pronouns have a significant effect in isolation (less DAT with nouns, more with the other two). By contrast, whereas it seems immediately clear that adjectives were less frequently marked by DAT than other parts of speech and this would intuitively make sense (adjectives representing potentially less individuated referents), this difference failed significance. The same is true but expected for `other`, which is a heterogeneous dustbin category and therefore shouldn't be significantly associated with DOM.

	NOM	DAT	significance
adj	103	8	no (p_{FT} = 0.11)
n	2327	246	yes (p_{FT} < 0.01)
dem	139	73	yes (p_{FT} < 0.01)
pro	0	47	yes (p_{FT} < 0.01)
other	219	21	no (p_{FT} = 0.08)

Table 3.9: Part of speech and DOM (p_{χ^2} < 0.01, CV = 0.38, R^2 = 0.17)

Looking at the proportions in Figure 3.7, it becomes clear that parts of speech can be arranged hierarchically: just as there is a clear step between nouns and demonstratives, there is another step between demonstratives and pronouns. Pronouns are marked by the dative in 100% of all cases, which is not beaten by any other variable value considered here.

The plot also illustrates nicely another correlation between DAT and value frequency. Most objects (81%) are nouns, wheras only 7% are demonstratives and 2% are pronouns. Thus, part of speech is another variable where DAT becomes the more likely the less frequent a value is for an object.

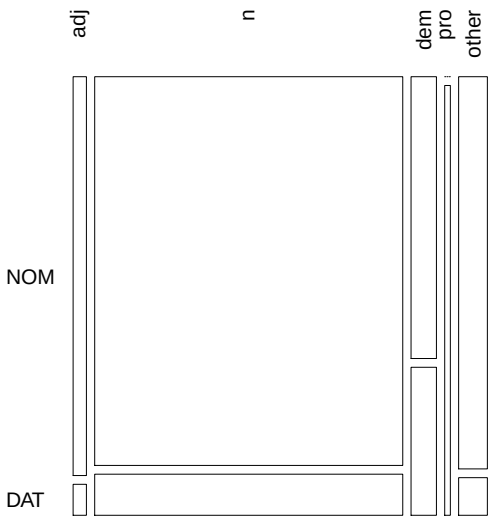

Figure 3.7: Proportions of part of speech and DOM

3.6.4.6 Modification

Modification also had a few values with low attestation that were fused with others. `latposs` (16 instances) was fused with `humposs`. `sortal` (13), `sortal.q` (3) and `interrog` (26) were all fused with `adj` because of their similarity to adjectives in their syntactic behaviour. This still leaves modification as the variable with most values, as can be seen in Table 3.10 and Figure 3.8.

	NOM	DAT	significance
none	1612	214	no (p_{FT} = 0.17)
num	109	7	yes (p_{FT} = 0.03)
relclause	178	14	yes (p_{FT} = 0.02)
poss	188	23	no (p_{FT} = 0.59)
adj	341	45	no (p_{FT} = 0.68)
humposs	93	27	yes (p_{FT} < 0.01)
dem	108	32	yes (p_{FT} < 0.01)
several	95	27	yes (p_{FT} < 0.01)
other	64	6	no (p_{FT} = 0.46)

Table 3.10: Modification and DOM (p_{χ^2} < 0.01, CV = 0.12, R^2 = 0.03)

Although the overall interaction is significant, this table clearly suggests that there are still too many distinctions within modification. Barely half of the values reach significance in isolation: `num` and `relclause` have an inhibitory effect (= fewer DAT than normal), and `humposs`, `dem`, and `several` have a positive effect.

The effect of `num` and `relclause` is unexpected. Numbered referents are necessarily quantifiable and relative clauses tend to contain much more specific information than a bare noun, so one would expect more DAT than NOM. There also seem to be no inhibitory values of other variables that could correlate with these two and thus explain their negative effect.

On the other hand, `dem` and `humposs` do behave as described in section 3.5.10. `dem` can only be used on referents which are at least specific and is an indicator of high topicality. The effect of `humposs` may be due to high possessors "rubbing off" on their possessums. The high proportion of DAT in `several` may be due to the fact that most modifier chains contain a demonstrative or a possessive pronoun.

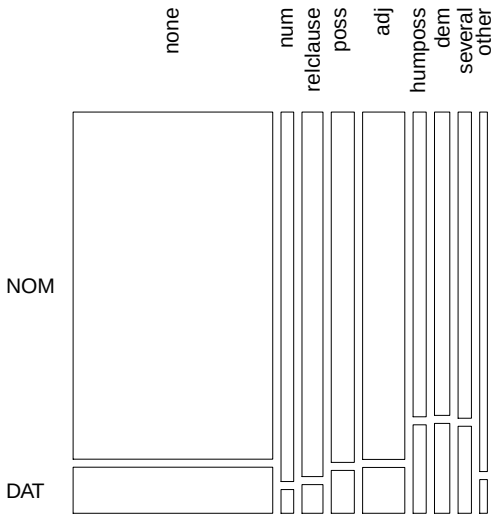

Figure 3.8: Proportions of modification and DOM

3.6.4.7 Focus

Focus as the most problematic variable barely stands the test of significance – see Table 3.11. The irrelevance of contrastive focus has already been discussed (section 3.5.11), so the failure of `contrast` vs other values is expected. The remaining two values are not highly significant but below the conventional level of 0.05. Note that these data are somewhat problematic because there are too few attestations for O-DAT with fragile focus (expected value: 1.73). Yate's correction was applied here.

	NOM	DAT	significance
nofoc	2736	380	yes (p_{FT} = 0.02)
contrast	43	10	no (p_{FT} = 0.14)
fragile	9	5	yes (p_{FT} = 0.02)

Table 3.11: Focus and DOM (p_{χ^2} = 0.01, p_{LR} = 0.03, CV = 0.05, R^2 < 0.01)

If one fuses `nofoc` and `contrast`, a problem arises: `fragile` is extremely sparsely attested. Fisher's exact test may not be as dependent on high observation numbers in all cells as the χ^2 test, but there is the deeper question of whether a value that is as rare as `fragile` can be assumed to be part of a speaker's knowledge at all. Sure enough `fragile` is not rare in absolute terms – any speaker's mental corpus is a thousand times bigger than the corpus annotated for this research. However, it is rare in relative terms – only 0.4% of all objects have this kind of focus. The proportion of DAT in a fused category `non-fragile` (12%) approximates the overall average. Thus, it would be highly unefficient for a speaker to test focus every time he produces a cased object because it's completely irrelevant in 99.6% of all cases.

As mentioned above in section 3.6.2, the role of unexpectedness only became clear after the annotation of focus had been in effect for some time. `fragile` can only be taken as an imperfect approximation of unexpectedness. A modified annotation system might yield more unexpected objects and more significant results.

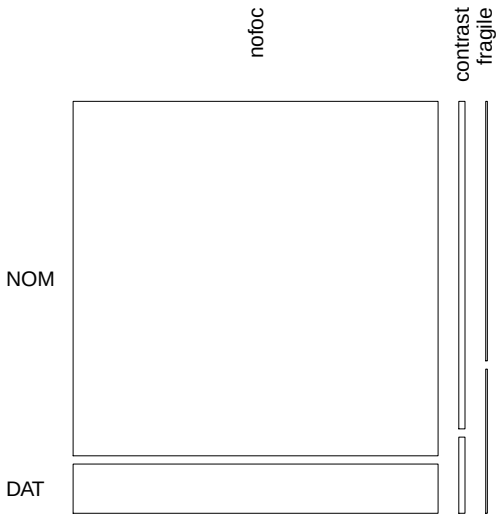

Figure 3.9: Proportions of focus and DOM

3.6.4.8 Diathesis

Diathesis is one of the few variables which do not significantly interact with DOM, as shown in Table 3.12.

	NOM	DAT	significance
active	93	11	no (= p_{TOTAL})
passive	2695	384	no (= p_{TOTAL})

Table 3.12: Diathesis and DOM (p_{FT} = 0.65, p_{LR} = 0.56, CV < 0.01, R^2 < 0.01)

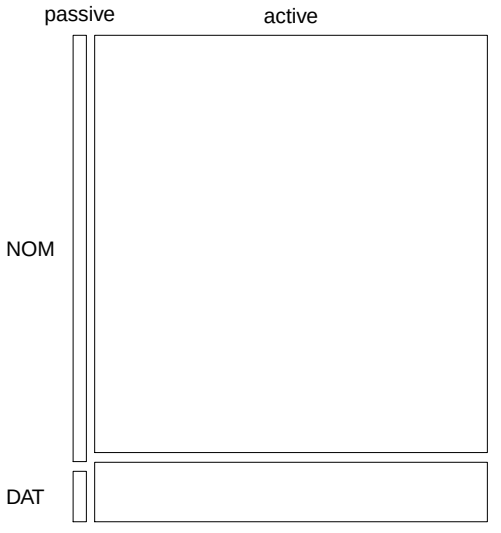

Figure 3.10: Proportions of diathesis and DOM

This could be taken as a blow to the probabilistic modelling of DOM because qualitative research clearly shows that there is interaction (cf. section 3.4.6). In particular, the passive is the only context

where pronouns and proper names may be marked by NOM. Upon second thought, however, this very property explains why the passive does not significantly interact with DOM. The passive itself does not have a strong preference for either O-NOM or O-DAT – its main characteristic is its interaction with other variables favouring DAT. It is therefore no wonder that it is insignificant when looked at in isolation. It will be shown in section 3.6.5 below that there is an effect when the passive is considered together with other variables.

3.6.4.9 Givenness

After a series of flawed or marginally relevant variables, givenness is another excellent candidate for predicting DOM. Its interaction with DOM is highly significant (given referents get more datives than expected by chance) and its CV and R^2 are comparable to those of quantifiability – see Table 3.13. This proves an important difference between Chintang and Nepali: in Nepali it is not only important whether it is *possible* to track a referent but also whether it *does* get tracked.

	NOM	DAT	significance
new	2191	181	yes (= p_{TOTAL})
given	597	214	yes (= p_{TOTAL})

Table 3.13: Givenness and DOM ($p_{\text{FT}} < 0.01$, CV = 0.25, $R^2 = 0.10$)

Figure 3.11 shows the corresponding proportions. Note that the majority of all object referents is new. Since all new referents become given as soon as they are mentioned a second time, this means that only a minority of object referents gets tracked. Although this is somewhat surprising, it fits well with the idea of the role of DAT as a marker of the exceptional, which was already brought up several times above.

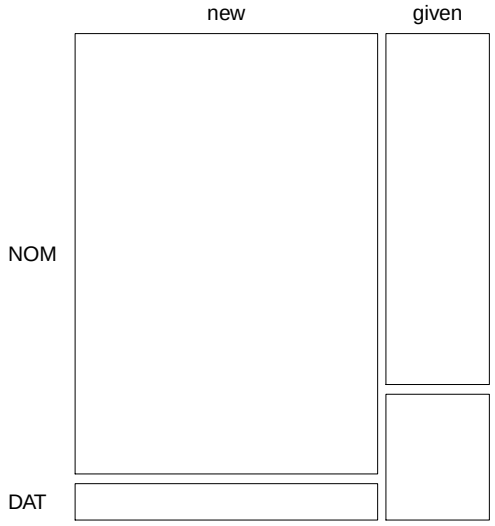

Figure 3.11: Proportions of givenness and DOM

3.6.4.10 Ranked frequency

Ranked frequency is the most important continuous variable. The r_{pbi} is 0.28 for ranked frequency up to the point where an object referent is mentioned ("ranked frequency so far") and 0.38 for ranked frequency summed over a whole text ("ranked frequency total"). Thus, the higher the frequency of a referent, the higher the probability for DAT. This fits with our assumption that ranked

frequency is a proxy for topicality. Looking at R^2, ranked frequency (total) fares again better than ranked frequency (so far) with R^2 = 0.17 vs 0.10. Both varieties are highly significant with p_{lr} < 0.01.

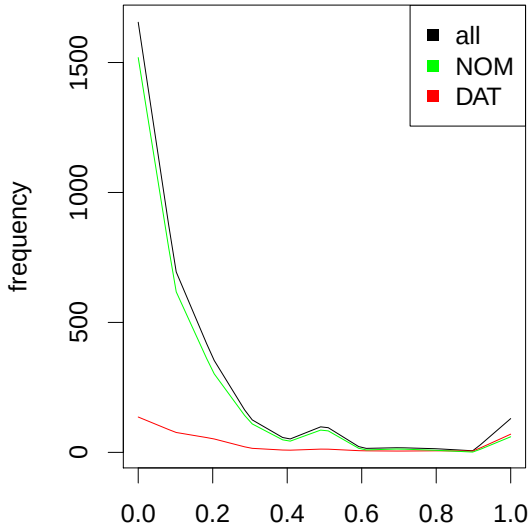

Figure 3.12: Ranked frequency (so far) and DOM

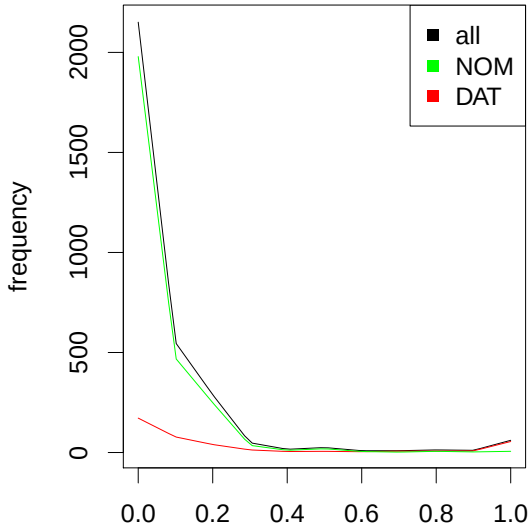

Figure 3.13: Ranked frequency (total) and DOM

The plots in Figure 3.12 and Figure 3.13 also show a small but important difference between ranked frequency (total) and ranked frequency (so far). Note that all values were rounded to one post-comma digit for these plots in order to smooth out single values with high frequency that are due to the dominance of some large corpus texts.

With ranked frequency (total), the frequency of DAT goes up in the high topicality region at the right end of the graph while the frequency of NOM stays on approximately the same level as before, thereby separating from the black average line. By contrast, the frequencies of both NOM and DAT rise together with the average in the same region for ranked frequency (so far), so the two do not behave notably different in this crucial stretch. Henceforth, only ranked frequency (total) will be considered and ranked frequency (so far) will be ignored.

3.6.4.11 Distance to last mention

The logged distance to the last mention of a referent fails significance with $p_{lr} = 0.20$. The r_{pbi} close to 0 (-0.04) and the low R^2 (< 0.01) confirm this picture. Also consider Figure 3.14, which does not reveal any interesting differences between NOM and DAT conditioned by this factor.

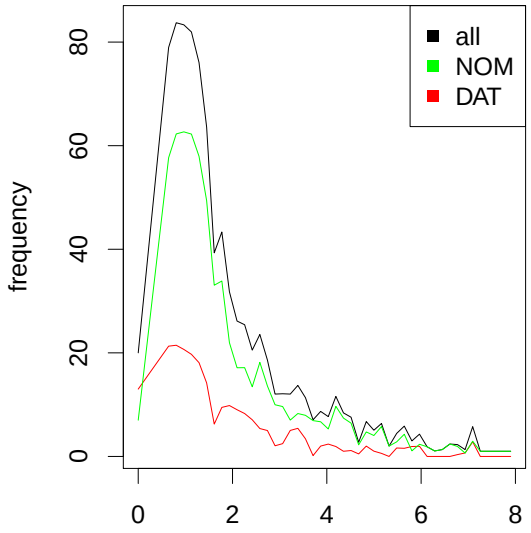

Figure 3.14: Distance to last mention and DOM

This result is a little surprising at first. Distance to last mention was, like ranked frequency, intended as a proxy for topicality: the closer the last mention of a referent, the more mentally present and thus topical should it be. However, any unimportant referent that gets mentioned twice within a sentence and then never again gets a higher value for this variable than a referent that is mentioned again and again but with several sentences in between. The distance to the last mention thus does not reflect the importance of a referent for a text as a whole, even though this is intuitively a more faithful indicator of discourse topicality than local phenomena. This assumption also fits with the better performance of ranked frequency (total) as compared to ranked frequency (so far).

3.6.4.12 Competitors

The number of competitors is a surprisingly good proxy for topicality. Its interaction with DOM is highly significant ($p_{lr} < 0.01$), its r_{pbi} is -0.23, and its R^2 is 0.11. The negative r_{pbi} indicates that, as expected, DAT is the less frequent the more competitors there are. It is maybe no coincidence that these values are comparable to that of ranked frequency (so far): part of the definition of competitors is that other referents are only viewed as competitors when their absolute frequency (so far) is higher than that of the referent in question. This entails a higher ranked frequency. Put differently, the more competitors a referent has, the lower its own ranked frequency (so far) has to be.

Figure 3.15 shows the connection between the number of competitors and DOM. In contrast to ranked frequency, `competitors` can only take on integer values with a limited spectrum (the

attested maximum being 14 competitors). This makes `competitors` much less prone to outliers so that the connection becomes visible much more clearly: whereas the line for NOM stays more or less parallel to the average line, the line for DAT steadily moves down as the number of competitors increases.

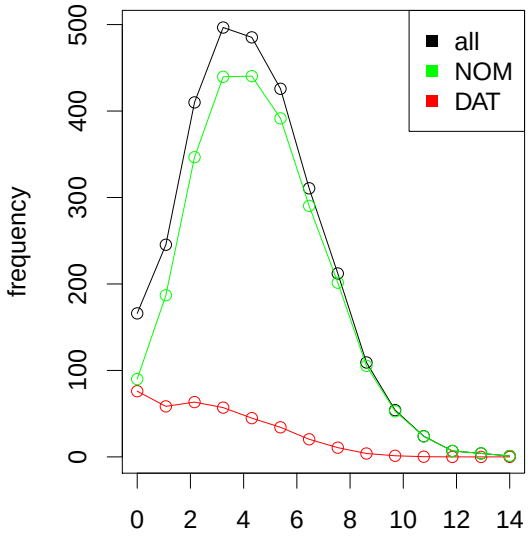

Figure 3.15: Competitors and DOM

3.6.4.13 Relative position

Relative position is a variable where one value is relatively informative whereas the others are neutral. The concerned value is `OA` (O before A), which triggers significantly more datives than usual in isolation. `AO` and `zero A/O` both fail this test. Although the variable as a whole is signficant by p_{χ^2}, the other evaluation criteria are on a low level – see Table 3.14.

	NOM	DAT	significance
AO	1016	134	no (p_{FT} = 0.34)
zero A/O	1534	202	no (p_{FT} = 0.16)
OA	238	59	yes (p_{FT} < 0.01)

Table 3.14: Relative position and DOM (p_{χ^2} < 0.01, CV = 0.07, R^2 = 0.01)

Figure 3.16 shows the higher proportion of DAT in `OA` and the similar behaviour of `AO` and `zero A/O`. Also note that relative position is another variable where DAT goes with the rare values – only 9% of all O precede A.

3.6.4.14 Distance from predicate

The distance of an object from the predicate is of medium relevance. Its interaction with DOM is highly significant (p_{lr} < 0.01) and reaches an r_{pbi} moderately far away from 0 (-0.15). Note that because distances in OV word order are expressed as negative numbers and positive numbers are only found in VO clauses, the r_{pbi} is negative even though a *greater* distance to the left correlates with more DAT. A greater distance to the right correlates with fewer DAT. In spite of its significance, distance from the predicate is a weak predictor with R^2 = 0.04.

It is interesting to note that distance from the predicate is more strongly associated with DAT than relative position, which one might rather have suspected to be the flagship of the relevance of

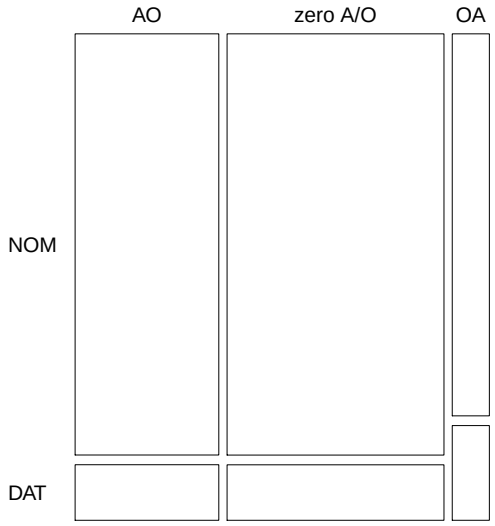

Figure 3.16: Proportions of relative position and DOM

disambiguation to DOM. A difference between the two that might be relevant is that an ambiguity created by A/O inversion does not necessarily last for a long time. By contrast, the time lag until the resolution of an ambiguity tends to grow with the distance to the predicate. Thus, ambiguity might be the more relevant to case marking the longer it holds. The fact that A case marking does not cancel the effect of inversion (cf. section 3.5.12) is another hint into the same direction.

The connection between distance and DAT becomes nicely visible in Figure 3.17.

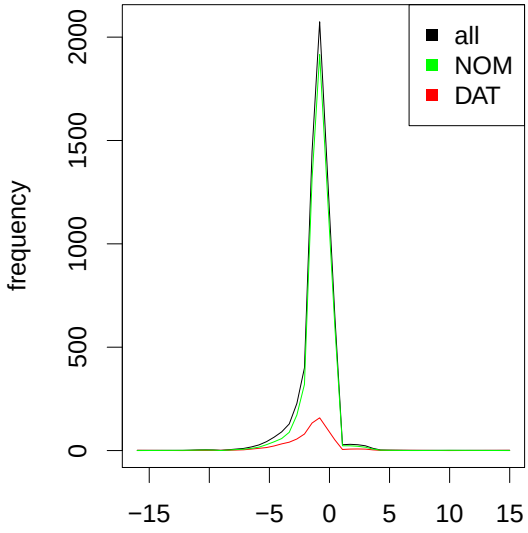

Figure 3.17: Distance from predicate and DOM

3.6.4.15 Co-argument case

As expected after what was said about double datives in section 3.4.4, co-argument case with ditransitives is highly significant with $p_{\chi^2} < 0.01$ and Cramer's V = 0.17. Nevertheless, it explains only a small proportion of the variation in object case marking ($R^2 = 0.06$) – although this is also expected if one considers that the role of T covers only a small portion of all O (about 16%). All

numbers are summarised in Table 3.15.

	NOM	DAT	significance
T with G-DAT	173	1	yes (p_{ft} < 0.01)
T with other G	71	34	yes (p_{ft} < 0.01)
T with zero G	238	5	yes (p_{ft} < 0.01)
P/G	2306	355	yes (p_{ft} < 0.01)

Table 3.15: Co-argument case and DOM (p_{χ^2} < 0.01, CV = 0.17, R^2 = 0.06)

T with G-DAT has a neglectable proportion of DAT – the only attestation in the annotated part of the NNC has already been cited in section 3.6.4.1, and elicited forms were given in sections 3.4.4 and 3.4.2. Interestinlgy, the same is true for T with zero G. A possible explanation for this is that most dropped G belong to the transfer ditransitive class Ia, where G would have been marked by DAT if it had been overt. This is plausible because class Ia contains two highly frequent items (*di-* 'give', *bhʌn-* 'say') that are used with covert G all the time (*di-* especially in metaphoric use, *bhʌn-* when an utterance is not directed to anybody in particular).

The most unexpected point is the extremely high proportion of DAT in T with other G (33%). Presently I do not have an explanation for this. The proportion of DAT in P/G approximates the average. That this value is nevertheless significant in isolation is due to the fact that the collapsed complementary category is still meaningful – it simply covers all T, so the test is equivalent to comparing T to the other object roles.

The proportions of NOM/DAT in the four values are visualised in Figure 3.18.

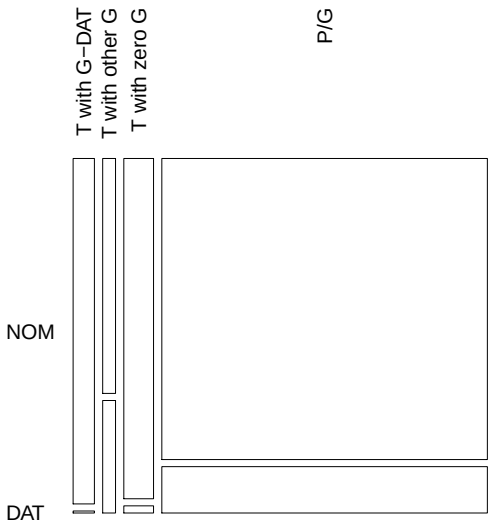

Figure 3.18: Proportions of co-argument case and DOM

3.6.4.16 Genre

Genre is another significant but weak predictor of DAT. DAT is more frequent in the written language, as shown in Table 3.16 and Figure 3.19. 9% of all O have DAT in the spoken language, vs 17% in the written language.

	NOM	DAT	significance
spoken	1531	142	yes ($p_{ft} < 0.01$)
written	1257	253	yes ($p_{ft} < 0.01$)

Table 3.16: Genre and DOM ($p_{FT} < 0.01$, CV = 0.12, $R^2 = 0.03$)

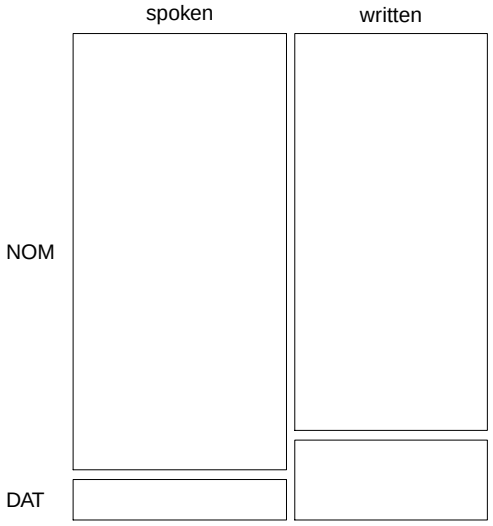

Figure 3.19: Proportions of genre and DOM

There are several functional variables that may contribute to the prevalence of DAT in the written language. First, written language makes it easier to construct long sentences. The fact that Nepali tends towards complex embedding structures where overt elements of a matrix clause frame subclauses on both sides makes such sentences often hard to parse, so DAT becomes more important as a role disambiguator. Second, written language may deal with more topics at the same time, especially including inanimate referents, so there will be more highly frequent referents in general and more highly frequent inanimate referents in particular. Third, topic shifts and the desire to point out unusual objects especially in journalistic and academic writing may lead to more instances of unexpectedness.

3.6.4.17 Speaker variables

This paragraph summarises some information on the role of identity, sex and age of speakers. As mentioned above, these variables could not be assessed in many cases because of gaps in the NNC metadata. They are therefore only discussed in this section and not in connection with the other variables.

Only the identity of speakers could be determined for all O because even where the name of a speaker was not known, speaker codes had been used in the spoken part of the NNC. For the written part, identity was equalled with the name of the file, making the idealised assumption that each file had exactly one author and that each author featured only once in the annotated corpus. Newspaper files were split into the articles contained in them. Speaker identity is not a

useful predictor *per se* because there are too many speakers of Nepali and for most applications the identity of the speaker is not given, anyway. It is, however, interesting to see how much variation in the use of DOM exists across of speakers.

The speaker with the highest proportion of datives used DAT on 43% of all O, whereas several speakers on the lower end did not use it at all. The mean was 11%, so there were slightly more speakers with unusually low proportions of DAT – otherwise the mean should have been 12%, the average proportion of DAT in all O. The standard deviation was also 11%, which is fairly high given the overall mean: it means that using no DAT at all or almost twice as many as normal are digressions that are still within average. The variation across speakers is summarised in Figure 3.20.

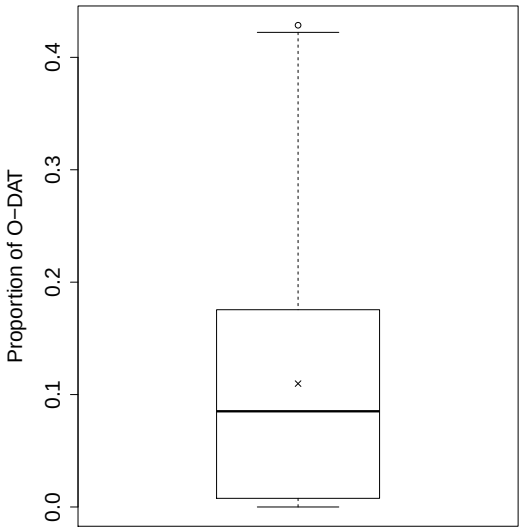

Figure 3.20: Variation in the frequency of O-DAT across speakers

The sex of speakers could be determined for 53% of all annotated O. Men and women produce about equally many DAT, so this variable is highly insignificant. The very low Cramer's V (0.005) and R^2 (0.0) are in accordance with this. Table 3.17 shows the numbers, and Figure 3.21 visualises them.

	NOM	**DAT**	**significance**
male	1005	95	no (p_{ft} = 0.78)
female	526	47	no (p_{ft} = 0.78)

Table 3.17: Speaker sex and DOM (p_{FT} = 0.78, CV = 0.12, R^2 = 0.03)

The third speaker variable that was investigated was age. This was known for 48% of all O. Ages ranged from 20 to 70 years with a mean of 38 years and a standard deviation of 17 years. Age got a moderately high r_{pbi} (0.10), indicating a positive relationship between age and DAT (more DAT in old age). However, age was not a significant predictor in logistic regression with p_{lr} = 0.38. Figure 3.22 also shows a rather chaotic picture, where the proportions of O-DAT within each age do not result in a smooth curve.

In addition, there are various problems with age: there are too many sentences by very young or very old speakers and no attestations at all for many ages, and some ages have enough attestations but all come from a single speaker. For these reasons, the results for age do not seem reliable until more data are available.

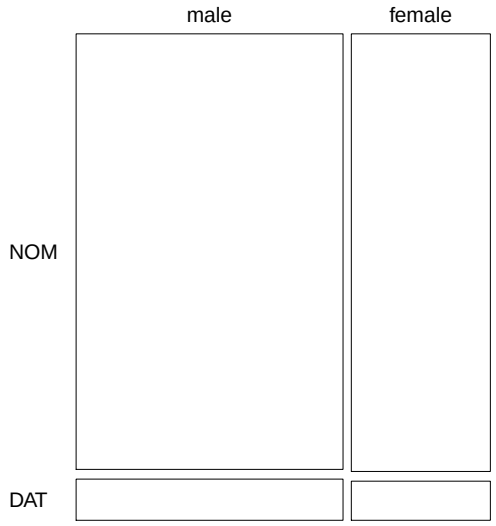

Figure 3.21: Proportions of speaker sex and DOM

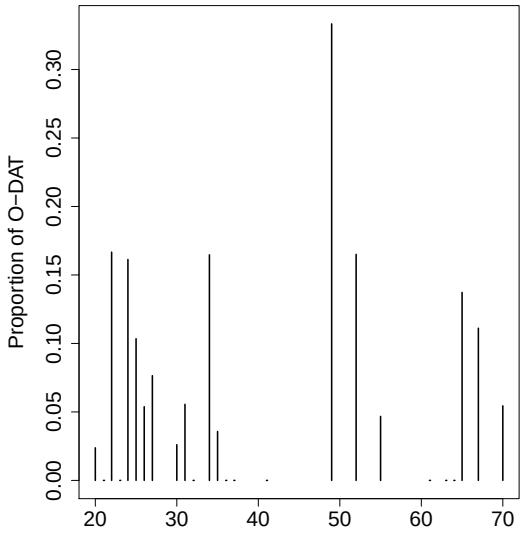

Figure 3.22: Age and DOM

3.6.4.18 Summary

The discussion above has shown that most variables that have been annotated in the NNC interact significantly with DOM. The only two variables that failed significance were distance to the last mention and diathesis. Focus got a markedly lower value than the other variables. Table 3.18 shows a summary of the statistics for all variables that is sorted by Cramer's V/r_{pbi}.

In general, the values for the measures of association CV/r_{pbi} and for the measure of predictive power R^2 point into the same direction, with a few minor exceptions that do not change much in the general picture. Animacy is clearly at the top of the list and does not have any serious competitors. It is followed by ranked frequency and part of speech, both proxies for topicality (with `pro` and `dem` being the most relevant parts of speech). The next few ranks are occupied by other proxies, by quantifiability, and by distance from the predicate as one indicator of ambiguity. Among the significant variables, role, co-argument case, modification, relative position, situation, and focus are the weakest.

	significance	CV/r_{pbi}	R^2
animacy	< 0.01	0.55	0.37
ranked frequency total	< 0.01	0.38	0.17
part of speech	< 0.001	0.38	0.17
ranked frequency so far	< 0.01	0.28	0.10
givenness	< 0.01	0.25	0.10
competitors	< 0.01	-0.23	0.11
quantifiability	< 0.01	0.23	0.12
co-argument case	< 0.01	0.17	0.06
distance from V	< 0.01	-0.15	0.04
modification	< 0.01	0.12	0.03
relative position	< 0.01	0.07	0.01
role	< 0.01	0.06	0.01
situation	< 0.01	0.06	0.01
focus	= 0.01	0.05	< 0.01
distance to last	= 0.20	-0.04	< 0.01
diathesis	= 0.65	0.01	< 0.01

Table 3.18: Summary of DOM variables in isolation

The preliminary picture that emerges from this is one of DOM as a strongly grammaticalised phenomenon. Animacy alone, which is a quasi-lexical factor, explains more than half of all the variation in case marking. The high rank of part of speech points into the same direction. Of course animacy and part of speech are both related to the more flexible variables approximating topicality, so their role in DOM is motivated. Nevertheless, both severely constrain the speaker's freedom of choice once he has made certain basic decisions. See the next two sections for a more detailed discussion.

Another point to note is that the highest CV for DOM (0.55 for animacy) is still far below what we observed for S/A detransitivisation (0.90 for quantifiability). This shows that DOM is functionally much more complex than S/A detransitivisation.

3.6.5 Interplay of variables

In the next step, all variables were fed into a single logistic regression model in order to investigate their cumulated effect. As a preparatory step, further value transformations had to be undertaken. Logistic regression tries to find a formula that predicts the log odds of the probability of the dependent variable based on input from the predictor variables. While the values of continuous variables can be easily integrated into such a formula, non-continuous variables have to be mapped to numbers first. This is unproblematic in the case of binary variables, whose values will simply be mapped to 0 and 1. However, in the case of variables with more than two values the transformation is not so easy. For instance, it was claimed in section 3.5.3 that the values of animacy form a hierarchy, that is, they are ordered. This means we could in principle transform them to a sequence of numbers. However, such a transformation presupposes that the variable is interval-scaled, and in reality we do not know whether the intervals between the values are of the same size. For instance, mapping `human` to 4, `concrete` to 3, and `state` to 2 entails that `concrete` is as far away (in terms of its strength as a predictor of case) from `human` as it is from `state` and that `human` is twice as strong as `state`.

Since these assumptions are not warranted, variables with more than two values have to be mapped to 0 and 1, too. This can be done in two ways. One possibility is to fuse several values so that only two categories are left in the end. However, this does not make equal sense for all variables. For instance, the last version of `ctag` (parts of speech) had the five values `adj`, `n`, `dem`, `pro`, and `other`. While `adj`, `n` and `other` could be easily fused to a single category, further fusions

would destroy crucial information. dem is likely to be a strong predictor of DAT, so it should not be fused with the default category. But pro is even stronger (allowing no NOM at all), so fusing dem and pro would water down the effect of the latter. In such cases it is a better solution to split the variable into two (in this case, pronominality and demonstrativeness). While this may seem conceptually odd because the two variables are so obviously coupled, this is not a problem for the regression formula: since a referent can never be a pronoun and a demonstrative at the same time and therefore never have the value 1 in both variables, one of the two will always get the value 0 and will accordingly be ignored.

A related problem is that many of the involved variables have a high degree of collinearity, that is, they are not only correlated with DOM but also with each other. For instance, ranked frequency and the number of competitors are related by definition, because only those referents are viewed as competitors of a given referent that have a higher frequency. A more subtle connection exists between animacy and all topicality-related measures: it is well known that highly animate referents are more prone to become topics than low referents (Iemmolo 2011:67 and references therein; see in particular Givón 1983 for several seminal papers).

In fact, there is hardly any pair of relevant variables in DOM where one could not suspect collinearity based on theoretical grounds. Collinearity is bad for logistic regression because it makes it hard to assess the effect of variables independently (Baayen 2008:181). On the other hand, there is no easy cure for collinearity because ignoring collinear variables is equivalent to ignoring data, however small their independent effect may be (Menard 2002:77). According to Harrell (2001:244), collinearity also is much less of a problem than non-linearity and overfitting because even though the individual contributions of collinear variables may get blurred to some degree, the cumulated effect of all variables is no less valid. I therefore inspected all variables in isolation (see section 3.6.4) in order to compare their individual effects but not did not take special measures against collinearity apart from that.

Here is an overview of the final transformations that were performed before regression analysis:

- Within role, P and G were fused. This is possible because G eligible for DOM are rather rare and do apparently not behave differently from P.
- Ctag (parts of speech) was split into two variables, pronoun and demonstrative.
- Modification still contained too many values. The most important values here are humposs and dem. All others are either not significant in isolation (none, poss, adj, several, other) or their behaviour is not well understood (inhibitory effect of num and relclause). humposs was split off into a separate variable of the same name. The same could have been done for dem, but since a conceptually very similar, dedicated variable already existed (demonstrative from former ctag dem), dem was fused with this variable so that eventually all referents got the value dem which were coded by a demonstrative or by a noun modified by a demonstrative.
- Animacy was split into three binary variables process (processes vs all other referent types), abstract (abstract referent types process and state vs the remaining, concrete referent types), and human (human including the earlier values human.group and human.prop vs the rest). Animacy proper as a distinction between animate and inanimate referents was ignored because of the rarity of anim and the resulting lack of distinct behaviour.
- Contrastive focus was suspected to be irrelevant in the qualitative discussion in section 3.5.11 and was confirmed to be non-significant in the last section, so it was fused with nofoc to a single category ordinary, standing in a binary opposition with fragile.
- Similarly, all values of relative position except OA were fused to ordinary, creating another binary opposition.
- All value of co-argument case except T with G-DAT were fused to ordinary O.

Two variables, givenness and ranked frequency (so far), had to be excluded from the analysis because they were so highly collinear with each other and with ranked frequency (total) that the `lrm()` function for creating logistic regression models wouldn't work when including them. The ranked frequency (so far) of a new referent is 1 divided by the highest frequency, whereas the same value for a given referent must be higher than that. This means that the values of givenness can be fully predicted from the values of ranked frequency (so far). There is additional collinearity between the two aspects of ranked frequency in that the values of ranked frequency (so far) approximate those of ranked frequency (total) the more the closer one comes to the end of a text. After the mention of the last referent, the values of the two must be equal.

Another variable, distance to the last mention, was excluded because it had too many missing values. Distance to the last mention is not defined for new referents, and since new referents make up the biggest part of all referents, distance gets `NA` in about 75% of all cases. Distance to the last mention failed significance in isolation, so it is unlikely that it would have made an important predictor.

The remaining variables were all fed into a single logistic regression model using `lrm()` in order to get a first impression of their performance. The resulting maximal model can be evaluated with respect to various test statistics and the significance of the contributions of the individual variables. These will be discussed in the following paragraphs. The full specifications for the model will only be given further below after a couple of amendments.

The R^2, whose meaning has already been described above (section 3.6.4), is 0.57 for the maximal model. This means that this model is about 57% better than the null model (or more precisely, it removes about 57% of the error implied by the null model). On the one hand, this is a substantial improvement, on the other, it is far from the ideal value of 1.00 (100% error deletion).

Another important statistic, Somer's D_{xy}, measures the ability of a model to discriminate between the outcomes of the dependent variable. It is based on another statistic, the index of concordance c, which is calculated by taking all possible pairs of observations such that the values of the dependent variable differ (i.e. in our case, where one has NOM and the other DAT) and checking whether the probability predicted for the value of interest (DAT) by the model is higher than that predicted for the other value (Harrell 2001:247). While c ranges from 1 (perfect prediction) to 0 (perfect prediction in the wrong direction) with 0.5 representing random prediction, D_{xy} is defined as 2(c - 0.5) and therefore ranges from -1 to 1.

Somer's D_{xy} for the maximal model is 0.86, so the model discriminates between NOM and DAT correctly in about 86% of all cases. This is good, but it should be kept in mind that D_{xy} does *not* necessarily say something about the predictive abilities of a model. For instance, if the model assigns a probability of 0.8 to DAT and of 0.1 to NOM in one case and 0.1 to DAT and 0.01 to NOM in another, D_{xy} considers both cases as successful discrimination. It ignores that the probability of 0.1 is once linked to a NOM and another time to a DAT, which means that given a probability of 0.1 we cannot say anything about which case is more likely to be chosen.

Another test statistic is the ratio of the likelihood of the data under the null model to the likelihood under the fitted model. This ratio is usually logged and multiplied by -2 because the resulting value (henceforth simply "L.R.") is approximately χ^2-distributed and can thus serve as the base for calculating significance (Harrell 2001:183).[5] L.R. has a minimum of 0 (null model and fitted model fare equally well) but no upper boundary: it gets the bigger the more likely the data are under the fitted model as compared to the null model. The L.R. for the present maximal model is 1119, which is not meaningful in itself but will be useful in comparing this model to others below.

A simple maximal model such as the one we are presently dealing with always runs the risk of overfitting the data, that is, it may be too heavily adapted to the data collected for this research and not perform well when provided with new data. The function `validate` in the `rms` package addresses precisely this problem. `validate()` resamples the data (i.e. it builds an arbitrary subset of the existing data) a number of times to be specified, refits the model, and performs fast backwards

[5] L.R. can also be defined as the difference between the two involved likelihoods, where each is again logged and multiplied by -2 (Menard 2002:21; also cf. Baayen's (2008:204) term "deviation"). The two definitions are equivalent since $-2log(\frac{LH_0}{LH_{\text{fit}}}) = -2log(LH_0) - -2log(LH_{\text{fit}})$.

	χ^2	retained?
humanness	115.89	yes
quantifiability	60.67	yes
process	47.09	yes
demonstrative	43.86	yes
genre	41.13	yes
competitors	22.79	yes
abstract	22.12	yes
distance from predicate	17.21	yes
diathesis	15.23	yes
human possessor	11.64	yes
co-argument case	8.79	yes
ranked frequency (total)	5.68	yes
situation	5.19	no
focus	5.13	yes
role	0.87	no
relative position	0.14	no
pronoun	0.04	no

Table 3.19: Significance of variables in the maximal model

elimination on the variables in the model. Fast backwards elimination drops variables that did not significantly improve models in a sufficient number of resampling steps. More details can be found in the documentation of `rms` (Harrell 2011:46) and in Lawless and Singhal (1978), the paper on which `validate()` is mainly based.

I used `validate` with 10,000 repetitions. In the first run, an extremely high threshold was chosen for keeping variables so that all variables were dropped. Since this happens in the order of importance, this has the side effect of producing a ranking. In the second run I used the default threshold. Table 3.19 shows all variables ranked according to their χ^2 values and whether they made it through fast backwards elimination in the second run or not.

Most variables got through, but four failed. Role, situation and relative position were significant in isolation but were among the variables with the lowest Cramer's V (sections 3.6.4.1, 3.6.4.4, 3.6.4.13). What is quite surprising, though, is the failure of pronominality. As stated in section 3.5.8, pronouns are one of the few referent types which strictly require DAT in object position. The best explanation for this is that pronominality is highly collinear with the two most significant predictors: pronominal referents are virtually always human (always in the annotated corpus), and they are almot always quantifiable (and always specific). Thus, what could be explained by reference to part of speech can normally just as well be explained by humanness and quantifiability.

Another surprising point is that diathesis made it into the final list even though it was *not* significant in isolation (section 3.6.4.8). However, this improvement makes sense if we consider that the main effect of the passive is to allow NOM where otherwise DAT would be very likely or even compulsory. This effect only becomes visible when diathesis is looked at in combination with other variables. An additional plus is that diathesis is the only variable that does not somehow correlate with the animacy/topicality complex – at least not in an obvious way. Its impact can therefore be easily separated from that of other variables.

One interesting fact about Table 3.19 is that two most significant predictors (animacy and quantifiability as a proxy for specificity) are also the two which have most frequently been mentioned in the literature on DOM in Nepali and in other Indo-Aryan languages (cf. section 3.5.2 and section 3.9 below). It thus seems that the most robust variables are also those which are easiest to see for the eye of the grammarian.

If this is true, it strengthens the role of descriptive work using more conservative methods for explaining DOM. Although logistic regression clearly brings with it a deeper and more realistic

	realistic model	minimal model
R^2	0.56	0.39
Somer's D_{xy}	0.85	0.64
L.R.	1102	741

Table 3.20: Realistic model and minimal model in contrast

understanding of DOM, the basics can be covered with animacy and quantifiability. This statement can be put into numbers by setting up a model using just humanness and quantifiability and comparing it to the model with all retained variables ("realistic model"). Table 3.20 contrasts the relevant test statistics.

All test statistics are notably better in the realistic model but far from bad in the minimal model. The improvement from the minimal to the realistic model is highly significant with $p < 0.01$[6]. So it may be said that simple explanations are not bad *per se* – all they should do is to admit that they are practical abstractions.

One last step to reduce the danger of overfitting is to introduce a penalty on coefficients. Coefficients are numbers that are multiplied with the values of the relevant variables in the regression formula in order to determine their contribution to the probability of DAT. As described in Baayen (2008:205), coefficients based on a single sample tend to be too large. In order to solve this problem, a penalty can be introduced to shrink the coefficients. The `rms` package provides the function `pentrace()` for this, which tries to estimate the difference between the real and the sample coefficients and outputs the best penalty for each coefficient. `pentrace()` was used on a model containing just the variables that made it through fast backwards elimination, and after that the same model was refitted once more using the provided set of penalties.

The resulting final model has an R^2 of 0.56, a Somer's D_{xy} of 0.85, and an L.R. of 1098. These values are more or less equal to those of the maximal and the realistic model, so the exclusion of the five bad predictors and the inclusion of a penalty came at a low price.

Table 3.21 shows the regression formula for the final model. The starting point of the formula is the intercept, to which the values of the various predictor variables are added after multiplying them with an appropriate coefficient. A positive coefficient indicates that a variable increases the probability of DAT, a negative coefficient does the opposite. Note, however, that the size of the coefficient does not necessarily indicate the strength of the predictor. Coefficients are only comparable for variables with equal ranges, which are most but not all. For instance, ranked frequency (total) ranges between 0 and 1 with its mean at 0.09, whereas there may be between 0 and 14 competitors, the mean being 4.26. An average number of competitors may therefore have a greater effect than an average ranked frequency, even if its coefficient is lower.

The values of quantitative variables can be multiplied and added directly. Non-quantitative variables are first mapped to 0 and 1 as described above. For these variables, the case corresponding to the value 1 is given along with the variable name.

The output of the formula is the logit L (the logarithm of the odds for DAT). The probability p(DAT) is then $\frac{1}{1+e^{-L}}$. As described above, non-quantitative variables had to be mapped to 0 and 1 in order to be integratable into the regression formula.

There are two points to be noted here. First, it has been repeatedly claimed above that the greater the distance from the predicate, the higher the probability for DAT. Now the negative coefficient makes it look as if it was the other way round. However, this is an artifact of the special definition of distance used here. Remember that O to the left of V get negative values (e.g. -3 = three words to the left of V) and only O to the right get positive values. This means that the higher the distance values are, the farther to the right is the O. The negative coefficient in Table 3.21 thus expresses that the probability of DAT sinks the farther to the right an O is, which conforms with the earlier claims.

The other point is that `abstract` (which was part of animacy before the regression analysis)

[6]The calculation of the significance of improvement was based on the difference between the log likelihoods of the two models, as for the L.R. measure explained above.

variable	coefficient
(intercept)	-12.20
+ **humanness = human**	· 4.03
+ **co-argument case = ordinary O**	· 3.03
+ **non-process**	· 2.07
+ **ranked frequency (total)**	· 1.64
+ **quantifiability = quantifiable**	· 1.90
+ **diathesis = active**	· 1.87
+ **demonstrative**	· 1.74
+ **focus = fragile**	· 1.46
+ **human possessor**	· 1.07
+ **abstract**	· 1.01
+ **genre = written**	· 0.95
+ **competitors**	· -0.14
+ **distance from predicate**	· -0.22

Table 3.21: A formula for predicting the logit of DAT

gets a positive coefficient. This is intuitively unexpected – concrete objects should get DAT more frequently than abstract ones because abstract referents are never animate and (possibly) less often specific. In fact, `abstract` does get a slightly negative coefficient (-1.10) in a model that is built only on that factor. However, as soon as `human` is taken in the coefficient becomes positive and stays so in more complex models including the final one. This can be explained if we assume that the correlation of `abstract` with NOM feeds on its correlation with `non-human` – all abstract referents are also non-human (if not the other way round). When `human` is taken in, the portion of the variance in O marking that could be imperfectly explained by `abstract` is taken over by the former and `abstract` only remains useful when combined with non-human reference. That `abstract` gets slightly more DAT there than `concrete` may be due to the fact that abstract concepts are harder to grasp, so that it is more often necessary to point them and their role status out by using DAT.

Figure 3.23 visualises the relation between the individual variables and the probability of DAT. The grey bands (quantitative variables) and the vertical blue lines (non-quantitative variables) mark confidence bands.

All the logistic regression models presented above have one thing in common: they ignore speaker-related variables. On the one hand this is justified – as discussed in section 3.6.4.17, age and gender are likely to be irrelevant for O case, and speaker identity is not a useful predictor because there are so many speakers of Nepali. On the other hand, as has also been mentioned, there is huge variation in the use of DAT between speakers. This variation is a hindrance to predicting the probability of DAT, so it would be good if it could be factored out. This is possible with the help of a mixed-effects model. Mixed-effects models take into account both fixed effects (variables with a well-defined set of values that may be used as predictors) and random effects (variables with a large, unknown set of values that are less useful as predictors). In the present case, all variables in the final model are fixed effects and speaker identity is a random effect.

The mixed-effects model was built using the `lmer` function from the `lme4` package (Bates et al. 2012). It is not trivial to compare the mixed-effects model to the simple models from above. There is no uniform definition for R^2 in mixed-effects models (Matuszewski 2011). Somer's D_{xy} is defined but not implemented in `lme4`. Since this statistic is of lesser interest, I refrained from implementing it myself. What is possible, though, is to calculate L.R. This value is 1156 for the mixed-effects model, which is slightly better than the 1098 of the final model. More importantly, the comparison of the log likelihoods of the two models shows that the improvement achieved by the mixed-effects model is significant with $p < 0.01$.

Table 3.22 shows the regression formula for the mixed-effects model. All variables that got

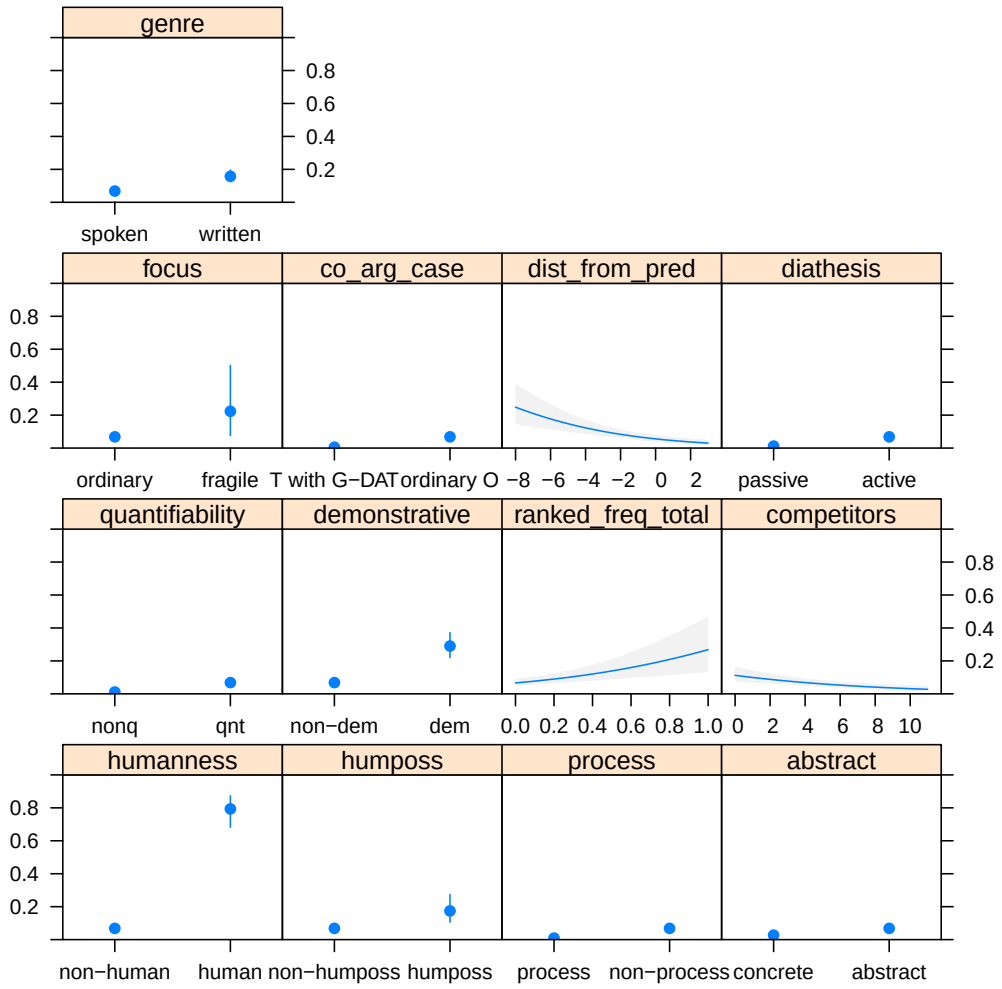

Figure 3.23: Relation between the final variables and the probability of DAT

through fast backwards elimination also stay significant in this model, so these variables may be said to be relevant across speakers. Most coefficients increase slightly, but there are no dramatic changes. The most important addition is the random effect of speaker identity. In the formula this takes the form of an additional summand modifying the intercept. No general value can be given because the summand is different for each speaker. For instance, for Dipaka Jaṅgama 1.43 would be added to the intercept, and for Pitāmbara Upadhyāya 0.48 would be subtracted. The range of the random effects is from -1.43 to 1.46, the standard deviation 0.89.

3.6.6 Predicting DOM by probability and rules

It is now time to compare the probabilistic models we built above (i.e. the final simple model and the mixed-effects model) with the kind of rule-based models implied by most existing work on DOM in Nepali and other Indo-Aryan languages.[7]

[7] Only a few grammarians have expressed doubts on whether DOM can be fully explained by rules. For instance, Bailey and Cummings (1912 [1994]:343) claim for Panjabi that "nothing but long practice will fully show when to insert and when to omit *nũ*" (the postposition corresponding to Nepali *-lai*). Greaves (1921 [1983]:96) makes even stronger claims for the Hindi object marker *-ko*:

variable	coefficient
(intercept)	-12.97
+ term for speaker identity	
+ **humanness = human**	· 4.42
+ **co-argument case = ordinary O**	· 3.27
+ **non-process**	· 2.04
+ **ranked frequency (total)**	· 1.97
+ **quantifiability = quantifiable**	· 1.96
+ **diathesis = active**	· 2.00
+ **demonstrative**	· 1.76
+ **focus = fragile**	· 1.53
+ **humposs**	· 1.08
+ **abstract**	· 0.75
+ **genre = written**	· 1.10
+ **competitors**	· -0.13
+ **distance from predicate**	· -0.22

Table 3.22: Formula for predicting the logit of DAT with speaker identity as random effect

The discussion in section 3.5 should have made it clear enough that a *simple* rule-based system based on one or two variables definitely does not exist. Even though there are a couple of form classes which mostly go together with a certain case, there are very few which do not allow for any exceptions. What's more, the strongly grammaticalised cases only cover a negligible share of all objects.

What might be possible is that there is a *complex* rule system, taking as many input variables as the logistic regression model or even more. However, the more variables one considers, the more combinations of values and the less instances of each combination one gets – the 13 final variables are attested in 995 different combinations even if one rounds all ranked frequencies to 0 or 1, and the combination with the highest frequency is still only attested 45 times. This makes complex rule systems less plausible because they require speakers to remember a lot of rules for more and more special cases (one for each combination that cannot be merged with other combinations based on shared values), some of which are extremely rare.

What's more, even highly complex systems do not yield a lot of value combinations that only occur with one O case. For instance, there are five referents in the annotated corpus which have the value combination {`role=P/G`, `co-argument case=ordinary O`, `relative position=OA`, `animacy=state`, `ctag=dem`, `modification=none`, `quantifiability=qnt`, `givenness=given`, `situation=non-concrete`, `focus=ordinary`, `diathesis=active`}. Three of these have NOM and two DAT, so yet another, unknown variable would be necessary to discriminate between them. These facts cannot rule out that DOM is governed by a rule-system – it is generally hard to prove that something is *not* the case. However, they make the existence of such a system rather unlikely.

One last objection is that even though DOM is probably not determined by rules in reality, rules might still do a better job at predicting O case. This is an empirical question, and two ingredients are needed to answer it. First, models have to be fixed that can predict case. We already have the final probabilistic model given in Table 3.21 above, but we yet need to define a rule-based model. Second, the outputs of the models have to be interpreted and compared to each other.

There are two ways in which a rule-based model can be set up. In order find the ideal model,

"To form a rule, or rules, by which it can be decided which form should be used in each individual instance is impossible. No rule exists on the subject, and not in all cases can it be said that the matter is regulated by idiomatic usage, for sentences could be given which in other respects thoroughly correspond, yet को is used in one, but not in the other."

As a more recent example, Mahapatra (2007:123) admits for Oriya that "it has not been possible to frame hard rules to predict" the occurrence of the dative marker *-ku*.

it would be necessary to draw all possible combinations of relevant variables and then all possible combinations of values of these variables. Each combination would have to be assumed to predict DAT, and the combination that yields most correct predictions would finally be chosen as the ideal model. Given the large computational effort required for this and the objections against complex rule systems that were made above I took a different approach: DAT was predicted for a few well-known and frequent cases and NOM for all others. The applied rules are as follows:

- Human, quantifiable referents get DAT.
- Quantifiable referents coded by a demonstrative get DAT.
- Pronouns get DAT.
- Highly topical referents get DAT. The mean ranked frequency (total) for DAT-marked O is 0.26, and high topicality was defined as anything above that threshold.

The next step is to interpret and evaluate the outputs of the models. There are several possibilities here. The most intuitive method is to assume that they predict cases and to compare their PREDICTIVE ACCURACY, that is, how many cases they predict correctly. In the case of the probabilistic models this method requires a mapping of probabilities to a binary distinction. We will therefore assume that when a model predicts a probability higher than 0.5 for a given value combination it predicts DAT and that it predicts NOM when the probability is equal to or lower than 0.5.

Note that this mapping is not without theoretical problems, the most important ones being that it reinterprets probabilities as activation values and that the cutoff point is arbitrary (Harrell 2001:248). Still, as predictive accuracy is an easy to grasp concept and most straightforward of all evaluation methods, I will use it below, keeping in mind that it makes strong assumptions. In addition to the overall predictive accuracies of the models, the accuracies within each case were also calculated. In order to reduce the danger of overfitting, the data were resampled 10,000 times and the mean accuracies were calculated in addition to the simple accuracies.

Another evaluation method that takes probability more literally is as follows. Strictly speaking, logistic regression does not predict which case will be used on a given object but how likely it is that it will be used. Applied to distributions this means that if DAT is found, for instance, in 60% of all instances of some combination of values, a perfect regression model will give a probability of 0.6 for this combination.[8] The rule-based model differs from the probabilistic one in that it can only predict 0% or 100% of DAT within each value combination.

The goodness of the model can then be measured by looking at the differences between the predicted and the attested proportions of DAT within each unique combination of values. Each difference is weighted by multiplying it with the number of attestations of its combination, and an average (hence MEAN DIGRESSION) is calculated by dividing the sum of weighted differences by the number of all attestations. The values for mean digression range from 0 to 1, with 0 marking a perfect model which exactly predicts the proportion of DAT in every combination of values. As with predictive accuracy, the mean digression was additionally calculated with 10,000 resampling runs.

Although this method is cleaner in the sense that it does not reinterpret probabilities, it is also harder to interpret from a cognitive point of view. The predictive accuracy method can be viewed as modelling production: when a speaker has to decide between two possible cases for an O, it doesn't help him much if he knows how probable each case is – rather, he will base his decision on whether the activation level for one value exceeds a certain threshold. Things become a bit more complicated when more than two cases are involved but can still be modelled based on the comparison of activation levels where none has to reach 100% as long as one is higher than all others.

Finally, the probabilistic models can also be evaluated by comparing their predictions to simulations (hence SIMULATIVE ACCURACY). For this, the data were resampled 10,000 times. In every resulting data set, a case was simulated for every observation based on the probability of DAT predicted for this observation. For instance, if the probability was 0.7, DAT was drawn with a

[8] For case prediction this would be a problem because if 0.5 is taken as the threshold beyond which DAT is predicted, the correct case would be predicted for only 60% of the instances of this combination.

probability of 0.7. It was then checked whether the simulated case equalled the observed case, and a mean was calculated across all repetitions.

In contrast to simple predictive accuracy, this method penalises less decisive predictions. For instance, probabilities of 0.7, 0.8, and 0.9 are all equally good for the simple method as long as the observed case is DAT. However, in simulation each probability yields a corresponding proportion of concordant cases on the long run, so 0.7 is worse than 0.8 and 0.9. The simple method reaches an accuracy of 100% as soon as the model predicts probabilities greater than 0.5 for all DAT and lower or equal to 0.5 for all NOM. By contrast, the simulation method only reaches 100% when the model predicts 1.0 for all DAT and 0.0 for all NOM. It is thus expected that the accuracy values in this method are lower than those for the simple method.

Similarly to the mean digression method, this method has no simple cognitive interpretation. In addition, it is only useful for probabilistic models. As mentioned above, a rule-based model can only ever predict 0% or 100% DAT within a unique combination of values. The simulation will accordingly yield either DAT for all observations belonging to one combination or for none and will therefore not be different from a simple rule-based prediction.

Table 3.23 shows the results of the three evaluation methods applied to the rule-based and the probabilistic models. Besides, the values for the null model are also shown for comparison. Table 3.24 shows the simulative accuracies of the probabilistic models.

	null model	**probabilistic simple**	**mixed-effects**	**rule-based**
accuracy all O	87%	92%	93%	89%
accuracy all O (mean)	73%	87%	89%	81%
accuracy DAT	0%	49%	56%	62%
accuracy DAT (mean)	0%	65%	71%	77%
accuracy NOM	100%	98%	98%	93%
accuracy NOM (mean)	100%	96%	96%	83%
mean digression	0.13	0.11	0.10	0.11
mean digression (mean)	0.27	0.17	0.15	0.19

Table 3.23: Predictive accuracy and mean digression for three models

	probabilistic simple	**mixed-effects**
all O (mean)	82%	85%
DAT (mean)	66%	67%
NOM (mean)	89%	91%

Table 3.24: Simulative accuracy for the probabilistic model

In general, resampling decreases the overall accuracy and the accuracy within NOM but increases the accuracy for DAT. This suggests that the original data contain a couple of unusual DAT that don't obey the rules and make the probabilistic model more careful than it would have to be. These DAT get easily lost when drawing a subset of the original data in resampling. The relation of the rule-based to the probabilistic models is constant through this variation: the probabilistic models are worse than the rule-based one in predicting DAT but better in predicting NOM, and since there are much more NOM than DAT they are also best in predicting all O cases. The accuracy of the rule-based model is not much better than that of the null model with the full data set (although the rule-based model of course covers more diverse cases); however, in resampling the rule-based model does come out as notably better than the null model. The mixed-effects model is better than the simple probabilistic model in all respects, but the higher percentage of correctly predicted DAT is particularly noteworthy.

Mean digression is not very telling when applied to the full data set. However, after resampling the inferiority of the null model becomes very clear, and the the probabilistic model is at least

slightly better than the rule-based model. Still, both are far from being perfect. A general problem with this evaluation method is that there are a lot of unique combinations of values that are only attested a single time in the annotated corpus (610 or 61% of all even if ranked frequency is rounded to 0 or 1). Within such combinations it is obviously odd to speak of distributions. The mean digression method may therefore be more useful for larger data sets.

As for simulative accuracy, it is mainly the accuracy for NOM that goes down as compared to predictive accuracy and draws the overall accuracy along. This means that less decisive predictions (predictions close to 0.5) are mainly found in the NOM area between 0.0 and 0.5. Thus, the probabilistic model would not only have to predict more DAT in order to improve but also to be more decisive with regard to NOM.

Figure 3.24 shows the densities of the predictive accuracies of the mixed-effects model, the mixed-effects model with simulation, and the rule-based model. The simple probabilistic model was omitted to keep the picture clear and because it was inferior to the mixed-effects model in all respects, anyway.

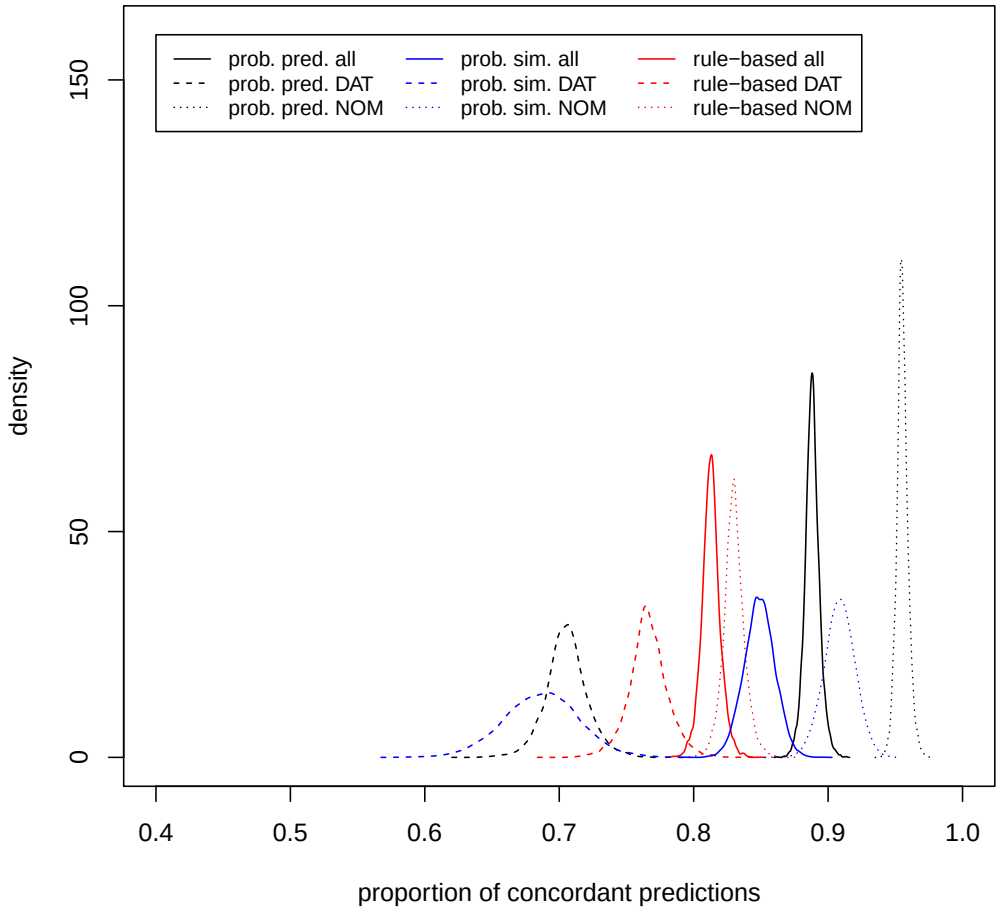

Figure 3.24: Predictive accuracy of various methods with resampling

Since neither the rule-based nor the probabilistic models fared exceptionally good, the possibility was considered that the best model might be one with hybrid characteristics. If we took the

rules predicting DAT from above and let the probabilistic model do the rest, too many DAT might be the result, so in addition to the DAT rules a couple of NOM rules were formulated:

- Processes get NOM.
- Non-quantifiable referents get NOM.
- T with an accompanying G-DAT get NOM.

The mixed-effects model as the stronger probabilistic model handled all observations that were not covered by one of the rules. The results are shown in Table 3.25.

The hybrid model is more or less equally good to the rule-based model in all respects. This has two important consequences. First and trivially, mixing probabilities and rules does not seem very promising – the method adds to the computational complexity of the prediction but not to its goodness. Second and more importantly, the quality of the probabilistic models does not lie in how they handle relatively marginal cases not covered by any rule. If that was the case, the hybrid model would represent an ideal division of labour and should yield better results. Instead, the probabilistic models are better in the core cases, precisely where one would expect rules to be operating.

	hybrid model
all O	89%
all O (mean)	82%
DAT	63%
DAT (mean)	79%
NOM	93%
NOM (mean)	83%
mean digression	0.12
mean digression (mean)	0.19

Table 3.25: Predictive accuracy of alternative methods

To summarise, we have seen that the probabilistic models we proposed in section 3.6.5 are theoretically more plausible than a rule-based model and also fare better in terms of predictive accuracy. A mixed-effects model yields better results than a simple probabilistic model. However, at the present state of knowledge the difference between to the rule-based model is still rather small so that the latter, which is much easier to implement, may be more appropriate for practical purposes such as machine translation or language learning.

If we take a cursory look at the observations where the probabilistic models are wrong, unexpectedness seems to be the most fruitful area for further investigations. The majority of DAT where too low probabilities were predicted have low animacy and topicality values but fall into one of the subclasses described in section 3.5.11. The big question is whether adding annotations for unexpectedness, fuzzy as it is, would on the whole produce more correct DAT than additional wrong NOM. It might well be the case that the present research methodology does and can not yet go deep enough into what is happening in the mind of the speaker in order to model unexpectedness and other underestimated factors behind DOM in a satisfying way.

3.7 Some notes on the history of DOM

3.7.1 The appearance of *-lai*

The history of linguistic forms often provides interesting hints to their present behaviour in form and function. Since Nepali, differently from Chintang, is a language with a documented history, it is easy to investigate this additional aspect of DOM. Large text collections spanning the time from the oldest inscriptions to the 19th century can be found in Pokharela (2031 V.S.) and Barāla (2046 V.S.). Although a detailed corpus study of the development of the dative yet remains to be done, a

good base is provided by Wallace (1981), who gives a qualitative overview of the rise and spread of *-lai* from the earliest inscriptions to modern time.

Below I will mainly summarise and discuss Wallace's data, adding some examples from the texts collected by Barāla. I will not give an overview of the history of Nepali because that has been done in several other works before (beside the ones mentioned, see e.g. Srivastava 1962, Hutt 1988) and because it is not important for the history of DOM. All examples have been transliterated in IAST rather than transcribed because nothing is known about the pronunciation of old Nepali. Dates for sources are given in the Christian era.

Only in the oldest texts it it still possible to find scattered remnants of old synthetic cases, as in the inscription in (184), where the anusvaras on *etikã devã* ultimately go back to the Sanskrit accusative suffix *-m*:

(184) *Vrahmā Viṣṇu Iśvara Vuddha Dharma Sãgha etikã devã ghal-e.*
Brahmā Viṣṇu Īśvara Buddha Dharma Sãgha such.ACC god.ACC mess.with-PTCP
'You will destroy such gods as Brahmā, Viṣṇu, Īśvara, Buddha, Dharma, and Sãgha.'
(Goldplate (1356), Wallace 1981:112)

Such examples are, however, very rare. Most of the oldest texts are characterised by a complete absence of structural case markers – both *-le* [ERG] and *-lai* [DAT] are missing, as also noted by Hutt (1988:79). Even after the emergence of the modern markers, optional zero marking continues for a long time. For instance, NOM is found in the much later example in (185) on the P *mahādeva* 'great god' even though several criteria (human specific referent, high degree of affectedness) favour DAT. In fact, NOM is no longer grammatical here in modern Nepali (elicitation NP 2012).

(185) *tyā mahādeva ghāt gar-yāko pāp*
MED.LOC great.god injury do-PST.PTCP sin
'the sin of having slaughtered the great god there' (Kāntipurakā rājā Lakṣmīnṛsiṃha Mallakō paśuvaliniṣedhādeśa (1642), Barāla 2046 V.S.:14)

The earliest instance of *-lai* stems, according to Wallace 1981, from the end of the 14th century:

(186) *Rāmadāsa Pādhyā lāhi vramavitrā mayā bha-i-ch-a.*
Rāmadāsa Pādhyā DAT Brahman.land gift be-PRF-NPST-3s
'Rāmadāsa Pādhyā received this Brahman land gift.'
(Copperplate, 1398, Wallace 1981:110)

Wallace classifies this as a dative subject. The base for this is dubious. The DAT-marked NP certainly is not S because there is at least one other argument (*vramavitrā* 'Brahman land'). It is also unlikely to be a "subject": semantically it is a recipient, i.e. a kind of G, and the argument that would trigger AGR in Modern Nepali is *vramavitrā*, the T. *bha-* is the perfective stem of the copula *hu-*, so *mayā bhaicha* can be viewed as a complex predicate in the light verb passive (see section 3.4.7). This also explains the absence of A. A more appropriate translation would then be 'This Brahman land gift was given to Rāmadāsa Pādhyā'. The first attested *-lai*, then, marks a human G. All other early examples cited by Wallace belong to this type, too – they are classified as dative subjects in his work but can more transparently be characterised as recipients of passivised ditransitive verbs.

One remarkable point here is that the function of *-lai* as used in (186) does not provide any hints to an earlier, less abstract semantics of this marker. Wallace (1981) does not talk about this issue, but when I searched through the texts provided in Barāla (2046 V.S.) I was not able to find a single instance of *-lai* which did not have one of the abstract functions known from present-day Nepali. Of course that does not mean that *-lai* entered the language as a full-fledged grammatical marker – that would contradict everything that is known about grammaticalisation. What seems most likely is that the grammaticalisation process took place in the spoken language and that the marker only entered the written registers when it was already established as a case marker.

Unfortunately Wallace is not very systematic about the further functional development of *-lai*. The earliest example for the use of *-lai* on an experiencer he gives is from 1770:

(187) *Tesa kurā-mā tā-lāi doṣa ch-ain-a.*
MED.OBL matter-LOC 2LH-DAT guilt be.there-NEG.NPST-3s
'You aren't guilty in this matter.'
(Pṛthvīnārāyaṇa Śāha, Letter to Abhimāna Siṃha (1770), Wallace 1981:118)

However, this use is at least a couple of decades older. The example in (188) was found in the texts provided by Barāla (2046 V.S.).

(188) *Āp-nu pāp paṃḍit-lāī lāg-aṃch-a.*
REFL-GEN sin scholar-DAT attach-NPST-3s
'His own$_i$ sin falls back to the scholar$_j$.'
(Premanidhi Panta, Prāyaścittavidhāna (1730), Barāla 2046 V.S.:69)

Although the DAT-marked NPs in these examples fit best into the category of dative experiencers if one is forced to classify them, they are not prototypical – guilt and sins are different from perceptions and emotions in that they exist to some degree independently of an experiencer. Sentences like these might thus form a bridge between the oldest uses of *-lai* on recipients and the use in more typical experiencer expressions in the modern language that was illustrated in section 3.5.1. Experiencers can be conceived of as the location of experiences, so it is not surprising that a marker that was first used on recipients (animate locations) should be extended to them. This may have been even easier if the first "experiences" were still akin to T in existing independently of an experiencer.

The first example for *-lai* marking an O comes from about the same period:

(189) *mitra-lāi mār-nu*
friend-DAT kill-INF$_1$
'to kill a friend' (Pṛthvīnārāyaṇa Śāha, Letter to Paṇḍita Rājīvalocana (1755), Wallace 1981:120)

Soon after this examples for O-DAT become more and more frequent:

(190) *Tava rājā-le sevaka-lāī ghara-mā na-mār-nu.*
then king-ERG servant-DAT house-LOC NEG-kill-INF$_1$
'Then, a king shall not kill (his) servant in (his) house.'
(Pṛthvīnārāyaṇa Śāha, Divya Upadēśa (1775?), Barāla 2046 V.S.:74)

(191) *Kāṭnā-māphika-lāī kāṭ-nu ḍanḍa hu-nyā-lāī ḍanḍa gar-nu.*
cutting-guilty-DAT cut-INF$_1$ penalty be-NPST.PTCP-DAT penalty do-INF$_1$
'Cut the one who is guilty of cutting, punish the one who is to be punished.'
(Vijita Kumāuṃmā praśāsanako savāla (1785), Barāla 2046 V.S.:44)

According to Wallace, *-lai* was from the beginnings competing with another marker, *-kʌnʌ*. This marker can still be used today in poetry or highly formal prose but has otherwise fallen out of use. Historically, O could initially only marked by *-kʌnʌ* (besides zero and remnants of the old accusative). *-lai* started as a marker of recipients in the light verb passive and similar constructions ("dative subjects" for Wallace) and then got extended to experiencers and O. It first became dominant over *-kʌnʌ* on "dative subjects" (where it's not clear whether this refers again to recipients or to dative experiencers), then on "indirect objects" (presumably recipients in active ditransitive frames), and lastly on "direct objects" (= O).

Wallace's own theory as to why experiencers, recipients, and O ended up being marked by a single marker is based on formal criteria. Recipients and O are connected via the fact that in transitive complex predicates such as *biha gʌr-* [marriage do] 'marry' the syntactic status of the O (the marriage partner) is ambiguous: it could either be analysed as a direct object (with N and V forming a single unit) or as an indirect object (with N functioning as direct object). Experiencers and recipients are linked by a construction that Wallace calls "O-V lexicalization" and that links clauses involving a complex predicate to clauses with only a verb. (192) shows an example for a normal transitive complex predicate, (193) an example for an experiential complex predicate:

(192) a. *Us-le aph-na pʌriwar-lai nas gʌr-y-o.*
DIST-ERG REFL-GEN family-DAT destruction do-PST-3s
'He destroyed his (own) family.'
b. *Us-le aph-na pʌriwar-lai nas-y-o.*
DIST-ERG REFL-GEN family-DAT destroy-PST-3s
'He destroyed his (own) family.' (Wallace 1981:115)

(193) a. *ʌbhagi-lai kha-ne bela-ma ris uʈh-ch-ʌ.*
unfortunate-DAT eat-IPFV.PTCP time-LOC anger rise-NPST-3s
'An unfortunate man gets angry at meal times.'
b. *ʌbhagi kha-ne bela-ma risaũ-ch-ʌ.*
unfortunate eat-IPFV.PTCP time-LOC get.angry-NPST-3s
'An unfortunate man gets angry at meal times.' (Wallace 1981:115)

Neither of these two connections between recipients, experiencers, and O is very convincing. In the case of recipients and O in complex predicates it's not clear why a purely syntactic ambiguity within one construction that does not have any semantic consequences should encourage speakers to give up a similar distinction across constructions. Further, it does not explain why until today only some O are marked by DAT, even though all O are similar to recipients by this criterion.

In the case of experiencers and recipients there are even more problems. First of all it is not clear to what extent O-V lexicalization is productive – it doesn't seem to be grammatical with the majority of complex predicates. Second, the effect of the transformation is rather different for words like *pʌriwar* in (192) (indirect objects in Wallace's analysis) and for words like *ʌbhagi* in (193) (dative experiencers). The former preserve their case marking, the latter lose it but take over AGR instead. In the case of dative experiencers the "fused" verb has an additional suffix *-au*, which is not there in the case of normal transitive complex predicates. Finally, it's not clear how a relatively marginal construction should have changed the case marking system of Nepali in such a fundamental way.

Instead of searching for a formal base for the gradual extension of *-lai*, it seems to be much simpler to look at functional criteria. As was already mentioned, experiencers can be conceived of as the locations of emotions, which makes them similar to recipients, which are the locations (or rather destinations) of a change in possession. Recipients and experiencers are also similar in typically being human or at least highly animate, specific, and highly topical. The latter commonality also makes the extension to O possible and at the same time explains why not all O are marked: only O with the mentioned properties are sufficiently similar to typical recipients and experiencers and therefore qualify for being marked by *-lai*.

The overview that was given of the historical development of *-lai* in this section leaves many questions open. More data and more diligent analyses are required. The most interesting but also least well attested period for the development of the modern case markers is the one before the rise of the Gorkhas. Only a careful review of all available texts from this period can answer questions such as whether experiencers were really marked by *-lai* before O or whether *-lai* got simultaneously extended from recipients to both. Another question concerns Wallace's "dative subjects" which are really recipients: is it true that DAT was first used in passive-like constructions, and if so why? A full account of the development of *-lai* would also have to present more quantitative data – otherwise all statements concerning the development of the dominance of certain functions of *-lai* and of the competition between *-lai* and *-kʌnʌ* must remain vague and impressionistic.

3.7.2 An etymology of *-lai*

The question of the origin of *-lai* is an important component of any historical treatment of DOM. Unfortunately, the etymologisation of *-lai* is hampered by the fact that it apparently only entered the written language when it was already grammaticalised to a considerable degree (cf. section 3.7.1 above). It is therefore impossible to trace its phonological development from an independent form or to observe its functional development from the very beginnings. Nevertheless, it is possible to make an educated guess that is formally possible and functionally plausible.

However, first I will have to argue against a couple of alternative suggestions that have been made by various scholars. All of these are either at odds with known sound laws, functionally implausible, or both. In order to make examples and languages comparable, IAST transliteration is again used below.

Hoernle (1880 [1975]:224) assumes a Sanskrit locative *labdhe* 'for the benefit of' as the etymon of *-lai* (*-lāī*). This cannot be right, since by the known sound laws (Beames 1872-79 [1966]a:283) *labdhe* would go to *lādhe* in the Prakrits. The form is also functionally odd because *labdha* is the past passive participle of the root *labh-* 'take', so the meaning of *labdhe* would rather be 'in/at the taken'. However, probably *labdhe* is not what Hoernle meant, anyway – Bloch (1914 [1970]:211) cites him as proposing Sanskrit *lābhe*. This is indeed the locative in *-e* of the noun *lābha* 'meeting, finding, getting, gain, profit' (Monier-Williams 1899 [1974]:897).

While this form fares better than *labdhe*, it still has its weak points. Although /b^{h}/ regularly goes to /h/ intervocalically (Beames 1872-79 [1966]a:268), /āhe/ > /āi/ is not a regular sound change, plausible though it may look – this word would be the only example where it is attested. Examples such as (186) above seem to present a point in favour of Hoernle's theory, but if he was right we would not expect *-lāhi* but *-lāhe* as the oldest form of the marker and later intermediate forms such as *-lāe*. Further, the diphthong resulting from /āe/ would probably have had a comparatively low second component in the beginning ([ae] or [aɪ]). Even if this later became the present [ai], it seems phonetically very unlikely that there should ever have been a variant [ai:]. Then, however, it's not clear why *-lāī* is spelt with the old long /i/ of Sanskrit, <ई>, in many old attestations and especially later in the 18th and 19th century.

Another proposal concerning the origin of *-lāī* is made by Beames (1872-79 [1966]b:252), who discusses the origin of case markers in several Indo-Aryan languages. He claims that all dative/accusative markers with initial /l/ (Marathi *-lā*,[9], Nepali *-lāī*) or /n/ (Panjabi *-nuṃ*, Gujarati *-neṃ*) as well as several other more peripheral markers (e.g. Old Hindi *-lau* 'up to, until') ultimately derive from the Sanskrit root *lag-* ('adhere, stick, cling or attach one's self to', Monier-Williams 1899 [1974]:893). He mentions that Marathi *-lā* has an older variant *-lāgīṃ* from which it is derived via (irregular) shortening and relates Nepali *-lāī* to the same form via elision of /g/. If *-lāī* really were directly related to *-lāgī*, however, nasalisation would be expected (*-lāīṃ*).

Beames also says that old Marathi *-lāgīṃ* is derived from a (apparently reconstructed) "participial form" *lagi*. Whereas for Marathi it is unclear how *lagi* got nasalised to *-lāgīṃ*, this form looks like a more promising antecedent of Nepali *-lāī*. The lengthening of the stem vowel can be easily explained: *lag-* already had a variant *lagy-* in Sanskrit (Monier-Williams 1899 [1974]:893), and /VCy/ regularly goes to /VCC/ and from there to /V:C/ via compensatory lengthening (Beames 1872-79 [1966]a:282), so *lagy-* > *lagg-* > *lāg-*. The *-i* can be traced back to the Sanskrit gerundive (Whitney 1889 [1974]:345) in *-ya*. According to Fahs (1989:182), *-ya* still exists in Pali, where it has a variant *-iya*. The sound change *-i(y)a* > *-i* exists, cf. the examples in Srivastava (1962:18), so **lagi/lāgi* is a possible Middle Indo-Aryan word form.

However, there are some other points that speak against this form being the ancestor of *-lāī*. First, dropping of intervocalic /g/ does only occur sporadically (e.g. Sanskrit *bhaginī* > Nepali *bahinī*, /h/ probably < /b^{h}/) and does not seem to be a regular sound change. Furthermore, this change is only attested for primary (old) /g/, not for secondary /g/ such as the one in **lāgi*, which must be assumed to have developed out of a geminate, as shown above. Second, a word of the form *lāgi* exists in modern Nepali – its function is [FIN$_1$] or, simpler, 'for'. If one assumes that *-lāī* is derived from **lāgi* via elision one gets difficulties in explaining the origin of *-lāgi* – it is easier to assume that the latter is the direct descendant of **lāgi*.

Turner (1931 [1990]:551) in his dictionary entry on *-lāī* first cites Hoernle and then suggests an "absol. or infinitive of Sk. *lāgayati*" (stem *lāg-*, a causative of *lag-*) as another possibility. This is similar to Beames' idea and has the same disadvantages.

The last proposal is by Srivastava (1962) and can be easily dispensed with. Srivastava (p. 93) claims that *-lāī* is directly derived from Sanskrit *laggati* via intermediate forms *laggai* and *laai*.

[9]It is interesting to note that like Nepali *-lāī*, this marker is absent from the oldest Marathi texts according to Bloch (1914 [1970]:210).

It is hard to tell whether he consciously ignores the fact that *laggati* is a third person singular indicative present, the common citation form for Sanskrit verbs (in this case probably the already mentioned *lag-*, though the geminate /gg/ anticipates the Pali form – cf. Childers 1875 [2005]:217). It is functionally very unlikely that a dative marker should have developed from a fully inflected form, and formally all involved changes are highly dubious.

To summarise, all existing etymologisations of *-lāī* have some obvious problems. To me the reason why so far no etymology could be found that works seems to be that scholars have been looking for an etymon in the wrong place. All proposals discussed above derive *-lāī* from a Sanskrit source, but that *-lāī* should go back so far is rather unlikely. Sanskrit detached itself from the normal, spoken language quite early. According to Masica (1991:55), Classical Sanskrit as a literary language had its greatest flowering in the first millenium AD but actually goes back to an Indo-Aryan variety that was spoken around the seventh century BC. Since the grammar and basic vocabulary of all modern Indo-Aryan languages including Nepali are undoubtedly not directly derived from literary Sanskrit but from its spoken equivalent, Sanskrit etymologies for case markers in general suggest a very early grammaticalisation. But this seems odd, first because Sanskrit with its rich case inflections had no need for additional case markers, and second because a marker that became grammaticalised so early would probably have shown up in texts earlier, e.g. in Middle Indo-Aryan. Of course one may argue that Sanskrit etymologies are only an imperfect replacement for etymologies from its spoken equivalent and that case markers such as *-lāī* could actually have come about much later – but then it's not clear why one should use Sanskrit etymologies at all rather than whatever comes closest to the spoken language.

Since Nepali *-lāī* appears relatively late – around the middle of the 14th century AD, as we have seen above – the best source for it seems to be New Indo-Aryan, i.e. Nepali itself. Here, there are two candidates for an etymology: the verbs *lāg-* 'be attached to, be at' (deriving from the Sanskrit *lag-* already discussed above) and *lā-* 'take' (< Sanskrit *lā-* 'take, receive, obtain', Monier-Williams 1899 [1974]:899), both in the converbial form in *-ī* (< Sanskrit *-ya*, compare above; modern *-i* [CVB_2]).

lāg- presents similar difficulties to Beames' **lāgi* – the late drop of /g/ requires a unique sound law which for some reason did not affect other words with intervocalic /g/, in particular *lāgi* [FIN_1]. Therefore, *lā-ī* [take-CVB_2] 'taking, having taken' seems the best candidate to me. The grammaticalisation of a concept meaning 'take' to an object marker is attested in several other languages, too, according to Heine and Kuteva (2002:289). A sentence like *suŋgur-lai mar-y-o* [pig-DAT kill-PST-3s] 'he killed the pig' would then have been derived from *suṃgur lā-ī mār-y-o* [...take-CVB_2...] 'taking the pig, he killed (it)'. Note, however, that the development of 'take' to a marker of recipients or experiencers is not attested, so the period during which *-lāī* was not yet used as an O marker is exceptional in terms of what is known about typologically common grammaticalisation paths.

3.8 Summary

This section summarises what has been said about DOM in this chapter.

DOM in Nepali is an alternation of two cases (zero-marked NOM and DAT marked by *-lai*) on the roles P, T, and G. Which P, T, and G are eligible for DOM can be determined based on the case-marking of co-arguments: P is O with A-NOM/ERG, T is O whenever it allows only NOM/DAT, G is O with T-ERG. The most notable formal constraint on P/T/G-DAT is a general tendency against double datives. DOM is thus impossible on P with A-DAT and extremely rare on T with G-DAT. T-DAT with a verb that normally governs G-DAT is much easier to get when G is covert, so here the tendency goes against two *overt* datives assigned by a single predicate.

It was examined whether O-NOM could be interpreted as incorporated, but that view was rejected on formal grounds. "Functionally incorporated" O do go together with NOM but do not cover its whole range. Other formal properties are that only O-NOM can trigger agreement in the passive, whereas O-DAT triggers dummy 3s-AGR, and that DAT is almost always ungrammatical on the N of complex predicates, even if an N is most conveniently interpreted as P. DOM is not possible in the light verb passive because this construction removes A from the valency and ac-

cordingly leaves O in S. Apart from DOM and the two passives, there are no other constructions that make reference to the grammatical relation of O.

The area where DOM is most challenging are its functional properties. Existing explanations of Nepali DOM haven't been able to capture the complexity of this phenomenon in several respects. First, many functional factors that are relevant had not been noticed before. This concerns topicality, unexpectedness, disambiguation, and the precise nature of the interplay of specificity and quantifiability. Second, distinctions within known factors were often described inadequately: for instance, the most important distinction within animacy is not animate/inanimate but human/non-human. Third, existing models of DOM are too simple. DOM is not monocausal and also cannot be described exhaustively by rule systems considering two or three variables. Most importantly, combinations of values that predict a certain case in 100% of cases are the exception rather than the rule. It is very often impossible to say for a given form which factor determined it. What is often possible, though, is to say which factors *contributed* to it.

The following functional factors were found to be relevant for DOM:

- **Animacy** or better humanness is highly relevant but does not produce as strict results as has sometimes been claimed – non-human referents can be marked by DAT and human referents by NOM. Evidence pointing to a hierarchy is mixed: whereas a hierarchy seems to be at work in elicitation, quantitative corpus data do not support this. A further important distinction that can either be integrated with animacy or be looked at separately is between static referents and processes.

- **Specificity** is also important, especially when taken together with animacy: it is impossible to mark double-high referents with NOM or double-low referents with DAT. Besides specificity in a strict sense (i.e. as identifiability), it is also relevant how much a speaker knows about a referent and how readily it can be accessed. Definiteness is completely irrelevant.

- **Quantifiability** is an important precondition for specificity. This holds especially true if quantifiability is viewed as the syntactic side of the individual/mass distinction – whereas individual concepts in O easily get DAT, mass concepts have to be made quantifiable first.

- **Topicality** was defined as mental presence. In many cases this can be approximated via discourse frequency. More topical O get more DAT.

- **Demonstratives and pronouns** – Pronouns (including reflexive *aphu*) and *tʌpaĩ* [2HH] must always be marked by DAT in O. Demonstratives must be marked by DAT when they have human reference. The demonstratives *u* [DIST] and *ko* 'who' are inherently human and therefore always have DAT.

- **Human proper names** must be marked by DAT in O, except when they do not refer to a person but to themselves.

- **Modification** does not have a notable effect, except for human possession, which has a weakly positive effect on DAT. This association holds independently of how it is marked (possessive pronoun or subclause).

- **Unexpectedness** can explain DAT in many cases where animacy and specificity can't. There are various reasons why a referent can be unexpected, e.g. because it is unlikely in a certain position or because it is in contrast with something in the discourse.

- **Disambiguation** means that O can be marked by DAT when its role would otherwise hard to see. This mainly happens in AO inversion or when O is far away from the associated predicate.

- **Affectedness** is marginally relevant. Strongly affected O are more likely to be marked by DAT, and DAT is impossible on effectuated O.

Most of these variables were directly annotated in a subcorpus of the NNC or calculated on the base of annotations (e.g. topicality-related variables from referential IDs). By and large the results of the quantitative analysis of the annotations confirmed the results of the qualitative discussion. The variables that correlated most strongly (measured by Cramer's V or the point-biserial correlation coefficient) with DOM in isolation were animacy, part of speech, and various topicality-related measures such as ranked frequency, givenness, and the number of competitors. Only distance to the last mention and diathesis failed significance.

A similar picture was produced by a logistic regression analysis. Some variables which had looked significant in isolation got dropped by fast-backwards elimination, viz. situation, position of O, role, and most surprisingly pronominality, whose explanatory share was completely taken over by humanness and topicality-related variables. On the other hand, diathesis made it into the list of final variables even though it had not shown a significant effect in isolation.

An evaluation of the predictive power of probabilistic models based on a logistic regression model and a mixed-effects model yielded the result that the probabilistic approach is, though by far not perfect, superior to a simple rule-based model, to a hybrid model, and to the null model always predicting NOM. A further alternative would have been a complex rule-based model, but this was judged to be unlikely because of the low frequency of repeating value combinations. The probabilistic model presently predicts case correctly in 92% of all cases (87% after resampling). The weak spot of the model is the correct prediction of datives. More research on additional factors is neaeded here.

The last section in this chapter discussed the history of the marker *-lai*. *-lai* appears around the middle of the 14th century in written sources. It starts out as a marker of recipients and only later gets extended to experiencers and O. Given obvious problems with all existing Sanskrit etymologies, a derivation from Old Nepali was deemed to be most likely. The proposed etymon is *lā-ī* [take-CVB_2] 'taking, having taken'.

3.9 DOM in other Indo-Aryan languages

3.9.1 Overview

Speaking about the Indo-Aryan languages as a whole is difficult for two reasons. One is their sheer number. The list of languages found in Masica (1991) suggests approximately ninety languages in this family, although he himself admits that this number is not without problems due to difficulties in distinguishing between dialects and languages, unclear genetic affiliation of marginal languages, and often a lack of reliable data. It is impossible to overlook such a big number of languages in anything else than a large-scale typological study, which the present work is not.

The second difficulty lies in the quality of the available linguistic descriptions. Although there is a large body of literature, much of it are school grammars or works that have been written in some other prescriptive or otherwise preoccupied framework such as Basic Linguistic Theory. Surprisingly enough, some of the best descriptions are found in the colonialist grammars of the 19th century. The wealth of examples that is typical of them is well worth translating their outdated terminology.

In order to mitigate these two problems, I will focus below on a few languages that are either big, particularly well described, or closely related to Nepali.

When looking at differential object marking in a number of Indo-Aryan languages, one thing that immediately springs to the eye is how widespread this feature is. In fact, the Indo-Aryan languages only form part of a much larger DOM area that includes as prominent members the Dravidian and Iranian language families (Masica 1982). Within Indo-Aryan, DOM seems to be found in more languages than other common-place characteristics of this family. For instance, The old distinction between dental and retroflex consonants has been lost in Assamese (Kakati 1941:59) and Romani (Matras 2002:37). Differential agent marking in tenses of the perfective aspect is not found in Bengali (cf. Mukherjee 1985) and Maithili (Bickel and Yadava 2000:345).

By contrast, so far I have not been able to find a single Indo-Aryan language without DOM. It

may not have the same flavour everywhere – for instance, Masica (1991:366) mentions that marked ("definitized") inanimate objects are less frequent in Gujarati and Marathi than in Hindi and Panjabi, and a few languages like Sinhala (Chandralal 2010) and Romani (Matras 2002) have separate dative and accusative cases, thereby changing the structural relations between the object case and the rest of the system. Nevertheless, all these languages have an equivalent to O that can be marked in two different ways, and in all cases the alternation seems to be conditioned by functional factors that are connected to animacy and identifiability.

Figure 3.25 shows an overview of dative markers in Indo-Aryan taken from Masica (1991). Note that this overview defines dative in a rather wide sense and therefore includes many markers that will not be discussed in the following sections, e.g. Nepali *-(ko)lagi* [FIN$_1$] 'for'.

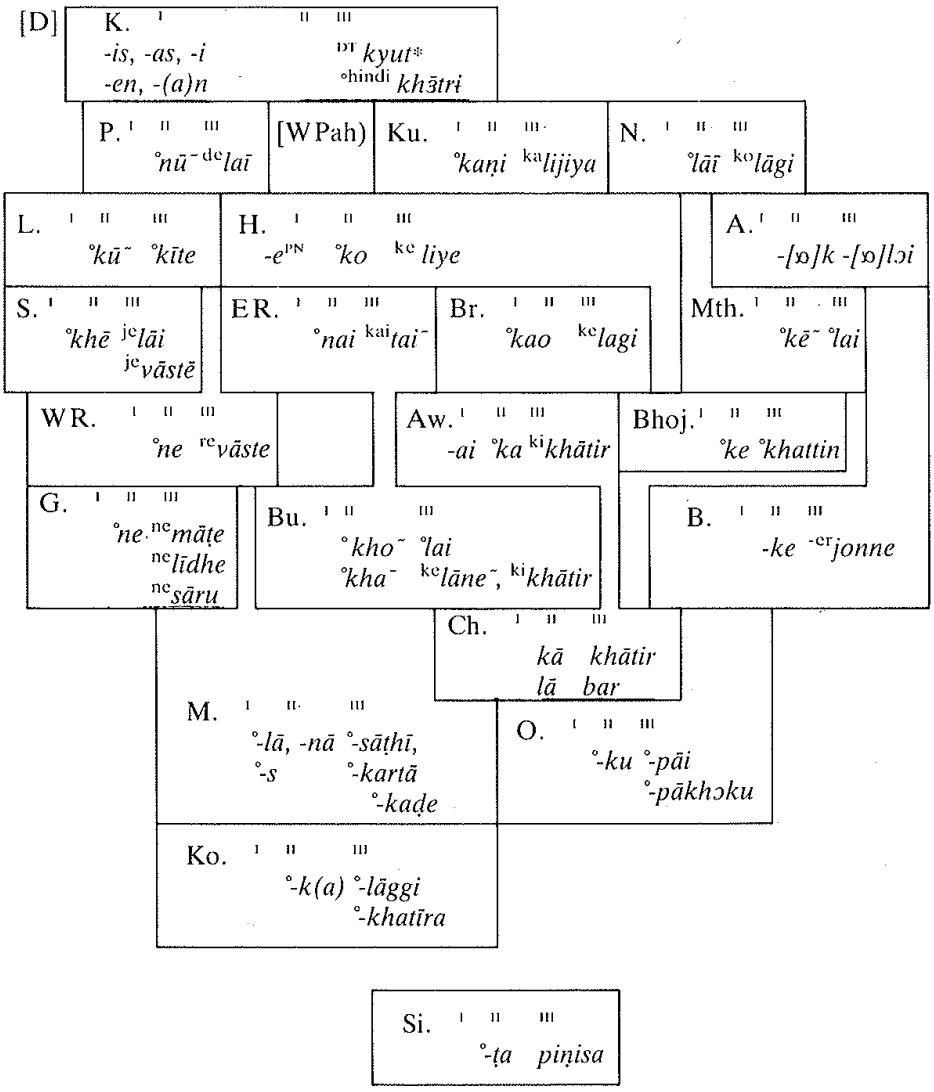

Figure 3.25: Dative markers in Indo-Aryan (Masica 1991:244). Abbreviations: **A**. Assamese, **Aw**. Awadhi, **B**. Bagheli, **Bhoj**. Bhojpuri, **Br**. Braj, **Bu**. Bundeli, **Ch**. Chhattisgarhi, **E R**. Eastern Rajasthani, **G**. Gujarati, **H**. Hindi, **K**. Kashmiri, **Ko**. Konkani, **Ku**. Kumauni, **L**. Lahnda, **M**. Marathi, **Mth**. Maithili, **N**. Nepali, **O**. Oriya, **P**. Panjabi, **S**. Sindhi, **Si**. Sinhala, **W R**. Western Rajasthani

The pervasiveness of DOM in Indo-Aryan becomes a little puzzling when one considers that this is a comparatively young feature – younger, for instance, than the already mentioned retroflex

consonants (which have been present since the oldest Vedic texts, cf. Macdonell 1916 [1941]:3) and differential agent marking (whose roots can be traced back to Sanskrit and which was fully present, albeit based on synthetic cases, as early as Pali according to Peterson 1998). Most importantly, DOM is much younger than the latest common ancestor of the modern Indo-Aryan languages.

This is easy to see for the presently employed object markers, all of which are hard to etymologize. For instance, none of the attempts in the otherwise groundbreaking works of Beames (1872-79 [1966]b) and Hoernle (1880 [1975]) makes use of regular sound laws for deriving these forms. Such difficulties would be unexpected if these markers had easy to identify equivalents in older stages of these languages. For a couple of markers we even know that they are younger than the oldest stage of the individual language they occur in. As mentioned in section 3.7.1, Nepali *-lai* is absent from the oldest inscriptions (Hutt 1988:79), and Marathi *-la* likewise does not occur in the oldest texts (Bloch 1914 [1970]:210). Hindi *-ko* only appeared after the 17th century according to Masica (1982:43). On top of such difficulties, there are no traces of DOM in Middle Indo-Aryan languages such as Pali (cf. Peterson 1998, Collins 2005:20).

It therefore seems that Indo-Aryan DOM is a shared innovation. The question is, then, whether this innovation is due to chance or whether it can be explained based on general principles. A possible simple answer is that DOM is there in all modern Indo-Aryan languages because object marking in general is a recent phenomenon in this family, at least in the present form.

In order to explain this in greater detail we first need some background from historical morphosyntax. Masica (1991) presents a useful descriptive system for Indo-Aryan case marking that is based on the concept of CASE LAYERS:

- layer I: inherited case suffixes directly attached to the stem, often involving morphophonology
- layer II: innovated case suffixes or postpositions with minimal morphophonology, often attached to an oblique case from layer I
- layer III: innovated forms with mostly transparent nominal etymology, always mediated by a layer II form (mostly genitive) and semantically more specific than both other layers

Markers in layer I are generally highly eroded and do not carry a great functional load. Often only one case is preserved and reinterpreted as oblique form of the stem by virtue of its being used before layer II forms. Less frequently other old cases like the locative are still in use, but apparently no modern Indo-Aryan language has preserved old object cases such as the Sanskrit accusative or dative[10]. This means that there was a stage in the development of Indo-Aryan where the old morphological system for object marking (and argument marking in general) became more and more dysfunctional. There are two possible ways out of such a state: a language can put up with the lost flagging by using word order instead to indicate roles, as it happened in the case of English. The Indo-Aryan languages never developed a rigid order but chose the other way out, that is, they created new cases.

According to Sinnemäki (forthcoming), differential object marking is typologically much more frequent than consistent object marking ("COM"). Sinnemäki therefore predicts that most languages with COM should change to DOM over the course of time and, more importantly, that a language without any kind of object marking is more likely to develop DOM than COM. If we apply this to the hypothetical stage in the history of Indo-Aryan where the old case system had become so eroded that it wasn't of much use any longer, it makes perfect sense that DOM should have been introduced first. The seed that explains the shared innovation of DOM in Indo-Aryan is then not some deep formal or functional property of the historical case system but the simple fact that this system was dysfunctional and that there was an urge to create a new one.

In the following sections, glosses have been adapted for all examples. In particular, the object marker is always glossed as DAT (except in Sinhala, where the object marker is ACC and DAT is a separate case). Where no glosses were given in the cited work I created glosses, placing question marks under forms that I was not able to parse. Transcriptions and translations are unchanged, but

[10] There are, however, a few languages that have reinterpreted other layer I markers as object markers and now use them for DOM, e.g. Romani (Matras 2002:86) and Kumauni (Apte and Pattanayak 1967:31).

transcriptions in indigenous scripts have been transliterated to IAST, and translations in languages other than English have been re-translated.

3.9.2 Panjabi

The information in this section is based on Bailey and Cummings (1912 [1994]), Tolstaya (1960), and Bhatia (1993). Panjabi has an object marker *nū̃/nüü*, which is employed in DOM:

(194) a. *uh ciṭṭhī paṛh-d-ā hâe*
3s letter(f) read-IPFV.PTCP-sm AUX.PRS.3s
'He reads a letter.'
b. *uh mâe nū̃ vekh-d-ā hâe*
3s 1s DAT see-IPFV.PTCP-sm AUX.PRS.3s
'He sees me.' (Tolstaya 1960:68)

The functional conditions for this marker are described in rather vague terms in all checked grammars but sound familiar enough. Bhatia is most explicit in claiming that all animate "direct objects" (both human and non-human) require the marker whereas it is optional on inanimates and "motivated by their definite reference" there. He gives the following examples to illustrate his claim:

(195) a. *aadmii nüü vekh-o*
man(m) DAT see-IMP.2p
'Look at a/the man.'
b. *kataab vekh-o*
book(f) look-IMP.2p
'Look at a book.'
c. *kataab nüü vekh-o*
book(f) DAT look-IMP.2p
'Look at the book.' (Bhatia 1993:88)

Bailey considers proper nouns as the only category where *nū̃* is obligatory and says that for all other nouns adding the marker "has the effect of making more definite or of indicating something already referred to or previously known" (Bailey and Cummings 1912 [1994]:343). Interestingly, the example he gives for a DAT-marked inanimate object involves an *as-for* topicalisation:

(196) a. *J̃hanḍū nū̃ kōḷ khal-hārkē ill mār-ī.*
Jhandu(m) DAT beside stand-? kite(f) kill-PRFV.PTCP.sf
'He put Jhandu standing beside him and killed a kite.'
b. *ill nū̃ te mār-chaḍḍ-eā*
kite(f) DAT TOP kill-COMPL-PRFV.PTCP.sm
'As for the kite he killed it.' (Bailey and Cummings 1912 [1994]:343)

Tolstaya does not have much to add; for her, specificity is the decisive factor for object marking, and animate nouns are "customarily" marked.

All three grammars confirm that the form in question also marks "indirect objects", that is, the typically human G of verbs such as 'give'. Bhatia spends some time on discussing this and also mentions that for most speakers T in combination with G-DAT is always NOM but some allow T-DAT G-DAT:

(197) *mãi mãã nüü kaake nüü ditt-aa*
1s mother(f) DAT child(?) DAT give-PRFV.PTCP.sm
'I gave the mother a/the child.' (Bhatia 1993:89)

He also shows that otherwise multiple *nüü* within one sentence are not a problem:

(198) *ó ne mãi nüü [raam nüü kataab de-N nüü] aakh-iaa.*
3s ERG 1s DAT Ram(m) DAT book(f) give-INF.OBL DAT say-PRFV.PTCP.sm

class	DIR.SG	OBL.SG	DIR.PL	OBL.PL
I	*-o*	*-a*	*-a*	*-a-n*
II	-X	-X	-X-*a*	-X-*a-n*
III	-X	-X	-X	-X-*ə-n*
IV	-u	-a	(only mass nouns, no PL)	

Table 3.26: Direct and oblique case in Kumauni

'He asked me to give a book to Ram.' (Bhatia 1993:174)

This is interesting because it shows that the dispreference of T-DAT G-DAT in Punjabi is not due to mere "dative jamming" but is a particularity of ditransitive verbs that might be connected to the hierarchical status of G.

Some other formal properties are mentioned by Bailey. He doesn't speak about double datives but mentions that double nominatives are occasionally possible. The examples he gives show the Panjabi equivalent to the equational ditransitive frame of Nepali:

(199) a. *tũ̃ ōn-nũ̃ hāfaj man-n-ā ẽ*
you 3s-DAT hafiz(m) consider-IPFV.PTCP-sm AUX.PRS.3s
'You believe him to be a Hafiz.'

b. *maĩ tē-r-ī beiztī ap-ṇ-ī beiztī samjh-n-ā wã̃*
1s 2s-GEN-sf dishonour(f) REFL-GEN-sf dishonour(f) consider-IPFV.PTCP-sm AUX.PRS.3s
'I consider your dishounour mine.' (Bailey and Cummings 1912 [1994]:343)

Further, he says that in combinations of nouns and verbs that contain a *figura etymologica* (e.g. *laṛāī laṛnī* 'fight a fight') or that "express a single idea" (e.g. *jhūṭh mārnā* 'tell a lie', probably complex predicates), the noun is never marked by DAT.

Information on the behaviour of DAT in the passive comes again from Bhatia. He explicitly denies the possibility of keeping DAT on the P of passivised predicates:

(200) *hakiim tõ mariiz (*nüü) vekh-iaa g-iaa.*
doctor(m) by patient(m) DAT see-PRFV.PTCP.sm go-PRFV.PTCP.sm
'The patient was examined by the doctor.' (Bhatia 1993:173)

Bhatia also mentions some other uses of *nũ̃/nüü*, among these marking experiencers as in (201). Possessors, however, are marked by the genitive, and destinations are marked by zero or layer III compound postpositions.

(201) *kuRii nüü gussaa a-iaa*
girl(f) DAT anger(m) come-PRFV.PTCP.sm
'The girl became angry.' (Bhatia 1993:87)

3.9.3 Kumauni

Kumauni is not a big language by Indo-Aryan standards but is of interest because it is the direct western neighbour of Nepali. Compared to most other Indo-Aryan languages, Kumauni is exotic because low objects are not zero-marked and because it makes use of inherited case suffixes. According to Apte and Pattanayak (1967), singular animate object nouns are marked by the suffix *-ac*, whereas plural and/or inanimate object nouns are in the bare oblique case. Although there are a couple of homophonies, the bare oblique case is formally clearly distinct from the direct case, which is used for subjects. Table 3.26 (adapted from Apte and Pattanayak 1967:32) shows the declension classes of Kumauni.

Here is a pair of examples illustrating the use of *-ac* (here in its allomorph *-j*):

(202) a. *u: cəll-a-j dekh-ən chə*
3 bird-OBL-DAT see-NPST.PTCP AUX.NPST.3sm
'He sees a bird.'
b. *u: cəll-a-n dekh-ən chə*
3 bird-OBL-PL see-NPST.PTCP AUX.NPST.3sm
'He sees the birds.' (Apte and Pattanayak 1967:33)

The singular/plural distinction holds even for pronouns:

(203) a. *u: twe-c chu:-n chə*
3 2s.OBL-DAT touch-NPST.PTCP AUX.NPST.3sm
'He touches you.'
b. *u: həmən chuma:n*
3 1p.OBL touch.NPST.3p
'They touch us.' (Apte and Pattanayak 1967:33)

The suffix *-ac* can also be used to mark animate G, as in (204a). However, there also is a dedicated dative suffix *-tɛ̃*, which seems to be more common than *-ac*, although it's not clear whether it is used with more verbs or with the same verbs as *-ac* but more often. This is another deviation from standard Indo-Aryan and is shown in (204b).

(204) a. *həm ghwaḍ-ac pani dinu*
1p horse-DAT water give.NPST.1p
'We give water to a horse.'
b. *wil bhalu-tɛ̃ kɔ:c ki...*
3s.ERG bear-DAT say.PST.3sm COMPL
'He said to the bear, (...).' (Apte and Pattanayak 1967:34)

3.9.4 Maithili

This section is based on the grammar by Yadav (1996). Maithili, too, has the familiar Indo-Aryan DOM pattern, where the object marker is *ke/kẽ*. Understanding Maithili examples is harder than for other Indo-Aryan languages because of its highly complex agreement system (Yadava 1996, Bickel 1999, Bickel and Yadava 2000). Verbal suffixes regularly index two arguments, and agreement is not tightly linked to roles but rather to case and social factors such as face and empathy (Bickel 1999): one set of agreement suffixes ("NOM-AGR") goes with NOM-marked S/A, the other ("non-NOM-AGR") with a variety of other referents that do not even have to be arguments. This includes various DAT-marked arguments such as O, experiencers, or G. Since DOM and DOI are independent phenomena in Maithili I well not dwell on the latter here.

Below is an example featuring A-NOM and P-DAT, both indexed on the verb. This example also illustrates one use of *-ke* according to Yadav (1996): it is obligatory with human proper names, kin terms, and personal pronouns with human reference.

(205) *həm jibəch kẽ dekh-əl-i-əinh*
1s Jibach DAT see-PST-1.NOM-3H.NNOM
'I saw Jibach.' (Yadav 1996:74)

With the remaining animate nouns, the use of *ke* is conditioned by definiteness:

(206) a. *əhã nokər tək-əit ch-i*
2H servant search-IPFV.PTCP AUX.NFUT-2H.NOM
'Are you looking for a servant?' (indefinite non-specific)

b. *əhã ek-ṭa nokər tək-əit ch-i*
2H one-CLF servant search-IPFV.PTCP AUX.NFUT-2H.NOM
'Are you looking for a servant?' (indefinite specific)

c. *əhã nokər ke tək-əit ch-i*
2H servant DAT search-IPFV.PTCP AUX.NFUT-2H.NOM
'Are you looking for the servant?' (definite) (Yadav 1996:79)

What "definiteness" means exactly remains unclear. Yadav mentions some examples where a high referent that should be definite under any possible definition is still not marked by *ke*, as in (207), but does not dwell on these:

(207) a. *u o-kər gai cərɔ-t-əik*
DIST DIST-GEN cow graze-FUT-3LH.NOM.3LH.NNOM
'He will graze his cow.' (Yadav 1996:77)

b. *tõ u admi dekh-l-əh-i(k)?*
2LH DIST man see-PST-2LH.NOM-3PROXLH.NNOM
'Did you see that man?' (Yadav 1996:78)

DAT marking is impossible with inanimates, no matter whether they are definite or not. Even when preceded by a demonstrative pronoun only NOM is grammatical:

(208) *u i gach kəṭ-l-ək*
DIST PROX tree cut-PST-3LH.NOM
'He felled the tree.' (Yadav 1996:78)

But just a little later Yadav himself says that any direct objects which are preceeded by the demonstratives *ehi* or *ohi* must be marked by DAT. He does not speak about the function of these demonstratives in contrast to *i* [PROX] and *o* [DIST], from which they seem to be derived. Davis (1984) calls these forms "emphatic", which suggests that they are similar in function to the cognate Nepali forms *ehi* [PROX.FOC] and *uhi* [DIST.FOC]:

(209) *həm ehi kitab kẽ pəṛh-l-əhũ*
1s PROX.FOC book DAT read-PST-1.NOM
'I read the book.' (Yadav 1996:73)

This fits with the fact that Yadav says that the examples for definite objects he cites may also bear "emphatic stress" – he does not make this any more explicit but says that, for instance, item (206c) could also be translated as "Are you looking for the *servant*?", where the italics seem to indicate contrastive focus. Yadav also cites two examples where according to him *ke* serves to make the identity of a referent clear. If one strips these examples of their illocutionary force, it becomes clear that once more object focus is involved:

(210) *kon nokər ke tək-əit ch-i?*
which servant DAT search-IPFV.PTCP AUX.NFUT-2H.NOM
'Which servant are you looking for?' (Yadav 1996:80)

Maithili is, apart from Nepali (cf. section 3.5.11), the only Indo-Aryan language in this sample for which focus is described to have an effect on DOM.

As for the formal properties of DOM, Yadav mentions that P-DAT are impossible in either of the two passives (p. 319) and that DAT-marked T are impossible in the presence of G-DAT (p. 81). (211) illustrates T-NOM G-DAT. Note that NNOM-AGR goes with G.

(211) *əhã jən kẽ jəlkhəi de-l-i-əik*
2H laborer DAT breakfast give-2H.NOM-3LH.NNOM
'You gave the laborer breakfast.' (Yadav 1996:82)

DAT is also used on experiencers, which once more trigger NNOM-AGR:

(212) *babu kẽ bokhar laig ge-l-əinh*
father DAT fever attach go-PST-3LH.NOM.3HH.NNOM
'Father caught fever.' (Yadav 1996:83)

3.9.5 Bhojpuri

Bhojpuri is another small language that is of interest because it is a direct neighbour of Nepali. I used two sources for Bhojpuri, Shukla (1981) and Verma (2003).

As in Maithili, the object marker is *ke* in Bhojpuri. According to Shukla (p. 98), this marker is optional with human objects but impossible with non-humans. Human objects marked with *ke* "carry some emphasis" (p. 98):

(213) a. *ham kita:bi paD^{h}-ab*
1s book read-FUT.1s
'I will read a book.'
b. *ham manai: (ke) de:k^{h}-ab*
1s man DAT see-FUT.1s
'I will see the man.' (Shukla 1981:97)

Verma is not any more precise about the conditions for *ke*. He sees the border for its use between animates and inanimates and says that NOM is possible with the former when they have "generic"

reference:

(214) a. *ham torā laṛki ke ṭhik se dekh-ab*
1s 2s.GEN girl DAT well ADV see-FUT.1s
'I will look after your daughter well.'
b. *ham laṛki dekh-e jā-t bān-i*
1s girl see-INF go-PRS.PTCP AUX.PRS-1
'I am going looking for a girl (for marriage).' (Verma 2003:533)

"Indirect objects" always get *ke* (Shukla 1981:98, Verma 2003:534), and "direct objects" that co-occur with them never do:

(215) *ham laika:-ke kita:bi de:-b*
1s boy-DAT book give-FUT.1s
'I will give the boy the book.' (Shukla 1981:98)

The frame T-NOM G-DAT is also fixed with "object complements", i.e. with the frame that has been called equational ditransitive here (cf. section 3.3.2.7):

(216) *tu i phul ke gulāb bujh-al-a*
2s PROX flower DAT rose think-PST-2m
'You thought this flower a rose.' (Verma 2003:534)

3.9.6 Hindi-Urdu

Hindi-Urdu is probably the Indo-Aryan language with most publications, among them also several works dedicated to syntax or even case. The most important work for our purposes is Mohanan (1994), on which most of the discussion below will be based. With the exception of the first example in (217), all data is drawn from publications on Hindi, so I will speak of Hindi (instead of Hindi-Urdu) for the sake of simplicity.

The DAT marker in Hindi is *ko*. (217) shows an example of how it is involved in differential object marking:

(217) a. *pānī kā nal xarāb hai, fauran nalvalē kō bulā-ō.*
water GEN.sm pipe broken COP.PRS.3s at.once plumber.OBL DAT call-IMP.2MH
'The water pipe is broken; call the plumber.' (Schmidt 1999:70)
b. *pānī kā nal xarāb hai, fauran nalvalā bulā-ō!*
water GEN.sm pipe broken COP.PRS.3s at.once plumber call-IMP.2MH
'The water pipe is broken; call a plumber at once!' (Schmidt 1999:71)

As the example shows, definiteness is once more involved. The other important factor for Hindi is animacy. Mohanan treats these two factors in a somewhat confusing way by making statements about their influence on case sometimes from one perspective ("inanimate nouns can only be DAT-marked when they are definite") and sometimes from the other ("NOM-marked nouns which are human must have generic reference"). Still, if one takes together all statements and examples from various pages, a very detailed picture of the function of DOM emerges. This information is represented in a more systematic way in Table 3.27. "Animate" refers to non-human animates and "specific" to non-definite specific referents.

	human	**animate**	**inanimate**
definite	DAT	DAT	DAT/NOM
specific	DAT	DAT	NOM
non-specific	DAT	NOM	NOM
incorporated[11]	NOM	NOM	NOM

Table 3.27: Factors behind DOM in Hindi

The table shows nicely that both variables have a strong influence on case assignment. Further, both influences seem to work together. Added low values (e.g. non-specific inanimate) or high values (e.g. animate specific) yield the corresponding cases (low NOM, high DAT), but a very high value (human/definite) has enough weight to be combinable with a low value (e.g. human+non-specific, inanimate+definite) and still yield a high case (DAT).

This relation can be very simply modelled by assigning values from 0 to 3 to the steps on the animacy scale and values from 0 to 4 to definiteness. DAT can then be said to be found in all cells where the values pointing to it add up to a value higher than 2. The only exception is the definite+inanimate cell, where NOM is also possible besides DAT.

An interesting functional twist to DOM in Hindi is that semantic verb classes play a role. According to Mohanan, verbs whose semantics only allow inanimate objects always mark them by NOM, whereas verbs whose semantics only allow animate objects require DAT. For instance, *likh-* 'write' only combines with inanimate objects, so DAT is impossible in (218) in spite of the definiteness of *khat* 'letter':

(218) *ilaa-ne yah kʰat(-*ko) likʰ-aa.*
Ila(f)-ERG PROX letter(m)-DAT write-PRFV.PTCP.sm
'Ila wrote this letter.' (Mohanan 1994:81)

Note, however, that this association does not hold in a hundred percent of cases (Aissen 2003:449), so it might be a tendency rather than a rule.

Statements about the function of Hindi DOM in other treatments are not near as detailed. It is still worth mentioning them in order to show that while they are all very similar on the one hand, there are also a lot of subtle differences on the other.

- Masica (1982) says that the most important factor for DOM is animacy. *ko* with inanimates marks specificity or definiteness, whereas NOM with animates is possible when they are indefinite or when "there is a desire to depersonalize them" (Masica 1982:17).

- Greaves (1921 [1983]) also notes the connection between DOM and animacy and adds a couple of details. According to him, *ko* is "generally" used with pronouns and "widely" with proper names. It also may indicate "stress and emphasis" – this again points to the direction of focus, although Greaves doesn't give examples to substantiate his claims. Greaves also has an equivalent to Mohanan's "incorporation" – he says that NOM is usual "where the connection between the verb and the noun is very close".

- In addition to animacy and definiteness/specificity, Jain (1995) also says that *ko* on animates may "emphasise" an object.

- Kachru (2006) requires *ko* for uniquely referring elements such as proper nouns and pronouns. Animate objects have *ko* optionally, and inanimate objects can only have it when they are definite.

- Montaut (2004) says that *ko* is used on human or specific inanimate patients. However, she makes clear that even definite inanimate patients will only be marked when both SAP have them on their minds and that even human patients may remain unmarked when they refer to a role or a function in a specific context rather than to an individuum.

- Sandahl (2000) in general does not go into details but claims that objects marked by *ko* are "particularized".

- Schmidt (1999) views animacy as the primary factor but admits that both NOM on animates and DAT on inanimates are possible in order to specify or despecify them, respectively.

[11] Mohanan's notion of incorporation is a mixed formal-functional concept that resembles but is not identical to incorporation in a more strict sense. See section 3.4.5 for some further discussion.

As for the formal properties of DOM, Mohanan claims that direct objects in the presence of DAT-marked indirect object are always in the nominative:

(219) *ilaa-ne mãã-ko baccaa(*-ko) diy-aa*
Ila(f)-ERG mother-DAT child-DAT give-PRFV.PTCP.sm
'Ila gave a/the child to the mother.' (Mohanan 1994:85)

Her analysis of this is that case assignment for the indirect object takes precedence over that for the direct object: the two objects compete for one available DAT, and the indirect object wins. Greaves (1921 [1983]) confirms this, although his interpretation is not that only one DAT is available but that several DAT within one sentence are to be avoided. He adds that T-NOM G-DAT is also fixed for verbs with "double accusative", by which he means verb senses employing a frame corresponding to the Nepali equational ditransitive frame (section 3.3.2.7). This connection is also noted by Sandahl (2000) and Kachru (2006), who speaks of the "double transitive construction" and gives the example in (220):

(220) *mɛ̃ is məkan ko səsta səməj^{h}-t-a h-ɦ̃ũ.*
1s PROX house(m) DAT cheap consider-IPFV.PTCP-sm AUX.PRS-1s
'I consider this house inexpensive.' (Kachru 2006:175)

In the passive, both O-NOM and O-DAT are possible, although O-DAT is not accepted by all speakers. Mohanan does not talk about agreement and her examples are ambiguous, but Sandahl (2000) confirms that agreement behaves as expected, i.e. it goes with O-NOM and is set to a dummy 3sm with O-DAT:

(221) a. *Ciṭṭhī ḍāk se bhejī th-ī.*
letter(f) mail(f) INST send-PRFV.PTCP.sf AUX.PST-sf
'The letter had been sent by mail.' (Sandahl 2000:102)
b. *J̌īp meṃ śrīmatī Gandhī ko biṭhā-yā ga-yā.*
Jeep(f) LOC wife(f) Gandhi(m) DAT seat-PRFV.PTCP.sm go-PRFV.PTCP.sm
'Mrs. Gāndhī was seated in the jeep.' (Sandahl 2000:103)

A formal peculiarity of Hindi DOM that has not been reported from other Indo-Aryan languages is that two conjoined NPs must have the same case. If the first NP has DAT, as in example (222), the second NP must have DAT, too, irrespective of its animacy:

(222) *raam-ne bacce-ko aur us-k-e juute*(-ko) uṭ^{h}aa-yaa.*
Ram(m)-ERG child(m)-DAT and 3s-GEN-sm.OBL shoe-DAT pick.up-PRFV.PTCP.sm
'Ram picked up the child and its shoes.' (Mohanan 1994:90)

As in the other Indo-Aryan languages, Hindi *ko* is polyfunctional. It has already become clear above that it is used to mark certain "indirect objects" (= animate G aka recipients). Interestingly, *ko* is also found on some inanimate G, which for instance in Nepali would be marked by NOM or LOC (Sandahl 2000:29, Schmidt 1999:72):

(223) *Maiṃ bāzār ko jā rah-ā h-ūṃ.*
1s market(m) DAT go PROG-sm AUX.PRS-1s
'I am going to the market.' (Sandahl 2000:29)

Marking certain experiencer S (224a) and A (224b) is another function of *ko*. As in Nepali, the P of transitive experiencer verbs which mark A by DAT must always be marked by NOM and agrees with the verb (Mohanan 1994:97):

(224) a. *tuṣaar-ko k^{h}ušii hu-ii.*
Tushar(m)-DAT happiness(f) happen-PRFV.PTCP.sf
'Tushar became happy.'

b. *tuṣaar-ko vah kahaanii yaad aa-yii.*
Tushar(m)-DAT DIST story(f) memory(f) come-PRFV.PTCP.sf
'Tushar remembered that story.' (Mohanan 1994:141)

DAT is further found on some time adverbs and on S/A in deontic sentences (Sandahl 2000:28) as well as after the infinitive to mark "impending events" (Schmidt 1999:140).

3.9.7 Bengali

For Bengali there is a dedicated work on case in Mukherjee (1985). Other works will only be cited below where they have something to add to his analyses. (225) shows an example for DOM. The object marker in Bengali has the form *-ke*.

(225) a. *ami ḍakṭar ḍak-b-o*
1s doctor call-FUT-1
'I will call (any) doctor.'
b. *ami ḍakṭar-ṭa-ke ḍak-b-o*
1s doctor-DEF-DAT call-FUT-1
'I will call the doctor.' (Mukherjee 1985:19)

(226b) shows a peculiarity of Bengali: there is a suffixed definite article *-ṭa*. Although the article frequently co-occurs with DAT, both can appear independently on O NPs (cf. examples (228a), (229a) below). This is remarkable because Mukherjee analyses definiteness as the most important factor in DOM.

However, not all grammarians are of this opinion. Ray et al. (1966:35) speaks of "particular" referents being marked by DAT, but Bykova (1981:57) and Smith (1997:38) take animacy as primary. Bykova (p. 57) admits that definiteness may also play a role but emphasises that this includes "the concept of totality and collectivity". "Collective, generalised notions" may be marked by NOM even when the noun is animate. This becomes best visible in the following minimal pair (cited by Bykova 1981:57 from Caṭṭopādhyāya 1966:242):

(226) a. *rakhal goru cɔra-y*
cowherd ox graze-PRS.3s
'The cowherd tends cows.'
b. *goru-ṭa-ke gohaler bhitore loiya aiʃ-o*
ox-DEF-DAT cowshed into drive ?-IMP.2MH
'Drive the cow into the shed!'

Smith adds another secondary factor, namely "emphasis", but his examples do not make clear what he means by this – it is definitely not contrastive focus. Mukherjee makes the similarly vague claim that the theme/rheme contrast may sometimes play a role so that thematic referents may be marked by DAT:

(227) *ṭaka-ṭa-ke har-ie-ch-o tumi?*
money-DEF-DAT lose-PRFV-PRS-2MH 2MH
'Was it you who lost the money?' (Mukherjee 1985:21)

Another final factor is what Mukherjee calls the "concreteness" of events. When looking at his examples, however, it rather looks as if he meant the relation between the event and the existence of the object referent. Objects which come into being through an event are less likely to be marked than objects which exist prior to it and are changed or destroyed by it. This resembles the role of affectedness in Nepali (section 3.5.13).

(228) a. *ami baṛi-ṭa ban-ie-c̱h-i*
1s house-DEF build-PRFV-PRS-1
'I built the house.'

b. *ami janla-ṭa-ke bhen-e-cẖ-i*
1s window-DEF-DAT break-PRFV-PRS-1
'I broke the window.' (Mukherjee 1985:21)

-ke also marks "indirect objects". With T which are high on both the animacy and the definiteness scale, DAT doubling is possible:

(229) a. *ca-ṭa de-b-o ami toma-ke*
tea-DEF give-FUT-1 2s-DAT
'I will give you the tea.'
b. *o toma-ke ama-ke bech-b-e*
3s 2s-DAT 1s-DAT sell-FUT-3s
'He will sell you to me.' or 'He will sell me to you.' (Mukherjee 1985:19)

Interestingly, there seem to be no special restrictions on T when G is not expressed overtly. This is again reminiscent of Nepali (cf. section 3.4.4):

(230) *ca-ṭa-ke de-b-o ami*
tea-DEF-DAT give-FUT-1 1s
'I will give the tea (away).' (Mukherjee 1985:19)

Passive O can be marked by *-ke* (231a) or not (231b):

(231) a. *toma-ke kẖun kôr-a ho-e-ch-e*
2s-DAT murder do-VN AUX-PRFV-PRS-3s
'You've been murdered.' (Mukherjee 1985:57)
b. *sap-ṭa-dara ami doŋs-ito ho-e-cẖ-i*
snake-DEF-by 1s bite-PASS.PTCP AUX-PRFV-PRS-1
'I've been bitten by the snake.' (Mukherjee 1985:67)

According to Smith (1997:39), the equivalents of the equational ditransitive frame of Nepali (e.g. 'call', 'consider as', 'know as') once more have the invariable frame T-NOM G-DAT.

Apart from marking O and some G, *ke* does not seem to have any other functions. This makes *-ke* one of the most specialised object markers of Indo-Aryan. Experiencers and possessors, which are otherwise frequently marked by DAT, are marked by the genitive *-r* in Bengali (Smith 1997:141).

A final point of interest concerns plural marking. Bengali has two plural markers, which are sensitive to animacy. One (*-gulo*) is used with low referents, the other (*-ra*, OBL *-der*) with high ones (Mukherjee 1985:4). Mukherjee claims that animacy is a relatively flexible concept in Bengali, so the use of *-gulo* and *-ra* is not lexically fixed. As a consequence, the marking of plural O is highly complex: they may have the article or not depending on definiteness, they may take *-gulo* or *-ra* depending on animacy, and they may be marked by zero or *-ke* depending on animacy *and* definiteness (and probably additional factors). Instead of the combination *-der-ke* [-PL.OBL-DAT], *-der* alone is more usual in the modern language (p. 17), but that doesn't touch the functional contrast.

3.9.8 Gujarati

Resources on Gujarati are scarce. Most information on DOM is found in Taylor (1908) and Mistry (1997).

The DAT marker of Gujarati is *-ne*:

(232) a. *Te potā-n-o pāṭh vāṃc-e ch-e.*
3s REFL-GEN-sm lesson(m) read-PRS.3s AUX.PRS-3s
'He reads his lesson.' (Gupta 1976:86)
b. *Te copaḍī ahīṃ lāv-o.*
3s book(f) PROX.LOC bring-IMP.2p
'Bring that book here.' (Gupta 1976:87)

Gupta (1976:87) says that *-ne* is used with definite and/or human referents.

A different, peculiar theory is put forward by Taylor (1908:132), who claims that NOM on objects expresses a particularly close connection between the noun and the verb. O-NOM are part of the "subject-matter" of the verb, whereas O-DAT are its "goal". This is somewhat reminiscent of Mohanan's (1994) concept of functional incorporation, but Taylor's examples ('I acknowledge my transgression' with O-NOM, 'I acknowledge (believe in) God' with O-DAT) rather point to the familiar variable of animacy. He admits himself on the same page that inanimate referents are mostly marked by NOM and animate referents mostly by DAT. Personal pronouns are, according to him, always marked by DAT when in object position.

Mistry (1997) criticises theories that try to explain DOM in Gujarati by a single functional factor. He proposes that two separate morphemes with the shape *-ne* should be assumed. One is an object marker that is lexicalised with a couple of verbs such as *karaḍ* 'bite', *maḷ* 'meet', or *vaḍh* 'rebuke'. Although Mistry does not notice, the verbs on his list all seem to require animate objects. Gujarati may thus exhibit a similar pattern as Hindi, for which it was claimed by Mohanan (1994) that verbs requiring an animate object require DAT, whereas verbs requiring an inanimate object require NOM.

The other, homophonous postposition *-ne* is analysed by Mistry as a marker of specificity. However, this interpretation seems to be short-sighted, since both *-ne* are clearly restricted to objects. A more standard interpretation would be to say that there is one object marker that is obligatory on certain verbs and that marks specificity on others, as exemplified in (233):

(233) a. *Principal caar šikšak-o nim-š-e.*
principal four teacher-PL appoint-FUT-3s
'The principal will appoint (any) four teachers.'
b. *Principal caar šikšak-o-ne nim-š-e.*
principal four teacher-PL-DAT appoint-FUT-3s
'The principal will select four (specific) teachers.' (Mistry 1997:433)

"Indirect objects" are marked by DAT, too:

(234) *mita-e lina-ne cɔpḍi api*
Mita-ERG Lina-DAT book(f) give.PST.PTCP.f
'Mita gave the book to Lina.' (Doctor 2004:76)

This is confirmed by Taylor (1908:130), who also adds (p. 132) that verbs equating T and G have the frame T-NOM G-DAT. Whether T-NOM G-DAT is fixed on verbs with "indirect objects", too, or whether T-DAT is occasionally possible there does not become clear from the checked grammars.

Besides marking O and recipients, *-ne* can also mark experiencers and deontic S/A (Gupta 1976:87). Below is an example for an DAT-marked experiencer.

(235) *Mā-r-ī vāt te-ne sarī na lāg-ī.*
1s-GEN-f talk(f) 3s-DAT satisfaction(f) NEG be.at-PST.PTCP.f
'My talk did not satisfy him.' (Gupta 1976:87)

The most remarkable feature of Gujarati on the formal side is that DAT-marked objects can control agreement. Mistry (1997:430) says that this is only possible with the *-ne* that marks specificity. If we again assume that this *-ne* is identical to the object marker *-ne*, we can instead say that only specific objects may get AGR (whereas objects that are marked by DAT because their predicate requires it may not). This means that Gujarati has two independent mechanisms, DOM and DOI. (236) shows an example for this phenomenon, which is highly unusual for Indo-Aryan (Deo and Sharma 2006:10):

(236) *Ugravaadi-o-e police-ni car-ne atkaav-i.*
militant-PL-ERG police-GEN car(f)-DAT stop-PST.PTCP.f
'The militants stopped the police car.' (Mistry 1997:436)

3.9.9 Oriya

There are two good grammars of Oriya, which are the base of the data in this section, Neukom and Patnaik (2003) and Mahapatra (2007).

The Oriya DAT marker is *-ku*. This suffix interacts with the suffix *-nkɔ*, which is used to mark honorificity and in the oblique plural of human nouns. Both functions frequently go together with object marking, and *-nkɔ-ku* is contracted to *-n-ku* (Neukom and Patnaik 2003:47).

(237) a. *mũ corɔ dhɔr-i-ch-i*
1s thief catch-PRFV-PRS-1s
'I have caught (some) thief.' (Mahapatra 2007:124)

b. *mũ corɔ-ku dhɔr-i-ch-i*
1s thief-DAT catch-PRFV-PRS.PRF-1s
'I have caught (the one, who is) the thief.' (Mahapatra 2007:124)

The most important factors behind DOM are once more animacy and definiteness. While Neukom (p. 51) formulates relatively concrete rules (DAT is obligatory with definite animates and impossible with indefinite inanimates), Mahapatra is more skeptical about rules (p. 123: "It has not been possible to frame hard rules to predict their [i.e. the cases', author's note] occurrence"). Neukom (p. 51) says that in the fuzzy cases (indefinite animates, definite inanimates), the use of DAT increases the specificity of the referent.

Like Bengali, Oriya has developed a definite article *-ṭa/-ṭi*. Neukom (p. 25) claims that the function of the article is to "ascribe communicative relevance for the discourse (or specificity)" to the marked noun. Both forms can occur in isolation. (237b) shows *-ku* without *-ṭa*, (238) the reversed case:

(238) *au thɔre cauḷɔ-ṭa dhu-ɔ*
once more rice-DEF wash-IMP.2p
'Wash the rice once more.' (Mahapatra 2007:122)

Some of Mahapatra's examples suggest that specificity is more important than definiteness for DOM, although he doesn't mention this himself:

(239) a. *Raja ghoṛa khoj-u-ch-oɔnti*
king horse search-IPFV-PRS-3p
'The king is looking for a (any) horse.'

b. *Raja ghoṛa-ku khoj-u-choɔnti*
king horse-DAT search-IPFV-PRS-3p
'The king is looking for a (particular) horse.' (Mahapatra 2007:123)

In another minimal pair for DOM, the addition of *-ku* has quite a remarkable effect:

(240) a. *mũ cauḷɔ dho-u-ch-i*
1s rice wash-IPFV-PRS-1s
'I am washing rice (normal).'

b. *mũ cauḷɔ-ku dho-u-ch-i*
1s rice-DAT wash-IPFV-PRS-1s
'I am washing rice (with some purpose/emphasis).'

Mahapatra does, however, not explain this effect in more detail but contents himself with stating that the distribution of the cases is "unstable".

-ku also marks recipients. DAT doubling is possible in Oriya:

(241) *mũ purbɔ dinɔ jẽũ jama-ṭi kiṇ-i-thil-i sei-ṭi-ku mo bhɔuṇi-ku*
1s before day which frock-DEF buy-PRFV-PST-1s that-DEF-DAT 1sPOR sister-DAT
de-l-i
give-PST-1s
'I gave my sister the frock which I had bought the day before.' (Neukom and Patnaik 2003:52)

DAT can also be retained in passives. As usual, AGR goes with O-NOM but O-DAT triggers dummy agreement:

(242) a. *pila-mane se lokɔ dwara khoj-a-gɔl-e*
child-PL DIST man by search-PASS-go.PST-3p
'The children were looked for by that man.'
b. *pila-manɔ-n-ku se lokɔ dwara khoj-a-gɔl-a*
child-PL-PL.OBL-DAT DIST man by search-PASS-go.PST-3s
'The children were looked for by that man.' (Neukom and Patnaik 2003:289)

Besides O and recipients, *-ku* also marks experiencers. In addition, Oriya is one of the few Indo-Aryan languages which also use DAT for inanimate G:

(243) aji mũ gã:-ku ja-u-ch-i
today 1s village-DAT go-IPFV-PRS-1s
'Today, I am going to the village.' (Mahapatra 2007:124)

In this function it alternates with NOM:

(244) mũ puri ja-u-ch-i
1s Puri go-IPFV-PRS-1s
'I am going to Puri.' (Mahapatra 2007:124)

More marginal functions mentioned by both Neukom and Mahapatra are the marking of propertions ('increase by twenty'), rates ('per day'), temporal locations ('at night') and cirumstances ('by chance').

3.9.10 Marathi

The sources used for Marathi are Pandharipande (1997) and Dhongde and Wali (2009).

The DAT marker of Marathi is *-la*. This puts Marathi closer to Nepali in this respect than all other Indo-Aryan languages, and as we have seen in section 3.7.2 it has been hypothesised that *-la* and *-lai* share the same origin. Below is a series of examples for the use of NOM and DAT on P.

(245) a. *mī dzhāḍ pāhi-l-a*
1s tree(n) see-PRFV-3sn
'I saw a tree.'
b. *mī mulī-lā pāhi-l-a*
1s girl(f)-DAT see-PRFV-3sn
'I saw a/the girl.'
c. *mī dzhāḍā-lā pāhi-l-a*
1s tree(n)-DAT see-PRFV-3sn
'I saw the tree.' (Pandharipande 1997:134)

These examples illustrate Pandharipande's claims about the function of DOM: animate nouns in object position must always be marked by *-la* (although NOM is optionally possible for some speakers), but inanimates can (and must) only be marked by *-la* when they are definite. Dhongde and Wali are less strict and only say that animates and inanimates are generally associated with DAT and NOM, respectively. They also give an example for a NOM-marked human object:

Verb class also seems to play a role for DOM in Marathi in that a couple of verbs require O-DAT. However, differently from Hindi and Gujarati, semantic generalisations over this verb class are not possible. Pandharipande (1997:289) mentions *sparśa karṇe* 'touch' and *bolawne* 'call' as examples.

"Indirect objects" are also marked by *-la*. T with G of this type usually get NOM, but DAT doubling is also possible:

(246) *ai-ni babu-la nat(i-la) dakhəw-l-i.*
mother(f)-ERG Babu(m)-DAT grand.daughter(f)-DAT show-PRFV-sf
'Mother showed her grand-daughter to Babu.' (Dhongde and Wali 2009:192)

O-DAT can be retained in passives but does not trigger agreement, in contrast to passivised O-NOM:

(247) a. *polisā-kaḍūn tsorø pakḍ-l-e ge-l-e*
policeman(m)-by thieve(m) catch-PRFV-pm go-PRFV-pm
'The thieves were caught by the policeman.'
b. *polisā-kaḍūn tsorə̃n-nā pakaḍ-l-a ge-l-a*
policeman(m)-by thieves(m)-DAT catch-PRFV-sm go-PRFV-sm
'The thieves were caught by the policeman.' (Pandharipande 1997:289)

-la also marks a range of other functions, which Pandharipande (p. 292) summarises as "purpose, goal, possession, location, etc." She also mentions that *-la* can mark "dative subjects", i.e. S/A experiencers. (248) shows an example for this:

(248) *Ti-lā rāg ā-l-ā.*
3sf-DAT anger come-PRFV-sm
'She got angry.' (Pandharipande 1990:161)

DAT in Marathi can also mark possessors, where a special split is found. According to Pandharipande, GEN is used with alienable and DAT with inalienable possessums:

(249) *Ma-lā tīn bhāū āhet.*
1s-DAT three brother be.3p
'I have three brothers.' (Pandharipande 1997:230), inalienable be

3.9.11 Sinhala

The main resource I used for Sinhala is Chandralal (2010).

Sinhala is an Indo-Aryan island in the Southern part of South Asia, which is otherwise dominated by Dravidian languages. This may explain why Sinhala has developed two non-zero object cases, which are commonly called accusative and dative – this is, according to Masica (1982:26), a typically Dravidian feature. In spite of the separation of these two cases, Sinhala still exhibits an alternation between the accusative and zero:

(250) *balla nayaa(-wə) hæp-u-wa*
dog cobra-ACC bite-PST-IND
'The dog bit the cobra.' (Chandralal 2010:127)

According to Chandralal (p. 81), *-wə* [ACC] is used on animate nouns that are found in "an unaccustomed role, i.e. as Undergoer". This does not explain, however, why not all animate undergoers must be marked by *-wə* – cf. the example above. Gair and Paolillo (1997:31) add to this that ACC is optional on all animate nouns and pronouns and impossible on inanimate nouns and pronouns but also do not make clear what exactly conditions the presence of DAT on animate NPs.

-wə is also used for disambiguation. When an object is put in a position that digresses from the default word order AOV, its syntactic status can be clarified by marking it (Chandralal 2010:127). In (251), *nayaa* must be marked in order to achieve the given meaning – otherwise the scenario gets reversed and it is the dog that is bitten:

(251) *nayaa-wə balla hæp-u-wa*
cobra-ACC dog bite-PST-IND
'The dog bit the cobra.' (Chandralal 2010:127)

Many verbs require the dative marker *-ʈa* instead of *-wə* on their P, for instance, *wandinəwa* 'worship', *baninəwa* 'scold, blame', and *gahanəwa* 'hit, beat'. All verbs of this type that are listed by Chandralal (2010:128) require an animate P, which is once more reminiscent of Mohanan's (1994) claim that in Hindi the P of verbs which require animate P must be marked by DAT. The difference is that in Sinhala the case used in DOM (*-wə* [ACC]) is distinct from the one required by verbs with inherently animate P (*-ʈa* [DAT]). DAT never seems to enter alternations with NOM and ACC.

The NOM/ACC alternation is also found in two unusual places. First, animate S of some non-volitional predicates are optionally marked by *-wə*:

(252) *lamea(-wə) wæʈe-nə-wa*
child-ACC fall-NPST-IND
'The child is falling.' (Chandralal 2010:102)

Second, where other Indo-Aryan languages would try to avoid double datives, Sinhala uses NOM/ACC on one argument and DAT on the other. For instance, with ditransitive verbs of the type 'give', T is marked by NOM/ACC and G by DAT:

(253) *Ranjit Chitra-ʈə leensu-ak de-nə-wa*
Ranjit Chitra-DAT handkerchief-IDF give-NPST-IND
'Ranjit gives Chitra a handkerchief.' (Chandralal 2010:113)

Similarly, NOM/ACC is possible on the P of an A-DAT. A-DAT are found, for instance, with some non-volitional transitive predicates:

(254) *Ranjit-ʈə puusa(-wə) pæænge-nə-wa*
Ranjit-DAT cat-ACC step.on.NVOL-NPST-IND
'Ranjit is accidentally stepping on the cat.' (Chandralal 2010:106)

Sinhala does not have a true passive (Chandralal 2010:154), so the question how O behaves there does not arise.

The Sinhalese ACC is not as polyfunctional as the object markers in other Indo-Aryan languages. For instance, experiencer S, possessors, and deontic S/A are all marked by DAT, not by ACC:

(255) *ma-ʈə unə*
1s-DAT fever
'I have a fever.' (Chandralal 2010:104)

(256) *Chitra-ʈə kaarek-ak tie-nə-wa*
Chitra-DAT car-IDF be-NPST-IND
'Chitra has a car.' (Chandralal 2010:106)

(257) *ma-ʈa heʈə Kolamba ya-nnə tie-nə-wa*
1s-DAT tomorrow Colombo go-INF be-NPST-IND
'I have to go to Colombo tomorrow.' (Chandralal 2010:139)

Beside these, the dative also takes over a couple of less usual functions. For instance, it can mark "external causes" (terminology by Chandralal):

(258) *huləngə-ʈə gas perəle-nə-wa*
wind-DAT tree fall.down-NPST-IND
'The trees are falling from the wind.' (Chandralal 2010:105)

(259) *kaɖuə-ʈə atə kæpe-nə-wa*
sword-DAT hand cut.NVOL-NPST-IND
'The sword is cutting his hand.' (Chandralal 2010:105)

DAT also regularly marks inanimate destinations, be they P as in (260) or G as in (261):

(260) *Ranjit pansələ-ʧə ya-nə-wa*
Ranjit temple-DAT go-NPST-IND
'Ranjit is going to the temple.' (Chandralal 2010:111)

(261) *taatta salli laachchuə-ʧə daa-nə-wa*
father money drawer-DAT put-NPST-IND
'Father puts money into the drawer.' (Chandralal 2010:114)

3.9.12 Summary

Table 3.28 summarises some properties of DOM in the languages discussed in this section. Where nothing is known about a feature a question mark is given. The keys for the abbreviations in the row "other factors" are as follows: **COLL** collective nouns, **DIS** disambiguation, **EMPH** emphasis, **FOC** focus, **INC** (semantic) incorporation, **KIN** kinship terms, **NUM** number, **PRP** proper nouns, **VSEM** verb semantics.

The discussion in this section has shown that DOM is extremely widespread in Indo-Aryan – in fact, from the present sample it looks as if it was present in all languages. The core feature of Indo-Aryan DOM is that the same form that can optionally mark O is also used on animate G. The only exception to this is Sinhala, which has developed separate ACC and DAT cases due to Dravidian influence. Several other features are also widespread: dative doubling is generally dispreferred or even banned, O-DAT is often possible in the passive, and O-DAT usually cannot trigger agreement, two notable exceptions being Maithili and Gujarati.

Animacy is described as relevant in all languages, and identifiability is in almost all. Other functional factors are less frequently recurring. However, the unsystematic character of most descriptions and the fact that almost every factor that is relevant in some language was found to be relevant in Nepali in the present work suggest that the range of factors may be more homogeneous than it looks at first sight, with differences rather to be found in how important the individual factors are and how they work together.

Two big descriptive flaws in almost all treatments of DOM in Indo-Aryan are that they do not admit that they are incomplete and do not refer to each other. A particularly impressive case is Hindi-Urdu, where a lot of grammars have made similar yet slightly different proposals with respect to the distribution and function of DOM. Obviously not all of these can be true at the same time, so much would have been gained if some treatments would have confessed that they were only adding hypotheses or if authors would have looked at existing hypotheses first.

	Panjabi	Kumauni	Nepali	Maithili	Bhojpuri	Hindi	Bengali	Gujarati	Oriya	Marathi	Sinhala
O marker	*nữ*	*ac*/OBL	*lai*	*ke*	*ke*	*ko*	*ke*	*ne*	*ku*	*la*	*wə*
doubling	+	?	+	-	-	-	+	-?	?	+	n.a.
marked O_{PASS}	-	?	+	-	?	+	+	?	+	+	n.a.
hum+pron	+	±	+	+	?	+	+	+	+	?	±
hum+def	+	±	+	+	±	+	+	?	±	+	+
hum+idf	+	±	±	-	±	±	-	?	-	+	+
inan+def	+	-	±	-	-	±	+	?	±	+	-?
inan+idf	-	-	-	-	-	-	-	?	-?	-	-?
other factors	PRP	NUM	(see section 3.5)	FOC, KIN, PRP	EMPH,	INC, PRP, VSEM	COLL, FOC, KIN	VSEM	-	VSEM	DIS
animate G	+	±	+	+	+	+	+	+	+	+	-
inanimate G	-	?	-	-	?	+	?	?	+	?	-
equational G	+	?	+	?	+	+	?	+	?	?	?
experiencers	+	?	+	+	?	+	-	+	+	+	-
possessors	-	?	-	-	?	-	-	-	-	+	-

Table 3.28: DOM in some Indo-Aryan languages

Chapter 4

Conclusions

4.1 Chintang vs Nepali

4.1.1 Commonalities

Both S/A detransitivisation in Chintang and differential object marking in Nepali are conditioned by properties of objects. This most basic commonality motivated the title of the present work. In both cases the relevant properties are related to referent accessibility in a wide sense – objects get O-AGR in Chintang when they are specific (identifiable for the speaker) and DAT in Nepali when they have features that are unusual for objects (among them several that are connected to topicality or topicworthiness). This also explains the special status of pronouns and demonstratives in both systems.

However, note that accessibility is frequently involved in various other kinds of object-conditioned patterns, too – for instance, it has been described to be relevant for pure DOI, antipassives, and noun incorporation. The similarity of Chintang and Nepali in this respect is thus nothing especially noteworthy.

The grammatical relation of object as defined by the phenomena in question themselves is another similar point: it encompasses the roles P, T, and G. In addition, O is tied to a differentially marked A to the effect that object-conditioned differential marking is always and only possible in verb classes that also allow A-ERG. The further details vary between the two languages.

Another interesting parallel between Chintang and Nepali is that an iconic relationship holds between the formal marking of O and its functional properties. In both S/A detransitivisation and DOM, the O lacking formal marking (i.e. the O without AGR or marked by NOM, respectively) is also less graspable functionally – it is unidentifiable, unimportant, or uninteresting. By contrast, the marked O (i.e. the O with AGR or marked by DAT) stands out formally and also deserves more attention on the functional side.

Apart from such rather basic commonalities, however, S/A detransitivisation and DOM are quite different. The differences will be summarised in the next section.

4.1.2 Differences

The most obvious differences between S/A detransitivisation and DOM are found on the formal surface. DOM is characterised by a single locus of marking, whereas S/A detransitivisation becomes visible in several places (thereby fulfilling the conditions for what has been called "differential framing" here). In DOM, the locus of marking is identical to the locus of conditions, the object. S/A detransitivisation is complementary to this: the object is the only core constituent which does not bear a marker and whose marking does not change between the frame, even though it is once more the locus of conditions. Instead, S/A detransitivisation affects A case and verbal agreement.

In accordance with this, S/A detransitivisation is deeply intertwined with many areas of morphosyntax, namely with all areas where transitive agreement plays a role: differential A marking,

agreement in non-finite forms (infinitive, purposive), raising to O- and S-AGR with infinitives, raising and choice of light verb with the converb *-saŋa*, the vector verb *-hat(t)* [AWAY]. Because S/A detransitivisation is marked in several places and interacts with a lot of other phenomena, it can also appear in quite different forms itself. For instance, in subclauses where a non-finite verb assigns case to A but has reduced possibilities for agreement, S/A detransitivisation may be expressed solely by the case of A. In contrast to this, DOM with its restriction to a single locus of marking does not interact with a lot of other phenomena. Two notable cases are the double DAT family of constraints and the inability of O-DAT to trigger AGR.

The grammatical relation of O is easy to define independently in Chintang – it is the argument linked to O-AGR (which is also always marked by NOM). In Nepali, it is impossible to define O without reference to DOM, and even then the definition is rather roundabout: O is the argument of a transitive verb with A-ERG/NOM whose case marking alternates between NOM and DAT (and sometimes GEN). The role sets covered by these definitions are largely congruent but diverge in one remarkable case: G corresponding to recipients are linked to O-AGR in Chintang and therefore subject to S/A detransitivisation, whereas in Nepali they have fixed DAT and DOM is only (marginally) possible on the associated T.

Both differential marking patterns feature a binary, privative opposition. However, the marking systems into which these oppositions are embedded are rather different. NOM and DAT in Nepali are only two of a variety of cases, many of which can be used to mark P/T/G as well. Chintang O-AGR is one of only three options (the others being S-AGR and A-AGR), and there are only two cases for A (ERG and NOM). Further, the lack of O-AGR that is found in S/A detransitivisation is not the same as the lack of an overt case marker on Nepali O-NOM: NOM is paradigmaticised with other case markers and therefore a true zero, whereas none of the affixes in a Chintang detransitivised verb form is necessarily in contrast with an affix in a transitive verb form. This is because agreement affixes in Chintang do not exhibit a uniform alignment pattern, so that a detransitivised form cannot simply be derived from a transitive verb form by replacing the O-AGR affixes by zeros. For these reasons, DOM may be said to be a mechanism that replaces a default case with another case from a wide range, whereas S/A detransitivisation cuts an agreement link so that AGR and A case have to be changed to the only other available choice.

This also has functional implications: in the case of DOM it may be asked why out of all cases DAT was chosen as an alternative O marker. The most likely answer is that the other main argument types marked by DAT – recipients and experiencers – historically shared important properties with certain O such as frequently being animate, specific, and highly topical, which made it possible to extend the use of *-lai*. Synchronically, too, the Nepali DAT is much less strongly associated with roles than Chintang O-AGR. DAT is most frequent on P/T/G but can in principle mark every role including S and A, whereas O-AGR is confined to O – the only exception are a handful of deponent verbs and constructions where O-AGR is formally present but not linked to any argument.

The oppositions in the two language are also different in another important respect, which is frequency. In Nepali, zero (NOM) is the default, but in Chintang "zero" (lack of O-AGR) is the exception. Put differently, the default in Nepali is marked by less morphological material than the exception, whereas in Chintang it's the exception that features less material. This has consequences for marking statements: one could say that DAT marks high O whereas S/A detransitivisation marks low O.

Finally, there are also deep functional differences between the phenomena in question. First of all, S/A detransitivisation in Chintang is functionally simple whereas Nepali DOM is highly complex. S/A detransitivisation involves a single main variable (specificity) with two values and a minimum degree of flexibility on the side of the speaker. The only major exception is arbitrary reference, where the speaker has some freedom to present a referent as arbitrary or not. Specificity is in turn almost always congruent with quantifiability. Nepali DOM does not only depend on many more factors (animacy, specificity, quantifiability, topicality, part of speech, modification, unexpectedness, disambiguation, affectedness); most of these are also much more fluid and therefore harder to assess than specificity. Furthermore, when looking at Nepali DOM one gets the impression that anything goes, whilst S/A detransitivisation is rather strict with respect to grammaticality

statements. S/A detransitivisation can be modelled in terms of rules, but rules are the exception in DOM and rather hamper a deeper understanding of the phenomenon than facilitate it. DOM is therefore better modelled in probabilistic terms.

The crucial factors for the two phenomena also come from rather different functional areas. Specificity is for the largest part a referential property on the level of the clause that looks neither back into the lexicon nor out into discourse. This restriction is the main reason why S/A detransitivisation is functionally so simple. By contrast, Nepali DOM is concerned with all three mentioned levels: lexical or semi-lexical properties such as animacy and part of speech lay out the base, clause-level properties such as specificity/quantifiability, modification, ambiguity, and affectedness modify the base, and discourse-level properties such as topicality and unexpectedness complete the picture. There is mutual influence between the levels, especially between the lexicon and discourse. For instance, a pronoun comes with a different lexical disposition from a noun and will therefore be used differently in discourse. On the other hand, usage patterns of pronouns in discourse can over time become entrenched and be fed back into the lexicon.

So all in all there are many more remarkable differences between Chintang S/A detransitivisation and Nepali DOM than there are remarkable similarities. There is the question whether this is pure chance or principled. I would like to claim that the majority of differences can be related to the fact that S/A detransitivisation is primarily expressed via agreement, whereas DOM is exclusively expressed via case.

This starts with the form. Agreement in general is connected to a lot more phenomena than case marking, especially in a language like Chintang, where agreement is potentially bipersonal and arguments have agreement of some form in almost all constructions. This accounts for the formal intricacies of S/A detransitivisation. The expression of S/A detransitivisation in several loci can be explained by the interaction of a differential agreement pattern with a language-specific rule stating that ERG-marked arguments may not be linked to S-AGR. The centrality of O-AGR in Chintang also explains why O is easy to define on its base. Finally, the different opposition types in the two phenomena are in line with the general background, too. Most languages with agreement only have a single agreement slot, and more than two (A/O-AGR aligning with various roles) are very rare. By contrast, having more than two cases is the rule rather than the exception for languages which do have case, and inventories with dozens of forms are nothing unusual, especially if one doesn't restrict the concept of case to affixes.

Similarly, the functional differences between S/A detransitivisation and DOM can be related to more general functional properties of agreement and case. Although these two phenomena are superficially similar in marking syntactic functions and frequently being subject to differential marking, a closer look reveals important differences. A useful summary of the functional literature is given in Iemmolo (2011:48ff.): case marking mainly serves to disambiguate roles (especially peripheral roles which cannot be easily inferred), whereas agreement is a referent-tracking device. This motivates the central role of specificity for S/A detransitivisation: only those O that *can* be tracked *are* tracked via O-AGR. By contrast, there is a plethora of functional factors that are unusual for O referents and hence make it harder to identify their role – this explains the functional complexity of Nepali DOM. Fluidity comes in as soon as a wider discourse window must be looked at and factors interact with each other. The sensitivity of agreement to referent tracking also motivates the different role sets covered by O as defined by S/A detransitivisation and DOM: highly animate recipients are on average easier and more interesting to track than their associated T (Dryer 1986:841). Secundative alignment (P=G) is also typologically slightly more frequent in agreement than indirective alignment (P=T), whereas indirective alignment is much more frequent than secundative alignment in case marking (Haspelmath 2005:5).

4.1.3 Mutual influence?

The question of mutual influence between Nepali and Chintang must be asked here for the sake of completeness but can be answered in the negative without much discussion. The profound differences between S/A detransitivisation and DOM make it *a priori* unlikely that they should be identified across languages even by fully bilingual speakers.

This is nicely illustrated by the case of Puma, a language closely related to Chintang. According to Bickel et al. (2007b), Puma has S/A detransitivisation parallel to Chintang and has in addition borrowed Nepali *-lai* together with the differential marking pattern. Although there is some interaction between the two (DAT is impossible on detransitivised O), the two phenomena coexist and do not seem to have influenced each other. Chintang does not even make regular use of Nepali *-lai*, and I have never noticed any parallelisms between the use of the two mechanisms in the speech of a single speaker. It's theoretically conceivable that older speakers with a bad command of Nepali would use DAT on all specific referents or that younger speakers whose Nepali is better than their Chintang would only track referents via O-AGR which are high in the sense of Nepali DOM. However, nothing of the sort is attested.

It is also highly unlikely that S/A detransitivisation should have been influenced by DOM historically. S/A detransitivisation stems from differential agreement, for which different diachronic sources must be assumed than for differential argument marking. Nepali *-lai* could be derived from an old converbial form (see section 3.7.2), but this is excluded for S/A detransitivisation, which rather looks like the intransitive agreement pattern had been extended to certain transitive clauses.[1]

Finally, influence from Chintang to Nepali is completely out of the question. As noted by Hutt (1988:29), "the Indo-Aryan immigrants invariably imposed their rule on such peoples and imbibed little of their culture". What's more, Chintang is much too small to have been able to influence any big, prestigious language such as Nepali.

4.2 Repercussions for general linguistics

4.2.1 Differential marking

Chintang S/A detransitivisation shows that it is fruitful to define differential marking in a broad way and to assume that the locus of conditions (in our case, the object) can serve as a *tertium comparationis* for various phenomena.

S/A detransitivisation is formally located between several other differential marking patterns. It is close to differential agreement but different from its pure form in that A- and O-AGR change at the same time and A case changes, too. It is also different from S/A ambitransitivity (it is not lexicalised), from antipassives (O is not removed from the valency and there is no verbal marker), and from noun incorporation (O stays syntactically independent). Functionally, however, S/A detransitivisation is rather similar to all these phenomena in marking non-specificity of O. Thus, ultimately it might be more useful for typology to treat the mentioned phenomena as different formal realisations of a broader functional category of object-conditioned differential marking. The antipassive as the construction with the widest functional coverage only partially belongs here since not all antipassives are object-conditioned.

There is the question where DOM would fall in this picture. On the one hand, DOM is just another object-conditioned differential marking pattern. On the other hand, DOM shows a simple but important difference to the mentioned phenomena: it clearly marks O. Whereas zero marking seems to be the default in most privative DOM systems, DOI and other object-conditioned differential marking patterns seem to be rather open with respect to this question, and sometimes (as in the case of Chintang S/A detransitivisation) the pattern even gets reversed so that zero-marking is the exception. Impressionistically I would thus tend to place DOM in a separate subclass, which would yield a dichotomy of differential marking patterns formally centered on O vs patterns affecting the formal relation between O and the predicate. Whether such a distinction is indeed useful for large-scale typology is an important open question.

[1] Of course it would in principle also be possible that S-AGR was initially used in all clauses and O-AGR only developed when specific referents got indexed by person clitics. This theory can, however, not deal with the fact that transitive agreement in Kiranti languages can usually not be constructed as S-AGR + X: just adding *-u* [3O] to an S-AGR form in Chintang produces ungrammatical forms in most cases.

4.2.2 Identifiability and quantifiability

For the description of the functional properties of S/A detransitivisation, a definition for definiteness and specificity was developed in section 2.5 that bases both of them on identifiability.

The most important prerequisite for specificity in Chintang is quantifiability, i.e. the possibility of determining the quantity a referent. Quantifiability is crucial for specificity (and identifiability in general), because in order to identify two or more referent ensembles with each other one has to know their boundaries – if it is not clear which individual referents belong to an intended group or which parts to an intended mass, it will not be possible to assess identity for all individual referents or parts. The close relation between specificity and quantifiability in Chintang is of typological interest for several reasons.

First, in the classic languages for research on identifiability (European languages with articles such as English or French), the role of quantification can only be observed indirectly because the articles tend to interact with other nominal modifiers. Most importantly, indefinite articles (English *a*, French *un(e)* etc.) may rarely ever co-occur with independent quantifiers, so it is hard to tell whether a phrase like *three houses* is (from a language-internal, structural perspective) specific or not. By contrast, specificity in Chintang is not expressed on the object NP but on the verb and A, so the effect of quantification can be observed more easily.

Second, quantifiability has so far been almost completely neglected in the study of differential marking patterns. The only exception is the well-known case of symmetrical DOM (alternation of ACC with PART or GEN) in Finno-Ugric and Slavic languages. Chintang presents a link between this pattern and the multitude of differential marking patterns where specificity and related notions are acknowledged to play a role.

Finally, quantifiability has proven to be especially useful for the description of mass concepts. Mass concepts are omnipresent in everyday discourse (e.g. in the form of food) but are usually ignored in the description of object-conditioned differential marking patterns, which rather focus on concrete individual concepts such as persons or apples. Talking about identifiability in this area is impossible without talking about quantifiability, so Chintang S/A detransitivisation offers a good starting point for exploring the behaviour of mass concepts in other languages, too.

There are two strands of research on quantification which were largely ignored in this work for reasons of space. One is the formal semantic tradition of explaining identifiability in terms of quantification in a more strict sense (i.e. via existential and universal quantification). The other is research concerned with relations between quantification and verbal properties such as aspect. Linking S/A detransitivisation to these areas would be another point for future research.

4.2.3 Non-reductionist explanation

Section section 3.5 has shed some light on the factors behind DOM in Nepali. It was shown that this pattern cannot be explained based on a single factor such as animacy or specificity and that if one considers several factors, integrating them into a probabilistic system yields better explanatory results than a classical rule system where each possible combination of values is linked to an unambiguous outcome.

These results are another small blow to reductionism in linguistics. Even though it is commonplace in most sciences dealing with complex systems that monocausal, deterministic explanations are rarely realistic, the belief that they mostly are is only recently and very slowly losing ground in the language sciences. The stance taken by the present work is that it is always worth taking a second glance. To be sure, there are phenomena like Chintang S/A detransitivisation which come close to the reductionist ideal, but there are also phenomena like Nepali DOM, which look simple at first sight (cf. the literature review in section 3.5.2) but turn out to be highly complex on closer scrutiny.

One technique that fosters reductionism is elicitation. The reason is that elicitation forces speakers to construct a context for a given sentence, with the consequence that many variables are not controlled by the linguist but by the speaker. What's more, since not all speakers are equally good at constructing contexts (especially rare or bizarre ones), many sentences will get

a simple rejection in elicitation which are actually attested in corpora (and accepted by the same speakers when viewed in context). This danger can be reinforced by too coarse grammaticality judgements – a binary opposition will produce the illusion of binary grammar. Elicitation keeps an important role because it can provide information on what is impossible and yields quick results when exploring a phenomenon. However, corpus data are more appropriate for finding out what *is* possible and for developing a more sophisticated understanding of language.

4.2.4 Theories of DOM

One of the central questions for theories of DOM has been what the function of DOM is. The main answers divide them into "distinguishing" and "indexing" theories. The present study gives a rather radical answer concerning the function of DOM in Nepali. First, the distinguishing and the indexing functions of DOM are not mutually exclusive – both play a role, although indexing factors are more numerous and more important.

Second, *the* function of DOM may not exist. There is a common denominator to most values favouring DAT in Nepali – they are unexpected for an O referent. However, this property is way too abstract and vague to serve as a case predictor. Thus, instead of interpreting it as a function, it may be more appropriate to view it as one of the pathways by which the use of DAT got extended diachronically.

Another point where the present study may contribute to a general theory of DOM is the problem of referential hierarchies. Such hierarchies are of some use for the description of DOM in Nepali – in particular, an ordinal scale of animacy yields good results in elicitation. However, most relevant variables are either categorical without any possibility of ranking values (for instance, modification) or even binary so that the distinction becomes irrelevant (for instance, specificity). What's more, the animacy hierarchy that was found in elicitation is not supported by quantitative corpus data.

Apart from animacy, a wide variety of factors were shown to be relevant, some of which haven't been observed in other Indo-Aryan languages or are even uncommon in a wider, typological perspective (quantifiability, unexpectedness, affectedness).

Appendix A

Annotation guidelines Chintang

A.1 How to tag files

All files to be tagged are in the Toolbox format, so every line begins with a field marker indicating the type of information (e.g. \mph for morphemes, \mgl for glosses). For annotation, an additional tier \anno has to be inserted and modified. This can either be done in Toolbox or in any text editor. The procedure in detail is as follows:

1. Check and copy the line \gw (**g**rammatical **w**ords)
2. Insert the copy below \lg (**lan**guage) or \id (**id**entifier, if existing). Set the field marker for the new tier to \anno.
3. Identify and tag domains (see below for details). If a core role is covert insert a zero 0 in its place.
4. Identify and tag core roles (see below for details).
5. Tag at least all P/T/G for identifiability and quantifiability (in this order).
6. Tag all verbs for lexical class and alternations.

These steps can be carried out in a different order as long as the order of tags is preserved. For instance, one could first look at all nouns, check whether they are core roles, assign to them a preliminary domain identifier and a role, and tag them for identifiability and quantifiability.
\gw lines lacking any taggable elements do not have to be copied and renamed.

Tags are separated by a dot (.).

A.2 Variables, carriers, and values: overview

A.2.1 Domains

carrier: predicates (word carrying semantic content in multi-word forms) and heads of argument NPs
values:

- 1 = belongs to domain 1
- 2 = belongs to domain 2
- 3, 4, 5… = …
- 1/2 = belongs to domain 2 under domain 1 (in complex sentences)
- 1+2 = belongs to domain 1 and 2 (shared arguments)
- 1+2* = belongs to domain 1 and 2 but is realised as argument of 1
- xdom = insecure, to be specified later

A.2.2 Role

carrier: heads of argument NPs
values:

- S
- A
- P
- T
- G
- CT = copulative theme
- CR = copulative rheme
- NEXP = experiencer noun
- BEN = beneficient
- CSR = causor
- xrol = insecure, to be specified later

A.2.3 Identifiability

carrier: heads of P/T/G NPs
values:

- def = definite
- spec = (indefinite) specific
- idf = indefinite (non-specific)
- xdef = insecure, to be specified later

A.2.4 Quantifiability

carrier: head of P/T/G NPs
values:

- qnt = quantifiable
- nonq = non-quantifiable
- xqnt = insecure, to be specified later

A.2.5 Verb class

carrier: predicates
values:

- aux = auxiliary
- dido = direct object ditransitive
- dioo = double object ditransitive
- dipo = primary object ditransitive
- expitr = intransitive experiential
- exptr = transitive experiential
- itr = intransitive
- other = other frame
- tr = monotransitive
- uninf = uninflected verboid
- xcla = insecure, to be specified later

A.2.6 Alternations

carrier: predicates
values:

- (empty if no alternation)
- ambrec = intransitive variant of reciprocal ambitransitive
- ben = benefactive
- caus = causative
- cop = copulative
- dumA = dummy A-AGR (in transitive experiential class)
- idt = indeterminate as to detransitivisation status
- OtoS = embedded O with matrix S-AGR
- pass = passive (with *-mayaŋ* [PASS.PTCP])
- poss = possessive
- recp = reciprocal (with *-ka* [RECP])
- refl = reflexive (with *-ce/cĩ/cɨ* [REFL])
- sad = S/A detransitivisation
- sod = S/O detransitivisation
- xalt = insecure, to be specified later

A.3 Variables, carriers, and values: details

A.3.1 Domains

A domain is a set of forms which interact morphosyntactically and can correspond to various syntactic concepts, e.g. a clause or a sentence. A minimal domain consists of a predicate only, but usually arguments are contained as well. Generally as sequences of more forms are considered, interaction tends to become less formal and more functional, which means that larger domains are much harder to define than small domains. For the present purpose it is not necessary to define domains larger than sentences.

All elements belonging to one domain have the same ID, and elements in other domains have different IDs. It is not necessary to assign IDs to all elements but only to those which are also tagged for other variables (verbs and core arguments).

Some cautions:

- If there is no overt verb, nevertheless try to identify domains (this is mostly possible when there are several arguments so one can easily guess their relative roles). For instance, in *sencak menuwa-ŋa* [mouse cat-ERG] the most probable frame is one where *sencak* is P and *menuwa* is A.

- If the domain for what looks like an argument cannot be determined use `xdom`.

- If there are isolated words of which it seems safe to say that they do not belong to any domain even though they look like they are (or were intended to be) arguments, rather give them a domain ID of their own. This most frequently happens with false starts, as in *Hana akka nis-u-ku-ŋ-nɨŋ* [2s 1s know-3[s]O-IND.NPST-1sA-NEG] 'You, uh, I don't know'. Complete repetitions which are only due to hesitation (*akka akka nisukuŋnɨŋ*) may be ignored.

- In complex NPs it is always the head which should bear the ID and all other tags. In Chintang, all forms that can be used adnominally can also be used referentially, so there is no reason to assume zero heads anywhere. This means that as soon as there is one overt element it will be considered the head. In case there is no overt head (e.g. *hana phorokŋa* 'you the frog' or names such as *Ram Bahadur*), simply tag the last element (*phorokŋa*, *Bahadur*).

- In most cases predicates are realised by verbs. However, there are also non-inflectable verboids (*manchi*, *maha?*, *phophei?*) which are morphosyntactically different from verbs but still have arguments (and thus constitute a domain together with them).
- There are subdomains and other special rules in complex sentences (see A.4.2).

A.3.2 Role

Which arguments are assigned which role? This project recognises five basic roles (S, A, P, T, G) and five auxiliary roles (CT, CR, NEXP, BEN, CSR) for special purposes.

- **S**: the only core role specified by an intransitive verb
- **A**: the most proto-agent-like core role specified by a mono- or ditransitive verb
- **P**: the most proto-patient-like core role specified by a monotransitive verb
- **T**: the most "proto-theme"-like core role specified by a ditransitive verb (= the non-A argument that is moving in space or metaphorically)
- **G**: the most "proto-goal"-like core role specified by a ditransitive verb (= the non-A argument that is stationary)
- **CT**: theme in copulative constructions (= that which is talked about)
- **CR**: rheme in copulative constructions (= that which is predicated)
- **NEXP**: the possessed emotion or organ found with most experiencer verbs
- **BEN**: beneficient (additional argument introduced by *-bid*
- **CSR**: causor (additional argument introduced by *-mett*

These definitions represent a Dowtyan/Bickelian approach to roles (Dowty 1991, Bickel 2011). See Haspelmath (2011) for an overview about this and other approaches. The most important characteristic of this approach for our purposes is that role depends on valency. Verb senses which have identical valencies use the same role set, and different role sets must be used for verbs with different valencies. This leads to a couple of unexpected role assignments. Here are the most important examples:

- Destinations count as G when there is an A and a moving object (T). For instance, in *Father sent me to Kathmandu*, the speaker is T and *Kathmandu* is G.
- Instruments count as T. For instance, in *Don't cut the meat with the penknife*, *knife* is T and *meat* is G.
- Experiencers count as S or A, depending on whether there is a stimulus. For instance, *you* is S in *Are you hungry?* and A in *Are you afraid of cakes?*
- Where possession is predicated possessors are viewed as A and possessums as P. This is independent of the construction, so even if possession is expressed as 'There is possessum of possessor' as in Chintang (e.g. *Huĩsako uchauce sumbhaŋ uyuŋno* 'He has three children') the roles remain unchanged.

Covert referents occupying core roles are to be represented in \anno by zeros 0, which can then be tagged just like an overt referent (e.g. 0.1.S = zero referent filling S in domain 1). The position of the zero should follow the default word orders S-V, A-P-V, and A-G-T-V. The zero is to be placed in the earliest possible position where it is in accordance with the default word order and does not break up phrase structures. For instance, if there is an overt frame adv-G-ptcl-V with covert

A and T, the zeros for these should be placed at the beginning of the domain (not after the adverb) and right after G (not after the particle, except where the particle belongs into the same NP as G): A-adv-G-T-ptcl-V. Note that there are no adjectives or other specialised adnominal structures in Chintang, so it is not necessary to posit zero heads for any NPs with at least one overt element.

When a covert referent has already been mentioned some time before it may sometimes not be clear whether it should be expressed as a zero or whether the domain of the current predicate should be added to the original mention. The convention in such cases is to use a zero if the sentence where the referent was last mentioned is closed. If the mention was within the same sentence the other method is to be chosen.
Here are some more tricky cases:

- When a referent is followed by *=mo* [CIT] all tags should be added to *=mo*. The reason is that in such cases the real argument of the verb is not the referent but the form of the expression coding it. For instance, in *a-ppa mo lud-a?-na* [1sPOSS-father CIT tell-IMP-ERGIST] 'say "my father"', 'father' is not an argument of *lud-* in the same way as, for instance, *katha* 'story' might be – the real argument is "appa".

- If there are **several NPs** on the same level in the same role, assign that role to all of them. **Appositions** should not be considered as NPs on the same level as the NP modified by them.

- The label ***xrol*** is to be used where one is not sure about the role of an NP. Sometimes one might find cases where even the number of roles is insecure (e.g. an S/P ambitransitive verb where only S/P is overt and the verb form is ambiguous as to transitivity). In such cases the maximal number of possible roles should be assumed and all roles should be tagged *xrol*. For instance, a sentence like *gilas od-e* [glass break-IND.PST] could either mean 'the glass broke' (S) or 'he broke the glass' (A-G-T). If the meaning does not become clear from the context the sentence would have to be role-tagged `0.xrol gilas.xrol 0.xrol ode.xfra`.

- Sometimes one referent may occupy two roles in the same domain without a reflexive marker being used, e.g. *lupmiŋa dube* 'the needle pricked him (with itself)' (*lupmi* is A and T) or *hali laktaŋsehẽ* 'I stained myself with blood' (1s is A and G). Treat such referents as in a reflexive frame (i.e. assign the same domain ID twice but with two different corresponding rules) but do not specify any alternation on the verb.

A.3.3 Identifiability

This is the most problematic variable. In the linguistic literature it is more usually referred to as definiteness. A good overview of the topic is Lyons (1999). Other more specific and theoretically more informed sources are Hawkins (1978), Chesterman (1991), Abbott (2010).

Identifiability is an assumption of the speaker as to whether at utterance time it is possible to uniquely identify a referent. Although unique identifiability is a common concept in the definiteness literature, some amendments have to be made to make it work well:

- The identified items (i.e. the referents) are not real-world entities but concepts. Without this assumption it is difficult to deal with irreal settings and fictitious referents.

- Concepts are located in mental spaces (Epstein 1999, 2002, term by Fauconnier 1984, 1994). Without this assumption it is difficult to deal with scenarios that are set up ad hoc in a conversation and with cases where the hearer has limited knowledge of a referent but still can identify within a certain mental space.

- The descriptive content of an NP coding a referent is only one component of identifiability. The other component are the cognitive abilities of the hearer. The hearer can make use of various sources to supplement the information provided by the speaker:

    - previous discourse (referent has already been mentioned, e.g. *That's **the guy** I told you about before.*)

  - situational context (referent is present in the context of the speech situation, e.g. *Could you open **the window**?*)
  - common background of speaker/hearer (both participants know the referent from earlier interactions, e.g. ***The mice** are still there (the ones in my apartment – you've seen them).*)
  - world knowledge (referent is so well-known that it can be assumed to be known to virtually anybody, e.g. ***The sun**'s shining.*)
  - bridging (= association based on world knowledge, e.g. ***The leader** of Al-Qaeda has been killed.*)

- A description counts as unique as soon as it is *sufficiently* unique, that is, as soon as it makes it possible to separate one referent (or one group of referents) from all other referents in the relevant mental space.

A couple of problems related to identifiability maybe summarised under the heading of referential scope (not to be confused with the narrow/wide scope distinction that is sometimes made within specific referents). Here is a list:

- Many referring expressions can signify either a type or a token (distinction originally by Peirce 1906, used to analyse identifiability e.g. by Jackendoff 1983). For instance, in *Nepali khana ca-ma les-u-ku-ŋ* [Nepali food eat-INF like-3[s]P-IND.NPST-1sA] 'I like (to eat) Nepali food', *Nepali khana* can be interpreted as a type (if the statement is about general preferences of the speaker) or as a token (if it refers to a concrete situation). In the first case it is clear that *Nepali khana* is uniquely identifiable (because there is only one 'Nepali cuisine'). However, in the second case there are various possibilities. For instance, if the sentence was said at a potluck party where several international specialties are available and among them there is one Nepali dish, *Nepali khana* would still be uniquely identifiable. On the other hand, if there are several Nepali dishes *Nepali khana* might only be uniquely identifiable to the speaker (he knows what he's talking about) but not to the hearer.

- Sometimes there is an identifiable whole with parts that cannot be easily told apart and it's not clear whether an expression refers to the whole or to the parts. For instance, imagine there is a beaker of water and somebody asks you *Cuwa a-thuŋ-no?* [water 2[s]S-drink-IND.NPST] 'Do you (want to) drink water?' What is most likely is that you want to drink some of the water, i.e. a non-identifiable part of the body of water in the beaker. However, it might also be the case that you are supposed to drink the whole beaker, in which case the object of drinking would be identifiable.

- Sometimes a referent is uniquely identifiable in one mental space but not in another and it's not clear which space is relevant. For instance, take the sentence *Yo?nibai?ni menuwa=yaŋ u-hɨk-no* [here.and.there cat=ADD 3pS-keep-IND.NPST] 'In some places they keep cats'. *menuwa* is associated with a non-identifiable referent (people or households) and is thus itself not identifiable. However, this sentence sets up a mental space in which one household is keeping one or several cats, and within that frame *menuwa* may eventually become identifiable. Suppose the next sentence was *Kina cha-ce-ŋa teī-saŋa u-khoŋs-o-ko* [SEQ child-ns-ERG beat-CVB.FGR 3pA-play.with-3[s]P-IND.NPST] 'And the children play with it by beating it' – now the relevant mental space is one arbitrary household in which there is one uniquely identifiable cat.

- A similar situation sometimes occurs with large groups of referents that oscillate between indefinite and definite construals. For instance, in English it is possible to say *I never got along with girls* or *I never got along with the girls.* In the first case the group of girls is construed as indefinite, and the sentence says that there were several girls the speaker did not get along with and which are representative of the set of all girls (but not identical to it – that would make the referential group definite). In the second sentence the group of girls is construed

as definite, resulting in a reading where *get along* is asserted to extend to all members of the relevant group (not only to representatives). Since there are obviously too many girls to be able to say whether one really gets along with all of them the most likely reading of this sentence is one where the set of girls is further restricted by covert information (for instance, it could become clear from the cotext that the sentence is about the girls in the speaker's former class). In a language like Chintang which does not code definiteness explicitly the difference between the two construals is often difficult to tell.

There is no method that helps in all these cases. Whenever you come across one or another type of ambiguity of referential scope, you have to decide which scope is the most likely one in the present context and judge identifiability accordingly. For instance:

- When *Nepali khana cama lesukuŋ* is uttered in a conversation on likes and dislikes, *Nepali khana* is most likely to refer to the type. When it is uttered on a potluck party the token reading is more likely.
- When somebody asks you whether you can drink up some water he is probably talking about the whole. When a host asks you as the guest whether you want water he is probably talking about a part of the available water.
- When you encounter the sentence *Chaceŋa menuwa teĩsaŋa ukhoŋsoko* in a text on the bad sides of children you can be pretty sure that you don't have to assume any particular mental space, so there is no particular cat and *menuwa* is not identifiable. However, when you see the same sentence in the description of a family, the relevant frame for identification is the household, so the cat becomes identifiable.
- When there are some obvious restrictions on a referential group (e.g. all the girls of the village) or when a group is being talked about as a prototype (as in *The girls always stick together*) the group is probably identifiable.

The following values for identifiability are used:

- **definite**: (It is assumed that) both speaker and listener can uniquely identify the referent(s).
- **indefinite specific**: Only the speaker can uniquely identify the referent(s).
- **indefinite non-specific**: The referent(s) cannot be uniquely identified.

It's hard to develop reliable tests for when to assign which identifiability tag. The best method is to understand what unique identification within a mental space means. An informal way that often helps to estimate identifiability is to try inserting the following phrases:

- "You know which one(s).". If this is possible the speaker assumes both he *and* the listener can uniquely identify the same referent, so the value is `def`.
- "In a minute you/I will know which one(s) I/you mean". Points to `spec`.
- "It doesn't matter which one(s)." If this is true it is neither possible to identify the referent at the moment nor is it important to do so in the following discourse. This means `idf`.

A.3.4 Quantifiability

Quantifiability is a term that is used in this project to mark whether the quantity of a referent can in principle be determined or not. This distinction correlates with the count/mass distinction in the domain of the lexicon.

There are two main types of nominal concepts that behave differently with respect to quantifiability:

- Homogeneous concepts (≈ masses, term by Rijkhoff 2002) are such that their category label can be applied to arbitrary partitions of referents they apply to. For instance, if there is a large body of *water* and one separates a subamount from it that amount will still be categorised as *water*. Concepts of this type are usually construed as non-quantifiable and can only be made quantifiable by special means, e.g. by containers (as in *a glass of water*).
- Heterogeneous (≈ count) concepts have parts that belong to a different category than themselves. For instance, the head of a *dog* cannot itself be called *dog*. Concepts of this type are usually construed as quantifiable but can be non-quantifiable where their parts do not matter (as in *Have some dog (meat)!*) or where they occur in large groups that cannot be easily overlooked (as in *People use to walk their dogs around here*).

Quantifiability is certainly the most unusual variable used in this project, so don't hesitate to ask questions and/or use the tag `xqnt` whenever you are insecure. Here is a couple of hints to one or the other value:

- Numbered or measured referents are always quantifiable.
- Referents with indefinite quantifiers such as *them-them, sapphi, baddhe* are usually non-quantifiable. An exception to this are universal quantifiers like *jammai*. Referents marked by these are always quantifiable. The quantity itself may be indefinite, but it is fixed in that there are comparatively strict rules for when to use these quantifiers. For instance, if you said *Akka jammai ca-ŋ-ci̱-h-ẽ* [1s all eat-1sA-3nsO-1sA-IND.PST] 'I ate them all' of three apples but one was left you will normally be considered a liar. By contrast, if you said *Akka seu ca-ma le-ŋa-ʔã-niŋ* [1s apple eat-INF like-1sS-IND.NPST-NEG] 'I don't like (to eat) apples' and it turns out that there *is* one sort you like that shouldn't be a problem.
- Referents in containers are usually quantifiable, especially masses (*glass of water*, *bowl of rice* etc.).
- The comparison of quantities triggers quantifiability, so *money* is quantifiable in *You have more money than I*. The reasoning behind this is that precision matters when comparing quantities (e.g. if you had 100 Rupees less the statement might no longer be true).
- Where referents occur in a large group that is difficult to overlook, that group tends to be non-quantifiable.
- With complex NPs involving possessors, the default for quantifiability is to get percolated from the innermost possessor to the outermost possessum. For instance, in *people's thoughts* the possessor *people* is non-quantifiable, and so is the possessum *thoughts*. By contrast, in *Henry's thoughts* both possessor and possessum are quantifiable. The reasoning behind this is that a quantifiable possessor can only have a finite number of possessions and that it is hard to isolate a quantifiable referent within an non-quantifiable area (as spanned by an non-quantifiable possessor). Exceptions are rare but possible in both directions. *Any offsprings of theirs would simply have won the great genes lottery!* is an example for a quantifiable possessor with an non-quantifiable possessum (there may be so many offsprings over such a long period of time that they cannot be easily kept track of any longer). In *That's the idea people have when they come here* there is a quantifiable possessum within an non-quantifiable possessor (many people but just one idea for all of them).

A.3.5 Verb class

Verb class is one of the two factors determining frames. For the purposes of tagging, verb classes are simply defined by the frames found in the \val field in the Chintang dictionary. The only exception are S/O ambitransitive verbs such as *ot-* 'break' which regularly have two different entries with different frames in the dictionary (intransitive and mono-/ditransitive) but should be assumed to belong to the class defined by the transitive frame here.

If you are sure that you have identified a class but it is so rare that it is not on the list of tags label it `other`. If you are not sure which of the established classes a given frame is an instance of or whether it is on the list at all use the label `xcla`.

- **aux** (auxiliary): some verbs sometimes do not express semantics of their own but merely serve to make full-fledged predicates out of non-finite verb forms. Verbs which do this are *lus-*, *lis-*, *numd-*, and *mett-*. See A.4.2 for a list of the constructions in which they are used as auxiliaries.
- **dido** (direct object ditransitive): A-ERG T-NOM G-LOC V-a(A).o(T). Cf. Bickel (2007, 2008b) for discussion of the various ditransitive frames.
- **dioo** (double object ditransitive): A-ERG T-NOM G-NOM V-a(A).o(G).
- **dipo** (primary object ditransitive): A-ERG T-ERG G-NOM V-a(A).o(G).
- **expitr** (intransitive experiential): S-GEN/NOM poss(S)-NEXP V-s(NEXP).
- **exptr** (transitive experiential): A-ERG P-NOM poss(A)-NEXP V-a(A/3s).o(P). Much rarer than *expitr*. Note that with 3sA it is not possible to distinguish the unmarked frame of this class `exptr.dumA`, which has dummy 3sA-AGR. Use bare `exptr` in that case.
- **itr** (intransitive): S-NOM V-s(S) without the possibility for bipersonal inflection. All in all there are not many intransitive verbs – just about 20% of all Chintang verbs. A couple of frequent verbs are transitive (because they license the monotransitive frame) but are rarely ever used transitively and do S/A detransitivisation instead. These are *cekt-* 'speak; say', *hatt-* 'wait; wait for', *kupt-* 'perch; brood', *khoŋs-* 'play; play with', *ŋed-* 'study; read, count', *pheŋs-* 'plough', *ratt-* 'make noises, shout; scold', *ya-hatt-* 'chant'. Be careful not to tag these as `itr` but as `tr.of` where used intransitively. Check the dictionary when you are not sure whether some verb is really intransitive.
- **tr** (monotransitive): A-ERG P-NOM V-a(A).o(P). This is the most frequent class. Take care, though: due to the wide functional definition of ditransitives adopted here (cf. A.3.2 above) and in the Chintang dictionary there are probably fewer monotransitives than you might expect. Always check the dictionary when you feel insecure.
- **uninf** (uninflected verboid): S-NOM V. Verboids are words which take a single argument in the nominative and are not inflected (hence the label, which is hopefully less confusing than `verb`). Only three verboids are known so far: *manchi*, *maha?* and *phophei?*. Especially the first two are very frequent, so this class is important.

A.3.6 Alternations

- **ambrec** (intransitive with S=A+P, reciprocal ambitransitive): the label for S-NOM V-S where the verb also allows A-ERG P-NOM V-a(A).o(P) and S is coreferential to both A and P (possible for instance with *tup-* 'meet' just as in English: *A meets B* `tr` or *A and B meet* `ambrec`).
- **ben** (benefactive): marked by the V2 *-bid* [BEN], which adds a beneficient with NOM and O-AGR. *-bid* cannot attach to intransitive verbs. Use the special role BEN for the beneficient; all other roles stay the same.
- **caus** (causative): marked by *-mett* [CAUS]. Adds a causor which lets S/A do what they do. The causor gets linked to A-AGR and O-AGR goes with the causee (there might be exceptions, but as far as we presently know they are rare and not regular). Use the special role CSR for the causor; all other roles stay the same.
- **cop** (copulative): CT-NOM CR-NOM V-s(CT). This frame is only used by a few intransitive verbs (mainly *yuŋ-* and *lis-*) and by verboids (*manchi*, *maha?*). Copulation in Chintang actually does not need a copula, so there will often be no verb at all or just *-kha* [NMLZ].

- **dumA** (dummy A-AGR in transitive experiential class): A-GEN/NOM P-NOM poss(A)-NEXP V-a(3s).o(P).

- **idt** (indeterminate): use this label where the verb is transitive but the crucial indicators for detransitivisation – A marking and O-AGR – are covert cannot be inferred from some other source. There are two frequent cases where this label is appropriate. One is where the marker *-u* [3O] gets dropped before another vocalic suffix so that agreement becomes ambiguous. The other are non-finite forms such as *-saŋa* [CVB.FGR], which do not have agreement at all.

- **OtoS** (embedded O with matrix S-AGR): This alternation only occurs with *kond-* in the sense 'must'. The function is not fully clear yet, but there is some preliminary evidence that this alternation is used when the obligation expressed by *kond-* is characteristic of A but rather of O (e.g. 'they are such that one must like them'). Note that with 3sO the intransitive base frame `itr` can't be distinguished from the one with raising (`itr.OtoS`). Since OtoS is rather rare, bare `itr` should be used as the default in this case. `OtoS` is not possible with detransitivised subclauses, so it may be taken as indirect evidence that the embedded frame is fully transitive.

- **pass** (passive): S-NOM V-s(S). This alternation was introduced for the passive participle *-mayaŋ*, but the infinitive *-ma* may also sometimes use it. The verb is lexically transitive; S corresponds to the element linked to O-AGR in the transitive frame.

- **poss** (possessive): A-NOM P-NOM V-s(P). This alternation is to be used with intransitive verbs and verboids that are used to code possession, e.g. *yuŋ-* 'be there', *manchi?* 'be not there'. The possessor is interpreted as A and the possessum as P.

- **recp** (reciprocal): this alternation involves several referents doing something to each other and thus occupying two roles each. The way to mark this is to use the `+` sign and identical domains (e.g. `chace.1+1.A+P` = *cha* is A and P at the same time). Which roles are joined depends on the valency of the verbal base: A is joined with the role that triggers O-AGR. The reciprocal suffix *-ka* is non-finite and thus requires the auxiliary *lus-* to mark TMA and person/number. Assign `recp` to the verb marked by *-ka* and `aux` to *lus-*.

- **refl** (reflexive): unambiguously marked by reflexive inflection on the verb. `refl` works parallel to `recp` in that one NOM-marked referent is regularly found in two roles in the same domain, the difference being that there need not be several referents and if there are they are each occupied with themselves instead of each other. Which roles are found together follows the same rules as with `recp` above. *appi* 'oneself' is not to be tagged as an argument as it acts more or less like an adverb.

- **sad** (S/A detransitivised): any polyvalent frame where A is marked by NOM and there is no O-AGR. Formally often identical to S-NOM V-s(S), so make sure the verb allows for transitive inflection.

- **sod** (S/O detransitivised): S-NOM V-s(S). Assign this tag when there is an alternative transitive frame whose P/T/G corresponds to S. Watch out for ambiguities: transitive frames with dropped arguments may look identical!

A.4 Additional helps for assessing values

A.4.1 Special cases for identifiability and quantifiability

For certain formal and functional types of NPs, assigning identifiability and/or quantifiability can be confusing. Since both variables are often closely connected, these cases are listed together below.

- **Qualities**: words designing qualities (e.g. *halachop-* 'red', *the-*= 'big') are special because qualities tend to be perceived as (Aristotelian) accidences and not as substances. This means that they refer less frequently than other nouns. However, this does not mean that they cannot refer in principle (*the dead*, *a black one*). When they do refer they can be treated just like other nouns.

- **Quantified NPs**: quantified NPs (no matter whether there is a head noun or whether the quantifier itself functions as the head) are special because they have fixed quantifiability values. In Chintang, all numerals (together with their numeral-classified variants) as well as *jamma* 'all', *etti – tetti/utti* 'this much – that much' and *batta – motta – totta – yotta* 'this big – as big as down/up/over there' are quantifiable by default [1]. Other quantifiers are non-quantifiable by default, the most important one being *baddhe* 'much, many'.

- **Deictic NPs**: All deictics are `def` by default. Deictics are almost always quantifiable, but there are some rare exceptions (e.g. *To cuwa athuŋno?* 'Do you drink from the water up there?').

- **Personal pronouns**: Always `def.qnt`.

- **Interrogatives**: It is not relevant that with interrogatives the speaker does not know who or what he is talking about by definition. Imagine one speaker saw another talking to somebody and asked him *Who were you talking to?*. Here, *who* is clearly `def` because both speaker and hearer can identify what it refers to. In a different situation *who* might be `idf`, for instance if one speaker asked another one whom he talked at a party to without intending anyone in particular.

- **Types**: Many nouns can be used to signify both tokens and types, cf. for instance *There are two beers left in the fridge* vs. *The've only got two beers at the supermarket.* What is potentially confusing about the use referring to a type is that a type can have many tokens, which suggests non-quantifiableness (for instance, there could be many instances of the two beer brands). Remember in such cases that what's referred to is the type itself, not its tokens (so *two beers* is quantifiable in either case).

- **Places**: Definiteness is not a problem with places. Locative deictics (*huŋgoi?*, *moba*, *uyuba* etc.) are `def.qnt` by default (the intuition behind this being that when there is a *there* it is implied that there also is a *here* and a tentative boundary between the two).

- **Manner**: Some verbs license manner adverbs in core roles (especially *mett-*). Try replacing the adverb by an equivalent noun (*in a ... way*); if that doesn't immediately help use `xdef` and `xqnt`.

- **Events and facts**: These references are special because they are mostly not in predefined but in ad-hoc categories. Where it's not immediately clear what the identifiability and quantifiability values should be it often helps to add 'the event/fact that' to the referring clause (e.g. *I know that he was in the park* > *I know the fact that he was in the park/I know one fact, and that's that he was in the park*) or, where possible, replace it by a functionally similar noun (e.g. *I start walking* > *I start a/the walk*, *I can swim* > *I can do swims*). Be especially careful with this referent class and rather use `xdef` and `xqnt` than to guess when you are not sure.

- **Non-referring NPs**: There are a couple of different cases where NPs do not refer. Such NPs should be tagged `idf.nonq`.

    - **Category NPs**: This use occurs mainly in the copulative construction. The NP builds a category, but it's not necessary to select a matching referent because the category itself is being talked about, as in *My mother is ill.*

[1] Note that the *-tta* series can be used as quantifiers but (less frequently) also as modifiers (*ba-tta=kha tika* [PROX-EXT=NMLZ$_2$ mark] 'a mark this big').

- **Citation**: NPs and whole clauses which are cited from a different context do not refer (e.g. *What does "ma?mi" mean?*, *He said "She's not at home"*.).
- **Names**: Names are often used to refer (*Dipe has stolen my chocolate!*), in which case they are almost always `def.qnt`. However, names do not refer when they are attributed (*My name is Dipe*).

A.4.2 Complex sentences

Complex sentences pose several problems for assessing domains, roles, and predicate frames. Consider as a complex sentence any combination of verbs whose frames interact syntactically on a regular base. This can either mean that one verb is the predicate of a clause which in the other occupies a position that could also be occupied by a nominal argument or that the frames share one or several arguments. Adverbial subordination is hard to define in Chintang and is not relevant for the moment, so temporal subclauses with conjunctions and the like should be treated as separate sentences.

Important notes for interns: Tagging zeros and roles easily becomes very complex in most relative and complement clauses, so you can ignore the internal structure of these clause types. Here is a list of forms marking them:

- *ka-* [ACT.PTCP]: always marks relative clauses
- *-mayaŋ* [PASS.PTCP]: always marks relative clauses
- *-ma* [INF]: marks complement or relative clauses in the majority of cases (the main exception is the negative converb *mai--ma* [NEG--INF] with the meaning 'without doing')
- *=go* [NMLZ$_1$]: marks complement or relative clauses when attached to a verb. *=go* also frequently co-occurs with deictics but does not mark any kind of subordination there.
- *=kha* [NMLZ$_2$] (and various other glosses): sometimes marks relative clauses but mostly fulfills other functions

Another complement marker is *=mo* [CIT]. *Mo* does not pose any formal difficulties, though, so you can tag its internal structure of the clause it marks.

Regardless of the internal structure of nominalised clauses you should always tag the role it plays in the matrix.

A.4.2.1 Complex predicates

Sometimes it's not clear whether a noun and a Chintang verb with abstract semantics (especially *numd-* 'do', *lis-* 'be, take place') stand in a argument-predicate relation or form a single complex predicate. Examples are *khela numd-* 'do a game/play', *bihe numd-* 'do marriage/marry', *kama numd-* 'do work/work', *stat lis-* 'start take place/start'.

Since it's difficult to find good criteria for when there is a complex predicate, the default should be to assume that the noun is an argument of the verb (P with *numd-*, S with *lis-*). Only if there is a clear competitor for the same role in the form of an overt argument in the same clause or if O-AGR unambiguously points to such a competitor should a complex predicate be assumed. In this case the same procedure should be followed as with *-e lis-/numd-/mett-* (cf. A.4.2.4 below), that is, the noun receives the predicate tags and the verb is tagged as `aux`. Examples are *phon numd-* 'phone (sb.)' (lit. 'do phone'), *prastut numd-* 'present' (lit. 'make ready'), *bihe numd-* 'marry'.

A.4.2.2 Domains and core roles

There are two crucial differences between domain IDs in simple and in complex sentences. One is that there are subdomains, the other that arguments can belong to several domains.

- Subdomains directly map the hierarchical structure of a complex sentence. The verb that is not embedded constitutes the top-level domain and is marked by a single number just

like any verb in a simple sentence would be. Any verb that is directly embedded into the clause constituting the top-level domain is in a first-order subdomain. Such subdomains are marked by the top-level domain ID, a slash, and an additional following number. Further subdomains require further slashes and subdomain IDs. The most deeply embedding sentence I have found so far is *ban-e num-ma ni-ma kon-no* [build-V.NTVZ do-INF know-INF must-IND.NPST] 'one must know how to build it'. The hierarchical structure of this sentence – [[[[bane] numma] nima] konno] – is expressed in the domain system as `bane.1/1/1/1 numma.1/1/1 nima.1/1 konno.1`.

- This syntax also makes it possible to subordinate several subdomains under one matrix domain on the same level. For instance, *yuŋ-ma kina haĩ-ma kond-a-ŋs-e* [DEM-NMLZ sit-INF SEQ talk-INF must-PST-PRF-IND.PST] 'one must sit and talk' is `yuŋma.1/1 haĩma.1/2 kondaŋse.1`.

- Purposives and especially infinitives are often used without matrix verbs (*Kok ca-si!* [rice eat-PURP] 'Time to eat!', *Abo aŋ num-ma?* [now what do-INF] 'What to do now?'). In these cases it is not necessary to assume a zero matrix verb; instead, the non-finite verb should be assumed to constitute the top-level domain. The covert argument (the one that would be shared if there was a matrix verb) should be represented as a zero just like anywhere else. In case it should still desirable to somehow mark that a form without a visible matrix is subordinated, use `xdom` in addition to the slash syntax, e.g. `xdom/1` ('first subdomain of an unknown domain').

- Many constructions regularly feature referent sharing. Such constructions require special ID assignment rules. In the first step zeros should be inserted for all arguments that could be overt. In constructions where the shared referent may only be expressed overtly once it should also be represented only once (be it in the form of an NP or a zero). After this each argument is assigned ID and role tags:

    - If an argument *could* belong to several domains, it gets all corresponding IDs/roles joined by pipes `|`. This symbol should not be used across records, so if an argument in record 1 belongs to three domains one of which is in record 2 it will be assigned only the domains of record 1. In record 2 a new 0 has to be created. The same is true for referents that are shared between a citation (marked by *=mo*) and another clause.
    - If the argument *must* belong to several domains, it gets all corresponding IDs/roles joined by pluses `+`. This symbol should be used even across records, wherever there is a grammatical rule involving referent sharing.
    - If in the latter case the referent of the argument is shared but the argument itself unambiguously belongs to one domain, the domain that is semantically but not syntactically linked to the argument is marked by a star `*`.

A.4.2.3 Hybrid elements

A special problem is brought about by nominalising affixes. Recall that there are no specialised adnominal forms in Chintang. This means that all nominalised forms can become arguments without any further transformations. But at the same time these forms usually keep their own arguments, so they have to be tagged for predicate and argument variables at the same time. For instance, in *ma-ce-ŋa u-kukt-a-ŋs-a-c-e-go c-o-ha?* [woman-ns-ERG 3S/A-bring.down-PST-PRF-PST-d-IND.PST-NMLZ eat-3[s]O-IMP] 'eat the one the women brought down!', *ukuktaŋsacego* contains both a predicate belonging to the embedded clause and a referent occupying the role of P in the matrix and shared by the two clauses. In such cases the hybrid form gets the predicate tags of the one clause *and* the referent tags of the other. The two groups of tags should be separated by a colon, as in `ukuktaŋsacego.1/1.dido:1+1/1.P+T*.def.qnt`.

Sometimes it happens that a hybrid element occupies a peripheral role in the matrix clause (often with relative clauses, e.g. 'at the place where…'). Peripheral roles do not receive any referential

tags, so in such cases the hybrid element should only be tagged as the predicate of the embedded clause.

A.4.2.4 Overview of constructions involving complex sentences

Below is an alphabetically ordered list of all known constructions involving complex sentences, together with referent sharing and other relevant properties.

- ***-e lis-/numd-/mett-*** [V.NTVZ be/do/do]: Chintang auxiliaries are necessary to make a predicate out of a Nepali verb marked by the nativiser *-e*. These auxiliaries are marked by the label `aux`. All referents are shared and assigned to the subdomain containing the Nepali verb. What makes this construction a little tricky is that the frame is actually not determined by the Nepali verb but by the auxiliary (`itr` with *lis-*, `tr` with *numd-*, `tr` or `dioo` with *mett-*). Regardless of that, class and alternation should be marked on the Nepali verb so it is easier to determine which verbs are used with which way. Arguments are also linked to the Nepali verb.
- ***=go*** [$NMLZ_1$] and ***=kha*** [$NMLZ_2$]: these mostly mark relative clauses. The shared referent may be overt, but mostly it's not. Although it can never be overtly expressed twice it should exceptionally be represented in both clauses. When the shared referent is not expressed in the main clause, the form hosting *=go* or *=kha* itself should be taken as the direct argument of the matrix verb. This is where the colon syntax becomes necessary to separate predicate and argument tags. Apart from this *=go* and *=kha* can also mark complement clauses, but this use is less problematic.
- ***ka-*** [ACT.PTCP]: describes a referent as the S/A of a clause. The referent is always embedded into a matrix clause and therefore shared. The referent may be overt, but normally there is just the dummy *pa* 'man' which is conventionally analysed as suffix *-pa* [REF]. The default therefore is to assume that the participle itself is an argument of the matrix and to use the colon syntax to separate predicate and argument tags.
- ***-ka- lus-*** [RECP]: *lus-* is an auxiliary here, so all referents are shared and the two verbs form a single predicate. *-ka* has `recp`, *lus-* has `aux`; all tags are linked to the verb marked by *-ka*.
- ***-ma*** [INF]: the infinitive is used in various constructions with varying coreferentiality constraints. Here is just an overview.
    - **Infinitives marking complement clauses as P**: this group is special in many ways. Most matrix verbs taking P complements can also occur with nominal P, so it is assumed here that these verbs are basically monotransitive. Infinitival P complements require a shared S/A in embedded clause and matrix. Since agreement can only be realised in the matrix, S/A can be assumed to belong there syntactically.
      If the embedded verb is transitive, its own P wins over the whole clause in the competition for becoming the P of the matrix verb. This means that the embedded P is linked to matrix O-AGR. In this construction the referent in P is thus necessarily shared by both verbs but realised in the matrix.
      If the embedded verb is intransitive, the whole clause functionally becomes the P argument of the matrix verb. Depending on the class of the matrix verb as well as on additional factors that are not completely understood yet, the matrix verb will then be either inflected intransitively or transitively with dummy 3sO-AGR. Since both variants can be assumed to be semantically transitive, `of` should be used for the intransitive variant and the default for the transitive variant.
    - **Infinitive marking complement clauses in other positions**: not much is known yet about coreferentiality constraints in this construction type. The default should be to assume none. There is one prominent construction, though, which does have a constraint: *pid-* 'allow to' takes an infinitival clause as its T. The allowee (G) must be coreferential

with the S/A of the embedded clause. The allowee is always realised as an argument of the matrix clause. Infinitival subclauses with *kond-* and *lis-* in the meaning 'must, should, need (to)' should be treated as S.
    - **Non-complement infinitives** generally seem to have no coreferentiality constraints. Known counterexamples are (probably!) *lemd-* 'convince (to)' and *phad-* 'help (to)' (matrix P = embedded S/A) and *kat-* 'come up' (used with feelings, e.g. *suma kat-* 'be lazy (to)' (literally 'laziness comes up'); matrix experiencer/S = embedded S/A).
    - Two verbs, ***kond-* and *lis-***, occur with a special infinitival construction. The more frequent option for the infinitival clause is to behave as S of the matrix verb, in which case it is just a normal complement. However, it is also possible to link embedded P/T/G to S-AGR of *kond-*/*lis-* (= agreement raising). In this case the matrix verb should be assigned the special alternation tag `OtoS` (on top of `itr`). The arguments should still be assigned to the domain of the embedded verb.

- ***-ma*** [INF] in nominal use: this use cannot realise arguments in the same way predicates can and thus should not be tagged as a predicate but as a referent. Nominal infinitives tend to get lexicalised and therefore are usually easy to recognise. Examples are *cama* 'food' and *hupma* 'lid'.

- ***mai--ma*** [NEG--INF]: this combination (meaning 'without doing') has different syntactic properties from the bare infinitive in that it can be embedded into any matrix and does not have any coreferentiality restrictions.

- ***-mayaŋ*** [PASS.PTCP]: describes a referent as the P/T/G of a clause. The referent is always embedded into a matrix clause and therefore shared. So far overt referents are only attested with an additional *=go* [$NMLZ_1$].

- ***=mo*** [CIT]: when modifying clauses, *mo* marks citations occupying the position of P or T in the matrix. There are no obligatorily shared arguments. If the *=mo* clause clearly occupies a role, tag it as if it were nominal but with `idf` and `nonq` as standard values for identifiability and quantifiability. *=mo* is often used without any recognisable superordinate verb, most frequently in the combinations *mo kina* [CIT SEQ] 'saying that, thinking that, in order to' and *mo para* [CIT COND] 'if, supposed that'. In such cases the verb before *=mo* should not be tagged as subordinate but as the top-level domain.

- ***-saŋa*** [CVB]: subclause and matrix share S/A (i.e. S/A in one clause must have the same referent as S/A in the other). The shared referent is almost always expressed in the matrix.

- ***-saŋa yuŋ-/numd-/mett-/khat(t)-/thap(t)-*** [CVB.FGR be/do/do/take]: in this construction the light verbs used together with the converb function as aspectual markers, e.g. *wei? ta-saŋa numd-a-ŋs-e* [rain come-CVB.FGR do-PST-PRF-IND.PST[.3sS]] 'it's been keeping raining'). If the matrix verb is intransitive it simply has S-AGR with the embedded S/A, so the alternation of the embedded verb is `idt`. However, if the matrix verb is transitive it may have S-AGR with embedded S (as in the example), but also A+O-AGR with embedded A and P/T/G. S/A detransitivisation also raises into the matrix. That means that in this case alternation of the embedded verb become visible in the matrix. The matrix itself always gets the class `aux` and no alternation. All arguments are assigned to the subdomain around the *-saŋa* form.

- ***-si*** [PURP]: embedded S/A must be coreferential with a moving argument in the matrix (S or T). The shared referent is almost always expressed in the matrix.

Appendix B

Annotation guidelines Nepali

B.1 How to tag files

B.1.1 Introduction

Each sentence is tagged in two steps:

1. The first step is to **identify referents** and **analyse syntactic structure** as described below. Each referent is assigned a referential ID, and syntactic dependencies are annotated using domains and roles. Empty elements are inserted for zero arguments.
2. The second step is to **identify objects** eligible for DOM and to **tag** them for seven referential variables: part of speech, modification, animacy, situation, quantifiability, and focus.

You can tag all variables at once or one variable at a time. The first method might be a little bit slower but is also less cumbersome because you won't have to read through the file several times.

Not all elements receive the same tags - most of them only need a few. Altogether there are five classes of taggable elements:

- All referents need a referential ID.
- If a referent is an argument it also needs a role and domain tag.
- If an argument is P, T, or G it also needs a DOM tag.
- Only if a P/T/G is eligible for DOM and is marked by NOM or DAT does it need the remaining referential tags.
- Predicates only need a domain and optionally an diathesis tag.

Most variables are to be tagged on the morphosyntactic head of NPs. So, for instance, if you see something like मेरो ठूलो बुवाको घर, only घर is to be tagged. In case of long names such as तोया नाथ भट्ट, only tag the last element of the group (the motivation being that it is this element that can receive case marking, as in तोया नाथ भट्टलाई). In case of multi-word verb forms such as भन्नु भएको छ, place the tag on the word carrying the semantic content (e.g. in this case भन्नु).

B.1.2 NNC XML

The parts of the Nepali National Corpus that we work with are in **XML**. For working with texts you don't have to know a lot about XML, but it's useful to know that words are embedded in word tags of the form `<w>...</w>` and sentences in sentence tags `<s>...</s>`. Properties of words are specified as **attributes** within the tags. For instance, `<w animacy="human">`राम`</w>` indicates a human referent marked by the word राम. Here is a short example for a fully annotated sentence:

```
<s>
   <w domain="27" role="A" identity="Peter">उसले</w>
```

```
   <w domain="27" role="P" DOM="DAT" identity="Sally"
      animacy="human" quantifiability="qnt" modification=
      "none" focus="nofoc" situation="concrete">मलाई</w>
   <w domain="27">चिनेन</w>
   <w>।</w>
</s>
```

In XML, possible values of attributes and other characteristics of a specific standard are stored in a special format called DTD. For this project, there is a DTD called `xcesDofRef.dtd`. When you work with oXygen (see below), you can use the DTD to make annotation easier and more consistent. However, for this the DTD has to be in the right place. The path that is expected by all corpus files complying to the DTD is `../system files/xcesDocRef.dtd` (i.e. "from where the file is lying, go up one folder and then into the folder called `system files`"). Make sure to place the DTD there.

B.1.3 The oXygen editor

Tagging is most easily done in the oXygen editor. Here is how to use it.

- Open the editor. Open a file by selecting File > Open...
- After you open the file several windows will appear. There are three important ones:
    - The **outline window** on the lower left shows the logical structure of the text. You can explore the structure of the text by clicking on the small arrows to the left of each element shown in this window. This way you will get from the topmost element `<text>` to `<body>`, `<div>`, `<p>` (paragraphs), `<s>` (sentences), and finally to `<w>` (words).
    - The **main window** in the middle shows the XML text contained in the file. It's not as easy to read as the outline because there is no automatic structuring, but you may sometimes prefer to read the text in this window because you don't have to expand every element by clicking on it but can simply read through the text from the beginning to the end.
    - The **attribute window** on the upper right shows the attributes of the currently selected element. Each tagging variable corresponds to an attribute, so you will use this window a lot.
- You can read the text in the outline or in the main window. If you read it in the outline you'll have to click on the mentioned small arrows one after another to get to the words of the text.
- The relevant part of the file starts under the element `<text>`. Search this element and start reading below it.
- Inside the text the relevant elements are `<s>` (sentences) and `<w>` (words). Work through the text sentence-wise and check every word in every sentence.
- If you want to tag a word you have to select it. You can do this by clicking on it in the outline or in the main window.After you select an element, its attribute list is displayed in the attribute window. Determine the value for each tagging variable by clicking on the name of its attribute and then double-clicking on the cell to the right to it in order to insert a value. Most variables provide a dropdown list from which you can choose a value, but others require you to type text. Note that this will only work if the DTD `xcesDocRef.dtd` is in the right place (see above).
- **Repeated utterances** should not be tagged if they don't add anything new to what's already been said. Typical cases are one speaker repeating what he has just said while thinking about what to say next, or one speaker repeating what the other said for confirmation. However,

for instance, cases where one speaker asks a yes/no question and the other gives the answer using exactly the same or very similar verbs do not count as mere repetitions because the second utterances adds new information (the answer to the question was not known before). In such cases both utterances should be tagged. Yes/no questions where the asker repeats part of what he has said only changing polarity ("Have you eaten rice or haven't you eaten rice?") do count as repetitions.

B.2 Variables, carriers, and values: overview

This section is an overview of all variables with their possible values and is meant as a quick reference. Each variable corresponds to an attribute type. See section B.3 for all details.

B.2.1 Domains

carrier: verbs (word carrying semantic content in multi-word forms) and heads of argument NPs
values:

- 1 = belongs to domain 1
- 2 = belongs to domain 2
- 3, 4, 5… = …
- 1/2 = belongs to domain 2 under domain 1 (in complex sentences)
- 1+2 = belongs to domain 1 and 2 (shared arguments)
- x = insecure

B.2.2 Referential identity

carrier: head of any NP (on zeros only if core argument)
values: arbitrary unique ID for every referent; can be number or code. "x" is reserved for 'insecure'.

B.2.3 Role

carrier: head of argument NP
values:

- S = intransitive subject
- A = agent
- P = patient
- T = theme
- G = goal
- CT = theme of copular sentence
- CR = rheme of copular sentence
- x = insecure

B.2.4 DOM

carrier: head of P/T/G NP
values:

- NOM = nominative (zero)
- DAT = dative (-लाई)
- NA = non-applicable
- GEN = genitive (-को)
- x = insecure

B.2.5 Part of speech

carrier: head of eligible DAT/NOM NP
values:

- n = noun
- adj = adjective
- pro = SAP pronoun
- dem = demonstrative
- poss = possessive
- other = other part of speech
- x = insecure

Note: the biggest part of the corpus already has already been automatically tagged for this variable (with more values, but they can be easily mapped to our system). The pertaining attribute is named “ctag”.

B.2.6 Modification

carrier: head of eligible DAT/NOM NP
values:

- none = no modifier
- adj = adjective
- relclause = relative clause
- poss = possessor
- humposs = human possessor
- latposs = latent human possessor
- num = numeral
- dem = demonstrative modifier
- interrog = interrogative modifier
- several = several modifiers
- sortal = sortal demonstrative modifier
- sortal.q = sortal interrogative
- other = other modifier
- x = insecure

B.2.7 Animacy

carrier: head of eligible DAT/NOM NP
values:

- human = human
- human.fam = human family member
- human.prop = human proper name
- human.group = group of human beings
- high.anim = non-human higher animate
- high.anim.prop = non-human higher animate proper name
- mid.anim = non-human middle animate
- low.anim = non-human lower animate
- thing = inanimate concrete
- state = inanimate abstract static
- process = inanimate abstract dynamic
- x = insecure

B.2.8 Situation

carrier: head of eligible DAT/NOM NP
values:

- concrete = concrete situation
- exemplary = exemplary situation
- general = general situation
- abstract = no situation
- x = insecure

B.2.9 Trackability

carrier: head of eligible DAT/NOM NP
values:

- qnt = determinable quantity
- nonq = non-determinable quantity
- x = insecure

B.2.10 Focus

carrier: head of eligible DAT/NOM NP
values:

- nofoc = focus not specified
- contrast = contrastive focus
- fragile = fragile utterance
- x = insecure

B.2.11 Diathesis

carrier: predicate or word carrying semantic content in multi-word forms
The only value is "passive". For the default ("active") no diathesis needs to be specified.

B.3 Variables, carriers, and values: details

B.3.1 Domains

A domain is a set of forms that interact morphosyntactically and can correspond to various syntactic concepts, e.g. a clause or a sentence. A minimal domain consists of a predicate only, but usually arguments are associated as well. In general the larger the domains one looks at become, the less interaction is grammaticalised, so larger domains are much harder to define than small domains. For the present purpose it is not necessary to define domains larger than those approximately corresponding to sentences.

All elements belonging to one domain have the same domain ID, and elements in other domains have different IDs. Note that the existing sentence tags (<s>) in the NNC have been assigned on the basis of orthographic criteria (punctuation marks) in the written corpus and (probably) on the basis of intonational criteria (pauses) in the spoken corpus. Both criteria are not necessarily congruent with morphosyntactic criteria, so one "sentence" (in the sense of the tag) may contain several domains, or one domain may span over severall sentences and end and start in the middle of sentences.

NPs without a pertaining verb should only receive separate domain IDs when it is clear that they do or could have belonged to a separate domain than the surrounding verbs. For instance, in त्यसपछि मैले त्यसलाई, अँ, म त्यसो गरेर... the first half could have been a separate domain - it's

only the verb that is missing. In such a case the elements मैले and म should be assigned to different domains (and both receive referential IDs). However, in a case like तर मलाई, अँ, मलाई त्यो कुरा थाहा थिएन it's clear that the two मलाई belong to one domain and have only been repeated due to hesitation. Here, only one domain should be used (and only one मलाई should be assigned an ID).

See section B.4 for a number of special cases of domain assignment occurring in complex sentences.

B.3.2 Referential identity

This variable is special because it does not have fixed values and applies to all NPs (not only arguments!). Whenever you encounter a new referent in a text you are free to assign any value to it that is not already in use for a different referent. However, in order to make IDs easier to memorise, it is advisable to use codes providing hints to the identity of the referent. For instance, some of the characters featuring in the Ramayana might be given the codes `ram` (Ram), `sit` (Sita), `han` (Hanuman), and so on. Whenever a known referent comes up again later, the same ID should be used as the one given to it first.

One tricky case is grouped referents. If a referent is first introduced in a group and then comes back alone or the other way round there is of course some intersection between the two sets; still, they are not fully identical. In order to solve this dilemma the following strategy is adopted:

- Groups can be referred to by normal IDs or by joining several individual IDs by plus signs. For instance, if Ram has the ID `ram` and Sita has `sit`, the couple of them could be referred to as `coup` (or any other unused ID) or as <ram+sit>.
- The second method should be used whenever the hearer is likely to know that certain individual referents are members of a group. This is almost always the case when a referent is first mentioned alone and then in a group. For instance, if Dipak (`dip`) is sitting alone in a cafe, is then joined by Nagendra (`nag`) for a small conversation, and they both go out together, 'they' should not be given a new tag but should be tagged `dip+ram`.
- The opposite is the default in the reverse case where a group is mentioned first and then a referent gets singled out. For instance, if the speaker tells a story how he saw several people waiting at the bus station and only then recognised Devi among them, 'people' and 'Devi' should have different non-composite tags.
- Remember that both methods are only defaults. There may be (rare) cases where a referent joins a group secretly without the hearer knowing it, and there are quite a few cases where the hearer does know some referent is a member of a group though that referent hasn't been mentioned before in isolation.

Referents are frequently covert. Covert (or "zero") referents should only be tagged for referential identity if they are arguments, that is, S, A, P, T, G, or either of the two arguments constituting a copular sentence. See B.3.3 below for definitions of these.

You can check for the covert presence of arguments by trying to insert them. For instance, तिमीले पनि देख्यौ ? looks like there is only a subject. However, it would also be possible to say तिमीले पनि त्यो मान्छेलाई देख्यौ ? without altering the meaning. This means that there is a zero object in the first version. In order to be able to keep track of referents, zero arguments have to be tagged to. For this you first have to insert an empty element (i.e. a word without any overt content) into the corpus text. There are two ways to achieve this:

- Right-click on the word next to which you want to insert the zero element and choose `Insert Before > w`.
- Click in the main window where you want to insert the zero element and type `<\textbackslash w>`.

You can then click into the opening `<w>` tag and assign tags to it as to any normal word. Note that in constructions with obligatory referent sharing no zeros should be inserted and the shared

referent should only be ID-tagged once. For instance, most complement verbs such as थाल्नु have obligatory S/A sharing (the one who starts to do something is also the one who does it). Thus, in a sentence like उनी रुन थाली no extra zero is necessary just because there are two verbs, and one referential ID for उनी is sufficient. See section B.4 for more details.

In case you feel that some referent is very unlikely to be tracked in discourse, you are free to choose a dummy ID (e.g. just N1, N2 for various NPs). Where one noun comes up again and again in subsequent sentences but signifies a different unimportant referent each time you should simply use the sentence ID to keep the referents apart. For instance, newspaper articles often make use of the phrase भन्ने कुरा बतायो. Each instance of कुरा is a distinct but rather unimportant referent that is unlikely to come up again. In order to avoid incidentally using the same ID to some of them it is easiest to use IDs such as `<kura27>` in sentence 27, `<kura28>` in the next, and so on.

Here are some special types of nouns/referents that have proved to be tricky in some respect:

- In some cases it is not clear whether nouns refer at all - some semantic theories would say they don't. This is, for instance, true of **indefinite non-specific** NPs as in *I like **dogs*** and **themes of copular sentences** as in *My father is a **policeman***. We work with a maximal conception of referentiality, i.e. we assume referents in these cases, too - so don't forget to assign referential IDs. Sometimes such special referents even get tracked, cf. e.g. *I like dogs, how about you? - I can't stand them.* or *Her father is a politician, and that politician is corrupt, but he himself is a good person.*), so having IDs for them does no harm.

- Sometimes it may be clear that there is a role slot but it's completely unclear who or what occupies that slot. This is, for instance, often the case with the A of passives, as in चोर हेरियो. Use IDs of the format "whoever" or "whatever" combined with the sentence ID (e.g. `whoever21`, `whatever178`).

- Don't let yourself get confused by **deictic reference**. One referent can be referred to in various ways, e.g. by a name or by a **pronoun**, but it should always have the same ID. For instance, if a radio moderator first introduces his guest by name and later on addresses him as तपाईं, the name and तपाईं should be connected by the same ID.
  This is also true for **interrogative and relative pronouns**. For instance, if A sees that B is listening to a song but does not know what song and thus asks के सुन्नुहुन्छ ?, के should have the same ID that is given to the song because it is the first NP that refers to that referent. Of course it is advisable in such cases to use the name of the song as the ID for के and not the other way round (i.e. not to refer to the song as `what`) in order to avoid confusion.

- **Adjectives** should only receive IDs when they occupy an argument role or in copular sentences, where you can use dummy IDs (e.g. कागति अमिलो छ in sentence 8 gives the values `lemon CT` and `amilo8 CR`).

- **Genitive NPs** should receive an ID (e.g. नेपालको in नेपालको बिकास), but the first parts of compounds should be ignored (e.g. गलैंचा in गलैंचा उद्योग '(the) carpet industry').

- Do not ID-tag **postpositions**.

- Combinations of adjectives and verbs such as तयार गर्नु or फरक पार्नु should always be interpreted as complex predicates.

- Where a noun or a full NP is **repeated** without information being added (e.g. because the speaker is hesitating or wants to emphasise a point) the repeated forms should not have referential IDs. For instance, in a sentence like अँ, त्यो नून... त्यो नून हामीले खाँदैनौं only one of the two त्यो नून needs an ID.

- You can ignore nouns in **fixed phrases** such as एउटा कुरा के हो भने.

- Do not tag NPs within text sequences that have to be interpreted in a different context. A couple of frequent cases:

- **Metalanguage words**. For instance, don't tag कोदो in कोदो भनेको के हो but do tag भनेको (= the word, not 'millet') and के.
- **Titles** of books and the like. For instance, don't tag मान्छे and माया in the song title एउटा मान्छेको मायाले कति. Instead, insert a zero element with the ID of the song and tag that element.
- **Cited speech**. For instance, don't tag अफिस in अफिसमा छैन भनेर भन्यो (but again do insert a zero for the T of भन्नु!).
- **Proverbs**.

B.3.3 Role

The semantic role assigned to arguments. The following roles exist:

- **S**: the only core role specified by an intransitive verb
- **A**: the most proto-agent-like core role specified by a mono- or ditransitive verb
- **P**: the most proto-patient-like core role specified by a monotransitive verb
- **T**: the most "proto-theme"-like core role specified by a ditransitive verb (= the non-A argument that is moving in space or metaphorically)
- **G**: the most "proto-goal"-like core role specified by a ditransitive verb (= the non-A argument that is stationary)
- **CT**: theme in copulative constructions (= what is talked about)
- **CR**: rheme in copulative constructions (= what is predicated)

These definitions represent a Dowtyan/Bickelian approach to roles (Dowty 1991, Bickel 2011). See Haspelmath (2011) for an overview about this and other approaches. The most important characteristic of this approach for our purposes is that role depends on valency. Verb senses which have identical valencies use the same role set, and different role sets must be used for verbs with different valencies. This leads to a couple of unexpected role assignments. Here are the most important examples:

- Most **destinations** count as P. For instance, 'bank' is P in ऊ बैंक(मा) गयो.
- **Destinations** count as G when there is an A and a moving object (T). For instance, in मैले यसलाई काठमाण्डु(मा) पठाएँ, 'he' is T and 'Kathmandu' is G.
- **Instruments** count as T when they form part of the valency. For instance, in मासु चक्कुले नकाट !, 'knife' is T and 'meat' is G.
- **Experiencers** count as S or A, depending on whether there is a stimulus. For instance, 'you' is S in तपाईंलाई भोक लाग्यो ? (while 'hunger' should be ignored as part of a complex predicate भोक लाग-).
- When **possession** is predicated, the possessor is A and the possessum is P. For instance, in उसको घर छ, ऊ is A and घर is P (with DOM="NA").

Some phenomena pose difficulties for the assessment of roles. Here is how to deal with them:

- Sometimes there is what seems to be arguments but there is **no verb**. The most frequent case in the NNC are headlines and similar structures. If there is a noun (e.g. प्रचार) or an adjective (e.g. प्राप्त) marking the predicate and it is still possible to assess arguments and objects, do so. However, if you are in any way insecure, ignore the sentence.
- Objects can behave like subjects in some respects when the pertaining verb is turned into a **passive**. However, even if the argument in question becomes S in this process it should still be tagged like a normal object. For instance, Hari is a normal object in रामले हरीलाई मार्यो ।. If we make a passive out of this (हरीलाई (रामद्वरा) मारियो), Hari may gain some S properties but should still be treated the same way as in the underived sentence, i.e. as object.

- **Causatives** introduce new A and thus change the set of arguments to be tagged. For instance, in बच्चाले ढिंडो खायो । there is only one object, 'porridge'. By contrast, there are two objects in the causative variant आमाले बच्चालाई ढिंडो खुवायो, viz. 'porridge' (T) and 'child' (G, formerly A).
- Some verbs, e.g. **verba dicendi**, may have a whole clause as their argument, as in *He told me he'd be coming* (*he* = A, *me* = G, subclause = T) or *I know that the earth is a disk* (*I* = A, subclause = P). In such cases, there are two options. If there is a marker of subordination (e.g. a complementiser as in मललाई थाहा छ **कि** पृथ्वी सम्म छ or a citation particle as in म आउँछु **भनेर** भन्यो), that marker should get the necessary tags (including `DOM="NA"`). If there is no marker, insert a zero instead.

B.3.4 DOM

The central question of this project is: when are arguments in P/T/G marked by the nominative (zero, as in उसले कम्प्यूटर किन्यो) and when are they marked by the dative (-लाई, as in उसले हरीलाई भेट्यो)? This phenomenon is called DOM ("differential object marking", going back to Bossong 1998). All P/T/G NPs have to be marked for whether DOM can be observed on them and if yes, what the case marking is. P/T/G which are eligible get one of the attribute values `NOM` (nominative), `DAT` (dative), or `GEN` (genitive, rare). P/T/G which are *not* eligible get `DOM="NA"` (= non-applicable). There is a variety of reasons why DOM can be non-applicable:

- All zeros are `NA` since they are ambiguous with respect to case marking.
- Most G do not participate in the NOM/DAT alternation and are therefore `NA`. The two most frequent G types also belong here: G of verbs of the type दिनु are invariably marked by DAT, and G of verbs of the type लानु are marked by NOM or LOC but never by DAT.
- Destinations of verbs of motion (default frame {A-NOM P-NOM V-s(A)}) do not participate in the alternation. An example for जानु has already been given above (ऊ बैंक गयो).
- Instrumental T marked by ERG are always `NA`. For instance, चक्कु in उसले मासु चक्कुले काट्यो is considered T here but cannot be marked by -लाई.
- Sentences with "dative subjects" can be ignored. For instance, in a sentence like मलाई रुघा लाग्यो, मलाई would have to be judged as A and रुघा as P. However, रुघा does not participate in the NOM/DAT alternation (nor does मलाई, whose DAT is fixed).
- Vocatives are `NA` because they often represent a certain referent but do not get case from any verb. For instance, in दाइ, म अझै दिऊँ ?, दाइ of course represents the G of दिनु, but it's invariably in the nominative.
- Infinitives often occupy a P-like position in complement clauses (e.g. गर्न सक्नु, गर्न थाल्नु). They should receive a dummy ID (e.g. `garnu95`) and a role marker (P), but the DOM value will always be `NA` (गर्नलाई सक्नु etc. being impossible). The same is true for complement clauses without referent sharing, as in मैले तिमीलाई सुतेको हेरें.
- A couple of verb senses trigger an invariable frame {A-ERG/NOM T-NOM G-DAT}. Since these verbs do obviously not participate in DOM, the value for both T and G should be `NA`. The verbs in this class that are known so far are:
  - ठान्नु 'consider G a T'
  - तुल्याउनु 'make G into a T'
  - बताउनु 'announce G to be a T'
  - बनाउनु 'make G into a T'
  - भन्नु 'call G a T'
  - मान्नु 'believe G to be a T'

- सम्झनु 'consider G a T'

In principle all cases for NA could be determined automatically, but this is not trivial due to technical problems. There is no syntactic parser or electronic valency dictionary for Nepali, and orthography is inconsistent. For this project it is therefore easier to simply specify the DOM value manually.
Below are a couple of examples for DOM annotations.

- रामले हरीलाई भेट्यो । 'Hari' is P/DAT.
- रामले गाईलाई घाँसा दियो । 'Grass' is T/NOM, 'cow' is G/NA.
- रामले हरीलाई लौरीले कुट्यो । 'stick' is T/NA, 'Hari' is G/DAT.
- रामले उसको छोरालाई स्कुल पठायो । 'son' is T/DAT, 'school' is G/NA.

The DOM variable is not only of central importance for the project but also for tagging: objects that have been tagged as NA do not need any further tags.

B.3.5 Part of speech

The most important part of speech is nominals (because most objects are nominals). Other parts of speech need not be distinguished internally. The biggest part of the corpus has already been tagged for parts of speech automatically, so if an element already has a tag you can leave it as it is. The predefined tags are those defined by the Nelralec project (see Hardie (2005) for documentation). The tagset below is much simpler. All Nelralec tags can be uniquely mapped to one of the values listed below (but not the other way round).

Note that parts of speech are understood as purely lexical categories here and not, as often in computational linguistics, as mixed categories between lexicon and syntax. This means that syntactic processes like nominalisation are irrelevant to part of speech (a nominalised adjective is still adj, a nominalised verb is still other).

- **n** (noun): a word that regularly takes case markers (e.g. -ले, -लाई, -मा) without derivation and that is not in any of the other nominal classes listed below, e.g. घर, केटा, खोकी, संस्कृति... Nelralec NN.
- **adj** (adjective): a word that can modify a noun without derivation and that is not a demonstrative or a possessive, e.g. राम्रो, ऐतिहासिक, बढी... Nelralec JM, JF, JO, JX, JT, MOM, MOF, MOO, MOX.
- **pro** (pronoun): one of the words म, हामी, तँ, तिमी, तपाईं, आफू, को. The following words also count as pronouns when they are used to refer to persons: हजुर, यहाँ, सर्कार, मौसुफ. Nelralec PMX, PTN, PTM, PTH, PXH, PXR, PRF.
- **dem** (demonstrative): one of the words यो, त्यो, ऊ, यिनी, तिनी, as well as their focussed forms यही, त्यही, उही and their oblique stems यस, त्यस, उस (as in यसमा, त्यसैले, उल्लाइ etc.). Nelralec DDX.
- **poss** (possessive): any of the following words: मेरो, हाम्रो, तेरो, तिम्रो, आफ्नो. A word followed by the genitive marker -को is also to be tagged as possessive. Nelralec PMXKM, PMXKF, PMXKO, PTNKM, PTNKF, PTNKO, PTMKM, PTMKF, PTMKO, PRFKM, PRFKF, PRFKO, PMXKX, PTNKX, PTMKX, PRFKX.
- **other** (other part of speech): all other words, including numerals, various deictic forms, verbs, adverbs, and conjunctions. These are rarely found in object position, so this tag will not be a frequent one. Nelralec D- (except DDX), V-, R-, I-, ML-, MM, C-, TT, QQ, UU, Y-, F-, NULL.

Don't forget that only the head of an NP needs to be tagged. So for instance, in an NP like त्यो सानो बिरालोलाई it is not necessary to tag त्यो and सानो. Such elements only have to be tagged if they are the head of an NP themselves, such as in तेललाई पनि हेर त or त्यो सानोलाई देऊ. We do not assume zero heads, with one exception: if the head of an NP is formed by a possessor as in मैले सीताकोलाई देखें 'I saw Sita's (daughter)', an ID clash results: should the ID be assigned according to the possessor (Sita) or the covert possessum (daughter)? In such cases you should create a zero possessum and assign matching IDs to both.

B.3.6 Modification

This variable describes whether there is a modifier in case the NP in question is complex. Several types have to be distinguished:

- **none** = no modifier
- **adj** = adjective, e.g. राम्रो, जपानी, अबोध्य...
- **relclause** = a relative clause preceding or following the head, e.g. थाहा हुने (मान्छे), मैले सिकेको (भाषा), (फूल) जसलाई पूजामा प्रयोग गरिन्छ... Relative clauses coding latent possession are not `relclause` but `latposs` (see below).
- **poss** = possessor. This is a non-human NP followed by the genitive marker -को.
- **humposs** = human possessor. A human NP followed by the genitive, or a possessive pronoun.
- **latposs** = latent human possessor. This is functionally similar to `humposs`, but there is no overt possessive pronoun or genitive NP. Instead, possession is expressed in alternative ways, e.g. आफूले अंश पाएको सम्पति or मसँग भएका क्वलिटिहरू.
- **num** = numeral. A numeral such as एक, एउटा, एकजना, बयालीस.
- **dem** = demonstrative modifier. One of the words यो, त्यो, सो.
- **interrog** = the interrogative modifier कुन.
- **sortal** = one of the sortal demonstrative modifiers यस्तो, त्यस्तो, उस्तो.
- **sortal.q** = the sortal interrogative कस्तो.
- **other** = other modifiers of various types, such as त्यत्रो, सब, केही, कुनै...
- **several** = several modifiers, e.g. a possessor, a demonstrative and a numeral in मेरा यी दुई आँखाहरू.

Simple NPs (no matter what their part of speech is) are always `none`!

B.3.7 Animacy

Animacy is based on the intrinsic power of things and living beings: how much can they do in this world? Can they easily do things to other things or are they powerless, can they move around freely or are they bound to a certain location? Here are the definitions for the animacy categories used in this project:

- **human** (human): to be used for human and humanoid referents (which are not `human.fam` or `human.prop` - see below). Humanoid referents are those which do not belong to the biological species homo sapiens but resemble its members in behaviour and abilities (and mostly also appearance). Typical examples are ghosts and deities. Speaking animals as they are often found in fairytales are also to be classified as humanoid. Apes are not humanoids but normal animals.

- **human.fam** (human family member): any relative (e.g. mother, brother, son) or in-law (e.g. father-in-law, husband).
- **human.prop** (human proper name): a proper name designating a human or humanoid referent.
- **human.group** (group of human beings): a noun designating a group of human beings (e.g. clan, company, government). There are rare cases where both `human.prop` and `human.group` apply, e.g. *The Red Cross*. Use `human.prop` in such cases.
- **high.anim** (non-human higher animate): this class is formed by mammals (e.g. dogs, horses, mice) and birds, unless they are designated by a proper name (> `high.anim.prop`).
- **high.anim.prop** (non-human higher animate proper name): non-human higher animates designated by a proper name.
- **mid.anim** (non-human middle animate): all other animals are in this class, so reptiles, amphibians, fish, insects, worms and similar animals all go here.
- **low.anim** (non-human lower animate): this class is for non-animals such as plants, mushrooms, bacteria, and viruses.
- **thing** (inanimate concrete object): any object that is in none of the various animate classes and can be touched (at least in theory), e.g. tables, gold, hands, stars, hard drives.
- **state** (inanimate abstract static "object"): referents that cannot be touched and that can be defined independently of time, e.g. jaundice, provision, vacuum, problem, liberty.
- **process** (inanimate abstract dynamic "object"): referents that cannot be touched and that can only be defined with reference to time, e.g. music, imagination, education, marriage, mistake. If a referent can be used with गर्नु it is usually a process.

B.3.8 Situation

The variable situation replaces the earlier "tricky" variable identifiability/definiteness. Identifiability is likely to be a composite concept. Important components that are not part of the standard definition are quantifiability (see below) and mental spaces, both of which serve as prerequisites for identifiability proper. The composite nature of identifiability and the strong tendency of speakers of article languages to identify identifiability with (linguistically marked) definiteness made this variable untenable. It is hoped that situation has a clearer definition and is less easy to mix up with linguistic phenomena.

Situation expresses to what degree the event coded by an utterance is embedded into a concrete situation. The basic distinction is between yes and no, but various intermediate degrees are recognised:

- **concrete**: the event in question has a clear time and place, e.g. *I phoned her yesterday.* Repeated events also belong here as long as they form a block (no other events intervening, e.g. *I tried and tried, but it wouldn't work*) or as each of them has a clear time and place (e.g. *I phoned her several times yesterday*). States belong here when a certain stretch of them is picked out (e.g. *The bread is on the table (now)*).
- **exemplary**: there is a single event that could have a clear time and place, but it is perceived as typical and thus representing a whole series of events, e.g. *A dog would start barking in such a situation..*
- **general**: the event may have a clear place, but it does not have a clear time. It is either regularly repeated (e.g. *He used to drink a glass of wine in the evening*) or applies in general and is thus timeless (e.g. *Cats catch mice*). States also belong here when they are viewed as a whole and not divided into stretches (e.g. *The river Rhine originates in the Alps*).

- **abstract**: the event does not "take place" (i.e. it has no specific place) but is located in the sphere of ideas, e.g. *Ice is cold.*

B.3.9 Quantifiability

Quantifiability indicates whether the quantity of a referent can in principle be determined or not. This quality is crucial for tracking referents because only referents whose borders are fixed can be identified across utterances.

There are two main types of nominal concepts that behave differently with respect to quantifiability:

- Homogeneous (= mass) concepts are such that their category label can be applied to arbitrary partitions of referents they apply to. For instance, if there is a large body of water and one separates a subamount from it that amount will still be categorised as *water*. Concepts of this type are usually non-quantifiable and can only be made quantifiable by special means, usually by containers (as in *a glass of water*).

- Heterogeneous (= count) concepts have parts that belong to a different category than themselves. For instance, the head of a dog cannot itself be called *dog*. Concepts of this type are usually quantifiable, especially if there is only one, but can be "massified" when their parts do not matter (as in *Have some dog (meat)!*) or where they occur in large groups that cannot be easily overlooked (as in *People use to walk their dogs around here*).

Quantifiability is certainly the most unusual variable used in this project, so don't hesitate to ask questions and/or use the tag x whenever you are insecure. Here is a couple of hints to one or the other value:

- Numbered or measured referents are always quantifiable.

- Referents in containers are usually quantifiable, especially masses (*glass of water*, *bowl of rice* etc.).

- Where referents occur in a large group that is difficult to overlook, that group tends to be non-quantifiable.

- Referents with indefinite quantifiers such as धेरै, प्राय are usually non-quantifiable as they mark large groups. However, universal quantifiers such as दुबै or सबै always have quantifiable reference. The reason for this is that they allow no digression from a certain number (viz. the complete number), even if that number is often not known. For instance, a sentence like मैले केही साथी भेटें could be true if I had seen three, four, or twenty friends. In contrast, मैले सबै साथीहरूलाई भेटें is only true if I really met all my friends, so the number gets fixed as soon as we know how many people there are in this category.

- The comparison of quantities triggers quantifiability, so पैसा is quantifiable in मभन्दा तपाईंको धेरै पैसा छ होला (inspite of the presence of धेरै). The reasoning behind this is that precision matters when comparing quantities (e.g. if you had 100 Rupees less the statement might no longer be true).

- Note that it is often possible to look at one referent in different ways. For instance, imagine a bowl of rice: if you look at it as a whole it is quantifiable, but if you look at indefinite parts of it it may be non-quantifiable. Be careful with determining the viewpoint that is relevant in an utterance to be tagged. For instance, in उसले मेरो भात खाइदियो it is likely that भात is quantifiable (i.e. he has eaten all of my rice, the complete bowl), whereas in ऊ भात खाँदै छ the rice is probably non-quantifiable, because even if it is located in a bowl only an non-quantifiable subamount of it is affected by the eating.

- The problem of masses and subamounts can be metaphorised into the domain of types and tokens. For instance, if you want a shop assistant to show you a certain shoe model you can again look at the model in two different ways (analogous to the rice bowl): either you can ask him to show you the model itself (that is, one shoe that is representative of the model), in which case the referent is quantifiable. The other option is to ask for some instances of the model, in which case the referent is non-quantifiable.

In contrast to quantifiability, arbitrariness is easy to recognise. It is given when a speaker makes clear that the identity of a referent is completely unimportant and that the referential expression he uses could be linked to any referent. This is often marked by words such as कुनै, केही-न-केही or जोसुकै.

B.3.10 Focus

All types of focus draw the hearer's attention to an element of an utterance that deserves special attention. There are various reasons for "deserving" attention: for instance, the presence of an element may be contrary to what one expected or to what one expects in general, or it may be crucial for the illocutionary value of the utterance. Good introductions to focus can for instance be found in Féry et al. (2007), Götze et al. (2007), Valin and LaPolla (1997).

Focus is one of the most difficult yet most interesting variables for Nepali DOM. Earlier versions of these guidelines provided a sophisticated tagging system with initially more than 10 values. This system had to be abandoned because it was too complex and because many of the distinctions reflected in it were irrelevant for the data. The new system has only 3 values:

- **nofoc** (focus not specified): this is the default value. It applies when an element is not focussed *or* when its focus type is irrelevant. Examples:
  - दिपकले **चिनी** बढी हाल्छ के, **चिया** बनाउँदाखेरि ।
  - अघि **फोन** उठाउन नभ्याएको है ।
  - सेनाले **सुरक्षित क्षेत्र** बढाउने **कार्य** शुरू गरेका छन् ।

- **contrast** (contrastive focus, = former `other.than.exp` + `diff.than.exp`): the speaker has a specific expectation regarding the identity or the quality of an element. This expectation is, however, wrong, and the speaker draws the hearer's attention to this fact. This focus type can be tested by adding a phrase of the type *not A but B* to the element in question (e.g. *I said I like **mangoes** (add: not bananas)!*). Examples:
  - ओबामालाई मारेको होइन, **ओसामालाई** मारेछन् ।
  - अब हामिले के पाएका छन् भन्ने **कुरालाई** खोज्निति गर्नु पर्छ ।
  - गरीबले चहीं **त्यो कोदो** नै खानुपर्थ्यो ।
  - **रातो बटन** प्रेस गर (पहेंलो चहीं होइन) ।
  - कमिलाहरूले **दराजमा राखेको चिनी** खाएको रहेछन्, भुइँमा राखेको चहीं छोड्छन् ।

- **fragile** (fragile utterance): this is an ad-hoc term for a phenomenon that seems to play an important role for Nepali DOM but that is not commonly described in the literature. Speakers may often feel that the illocutionary force of an utterance (i.e. what makes it an assertion, an order, or a question) crucially depends on a single element in it, thus making it "fragile" in the sense that it applies only under very special circumstances. Of course in a sense, every utterance is fragile in that it contains precisely those elements that fit to its illocutionary force - for instance, *I am going to Zurich* wouldn't be true if one replaced any of the contained referents by another one. Fragility is thus only given when a speaker feels that an element is especially likely to fall away, e.g. because it is generally rare, because its coming together with other elements is perceived as unlikely, or because it is hard to achieve. Examples for fragile assertions:

- त्यसकारण चाहीं अब **हामीले पाएको एउटा अधिकारलाई** हामिले प्रयोग गर्न सक्नु पर्यो । 'Therefore we have to be able to use the one right we've got.' Here, the crucial element (अधिकार) was hard to achieve. The success of the young Nepalese women's rights movement depends on it: if it wasn't for this one right there'd be (subjectively) no way of demonstrating women's rights.
- तिनीहरूको विचार थियो, त्यहाँ त्यस्ता केही बस्तुहरू भेटिने सम्भावना छ जुन भेटियो भने सँसारैले मानवीय सभेताको इतिहास वा विकासलाई अर्को किसिमले व्याख्या गर्नु पर्छ । 'They believed that there was the possibility of discovering some things, and if these were discovered the world would have to rewrite the history or development of human civilisation in another form.' The history of human civilisation is such a big referent that it is very unlikely that it should be affected as a whole by an event.
- भनौं **दुइटैलाई** समेटेर चाहिने काभ्रेपलान्चोक भएको हो । 'One could say Kabhrepalancok is there to bring together the two (areas).' The two areas (Kabhre and Palancok) are crucial for the truth of the utterance because they constitute its name. If one changed the referent in object position (two other regions) not only would the utterance become wrong - the name of Kabhrepalancok itself would no longer make sense.

Illocutionary values other than assertion may be fragile, too. If you have difficulties asserting the focus value for orders and questions try to convert them to assertions and see what happens. One strong indicator of fragility in questions is कुन because it usually selects one specific referent or one specific set of referent that the speaker presupposes. If there was no such referent the whole question would become uninteresting, as in:

- अहिले तपाईंहरूले **कुन गीतलाई** सर्बश्रेष्ट मान्नुहुन्छ ? 'Which song do you presently like best?'

B.3.11 Diathesis

Use the value "passive" for passive verb forms such as गरिनु, कुटिनु, खाइनु. Active/underived verb forms don't need an diathesis tag. Keep in mind that we assume that Nepali passives normally do not change the role set of verbs, so even if a form is passive there should be the same roles present as if it were active.

One special case are spontaneous passives. Spontaneous passives use the normal passive morphology (*-i* [PASS]) but describe an event that does not have an agent. In that case the patient is the only remaining core role, and accordingly it should be annotated as S.

Note that there is a difference between unknown and non-existing agents. For instance, in भात खाइयो the agent of खानु is unknown but must exist because of the semantics of the predicate, but in पानी रोकियो there definitely is no agent - the rain just stopped by itself. In many cases the absence or presence of an agent has to be judged by looking at the context. For instance, हाँगा भाँचियो could have a covert agent (e.g. बच्चाद्वारा) or not.

B.4 Problems in complex sentences

Complex sentences pose several problems for assessing domains, roles, and referential IDs. Consider as a complex sentence any combination of verbs whose frames interact syntactically on a regular base. This can either mean that one verb is the predicate of a clause which in the other occupies a position that could also be occupied by a nominal argument or that the frames share one or several arguments.

B.4.1 Complex predicates

Nepali makes frequent use of **Complex predicates**. A complex predicate consists of a noun ("N") coding an action or an event and a semantically near-empty light verb ("V") such as हुनु, गर्न, पर्नु. For the question of how to tag complex predicates valency is crucial. If an N-V combination has arguments occupying the usual roles, it is considered as a complex predicate and N (as the carrier

of the main semantic content) gets all predicate tags. However, if N itself can be mapped to a role (most usually P), the predicate tags should be assigned to the verb and N should be tagged like a normal referent.

For instance, the N विस्फोट can be combined with the light verbs हुनु (composite meaning = 'explode (itr.)') or गर्नु ('explode (tr.)'). In both cases there are easy to recognise role sets, {S} for विस्फोट हुनु and {A P} for विस्फोट गर्नु. Thus, both variants are considered as complex predicates and N gets predicate tags but no referent tags. By contrast, in काम गर्नु the N काम can be mapped to P and the worker to A, so गर्नु should get the predicate tags and काम referent tags.
This procedure may seem unnecessarily complicated, but it is motivated. The N of combinations like काम गर्नु are different from those of the other two main types in that they can be modified: compare fully grammatical यो काम तिमीले गर with ब्याटेरी (*यो) विस्फोट भयो. That the additional argument in combinations of the काम गर्नु type is really A and not S (of the complex predicate) can also be seen from the ergative that is obligatory in past tense, just as with other transitive verbs (मैले/*म काम गरें).

B.4.2 Domains and core roles

There are two important differences between domain IDs in simple and in complex sentences. One is that there are subdomains, the other that arguments can belong to several domains.

- Subdomains directly map the hierarchical structure of a complex sentence. The predicate that is not embedded constitutes the top-level domain and is marked by a single number just like any verb in a simple sentence would be. Any predicate that is directly embedded into the clause constituting the top-level domain or that directly depends on it constitutes a first-order subdomain. Such subdomains are marked by the top-level domain ID, a slash, and an additional following number. Further subdomains require further slashes and subdomain IDs. For instance, in तिमी त्यो काम गर्न सक्छौ ? there are two domains: सक्छौ is on the top-level and could, for instance, receive the ID 34; गर्न is its first subdomain and should accordingly have the ID 34/1.

- This syntax also makes it possible to subordinate several subdomains under one matrix domain on the same level. For instance, in म किन्मेल गर्न र साथीलाई भेट्न जान्छु the top-level predicate जान्छु (e.g. domain 101) has two subdomains (गर्न in 101/1 and भेट्न in 101/2).

- Some non-finite forms may be used without matrix verbs (e.g. के गर्ने ? 'What to do?'). In these cases it is not necessary to assume a zero matrix verb; instead, all elements should be assigned to the same domain without distinguishing subdomains. The covert argument (the one that would be shared if there was a matrix verb) should be represented as a zero just as anywhere else. In case it is still desirable to somehow mark that a form without a visible matrix is subordinated, use x in addition to the slash syntax, e.g. x/1 ('first subdomain of an unknown domain').

- Within domains referents are frequently shared. For instance, in गणेशले सीतालाई हेरेर उनलाई बोलायो, गणेश is the A of both हेर्नु and बोलाउनु. In गणेश गाउँमा घुम्दा सीतालाई भेट्यो it is the S of घुम्नु but the A of भेट्नु. When a shared referent can only be overtly realised once it should also only be tagged once, that is, if there is an overt realisation it should be tagged but no additional zero should be inserted, and if there is no overt realisation only one zero should be inserted. In this special but rather frequent case it becomes necessary to assign several domain and role values to a single element. For this the plus sign is to be used. For instance, if in the last sentence भेट्यो is 56 and घुम्दा is 56/1, then गणेश has the ID 56+56/1 and the role A+S. How many times a shared referent can be realised depends on the construction.

- A special kind of referent sharing is found in relative clauses. On the one hand relative clauses have obligatory sharing, but on the other hand it is always clear which domain the head of the relative clause belongs to syntactically, namely to that of the main clause (because

it is the main clause that governs its case marking). For this reason the head should always be assigned the ID of the main clause domain and the referent in the subclause should be exceptionally represented by a separate zero.

B.4.3 Hybrid elements

A special problem is brought about by nominalising suffixes because they may need predicate and argument IDs at the same time. For instance, in काठमाण्डु जानेहरूलाई सघाउनु पर्छ, जानेहरू is a predicate in the subdomain but also the P of सघाउनु. Separate the predicate and argument IDs by a colon, e.g. `24/1:24`. After that insert all values as usual.

B.4.4 Overview of constructions involving complex sentences

Below is an alphabetically ordered list of all known constructions involving complex sentences.

- -नः e.g. गर्न खोज्छ. Shared S/A which can only be realised once (> tag shared referent only once, using the "+" syntax for domain and role).
- -नुः do not tag as separate predicate when used as auxiliary (गर्नु हुन्छ, गर्नु पर्छ). Tag as nominal constituent when nominalised, e.g. in मैले गीत गाउनुलाई माइनस गरेको होइन (where गाउनु = P-DAT).
- -नेः do not tag as separate predicate when used as auxiliary गर्ने छ, गर्ने गर्छ. Tag as nominal constituent when nominalised, e.g. in त्यसो गर्नेहरूलाई सजा दिनु पर्छ (where गर्नेहरूलाई = G-DAT).
- -दा(खेरि): e.g. नेपालमा हुँदा(खेरि) धेरै भात खान्थें. No obligatory coreference (> always tag shared referent twice). Do not tag as separate predicate in lexicalised expressions such as त्यसले गर्दाखेरि.
- -दैः e.g. पुस्तक पढ्दै कुरा गर्छ. No obligatory coreference (> always tag shared referent twice). Do not tag as separate predicate when used as auxiliary (बस्दै छ, बस्दै गर्छ, रित्तिँदै जान्छ). Tag only one predicate where doubled as in हिंड्दा हिंड्दै.
- -इः e.g. भित्र गई रुन थाल्यो. Usually but not necessarily shared S/A. Where S/A is shared it can only be overtly mentioned once, so tag the shared referent only once. Do not tag as separate predicate in combination with vector verbs such as -दिनु (गरिदिनु) or -सक्नु (गरिसक्नु).
- -इकनः e.g. चिया नखाइकन जानु हुँदैन. No obligatory coreference (> always tag shared referent twice).
- -एरः e.g. हाँसेर सुत्यो. Usually but not necessarily shared S/A. Where S/A is shared it can only be overtly mentioned once, so tag the shared referent only once. Interpret as auxiliary in connection with purely aspectual light verbs जानु/आउनु (बदलिएर जान्छ) and do not tag as separate predicate there.
- -एकोः e.g. तिमीले स्याउ चोरेको मैले हेरें. No obligatory coreference (> always tag shared referent twice). Do no tag as separate word when used as auxiliary (गरेको छ).
- -एः e.g. पानि परे घरमा बसौं. No obligatory coreference (> always tag shared referent twice). Same treatment in combination with particles and case markers (e.g. गरेपनि, गरेसम्म).
- -उन्जेलः e.g. यहाँ बसुन्जेल आराम गर्नुस्. No obligatory coreference (> always tag shared referent twice).
- ज-: e.g. जो माग्छ त्यसलाई दिइन्छ. No obligatory coreference (> always tag shared referent twice). Correlative pronouns starting with ज- are also often used where no two sentences can be identified, in which case no special tagging is necessary.

- क-: e.g. मलाई ऊ के गर्छ थाहा छैन. No obligatory coreference (> tag shared referent twice).
- भनेर/भनी/भन्नेः e.g, मैले राखें भनेर भन्यो. These particles mark citations, so you can ignore the complete clause preceding them (cf. comment on metalanguage in B.3.2). They can, however, also be used as normal predicates, in which case they are to be tagged as such (e.g. उनले नमस्ते भनेर हासी).

Nepali also has a few conjunctions such as तर, तैपनि, त्यसकारणले, त्यसपछि etc. These do not set up complex sentences in the above sense, so you should tag the sentences marked by them with an independent domain.

Appendix C

Scripts

C.1 sad-parse.pl

```
#!/usr/bin/perl -w
# Parses all CLC Toolbox files with syntactic annotation within a directory, builds an internal representation and outputs R-compatible tables
# Usage: perl sad-parse.pl path/to/directory
use strict;
use utf8;
binmode (STDIN, ":utf8");
binmode (STDOUT, ":utf8");


######################
### PRELIMINARIES ###
######################

# initialise global variables, create filehandles
my $directory = $ARGV[0] or die "No corpus directory specified!\n";
opendir(DIR,$directory) or die "Could not find $directory!\n";
my @files = readdir(DIR);
my ($last_id, $current_id, %all_domains, %analysis);
my $saddataframe = '"recordstretch","verb","verb class","alternation","form","role","identifiability","quantifiability"'."\n";
my $altdataframe = '"recordstretch","verb","verb class","alternation","role set","role order"'."\n";

my (@mph, @mgl, @gw, @anno);
# value of \ref has to be global so it can be attached to domain IDs (in order to disambiguate between files)
my $ref = '';
my $record = 0;


####################
### PARSE FILES ###
####################

# go through files in directory
foreach my $file(@files){
   # open text file
```

```
35    if($file =~ /(.*)\.txt$/){
36       my $filename = $1;
37       open(INPUT,'<:encoding(utf8)',"$directory/$file") or die "Could not
              open $file!\n";
38
39       while(<INPUT>){
40          chomp(my $line = $_);
41          $line =~ s/\r//g;
42
43          # build arrays from relevant tiers, check \gw and \anno
44          # sometimes Toolbox wraps records so that one record has several
                \gw etc. lines > split line into temporary array and then
               add it to the record array
45          if($line =~ /^\\mph\s+(.*)/){ my @mph_temp = &build_words($1);
               @mph = (@mph, @mph_temp); }
46          if($line =~ /^\\mgl\s+(.*)/){ my @mgl_temp = &build_words($1);
               @mgl = (@mgl, @mgl_temp)}
47          if($line =~ /^\\gw\s+(.*)/){ my @gw_temp = split(/\s+/,$1); @gw
               = (@gw, @gw_temp); }
48          if($line =~ /^\\anno\s+(.*)/){ my @anno_temp = split(/\s+/,$1);
               @anno = (@anno, @anno_temp); }
49
50          # at \ref: map tiers in preceding record to data structure,
               check consistency, empty variables
51          if($line =~ /^\\ref/){
52             my $overt_index = -1; # stores the position of words,
                  ignoring zeros
53
54             # put tagged elements into %all_domains
55             for(my $i=0; $i<=$#anno; $i++){
56
57                # update word position (except when element is zero)
58                unless($anno[$i] =~ /^0/){ $overt_index++; }
59
60                # tagged elements are recognised by dots
61                if($anno[$i] =~ /\./){
62                   my (@split1,@split2);
63
64                   # First step: split hybrid elements with a <:> into two
                        elements
65                   if($anno[$i] =~ /^([^\.]+\.)([^:]+):([^:]+)$/){
66                      push(@split1,$1.$2);
67                      push(@split1,$1.$3);
68                   }
69                   else{ push(@split1,$anno[$i]); }
70
71                   # Second step: split elements with <+> or </> into
                        several elements
72                   foreach my $split1 (@split1){
73                      if($split1 =~ /^([^\.]+)\.([^\.]+[\+\|][^\.]+)
                           \.([^\.]+)(\..*)?/ || $split1 =~ /^([^\.]+\.)
                           ([^\.]+)\.([^\.]+[\+\|][^\.]+)(\..*)?/){
74                         my ($name, $ids, $roles) = ($1, $2, $3);
75                         my $rest = '';
76                         if($4){ $rest = $4; }
77                         $ids =~ s/\*//g;
78                         $roles =~ s/\*//g;
79                         my (@ids,@roles);
```

```
80                      @ids = split(/[\+\|]/,$ids);
81                      @roles = split(/[\+\|]/,$roles);
82                      if($#ids ne $#roles){ print "domain ids and roles
                            do not correspond on $anno[$i] ($ref)\n"; }
83                      else{
84                          for(0..$#ids){
85                              my $split2 = $name.'.'.$ids[$_].'.'.$roles[
                                    $_];
86                              if($rest){ $split2 = $split2.$rest; }
87                              push(@split2,$split2);
88                          }
89                      }
90                  }
91                  else{ push(@split2,$split1); }
92              } # EOF split elements
93
94              # Third step: put elements in hash
95              foreach my $split2 (@split2){
96
97                  # characteristics of elements in \anno
98                  # fully annotated: ARG.domain.role.def.qnt , PRED.
                        domain.class(.alternation)
99                  # syntax only: ARG.domain.role, PRED.domain.class(.
                        alternation)
100
101                 # stem + domain + role: argument
102                 if($split2 =~ /^([^\.]+)\.([^\.]+)\.([SAPMGT]|CT|CR|
                        EMO|BEN|CSR)(\.([^\.]+))?(\.([^\.]+))?$/){
103                     my ($name, $id, $role) = ($1,$filename."."$2,$3)
                            ;
104                     my ($def, $qnt);
105                     if($5){ $def = $5; }
106                     else{ $def = 'missing'; }
107                     if($7){ $qnt = $7; }
108                     else{ $qnt = 'missing'; }
109
110                     if($name =~ /^0/){
111                         $all_domains{$id}{'roles'}{$role}{'name'} =
                                $name;
112                         push(@{$all_domains{$id}{'ordered_roles'}},
                                $role.'0');
113                     }
114                     else{
115                         # take element in @mph corresponding to
                                present element in anno and isolate stem
116                         $mph[$overt_index] =~ /:([^:]+):/;
117                         push(@{$all_domains{$id}{'ordered_roles'}},
                                $role);
118                         if($1){ $all_domains{$id}{'roles'}{$role}{'
                                name'} = $1; }
119                         else{ $all_domains{$id}{'roles'}{$role}{'name'
                                } = 'UNDEFINED'; }
120                     }
121                     $all_domains{$id}{'roles'}{$role}{'
                            identifiability'} = $def;
122                     $all_domains{$id}{'roles'}{$role}{'
                            quantifiability'} = $qnt;
123                     $all_domains{$id}{'records'}{$ref}++;
```

```
124                         }
125                         # stem + domain + non-role: predicate
126                         elsif($split2 =~ /^([^\.]+)\.([^\.]+)\.([^\.]+)
                                (\.([^\.]+))?$/){
127                            my ($name, $id, $verbclass) = ($1,$filename.".".
                                   $2,$3);
128                            my $alternation;
129                            if($5){ $all_domains{$id}{'alternation'} = $5; }
130                            else{ $all_domains{$id}{'alternation'} = 'default
                                   '; }
131                            # if(!$mph[$overt_index]){ print "$id
                                   $overt_index (@mph)\n"; }
132                            $mph[$overt_index] =~ /:([^:]+):/; # take element
                                    in @mph corresponding to present element in
                                   anno and isolate stem
133                            # if(!$mph[$overt_index]){ print "$id
                                   $overt_index (= $name) (@mph)\n"; }
134                            if($1){ $all_domains{$id}{'verbname'} = $1; }
135                            else{ $all_domains{$id}{'verbname'} = 'UNDEFINED'
                                   ; }
136                            push(@{$all_domains{$id}{'ordered_roles'}}, 'V');
137                            $all_domains{$id}{'verbclass'} = $verbclass;
138                            $all_domains{$id}{'records'}{$ref}++;
139                         }
140                      } # EOF  loop for putting elements in hash
141
142                   } # EOF loop on tagged elements in present line
143                } # EOF loop on @anno
144
145                # check number of elements in relevant tiers
146                if(@gw && @mph && @mgl && @anno){
147                   my @overt_elem;
148                   foreach my $elem(@anno){ if($elem !~ /^0/){ push(
                          @overt_elem,$elem); } }
149                }
150
151                # clear variables
152                if($line =~ /^\\ref\s+(.*)/){ $ref = $1; }
153                else{ print "no \\ref somewhere past $ref\n"; $ref = $ref.''
                       ; }
154                $record++;
155                @gw = ();
156                @mph = ();
157                @mgl = ();
158                @anno = ();
159             } # EOF emptying variables
160          } # EOF reading INPUT
161
162          close(INPUT);
163
164       } # EOF file extension check
165    } # EOF reading DIR
166
167
168    ######################
169    ### INTERPRET DATA ###
170    ######################
171
```

```
172 # go through %all_domains and build subdomains
173
174 # structure of %all_domains:
175 # ID -> 1 'records' -> n reference -> 1 frequency
176 #     1 'roles' -> n roles -> 1 'identifiability'
177 #                         1 'quantifiability'
178 #     1 'verbclass' -> 1 'verbclass'
179 #     1 'alternation' -> 1 'alternation'
180
181 foreach my $domain(sort keys %all_domains){
182    # get all information in present domain
183    my $recordstretch = join('/', sort keys %{$all_domains{$domain}{'
       records'}});
184    my @roles = keys %{$all_domains{$domain}{'roles'}};
185    my $roleset = join(' ', @roles);
186    my @ordered_roles = @{$all_domains{$domain}{'ordered_roles'}};
187    my $ordered_roles = join(' ', @ordered_roles);
188    my ($verbname, $verbclass, $alternation);
189    if($all_domains{$domain}{'verbname'}){ $verbname = $all_domains{
       $domain}{'verbname'}; }
190    else{ $verbname = '0'; }
191    if($all_domains{$domain}{'verbclass'}){ $verbclass = $all_domains{
       $domain}{'verbclass'}; }
192    elsif(grep(/^(CT|CR)$/,@roles)){ $verbclass = 'zerocop'; }
193    else{ $verbclass = 'missing'; }
194    if($all_domains{$domain}{'alternation'}){ $alternation = $all_domains{
       $domain}{'alternation'}; }
195    else{ $alternation = 'missing'; }
196
197    $altdataframe .= '"'.$recordstretch.'","'.$verbname.'","'.$verbclass.'
       ","'.$alternation.'","'.$roleset.'","'.$ordered_roles.'"'."\n";
198
199    $analysis{'allframes'}++;
200    if(@roles){
201       foreach my $role(@roles){
202          # get referential values for role
203          my $rolename = $all_domains{$domain}{'roles'}{$role}{'name'};
204          my $identifiability = $all_domains{$domain}{'roles'}{$role}{'
              identifiability'};
205          my $quantifiability = $all_domains{$domain}{'roles'}{$role}{'
              quantifiability'};
206
207          # check S/A detransitivisation
208          my %objects = ('tr' => 'P', 'dido' => 'T', 'dipo' => 'G', 'dioo'
                => 'G', 'exptr' => 'P');
209          if($verbclass && $alternation){
210             # if($verbclass
211             if(($verbclass =~ /^(tr|di[dpo]o|exptr)$/ && $role eq
                   $objects{$verbclass} && $alternation =~ /^(default|sad)$/)
212             # || ($verbclass =~ /^(aux|dumO|dumS|expitr|itr|other|xcla)$
                   /) # for including all valency classes
213             ){
214
215                # add data to CSV table
216                $saddataframe .= '"'.$recordstretch.'","'.$verbname.'","'.
                      $verbclass.'","'.$alternation.'","'.$rolename.'","'.
                      $role.'","'.$identifiability.'","'.$quantifiability.'"'
                      ."\n";
```

```
217
218                 # search for unexpected combinations sad+qnt, default+nonq
                        for double-checking annotations
219                 # if(($alternation eq 'sad' && $quantifiablity eq 'qnt')
220                 # || ($alternation eq 'default' && $quantifiability eq '
                        nonq')){
221                 #  print "unexpected $alternation in $recordstretch\n";
222                 # }
223             }
224         }
225     }
226   }
227 } # EOF domain loop
228
229
230 ####################
231 ### OUTPUT DATA ###
232 ####################
233
234 open(OUTPUT,'>:encoding(utf8)',"sad.csv");
235 print OUTPUT $saddataframe;
236 print "\nDetransitivisation data written to sad.csv\n";
237 close(OUTPUT);
238
239 open(OUTPUT,'>:encoding(utf8)',"alt_analysis.csv");
240 print OUTPUT $altdataframe;
241 print "Alternation data written to alt_analysis.csv\n\n";
242 close(OUTPUT);
243
244
245 ####################
246 ### SUBROUTINES ###
247 ####################
248
249 # read in one line from a toolbox text file and return list of elements
250 sub build_words {
251    my $heap = shift;
252    my $word = '';
253    my @clause = ();
254    my $it_could_have_ended_there = 0;
255
256    while($heap =~ /(\S+)/g){
257       my $morpheme = $1;
258
259       # endoclitics are marked by a preceding hyphen and space (e.g. =ta
              = < - ta>) which makes special treatment necessary
260       if($morpheme eq '-'){
261          $it_could_have_ended_there = 0;
262          $word = $word.'=';
263       }
264
265       # if morpheme could be the beginning of a word form (= prefix or
             stem)
266       if($morpheme =~/^[^-].*?(\-)?$/){
267          # [^-]+ is not valid because even glosses can contain hyphens in
                their middle (sister-in-law, thirty-five...)!
268          # [^-].* is not valid because * is so greedy that it devours the
                hyphens in prefixes such as a-, which makes $1 = 0 and thus
```

```
            makes prefixes words
269
270         # if morpheme is stem, mark by ::
271         if(!$1){ $morpheme = ":".$morpheme.":"; }
272
273         # if last morpheme could be the end of a word form: add word to
                clause
274         if($it_could_have_ended_there){
275            push(@clause,$word);
276            $word = '';
277         }
278         # add current morpheme to word and set boundary status
279         $word = $word.$morpheme;
280         if($1){ $it_could_have_ended_there = 0; }
281         else{ $it_could_have_ended_there = 1; }
282      }
283
284      # if morpheme is definitely not the beginning of a word form (=
            suffix)
285      elsif($morpheme =~/^\-.+/){
286         # add current morpheme to word
287         $word = $word.$morpheme;
288         $it_could_have_ended_there = 1;
289      }
290   }
291   # add last word form
292   push(@clause,$word);
293   return @clause;
294 }
```

C.2 sad-consistency.pl

```
1  #!/usr/bin/perl -w
2  # Checks an annotated CLC Toolbox file for internal consistency and
       compliance with the guidelines
3  # Usage: perl sad-consistency.pl path/to/file
4  use strict;
5  use utf8;
6  binmode (STDIN, ":utf8");
7  binmode (STDOUT, ":utf8");
8
9
10 ######################
11 ### PRELIMINARIES ###
12 ######################
13
14 # initialise global variables, create filehandles
15 my $input;
16 $input = $ARGV[0] or die "No input file specified!\n";
17 open(INPUT,'<:encoding(utf8)',$input) or die "Could not find $input!\n";
18 my ($last_id, $current_id, %all_domains);
19 my (@mph, @mgl, @gw, @anno);
20 my $ref = '';
21 my $record = 0;
22
23
24 ##################
```

```
25 ### PARSE FILE ###
26 ####################
27
28 while(<INPUT>){
29    chomp(my $line = $_);
30    $line =~ s/\r//g;
31
32    # build arrays from relevant tiers, check \gw and \anno
33    # sometimes Toolbox wraps records so that one record has several \gw
          etc. lines > split line into temporary array and then add it to the
           record array
34    if($line =~ /^\\mph\s+(.*)/){ my @mph_temp = &build_words($1); @mph =
          (@mph, @mph_temp); }
35    if($line =~ /^\\mgl\s+(.*)/){ my @mgl_temp = &build_words($1); @mgl =
          (@mgl, @mgl_temp)}
36    if($line =~ /^\\gw\s+(.*)/){
37       my @gw_temp = split(/\s+/,$1);
38       @gw = (@gw, @gw_temp);
39       # check \gw for suspicious characters indicating it might really be
             \anno
40       if($1 =~ /\.(\d+)\*?(\/\d+\*?)*([\+\|]\d+\*?(\/\d+\*?)*)*./){ print
            "forgot to rename \\gw ($ref)?\n"; }
41    }
42    if($line =~ /^\\anno\s+(.*)/){
43       my @anno_temp = split(/\s+/,$1);
44       @anno = (@anno, @anno_temp);
45       # check whether there are any tags at all. Use with care: retrieves
            many false positives (records without core arguments/verbs)
46       # if($1 !~ /\.\d+(\/\d+)?\./){ print "Forgot to tag \\anno ($ref)?\
            n"; }
47       # collect tagged elements in %all_domains
48       foreach my $elem (@anno_temp){
49          # tagged elements are recognised by dots
50          if($elem =~ /\./){
51             my (@split1,@split2);
52
53             # First step: split elements with a <:> into two elements
54             if($elem =~ /^([^\.]+\.)([^:]+):([^:]+)$/){
55                push(@split1,$1.$2);
56                push(@split1,$1.$3);
57             }
58             else{ push(@split1,$elem); }
59
60             # Second step: split elements with <+> or <|> into several
                  elements
61             foreach my $split1 (@split1){
62                if($split1 =~ /^([^\.]+)\.([^\.]+[\+\|][^\.]+)\.([^\.]+)
                      (\..*)?/ || $split1 =~ /^([^\.]+\.)([^\.]+)
                      \.([^\.]+[\+\|][^\.]+)(\..*)?/){
63                   my ($name, $ids, $roles) = ($1, $2, $3);
64                   my $rest = '';
65                   if($4){ $rest = $4; }
66                   $ids =~ s/\*//g;
67                   $roles =~ s/\*//g;
68                   my (@ids,@roles);
69                   @ids = split(/[\+\|]/,$ids);
70                   @roles = split(/[\+\|]/,$roles);
```

```
71                  if($#ids ne $#roles){ print "domain ids and roles do
                        not correspond on $elem ($ref)\n"; }
72                  else{
73                     for(0..$#ids){
74                        my $split2 = $name.'.'.$ids[$_].'.'.$roles[$_];
75                        if($rest){ $split2 = $split2.$rest; }
76                        push(@split2,$split2);
77                     }
78                  }
79               }
80               else{ push(@split2,$split1); }
81            } # EOF split elements
82
83            # Third step: put elements in hash
84            foreach my $split2 (@split2){
85               # stem + 4 elements: argument
86               if($split2 =~ /^([^\.]+)\.([^\.]+)\.([^\.]+)\.([^\.]+)
                     \.([^\.]+)$/){
87                  my ($name, $id, $role, $def, $qnt) = ($1,$2,$3,$4,$5);
88                  $all_domains{$id}{'roles'}{$role}{'identifiability'} =
                        $def;
89                  $all_domains{$id}{'roles'}{$role}{'quantifiability'} =
                        $qnt;
90                  $all_domains{$id}{'records'}{$ref}++;
91               }
92               # stem + 2 - 3 elements: predicate
93               elsif($split2 =~ /^([^\.]+)\.([^\.]+)\.([^\.]+)(\.([^\.]+)
                     )?$/){
94                  my ($name, $id, $verbclass) = ($1,$2,$3);
95                  $all_domains{$id}{'verbclass'} = $verbclass;
96                  if($5){ $all_domains{$id}{'alternation'} = $5; }
97                  $all_domains{$id}{'records'}{$ref}++;
98               }
99               elsif($split2 =~ /^([^\.]+)\.([^\.]+)?$/){ print "too few
                     tags on $elem ($ref)\n"; }
100              elsif($split2 =~ /^([^\.]+\.){5,}[^\.]+$/){ print "too
                     many tags on $elem ($ref)\n"; }
101              else{ print "something's wrong with $split2 ($ref) - dots
                     in wrong place?\n"; }
102           } # EOF loop on tagged elements in present line
103
104        } # EOF loop on tagged elements in @anno_temp
105     } # EOF loop on @anno_temp
106  } # EOF tier analysis \anno
107
108  # at \ref: check preceding record, empty variables
109  if($line =~ /^\\ref/){
110
111     # check number of elements in relevant tiers
112     if(@gw && @mph && @mgl && @anno){
113        my @overt_elem;
114        foreach my $elem(@anno){ if($elem !~ /^0/){ push(@overt_elem,
              $elem); } }
115        if($#gw != $#overt_elem){
116           print "different element numbers in gw ("; print ($#gw+1);
                 print ") and anno ("; print ($#overt_elem+1); print "
                 overt) ($ref)\n";
117        }
```

```
118             if(($#gw != $#mph) || ($#gw != $#mgl)){
119                 print "different element numbers in gw (".($#gw+1).") and
                        gloss lines (".($#mph + 1).", ".($#mgl+1).") in $ref.\n";
120             }
121             if(($#overt_elem != $#mph) || ($#overt_elem != $#mgl)){
122                 print "different element numbers in anno (".($#overt_elem+1).
                        " overt) and gloss lines (".($#mph+1).", ".($#mgl+1).") in
                         $ref\n";
123             }
124         }
125         # only include this line if all records should have an \anno line (
                even those without predicates or core arguments)
126         # elsif(!@anno && $record > 0){ print "missing \\anno in $ref\n"; }
127 
128         # clear variables
129         if($line =~ /^\\ref\s+(.*)/){ $ref = $1; }
130         else{ print "no \\ref somewhere past $ref\n"; $ref = $ref.''; }
131         $record++;
132         @gw = ();
133         @mph = ();
134         @mgl = ();
135         @anno = ();
136     } # EOF record check
137 }
138 
139 close(INPUT);
140 
141 
142 ##########################
143 ### CHECK CONSISTENCY ###
144 ##########################
145 
146 # go through %all_domains, build subdomains and check each for
        consistency
147 # within each domain, the top-level domain and all subdomains are
        assembled horizontally (the top-level domain ID thereby being repeated
        ) - this is possible because there is no need to relate superordinate
        to subordinate domains explicitly
148 
149 foreach my $domain(sort{&get_top($a) <=> &get_top($b)} keys %all_domains)
        {
150     my $recordstretch = join('/', sort keys %{$all_domains{$domain}{'
            records'}});
151 
152     # check ID format
153     if($domain !~ /^(xdom|[\d\/])+$/){ print "wrong ID format ""$domain (
            $recordstretch)\n"; }
154 
155     # check format of tags
156     # verbclass
157     if($all_domains{$domain}{'verbclass'} && $all_domains{$domain}{'
            verbclass'} !~ /^(aux|dido|dioo|dipo|expitr|exptr|itr|other|tr|
            uninf|xcla)$/){
158         print "verbclass "$all_domains{$domain}{'verbclass'"} 'doesnt exist
                ($domain, $recordstretch)\n";
159     }
160 
161     # check alternation
```

```
162     if($all_domains{$domain}{'alternation'} && $all_domains{$domain}{'
            alternation'} !~ /^(sod|ambrec|ben|caus|cop|dumA|idt|sad|OtoS|pass|
            recp|refl|xalt)$/){
163        print "alternation "$all_domains{$domain}{'alternation'"} 'doesnt
               exist ($domain, $recordstretch)\n";
164     }
165
166     # check roles
167     if($all_domains{$domain}{'roles'}){
168        foreach my $role(keys %{$all_domains{$domain}{'roles'}}){
169           if($role !~ /^([SAPGT]|C[RT]|NEXP|BEN|CSR|xrol)$/){
170              print "role value ""$role on 'doesnt exist ($domain,
                     $recordstretch)\n";
171           }
172           # identifiability
173           if($all_domains{$domain}{'roles'}{$role}{'identifiability'} !~
                /^(def|spec|idf|xdef)$/){
174              print "identifiability value "$all_domains{$domain}{'roles'}{
                     $role}{'identifiability'"} on 'doesnt exist ($domain,
                     $recordstretch)\n";
175           }
176
177           # quantifiability
178           if($all_domains{$domain}{'roles'}{$role}{'quantifiability'} !~
                /^(qnt|nonq|xqnt)$/){
179              print "identifiability value "$all_domains{$domain}{'roles'}{
                     $role}{'quantifiability'"} on 'doesnt exist ($domain,
                     $recordstretch)\n";
180           }
181        }
182     }
183     else{ unless($all_domains{$domain}{'verbclass'} eq 'aux' || $domain =~
             /xdom/){ print "no roles in $domain ($recordstretch) - domains
            without roles are impossible!\n"; }}
184
185     # check compatibility of tags
186     if($all_domains{$domain}{'verbclass'} && $all_domains{$domain}{'roles'
            } && $all_domains{$domain}{'verbclass'} !~ /^(xcla|other)$/){
187        my ($verbclass, @roles) = ($all_domains{$domain}{'verbclass'}, keys
              %{$all_domains{$domain}{'roles'}});
188        my $alternation = '';
189        if($all_domains{$domain}{'alternation'}){ $alternation =
              $all_domains{$domain}{'alternation'}; }
190
191        # missing roles in verb classes
192        if(($verbclass =~ /^(itr|uninf)$/ && !grep(/^S$/,@roles) &&
              $alternation ne 'cop')
193        || ($verbclass =~ /^(tr)$/ && (!grep(/A/,@roles) || !grep(/P/,
              @roles)) && $alternation !~ /^(sod|ambrec|pass)$/)
194        || ($verbclass =~ /^di[dpo]o$/ && (!grep(/^A$/,@roles) || !grep(/^
              T$/,@roles) || !grep(/G/,@roles)) && $alternation !~ /^(sod|
              ambrec|pass)$/)
195        || ($verbclass eq 'expitr' && (!grep(/^S$/,@roles) || !grep(/NEXP/,
              @roles)))
196        || ($verbclass eq 'exptr' && (!grep(/A/,@roles) || !grep(/P/,@roles
              ) || !grep(/NEXP/,@roles)))
197        # verbclasses which do not require roles: aux|other|xcla
198        ){
```

```
199             print "role missing in set {@roles} for $verbclass ($domain,
                    $recordstretch)\n";
200         }
201
202         # missing roles in alternations
203         if($alternation
204         && (($alternation =~ /^(sod|ambrec|pass)$/ && !grep(/^S$/,@roles))
205         || ($alternation eq 'ben' && !grep(/BEN/,@roles))
206         || ($alternation eq 'caus' && !grep(/CSR/,@roles))
207         || ($alternation eq 'cop' && (!grep(/CT/,@roles) || !grep(/CR/,
                @roles))))
208         ){
209             print "role missing in set {@roles} for $alternation ($domain,
                    $recordstretch)\n";
210         }
211
212         # redundant roles
213         foreach my $role(@roles){
214             if(($role eq 'S' && $verbclass !~ /^(itr|expitr|uninf)$/ &&
                        $alternation !~ /^(sod|ambrec|pass)$/)
215             || ($role eq 'A' && $verbclass !~ /^(tr|exptr|dido|dipo|dioo)$/
                        && $alternation ne 'poss')
216             || ($role eq 'P' && $verbclass !~ /^(tr|exptr)$/ && $alternation
                         ne 'poss')
217             || ($role =~ /^[GT]$/ && $verbclass !~ /^(dido|dipo|dioo)$/)
218             || ($role eq 'NEXP' && $verbclass !~ /^exp(itr|tr)$/)
219             || ($role =~ /^C[TR]$/ && (!$alternation || ($alternation ne '
                        cop')))
220             || ($role eq 'BEN' && $alternation ne 'ben')
221             || ($role eq 'CSR' && $alternation ne 'caus')
222             ){
223                 print "role ""$role not defined by frame $verbclass";
224                 if($alternation){ print ".$alternation"; }
225                 print " ($domain, $recordstretch)\n";
226             }
227         }
228
229         # class and alternation
230         if($alternation){
231             if(($alternation =~ /^(sod|ambrec|sad|OtoS|pass|recp|refl)$/ &&
                        $verbclass !~ /^di[dpo]o$|^tr$/)
232             || ($alternation eq 'cop' && $verbclass !~ /^itr$|^uninf$/)
233             || ($alternation eq 'dumA' && $verbclass ne 'exptr')
234             ){
235                 print "alternation ""$alternation not possible with verbclass
                            ""$verbclass ($domain, $recordstretch)\n";
236             }
237         }
238
239     } # EOF comparison roles/verbclasses
240
241
242 } # EOF loop on domains
243
244 print "Done.\n";
245 <>;
246
247
```

```
248 ####################
249 ### SUBROUTINES ###
250 ####################
251
252 # read in one line from a toolbox text file and return list of elements
253 sub build_words {
254     my $heap = shift;
255     my $word = '';
256     my @clause = ();
257     my $it_could_have_ended_there = 0;
258
259     while($heap =~ /(\S+)/g){
260         my $morpheme = $1;
261
262         # endoclitics are marked by a preceding hyphen and space (e.g. =ta
                = < - ta>) which makes special treatment necessary
263         if($morpheme eq '-'){
264             $it_could_have_ended_there = 0;
265             $word = $word.'=';
266         }
267
268         # if morpheme could be the beginning of a word form (= prefix or
                stem)
269         if($morpheme =~/^[^-].*?(\-)?$/){
270             # [^-]+ is not valid because even glosses can contain hyphens in
                    their middle (sister-in-law, thirty-five...)!
271             # [^-].* is not valide because * is so greedy that it devours
                    the hyphens in prefixes such as a-, which makes $1 = 0 and
                    thus makes prefixes words
272
273             # if last morpheme could be the end of a word form: add word to
                    clause
274             if($it_could_have_ended_there){
275                 push(@clause,$word);
276                 $word = '';
277             }
278             # add current morpheme to word and set boundary status
279             $word = $word.$morpheme;
280             if($1){ $it_could_have_ended_there = 0; }
281             else{ $it_could_have_ended_there = 1; }
282         }
283
284         # if morpheme is definitely not the beginning of a word form (=
                suffix)
285         elsif($morpheme =~/^\-.+/){
286             # add current morpheme to word
287             $word = $word.$morpheme;
288             $it_could_have_ended_there = 1;
289         }
290     }
291     # add last word form
292     push(@clause,$word);
293     return @clause;
294 }
295
296 # transform domain IDs so they become numerically comparable
297 sub get_top {
298     my $id = shift;
```

```
    $id =~ s/\//\./g;
    if($id =~ /^(\d+(\.\d+)?)/){ $id = $1; }
    else{ $id = 0; }
    return $id;
}
```

C.3 sad-analysis.R

```
# Does a basic statistical analysis of syntactic annotations in the CLC.
    Expected input format is CSV.
# Usage: source("path/to/sad-analysis.R")


#####################
### PRELIMINARIES ###
#####################

# read in S/A detransitivisation table and preprocess
sad <- read.csv(file.choose(), header=TRUE)
sad <- sad[sad$quantifiability!="xqnt" & sad$identifiability!="xdef",]
sad$quantifiability <- factor(sad$quantifiability, levels=c("nonq","qnt")
    , order=TRUE)
sad$identifiability <- factor(sad$identifiability, levels=c("idf","spec",
    "def"), order=TRUE)
library(rms)
datadist(sad) -> saddata
options(datadist = "saddata")
options(contrasts=c("contr.treatment","contr.treatment")) # needed to
    deal with ordered factors


#################
### MAIN PART ###
#################

# print numbers and proportions of DAT/NOM in all objects
cat("\n--- Summary: ---\n\n")
print(table(sad$alternation))
print(prop.table(table(sad$alternation)))
cat("\n");

# check two central variables
categorial_variables <- c("quantifiability", "identifiability")
for(var in categorial_variables){
    cat("\n\n
        ------------------------------------------------------------------\
        n\n--- Variable:", var, "---\n\n")
    # simple contingency table
    xtabs(~ sad[,var] + sad$alternation) -> cont_table
    print(cont_table)
    cat("\n")
    # contingency table with values proportional to row sums
    prop.table(cont_table,1) -> prop_table
    print(prop_table)
    cat("\n")
    # Chi square test with Yate's correction or Fisher's exact test, +
        Pearson's C (between 0-1)
```

```
43      chisq.test(cont_table) -> chi
44      if(nrow(cont_table) == 2){
45         fisher.test(cont_table) -> ftest
46         print(ftest)
47         cat("\n")
48      }
49      else{ print (chi) }
50
51      # calculate and print coefficients
52      pearsons_c <- sqrt(chi$statistic/(sum(cont_table)+chi$statistic))
53      dim_min <- min(nrow(cont_table),ncol(cont_table))  # will normally
            always be 2 because of NOM/DAT
54      # corrected_c <- sqrt(dim_min/(dim_min-1)) * pearsons_c
55      cramers_v <- sqrt(chi$statistic/(sum(cont_table)*(dim_min-1)))
56      cat("Pearson's contingency coefficient:", pearsons_c, "\nCramer's V:",
             cramers_v, "\n\n")
57
58      # try logistic regression
59      tmp.lrm = lrm(sad$alternation ~ sad[,var])
60      print(anova(tmp.lrm))
61      print(tmp.lrm)
62
63      # check impact of individual values
64      cat("--- Individual values ---\n\n")
65      for(val in levels(sad[,var])){
66         # for each value of the variable, collapse all others and do a
               Fisher's exact test
67         collapsed <- sad[,c(var,"alternation")]
68         levels(collapsed[,var])[levels(collapsed[,var]) != val] <- "the_
               rest"
69         xtabs(~ collapsed[,var] + collapsed$alternation) -> single_val_
               table
70         print(prop.table(single_val_table,1))
71         cat("\n")
72         cat("p-value < Fisher's exact test for \"", val, "\" being relevant
               : ", sep="")
73         print(fisher.test(single_val_table)$p.value)
74         cat("\n")
75         chisq.test(single_val_table) -> chi
76         cramers_v <- sqrt(chi$statistic/(sum(single_val_table))) # dim_min
               is 2, so (dim_min-1) = 1
77         cat("Cramer's V:", cramers_v, "\n\n")
78      }
79
80      # plot mosaic plot for every variable to files in graphics folder
81      pdf(paste("/Users/robertschikowski/ /The core of all things/graphics/
            chintang-",var,".pdf", sep=""))
82      par(mar=c(0.1,0.1,0.1,0.1), cex=2.5)
83      mosaicplot(as.matrix((xtabs(~ sad[,var] + sad$alternation))), color=F,
             main="", xlab="", ylab="", las=1)
84      dev.off()
85  }
86
87
88  ######################################
89  ### INSPECT VARIABLE COMBINATIONS ###
90  ######################################
91
```

```
# unusual combinations
cat("Quantifiable specific/definite referents with S/A detransitivisation:\n\n")
print(sad[sad$quantifiability=="qnt" & (sad$identifiability=="spec" | sad$identifiability=="nonq") & sad$alternation=="of",])
cat("\nNon-quantifiable indefinite referents with default:\n\n")
print(sad[sad$quantifiability=="nonq" & sad$identifiability=="idf" & sad$alternation=="default",])
```

C.4 dom-parse.pl

```
#!/usr/bin/perl -w
# Parses an NNC XML file with syntactic annotations related to DOM, calculates secondary attributes on their base, and outputs three files:
# (1) dom.csv: table with relevant objects in rows and various properties of theirs in columns
# (2) referents.csv: table with all referents (grouped by files) and their topicality values
# (3) roles.csv: table with all role sets and role orders
# Usage: perl parse-nnc-dom.pl (directory/with/files) - if no directory is given, the working directory is assumed
use Cwd;
use strict;
use utf8;
binmode (STDIN, ":utf8");


######################
### PRELIMINARIES ###
######################

# define attributes (primary and secondary)
my @possible_attributes = ('file', 'domain', 'identity', 'form', 'role', 'DOM', 'predicate', 'ctag', 'modification', 'animacy', 'situation', 'quantifiability', 'focus', 'givenness', 'relative_position', 'distance_from_predicate', 'absolute_frequency_sofar', 'relative_frequency_sofar', 'ranked_frequency_sofar', 'absolute_frequency_total', 'relative_frequency_total', 'ranked_frequency_total', 'distance_to_last', 'competitors', 'co_argument_case', 'diathesis', 'genre', 'speaker_name', 'speaker_sex', 'speaker_age');
# define header of DOM table
my $domtable = '';
foreach my $att(@possible_attributes){
   $domtable .= '"'.$att.'"';
   unless($att eq 'speaker_age'){ $domtable .= ','; }
}
$domtable .= "\n";
# define header of referent table
my $referenttable = '';
foreach my $att('identity', 'absolute_frequency_total', 'ranked_frequency_total', 'relative_frequency_total'){ $referenttable .= '"'.$att.'",'; }
$referenttable .= "\n";
# define header of role table
my $roletable = "file,domain,role_set,role_order\n";
```

```
32
33 # define attributes values (where predefined values exist)
34 my %possible_values;
35 $possible_values{'role'} = ['S','A','P','G','T','CR','CT','x'];
36 $possible_values{'DOM'} = ['NOM','DAT','NA','GEN','x'];
37 $possible_values{'ctag'} = ['n','adj','pro','poss','dem','o','x','NN','NP
    ','JM','JF','JO','JX','JT','MOM','MOF','MOO','MOX','DDX','PMX','PTN','
    PTM','PTH','PXH','PXR','PRF','PMXKM','PMXKF','PMXKO','PTNKM','PTNKF','
    PTNKO','PTMKM','PTMKF','PTMKO','PRFKM','PRFKF','PRFKO','PMXKX','PTNKX'
    ,'PTMKX','PRFKX','DDM','DDF','DDO','DKM','DKF','DKO','DKX','DJM','DJF'
    ,'DJO','DJX','DGM','DGF','DGO','DGX','MM','TT','QQ','UU','NULL','VI','
    VDM','VDF','VDO','VDX','VE','VN','VQ','VCN','VCM','VCH','VS','VR','
    VVMX1','VVMX2','VVTN1','VVTX2','VVYN1','VVYX2','VVTN1F','VVTM1F','
    VVYN1F','VVYM1F','VOMX1','VOMX2','VOTN1','VOTX2','VOYN1','VOYX2','RR',
    'RD','RK','RJ','II','IH','IE','IA','IKM','IKF','IKO','IKX','MLM','MLF'
    ,'MLO','MLX','CC','CSA','CSB','YF','YM','YQ','YB','FF','FS','FB','FO',
    'FZ','FU'];
38 $possible_values{'modification'} = ['none','adj','relclause','humposs','
    latposs','poss','num','dem','interrog','several','sortal','sortal.q','
    other','x'];
39 $possible_values{'animacy'} = ['human','human.fam','human.prop','human.
    group','high.anim','high.anim.prop','mid.anim','low.anim','thing','
    state','process','x'];
40 $possible_values{'quantifiability'} = ['qnt','nonq','x'];
41 $possible_values{'focus'} = ['nofoc','contrast','fragile','x'];
42 $possible_values{'situation'} = ['concrete','general','abstract','
    exemplary','x'];
43 $possible_values{'diathesis'} = ['passive'];
44
45 # read all files in directory
46 my $directory;
47 if($ARGV[0]){ $directory = $ARGV[0]; } else{ $directory = getcwd(); }
48 opendir(DIR,$directory) or die "No or wrong directory specified!\n";
49 my @files = readdir(DIR);
50
51 # counter for some simple statistics
52 my %global_counter = ('overt_core_referents' => 0, 'overt_PTG' => 0, '
    overt_objects' => 0, 'overt_G' => 0, 'dat_core_referents' => 0, '
    dat_PTG' => 0, 'dat_objects' => 0, 'dat_G_nonobjects' => 0, '
    nom_objects' => 0);
53
54
55 ####################
56 ### PARSE FILES ###
57 ####################
58
59 # go through files in directory
60 foreach my $file(@files){
61    # open XML file
62    if($file =~ /(.*)\.xml$/){
63       print STDERR "analysing $file...";
64       open(IN,'<:encoding(utf8)',"$directory/$file");
65       local $/;
66       my $text = <IN>;
67       my %counter = ('words' => 0, 'referents' => 0, 'top_frequency' =>
             0); # counter hash
68       my (%all_domains,%all_IDs,$speaker_name,$speaker_sex,$speaker_age);
69
```

```
70          # determine genre by looking at file name
71          my $genre;
72          if($file =~ /^[AV]00/){ $genre = 'spoken'; }
73          else{ $genre = 'written'; }
74
75          close(IN);
76
77          # parse XML by going through <w> elements; collect data in %
                all_domains and %all_IDs
78          while($text =~ /<w\s*([^>]+\"\s*)?\/>|<w\s*([^>]+\"\s*)?>([^<]+)
                ?<\/w>|<sp\s*([^>]+)>/sg){
79          #           empty tag + attr.    | content tag + attr. + content  |
                 speaker tag
80
81              # get form and attributes from <w>
82              my($form,$attributes,%attributes,$speakerattributes);
83              foreach($1,$2,$3){
84                  if($_){
85                      my $match = $_;
86                      $match =~ s/\s+/ /sg;
87                      if($match =~ /(\".*\")/s){ $attributes = $match; }
88                      elsif($match =~ /^([\s\w\+]+)$/s){ $form = $match; }
89                  }
90              }
91              if($4){ $speakerattributes = $4; }
92
93              # if element has no form but attributes, make it a zero (
                    elements with no form AND no attributes are NNC artefacts and
                     are ignored)
94              if(!$form && $attributes){ $form = '0'; }
95
96              # parse attributes
97              if($attributes){
98                  while($attributes =~ /(\w+)=\"([^\"]+)\"/sg){
99                      $attributes{$1} = $2;
100                 }
101             }
102             if($speakerattributes){
103                 if($speakerattributes =~ /who=\"([^\"]+)\"/){ $speaker_name=
                        $1; }
104                 if($speakerattributes =~ /gender=\"([^\"]+)\"/){ $speaker_sex
                        =$1; }
105                 if($speakerattributes =~ /age=\"([^\"]+)\"/){ $speaker_age=$1
                        ; }
106             }
107
108             # various counts
109             if($form and $form ne '0'){
110                 # count up overt words
111                 if($form ne '0'){ $counter{'words'}++; }
112                 # for each referent, count up distance to last mention iff it
                        has at least been mentioned once (otherwise it's not
                        defined)
113                 # note: zero referents are not counted because their position
                        is undefined, so A-0-A and 0-A-A should yield the same
                        distance for A-A
114                 foreach my $ID(keys %all_IDs){
115                     if($all_IDs{$ID}{'frequency'} > 0){
```

```
116                     $all_IDs{$ID}{'distance_to_last'}++;
117                     # remember present distance to last mention so that
                            distance_to_last can be set back to zero later
118                     $all_IDs{$ID}{'present_distance_to_last'} = $all_IDs{
                            $ID}{'distance_to_last'};
119                 }
120             }
121         }
122         # if there is an ID, update referent frequencies
123         my @IDs;
124         if($attributes{'identity'}){
125             # "identity" can contain several values concatenated by "+"
                    -> split
126             @IDs = split(/\+/,$attributes{'identity'});
127             foreach my $ID(@IDs){
128                 # count up frequency of referent
129                 $all_IDs{$ID}{'frequency'}++;
130                 # count up number of all referents
131                 $counter{'referents'}++;
132                 # update highest frequency in present text
133                 if($all_IDs{$ID}{'frequency'} > $counter{'top_frequency'})
                        {
134                     $counter{'top_frequency'} = $all_IDs{$ID}{'frequency'};
135                 }
136                 # set distance to last mention back to zero
137                 # note: present distance to last is still remembered as
                        present_distance_to_last so it can be used for
                        secondary attributes!
138                 $all_IDs{$ID}{'distance_to_last'} = 0;
139             }
140         }
141 
142         # for each domain, associate predicate, core referents, and
                roles
143         if($attributes{'domain'}){
144 
145             # (1) core referent = sth that has a domain plus an ID and a
                    role value
146             if($attributes{'identity'} && $attributes{'role'}){
147                 # count up number of overt and DAT marked core referents
148                 if($form ne '0'){
149                     $global_counter{'overt_core_referents'}++;
150                     if($attributes{'role'} =~ /^[PTG]$/){ $global_counter{'
                            overt_PTG'}++; }
151                     if($attributes{'role'} eq 'G'){ $global_counter{'
                            overt_G'}++; }
152                     if($form =~    /[][]?$/){ $global_counter{'
                            dat_core_referents'}++; }
153                     if($attributes{'role'} =~ /^[PTG]$/ && $form =~
                               /[][]?$/){ $global_counter{'dat_PTG'}++; }
154                 }
155 
156                 # "domain" and "roles" can contain several values
                        concatenated by "+" -> split
157                 my @domains = split(/\+/,$attributes{'domain'});
158                 my @roles = split(/\+/,$attributes{'role'});
159 
```

```
160                 # hybrids (words functioning as referent and predicate at
                        the same time): use first domain for predicate
161                 # use second domain in X:Y for all referential attributes
162                 if($attributes{'domain'} =~ /(.*):(.*)/){
163                     $all_domains{$1}{'predicate'}{'form'} = $form;
164                     $all_domains{$1}{'predicate'}{'absolute_position'} =
                            $counter{'words'};
165                     if($attributes{'diathesis'}){ $all_domains{$1}{'
                            predicate'}{'diathesis'} = $attributes{'diathesis'};
                             }
166                     @domains = $2;
167                 }
168
169                 # domain number must match argument number; otherwise
                        ignore
170                 if($#domains == $#roles){
171                     # go through domain/role pairs with same index and
                            assign attributes
172                     for(my $i=0; $i<=$#domains; $i++){
173                         # assign primary attributes to domain/role pair
174                         foreach my $att(keys %attributes){
175                             # don't look at domain/role themselves, which are
                                    higher-level keys
176                             unless($att eq 'domain' || $att eq 'role'){
177                                 # convert POS from NNC format
178                                 if($att eq 'ctag'){ $attributes{$att} = &
                                        convert_pos($attributes{$att}); }
179                                 $all_domains{$domains[$i]}{'roles'}{$roles[$i
                                        ]}{$att} = $attributes{$att};
180                             }
181                         }
182
183                         # remember role order for analysing word order
184                         if($form ne '0'){ push(@{$all_domains{$domains[$i]}{
                                'ordered_roles'}}, $roles[$i]); }
185                         else{ push(@{$all_domains{$domains[$i]}{'
                                ordered_roles'}}, $roles[$i].'0'); }
186
187                         # assign form used to code argument
188                         $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                                form'} = $form;
189
190                         ### calculate and assign secondary attributes
                                depending on the present position ###
191                         # absolute position in text: zeros have no position,
                                for overt words position equals present word
                                count
192                         if($form eq '0'){ $all_domains{$domains[$i]}{'roles'
                                }{$roles[$i]}{'absolute_position'} = 'none'; }
193                         else{ $all_domains{$domains[$i]}{'roles'}{$roles[$i
                                ]}{'absolute_position'} = $counter{'words'}; }
194                         # most secondary attributes are only relevant for
                                objects
195                         if($roles[$i] =~ /^[PTG]$/){
196
197                             # absolute and relative frequencies
198                             # in case of complex referents (several
                                    concatenated IDs), use the highest frequency
```

```
                involved as the benchmark
199             my $complex_frequency = 0;
200             foreach my $ID(@IDs){
201                if($all_IDs{$ID}{'frequency'} >
                       $complex_frequency){ $complex_frequency =
                       $all_IDs{$ID}{'frequency'}; }
202             }
203             $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                    absolute_frequency_sofar'} =
                    $complex_frequency;
204             # relative frequency can be calculated against
                    various benchmarks
205             $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                    ranked_frequency_sofar'} = $complex_frequency
                    / $counter{'top_frequency'};
206             $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                    relative_frequency_sofar'} =
                    $complex_frequency / $counter{'referents'};
207             # $all_domains{$domains[$i]}{'roles'}{$roles[$i
                    ]}{'relative_frequency_sofar'} = $all_IDs{$ID
                    }{'frequency'} / $counter{'words'};
208             # if frequency is minimal (1 mention only)
                    referent is new, otherwise given
209             if($complex_frequency == 1){ $all_domains{
                    $domains[$i]}{'roles'}{$roles[$i]}{'givenness'
                    } = 'new'; }
210             elsif($complex_frequency > 1){ $all_domains{
                    $domains[$i]}{'roles'}{$roles[$i]}{'givenness'
                    } = 'given'; }
211
212             # distance to last mention; only starts to get
                    counted after the first mention
213             # in case of complex referents, the distance of
                    all involved referents is checked and the
                    shortest one is used
214             my $complex_distance = 0;
215             foreach my $ID(@IDs){
216                if($all_IDs{$ID}{'present_distance_to_last'}
                       && $all_IDs{$ID}{'present_distance_to_last'
                       } > $complex_distance){
                      $complex_distance = $all_IDs{$ID}{'
                          present_distance_to_last'};
218                }
219             }
220             # if $complex_distance is still zero after
                    checking all participating IDs, none of the
                    referents has been mentioned so far
221             if($complex_distance == 0){ $complex_distance = '
                    NA'; }
222             $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                    distance_to_last'} = $complex_distance;
223
224             # computation of competing topics (frequency
                    above threshold A and distance to last mention
                     below threshold B)
225             my $competitors = 0;
226             foreach my $referent(keys %all_IDs){
```

```
227                             if(((($all_IDs{$referent}{'frequency'} /
                                    $complex_frequency) > 1)  # A = frequency
                                    of observed referent
228                             && $all_IDs{$referent}{'
                                    present_distance_to_last'}
229                             && ($all_IDs{$referent}{'
                                    present_distance_to_last'} < 50)){       #
                                    B = 50 words
230                                $competitors++;
231                             }
232                          }
233                          $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                                competitors'} = $competitors;
234                       } # EOF secondary attribute calculation
235                    } # EOF attribute assignment
236                 } # EOF loop on roles
237              } # EOF case "core referent"
238
239              # (2) predicate = sth that has a domain but no ID or role
                     value
240              elsif(!$attributes{'identity'} || !$attributes{'role'}){
241                 # remember V in role order
242                 push(@{$all_domains{$attributes{'domain'}}{'ordered_roles'
                        }}, 'V');
243
244                 # assign attributes to predicate (except domain itself,
                        which is a higher-level key)
245                 foreach my $att(keys %attributes){
246                    unless($att eq 'domain'){
247                       $all_domains{$attributes{'domain'}}{'predicate'}{
                              $att} = $attributes{$att};
248                    }
249                 }
250                 # assign form used to code predicate
251                 $all_domains{$attributes{'domain'}}{'predicate'}{'form'} =
                        $form;
252                 $all_domains{$attributes{'domain'}}{'predicate'}{'
                        absolute_position'} = $counter{'words'};
253
254              } # EOF case "predicate"
255
256              # (3) add speaker data to domain
257              my @domains = split(/\+/,$attributes{'domain'});
258              foreach(@domains){
259                 if($speaker_name){ $all_domains{$_}{'speaker_name'} =
                        $speaker_name; }
260                 else{
261                    if($genre eq 'written'){ $all_domains{$_}{'speaker_name
                           '} = $file; } # written files have only one speaker
                           each
262                    elsif($genre eq 'spoken'){ $all_domains{$_}{'
                           speaker_name'} = 'unknown'; }
263                 }
264                 if($speaker_sex){ $all_domains{$_}{'speaker_sex'} =
                        $speaker_sex; }
265                 else{ $all_domains{$_}{'speaker_sex'} = 'unknown'; }
266                 if($speaker_age){ $all_domains{$_}{'speaker_age'} =
                        $speaker_age; }
```

```
267                 else{ $all_domains{$_}{'speaker_age'} = 'unknown'; }
268             }
269
270          } # EOF core element check
271
272       } # EOF parsing XML structure
273
274       # for objects: calculate and assign secondary attributes that can
              only be seen in whole domains or the whole text
275       foreach my $domain(keys %all_domains){
276          foreach my $role(keys %{$all_domains{$domain}{'roles'}}){
277             if($role =~ /^[PTG]$/){
278                # mark O as passive when predicate is passive
279                if($all_domains{$domain}{'predicate'}{'diathesis'} &&
                       $all_domains{$domain}{'predicate'}{'diathesis'} eq '
                       passive'){
280                   $all_domains{$domain}{'roles'}{$role}{'diathesis'} = '
                          passive';
281                }
282                else{ $all_domains{$domain}{'roles'}{$role}{'diathesis'} =
                        'active'; }
283
284                # determine position of O relative to A and distance from
                       predicate
285                # when there is no A, relative position is not applicable
286                if(!$all_domains{$domain}{'roles'}{'A'}){ $all_domains{
                       $domain}{'roles'}{$role}{'relative_position'} = 'no A';
                        }
287                else{
288                   my $formA = $all_domains{$domain}{'roles'}{'A'}{'form'
                          };
289                   my $formO = $all_domains{$domain}{'roles'}{$role}{'form
                          '};
290                   # when either A or O is zero, relative position is not
                          applicable
291                   if($formA eq '0' || $formO eq '0'){ $all_domains{
                          $domain}{'roles'}{$role}{'relative_position'} = '
                          zero A/O'; }
292                   else{
293                      my $posA = $all_domains{$domain}{'roles'}{'A'}{'
                             absolute_position'};
294                      my $posO = $all_domains{$domain}{'roles'}{$role}{'
                             absolute_position'};
295                      if($posA < $posO){ $all_domains{$domain}{'roles'}{
                             $role}{'relative_position'} = 'AO'; }
296                      else{ $all_domains{$domain}{'roles'}{$role}{'
                             relative_position'} = 'OA'; }
297                   }
298                }
299                # distance from predicate only meaningful for non-zeros
300                if($all_domains{$domain}{'roles'}{$role}{'form'} ne '0' &&
                        $all_domains{$domain}{'predicate'}{'form'} ne '0'){
301                   $all_domains{$domain}{'roles'}{$role}{'
                          distance_from_predicate'} =
302                   -1 * ($all_domains{$domain}{'predicate'}{'
                          absolute_position'} - $all_domains{$domain}{'roles'
                          }{$role}{'absolute_position'});
```

```
303                     # print STDERR "distance: $all_domains{$domain}{'roles'
                            }{$role}{'distance_from_predicate'}, predicate:
                            $all_domains{$domain}{'predicate'}{'
                            absolute_position'}, ";
304                     # print STDERR "object: $all_domains{$domain}{'roles'}{
                            $role}{'absolute_position'}, ";
305                     # print STDERR "domain: $domain\n";
306                 }
307                 else{ $all_domains{$domain}{'roles'}{$role}{'
                        distance_from_predicate'} = 'NA'; }
308
309                 # determine relative frequencies in complete text
310                 # in case of complex referents (several concatenated IDs),
                        use the highest frequency involved as the benchmark
311                 my @IDs = split(/\+/,$all_domains{$domain}{'roles'}{$role
                        }{'identity'});
312                 my $complex_frequency = 0;
313                 foreach my $ID(@IDs){
314                     if($all_IDs{$ID}{'frequency'} > $complex_frequency){
                            $complex_frequency = $all_IDs{$ID}{'frequency'}; }
315                 }
316                 $all_domains{$domain}{'roles'}{$role}{'
                        absolute_frequency_total'} = $complex_frequency;
317                 # relative frequency option 1: relative to the most
                        frequent referent
318                 $all_domains{$domain}{'roles'}{$role}{'
                        ranked_frequency_total'} = $complex_frequency /
                        $counter{'top_frequency'};
319                 # relative frequency option 2: relative to all referents
320                 $all_domains{$domain}{'roles'}{$role}{'
                        relative_frequency_total'} = $complex_frequency /
                        $counter{'referents'};
321                 # relative frequency option 3: relative to all words
322                 # $all_domains{$domain}{'roles'}{$role}{'
                        relative_frequency_total'} = $all_IDs{$ID}{'frequency'}
                         / $counter{'words'};
323
324                 # determine co-argument case for ditransitive T
325                 if($role eq 'T'){
326                     if(!$all_domains{$domain}{'roles'}{'G'}){ $all_domains{
                            $domain}{'roles'}{$role}{'co_argument_case'} = 'no G
                            '; }
327                     else{
328                         if($all_domains{$domain}{'roles'}{'G'}{'form'} eq '0
                                '){
329                             $all_domains{$domain}{'roles'}{$role}{'
                                    co_argument_case'} = 'T with zero G';
330                         }
331                         elsif($all_domains{$domain}{'roles'}{'G'}{'form'} =~
                                 /[][]?$/){
332                             $all_domains{$domain}{'roles'}{$role}{'
                                    co_argument_case'} = 'T with G-DAT';
333                         }
334                         else{
335                             $all_domains{$domain}{'roles'}{$role}{'
                                    co_argument_case'} = 'T with other G';
336                         }
337                     }
```

```
338                 }
339                 else{ $all_domains{$domain}{'roles'}{$role}{'
                        co_argument_case'} = 'P/G'; }
340
341             } # EOF object check
342          } # EOF role check
343       } # EOF calculating secondary attributes
344
345       # append DOM data to $domtable
346       foreach my $domain(sort keys %all_domains){
347          foreach my $role(keys %{$all_domains{$domain}{'roles'}}){
348             my %attributes = %{$all_domains{$domain}{'roles'}{$role}};
349
350             # count up G with fixed DAT
351             if($attributes{'form'} ne 'O' && $role eq 'G' && $attributes{
                    'DOM'} && $attributes{'DOM'} eq 'NA'){ $global_counter{'
                    dat_G_nonobjects'}++; }
352
353             # only consider overt objects that are eligible for DOM and
                    that have all primary attributes specified
354             if($attributes{'form'} ne 'O' && $role =~ /^[PGT]$/ &&
                    $attributes{'DOM'} && $attributes{'DOM'} =~ /^(NOM|DAT)$/)
                    {
355                my $missingattflag = 0;
356                foreach my $variable("ctag","animacy","DOM","situation","
                       focus","quantifiability","modification"){
357                   if(!$attributes{$variable}){ $missingattflag = 1; print
                          "$variable missing in $file.$domain\n"; }
358                }
359                if($missingattflag == 0){
360                   # count up
361                   $global_counter{'overt_objects'}++;
362                   if($attributes{'DOM'} eq 'DAT'){ $global_counter{'
                          dat_objects'}++; }
363                   elsif($attributes{'DOM'} eq 'NOM'){ $global_counter{'
                          nom_objects'}++; }
364                   # add line to CSV table
365                   foreach my $att(@possible_attributes){
366                      my $val;
367                      if($att eq 'file'){ $val = $file; }
368                      elsif($att eq 'domain'){ $val = $domain; }
369                      elsif($att eq 'role'){ $val = $role; }
370                      elsif($att eq 'predicate'){ $val = $all_domains{
                             $domain}{'predicate'}{'form'}; }
371                      elsif($att eq 'genre'){ $val = $genre; }
372                      elsif($att eq 'speaker_name'){ $val = $all_domains{
                             $domain}{'speaker_name'}; }
373                      elsif($att eq 'speaker_sex'){ $val = $all_domains{
                             $domain}{'speaker_sex'}; }
374                      elsif($att eq 'speaker_age'){ $val = $all_domains{
                             $domain}{'speaker_age'}; }
375                      else{ $val = $attributes{$att}; }
376                      $domtable .= '"'.$val.'"';
377                      unless($att eq 'speaker_age'){ $domtable .= ','; }
378                   }
379                   $domtable .= "\n";
380                }
381
```

```
382             } # EOF printing relevant information
383          } # EOF loop on roles
384       } # EOF loop on domains
385
386       # append referent data to $referenttable
387       $referenttable = $referenttable."\n$file\n\n";
388       foreach my $ID(sort{$all_IDs{$b}{'frequency'} <=> $all_IDs{$a}{'
              frequency'}} keys %all_IDs){
389          $referenttable .= '"'.$ID.'","'.$all_IDs{$ID}{'frequency'}.'","'
                 .($all_IDs{$ID}{'frequency'} / $counter{'top_frequency'}).'",
                 "'.($all_IDs{$ID}{'frequency'} / $counter{'referents'}).'"'."
                 \n";
390       }
391       $referenttable = $referenttable."\n";
392
393       # append role data to $roletable
394       foreach my $domain(sort keys %all_domains){
395          $roletable .= $file.','.$domain.',';
396          $roletable .= join(' ', keys %{$all_domains{$domain}{'roles'}});
397          $roletable .= ',';
398          if($all_domains{$domain}{'ordered_roles'}){
399             $roletable .= join(' ', @{$all_domains{$domain}{'
                    ordered_roles'}});
400          }
401          $roletable .= "\n";
402       }
403
404
405
406       print STDERR "done.\n";
407    } # EOF loop on file
408
409 } # EOF loop on directory
410
411
412 ####################
413 ### OUTPUT DATA ###
414 ####################
415
416 open(OUT,'>:encoding(utf8)',"$directory/dom.csv");
417 print OUT $domtable;
418 close(OUT);
419
420 open(OUT,'>:encoding(utf8)',"$directory/referents.csv");
421 print OUT $referenttable;
422 close(OUT);
423
424 open(OUT,'>:encoding(utf8)',"$directory/roles.csv");
425 print OUT $roletable;
426 close(OUT);
427
428
429 # Print some basic statistics
430 select(STDOUT);
431 print "\nChecking completed, output written to files dom.csv and
       referents.csv in $directory. Global statistics:\n";
432
```

```
433 print "\t- out of $global_counter{'overt_core_referents'} overt core
        referents, $global_counter{'dat_core_referents'} have DAT (".sprintf("
        %.2f",(100*($global_counter{'dat_core_referents'}/$global_counter{'
        overt_core_referents'}))."%)\n";
434 print "\t- out of $global_counter{'overt_PTG'} overt P/T/G,
        $global_counter{'dat_PTG'} have DAT (".sprintf("%.2f",(100*(
        $global_counter{'dat_PTG'}/$global_counter{'overt_PTG'}))."%)\n";
435 print "\t- out of $global_counter{'overt_objects'} overt objects,
        $global_counter{'dat_objects'} have DAT (".sprintf("%.2f",(100*(
        $global_counter{'dat_objects'}/$global_counter{'overt_objects'}))."%)
         and $global_counter{'nom_objects'} have NOM (".sprintf("%.2f",(100*(
        $global_counter{'nom_objects'}/$global_counter{'overt_objects'}))."%)
        \n";
436 print "\t- out of $global_counter{'overt_G'} overt G, $global_counter{'
        dat_G_nonobjects'} have fixed DAT (".sprintf("%.2f",(100*(
        $global_counter{'dat_G_nonobjects'}/$global_counter{'overt_G'}))."%)\
        n";
437 print "\t- out of $global_counter{'dat_core_referents'} DAT-marked core
        referents, $global_counter{'dat_PTG'} are P/T/G (".sprintf("%.2f"
        ,(100*($global_counter{'dat_PTG'}/$global_counter{'dat_core_referents'
        }))."%)\n";
438 print "\t- out of $global_counter{'dat_PTG'} DAT-marked PTG,
        $global_counter{'dat_objects'} have DAT because of DOM (".sprintf("%.2
        f",(100*($global_counter{'dat_objects'}/$global_counter{'dat_PTG'}))).
        "%)\n";
439 print "\t- out of $global_counter{'dat_core_referents'} DAT-marked core
        referents, $global_counter{'dat_objects'} have DAT because of DOM (".
        sprintf("%.2f",(100*($global_counter{'dat_objects'}/$global_counter{'
        dat_core_referents'}))."%), $global_counter{'dat_G_nonobjects'} are
        fixed G-DAT (".sprintf("%.2f",(100*($global_counter{'dat_G_nonobjects'
        }/$global_counter{'dat_core_referents'}))."%), and $global_counter{'
        dat_G_nonobjects'} have DAT for other reasons (".sprintf("%.2f"
        ,(100*(($global_counter{'dat_core_referents'} - $global_counter{'
        dat_objects'} - $global_counter{'dat_G_nonobjects'})/$global_counter{'
        dat_core_referents'}))."%)\n";
440 print "Press enter.\n";
441 <>;
442
443
444 ######################
445 ###  SUBROUTINES  ###
446 ######################
447
448 # converts NNC POS to the format in the guidelines
449 sub convert_pos {
450    my $pos = shift;
451    if(!grep(/^$pos$/, ('n', 'adj', 'pro', 'dem', 'poss', 'other'))){
452       # first remove all case and other suffixes (separated from stem by
            "+")
453       $pos =~ s/\+.*//g;
454       # then convert the POS of the stem
455       if($pos eq 'NN'){ $pos = 'n'; }
456       elsif($pos =~ /^(J|MO)/){ $pos = 'adj'; }
457       elsif($pos =~ /^(PMX|PT[NMH]|PX[HR]|PRF)$/){ $pos = 'pro'; }
458       elsif($pos eq 'DDX'){ $pos = 'dem'; }
459       elsif($pos =~ /^P[MTR]/){ $pos = 'poss'; }
460       else{ $pos = 'other'; }
461    }
```

```
462
463     return $pos;
464 }
```

C.5 dom-consistency.pl

```
1  #!/usr/bin/perl -w
2  # Parses an NNC XML file with syntactic annotations and checks it for
       various mistakes output to mistakes.txt: existence of attributes and
       values, completeness of attributes, domain format, completeness and
       consistency of role sets
3  # Usage: perl parse-nnc-mistakes.pl (directory/with/files) - if no
       directory is given, the working directory is assumed
4  use Cwd;
5  use strict;
6  use utf8;
7  binmode (STDIN, ":utf8");
8
9
10 ######################
11 ### PRELIMINARIES ###
12 ######################
13
14 # define possible values for all attributes
15 my %possible_values;
16 $possible_values{'diathesis'} = ['passive'];
17 $possible_values{'animacy'} = ['human','human.fam','human.prop','human.
       group','high.anim','high.anim.prop','mid.anim','low.anim','thing','
       state','process','x'];
18 $possible_values{'ctag'} = ['n','adj','pro','poss','dem','other','x','NN'
       ,'NP','JM','JF','JO','JX','JT','MOM','MOF','MOO','MOX','DDX','PMX','
       PTN','PTM','PTH','PXH','PXR','PRF','PMXKM','PMXKF','PMXKO','PTNKM','
       PTNKF','PTNKO','PTMKM','PTMKF','PTMKO','PRFKM','PRFKF','PRFKO','PMXKX'
       ,'PTNKX','PTMKX','PRFKX','DDM','DDF','DDO','DKM','DKF','DKO','DKX','
       DJM','DJF','DJO','DJX','DGM','DGF','DGO','DGX','MM','TT','QQ','UU','
       NULL','VI','VDM','VDF','VDO','VDX','VE','VN','VQ','VCN','VCM','VCH','
       VS','VR','VVMX1','VVMX2','VVTN1','VVTX2','VVYN1','VVYX2','VVTN1F','
       VVTM1F','VVYN1F','VVYM1F','VOMX1','VOMX2','VOTN1','VOTX2','VOYN1','
       VOYX2','RR','RD','RK','RJ','II','IH','IE','IA','IKM','IKF','IKO','IKX'
       ,'MLM','MLF','MLO','MLX','CC','CSA','CSB','YF','YM','YQ','YB','FF','FS
       ','FB','FO','FZ','FU'];
19 $possible_values{'modification'} = ['none','adj','relclause','humposs','
       latposs','poss','num','dem','interrog','several','sortal','sortal.q','
       other','x'];
20 $possible_values{'DOM'} = ['NOM','DAT','NA','GEN','x'];
21 $possible_values{'domain'} = [];
22 $possible_values{'focus'} = ['nofoc','contrast','fragile','x'];
23 $possible_values{'identity'} = [];
24 $possible_values{'role'} = ['S','A','P','G','T','CR','CT','x'];
25 $possible_values{'situation'} = ['concrete','general','abstract','
       exemplary','x'];
26 $possible_values{'quantifiability'} = ['qnt','nonq','x'];
27
28 # read all files in directory
29 my $directory;
30 if($ARGV[0]){ $directory = $ARGV[0]; } else{ $directory = getcwd(); }
31 opendir(DIR,$directory) or die "No or wrong directory specified!\n";
```

```
32 my @files = readdir(DIR);
33 open(OUT,'>:encoding(utf8)',"$directory/mistakes.txt");
34 select(OUT);
35
36
37 ####################
38 ### PARSE FILES ###
39 ####################
40
41 # go through files in directory
42 foreach my $file(@files){
43    # open XML file
44    if($file =~ /(.*)\.xml$/){
45       print STDERR "checking $file...";
46       open(IN,'<:encoding(utf8)',"$directory/$file");
47       local $/;
48       my $text = <IN>;
49       my %all_domains;
50       close(IN);
51
52       # parse and check XML by going through <w> elements; collect data
             in %all_domains
53       while($text =~ /<w\s*([^>]+\"\s*)?\/>|<w\s*([^>]+\"\s*)?>([^<]+)
             ?<\/w>/sg){
54       #           empty tag + attr.     | content tag + attr. + content
55
56          # get form and attributes from <w>
57          my($form,$attributes,%attributes);
58          foreach($1,$2,$3){
59             if($_){
60                my $match = $_;
61                $match =~ s/\s+/ /sg;
62                if($match =~ /(\".*\")/s){ $attributes = $match; }
63                elsif($match =~ /^([\s\w\+]+)$/s){ $form = $match; }
64             }
65          }
66
67          # parse and check attributes
68          if($attributes){
69             # an element with attributes but without form is a zero (
                   elements without form AND without attributes are NNC
                   artefacts that are ignored)
70             if(!$form){ $form = '0'; }
71             # first get attribute/value pairs (domain is important for
                   locating mistakes)
72             while($attributes =~ /(\w+)=\"([^\"]+)\"/sg){ $attributes{$1}
                    = $2; }
73             # then check well-formedness
74             # first step: are there any attributes beside ID and POS?
75             # (only ID = isolated referent, only POS = word tagged
                   automatically during NNC creation)
76             my $attflag = 0;
77             foreach my $att(keys %attributes){
78                if($att ne 'identity' && $att ne 'ctag'){ $attflag = 1; }
79             }
80             # if there are other attributes, domain must be there - error
                    message if not
```

```
81              if($attflag == 1 && !$attributes{'domain'}){ print "$file:
                    domain missing for $form (unknown place)\n"; }
82              # ignored: $attflag==0 && !$attributes{'domain'}: only ID and
                    /or POS are there; ID has no predefined values, POS must
                    have been tagged automatically
83              # if domain is there, check other attributes
84              elsif($attributes{'domain'}){
85                 foreach my $att(keys %attributes){
86                    # check existence of attribute
87                    if(!$possible_values{$att}){ print "$file.$attributes{'
                          domain'}: variable \"$att\" doesn't exist\n"; }
88                    # if attribute exists and has predefined values, check
                          existence of value(s)
89                    elsif($possible_values{$att} && @{$possible_values{$att
                          }}){
90                       # if value consists of several values joined by "+",
                             split
91                       my @single_values = split(/\+/,$attributes{$att});
92                       foreach my $sval(@single_values){
93                          if(!grep(/^$sval$/,@{$possible_values{$att}})){
                                print "$file.$attributes{'domain'}: value \"
                                $sval\" of $att doesn't exist\n"; }
94                       }
95                    }
96                 }
97
98              } # EOF attribute/value check
99           } # EOF parsing attributes
100
101          # for each domain, associate predicate, core referents, and
                roles
102          if($attributes{'domain'}){
103             my $domain = $attributes{'domain'};
104
105             # (1) core referent = sth that has a domain plus an ID and a
                    role value
106             # anomaly: ID or role is there but not the other value
107             if(($attributes{'identity'} && !$attributes{'role'})
108             || (!$attributes{'identity'} && $attributes{'role'})){
109                print "$file.$domain: $form looks like a core referent,
                       but identity or role is missing (or redundant domain/
                       identity/role tag)\n";
110             }
111             elsif($attributes{'identity'} && $attributes{'role'}){
112                # "domain", "identity" and "roles" can contain several IDs
                       concatenated by "+" -> split
113                my @domains = split(/\+/,$domain);
114                my @IDs = split(/\+/,$attributes{'identity'});
115                my @roles = split(/\+/,$attributes{'role'});
116
117                # hybrids (words functioning as referent and predicate at
                       the same time): use first domain for predicate
118                # use second domain in X:Y for all attributes
119                if($domain =~ /(.*):(.*)/){
120                   $all_domains{$1}{'predicate'}{'form'} = $form;
121                   @domains = $2;
122                }
123
```

```
124                 # domain number must match argument number
125                 if($#domains != $#roles){ print "$file.$domain: number of
                        domains doesn't match number of roles\n"; }
126                 elsif($#domains == $#roles){
127                    # go through domain/role pairs with same index
128                    for(my $i=0; $i<=$#domains; $i++){
129
130                       # assign attributes to domain/role pair (except
                           domain/role themselves, which are higher-level
                           keys)
131                       foreach my $att(keys %attributes){
132                          unless($att eq 'domain' || $att eq 'role'){
133                             $all_domains{$domains[$i]}{'roles'}{$roles[$i
                                 ]}{$att} = $attributes{$att};
134                          }
135                          # check domain format
136                          if($att eq 'domain'){
137                             if($attributes{$att} !~ /^[\d\/\+:a-z]+$/){
138                                print "$file.$domain: domain \"$attributes{
                                    $att}\" has wrong format\n";
139                             }
140                          }
141
142                       }
143                       # assign form used to code argument
144                       $all_domains{$domains[$i]}{'roles'}{$roles[$i]}{'
                           form'} = $form;
145                    }
146                 } # EOF loop on roles
147
148              } # EOF case "core referent"
149
150              # (2) predicate = sth that has a domain but no ID or role
                    value
151              elsif(!$attributes{'identity'} || !$attributes{'role'}){
152                 # assign attributes to predicate (except domain itself,
                        which is a higher-level key)
153                 foreach my $att(keys %attributes){
154                    unless($att eq 'domain'){
155                       $all_domains{$domain}{'predicate'}{$att} =
                           $attributes{$att};
156                    }
157                    # predicates only allow a few attributes; error if
                        others are present
158                    if($att !~ /^(diathesis|ctag|domain)$/){
159                       print "$file.$domain: $form looks like a predicate,
                           but $att is not allowed\n";
160                    }
161                 }
162                 # assign form used to code predicate
163                 $all_domains{$domain}{'predicate'}{'form'} = $form;
164              } # EOF case "predicate"
165
166           } # EOF core element check
167
168        } # EOF parsing XML structure
169
170        # check overall syntactic structure
```

```
171         foreach my $domain(sort keys %all_domains){
172            # check completeness
173            if(!$all_domains{$domain}{'predicate'}){ print "$file.$domain:
                   no predicate\n"; }
174            if($all_domains{$domain}{'roles'}){
175               my @roles = keys %{$all_domains{$domain}{'roles'}};
176
177               # check completeness of role sets
178               if(!grep(/^x$/,@roles) && (
179                  (grep(/^A$/,@roles) && !grep(/^[PGT]$/,@roles))
180               || (grep(/^P$/,@roles) && !grep(/^A$/,@roles))
181               || (grep(/^G$/,@roles) && !grep(/^A$/,@roles))
182               || (grep(/^G$/,@roles) && !grep(/^T$/,@roles))
183               || (grep(/^T$/,@roles) && !grep(/^A$/,@roles))
184               || (grep(/^T$/,@roles) && !grep(/^G$/,@roles))
185               || (grep(/^CT$/,@roles) && !grep(/^CR$/,@roles))
186               || (grep(/^CR$/,@roles) && !grep(/^CT$/,@roles))
187               )){ print "$file.$domain: role missing in set {@roles}\n"; }
188
189               # check consistency of role sets
190               if((grep(/^S$/,@roles) && ($#roles > 0))
191               || (grep(/^A$/,@roles) && grep(/^(S|CT|CR)$/,@roles))
192               || (grep(/^P$/,@roles) && grep(/^(S|G|T|CT|CR)$/,@roles))
193               || (grep(/^[GT]$/,@roles) && grep(/^(S|P|CT|CR)$/,@roles))
194               || (grep(/^(CT|CR)$/,@roles) && grep(/^[SAGT]$/,@roles))
195               ){ print "$file.$domain: set {@roles} is inconsistent\n"; }
196
197               # check completeness of DOM tags
198               foreach my $role(@roles){
199                  my %attributes = %{$all_domains{$domain}{'roles'}{$role}};
200                  if($role =~ /^[PGT]$/ && !$attributes{'DOM'}){ print "
                       $file.$domain: DOM tags missing\n"; }
201                  # objects selected by DOM
202                  elsif($role =~ /^[PGT]$/ && $attributes{'DOM'} &&
                       $attributes{'DOM'} =~ /^(NOM|DAT)$/){
203                     foreach my $variable("animacy","ctag","modification","
                          DOM","focus","situation","quantifiability"){
204                        if(!$attributes{$variable}){ print "$file.$domain:
                             missing DOM variable $variable\n"; }
205                     }
206                  }
207               } # EOF loop on roles
208
209            } # EOF role checks
210
211         } # EOF checking structure
212
213         print STDERR "done.\n";
214
215      } # EOF loop on file
216
217   } # EOF loop on directory
218
219
220   close(OUT);
221
222
223   select(STDOUT);
```

```
print "\nChecking completed, output written to file mistakes.txt. Press enter.\n";
<>;
```

C.6 dom-analysis.R

```
# Does some statistical analyses of DOM annotations in the NNC. Expected input format is CSV.
# Usage: source("path/to/dom-analysis.R")


#####################
### PRELIMINARIES ###
#####################

# read in DOM table and preprocess
source("dom-analysis-preprocessing.R")
library(ltm)
library(rms)
datadist(dom) -> domdata
options(datadist = "domdata")
# needed to deal with ordered factors
options(contrasts=c("contr.treatment","contr.treatment"))


############################
### CATEGORIAL VARIABLES ###
############################

# print numbers and proportions of DAT/NOM in all objects
cat("\n--- Summary: ---\n\n")
print(table(dom$DOM))
print(prop.table(table(dom$DOM)))
cat("\n");

# check all categorial variables
categorial_variables <- c("role", "ctag", "modification", "animacy", "situation", "quantifiability", "focus", "givenness", "relative_position", "co_argument_case", "diathesis", "genre")
for(var in categorial_variables){
   cat("\n\n-----------------------------------------------------------------------\n\n--- Variable:", var, "---\n\n")
   # simple contingency table
   xtabs(~ dom[,var] + dom$DOM) -> cont_table
   print(cont_table)
   cat("\n")
   # contingency table with values proportional to row sums
   prop.table(cont_table,1) -> prop_table
   print(prop_table)
   cat("\n")
   # Chi square test with Yate's correction or Fisher's exact test, + Pearson's C (between 0-1)
   chisq.test(cont_table) -> chi
   if(nrow(cont_table) == 2){
      fisher.test(cont_table) -> ftest
      print(ftest)
```

```
46        cat("\n")
47    }
48    else{ print (chi) }
49
50    # calculate and print coefficients
51    pearsons_c <- sqrt(chi$statistic/(sum(cont_table)+chi$statistic))
52    # will normally always be 2 because of NOM/DAT
53    dim_min <- min(nrow(cont_table),ncol(cont_table))
54    # corrected_c <- sqrt(dim_min/(dim_min-1)) * pearsons_c
55    cramers_v <- sqrt(chi$statistic/(sum(cont_table)*(dim_min-1)))
56    cat("Pearson's contingency coefficient:", pearsons_c, "\nCramer's V:",
          cramers_v, "\n\n")
57
58    # logistic regression with single variables
59    tmp.lrm = lrm(dom$DOM ~ dom[,var])
60    print(anova(tmp.lrm))
61    print(tmp.lrm)
62
63    # check impact of individual values
64    cat("--- Individual values ---\n\n")
65    for(val in levels(dom[,var])){
66       # for each value of the variable, collapse all others and do a
             Fisher's exact test
67       collapsed <- dom[,c(var,"DOM")]
68       levels(collapsed[,var])[levels(collapsed[,var]) != val] <- "the_
             rest"
69       xtabs(~ collapsed[,var] + collapsed$DOM) -> single_val_table
70       cat("p-value < Fisher's exact test for \"", val, "\" being relevant
             : ", sep="")
71       print(fisher.test(single_val_table)$p.value)
72    }
73
74    # plot mosaic plot for every variable to files in graphics folder
75    pdf(paste("/Users/robertschikowski/ /The core of all things/graphics/"
          ,var,".pdf", sep=""))
76    par(mar=c(0.2,0.2,0.2,0.2), cex=2.5)
77    if(nlevels(dom[,var]) == 2 | var %in% c("role","relative_position")){
78       mosaicplot(as.matrix((xtabs(~ dom[,var] + dom$DOM))), color=F, main
             ="", xlab="", ylab="", las=1)
79    }
80    else{ mosaicplot(as.matrix((xtabs(~ dom[,var] + dom$DOM))), color=F,
          main="", xlab="", ylab="", las=2) }
81    dev.off()
82 }
83
84
85 ####################################
86 ### INSPECT VARIABLE COMBINATIONS ###
87 ####################################
88
89 # unusual combinations
90 cat("Human quantifiable referents with NOM (excluding passives):\n\n")
91 print(dom[dom$animacy=="human" & dom$quantifiability=="qnt" & dom$
       diathesis!="passive" & dom$DOM=="NOM",])
92 cat("\nNon-human non-quantifiable referents with DAT:\n\n")
93 print(dom[dom$animacy!="human" & dom$quantifiable=="nonq" & dom$DOM=="DAT
       ",])
94 # number and frequency of distinct combinations
```

```
95  combi <- sort(table(paste(dom[,"DOM"], dom[,"role"], dom[,"ctag"], dom[,"
        animacy"], dom[,"situation"], dom[,"quantifiability"], dom[,"givenness
        "], dom[,"relative_position"], dom[,"co_argument_case"], sep=":")),
        decreasing=TRUE)
96  combi_all <- sort(table(paste(dom[,"DOM"], dom[,"role"], dom[,"ctag"],
        dom[,"modification"], dom[,"animacy"], dom[,"situation"], dom[,"
        quantifiability"], dom[,"givenness"], dom[,"relative_position"], dom[,
        "focus"], dom[,"co_argument_case"], dom[,"diathesis"], sep=":")),
        decreasing=TRUE)
97  cat("\nNumber of distinct value combinations in all categorial variables:
        ", nrow(combi_all))
98  cat("\nNumber of distinct value combinations in all relevant categorial
        variables:", nrow(combi))
99  cat("\nNumber of distinct value combinations with frequency higher than
        25 (= approximately half of all combinations) in all relevant
        categorial variables:", nrow(combi[combi>25]))
100 cat("\nAll distinct value combinations with frequency higher than 25 in
        all relevant categorial variables::\n")
101 print(combi[combi>25])
102
103
104 ##############################
105 ### QUANTITATIVE VARIABLES ###
106 ##############################
107
108 quantitative_variables <- c("distance_from_predicate", "ranked_frequency_
        sofar", "ranked_frequency_total", "relative_frequency_sofar", "
        relative_frequency_total", "distance_to_last", "competitors")
109 dom$DOM <- factor(dom$DOM, levels=c("NOM","DAT"), order=TRUE)
110
111 # check all quantitative variables
112 for(var in quantitative_variables){
113    cat("
           ------------------------------------------------------------------------\
           n\n--- Variable:", var, "---\n\n")
114    print(summary(dom[,var]))
115    combi <- matrix(c(dom[,var], dom$DOM), ncol=2)
116    combi <- na.omit(combi)
117
118    # point-biserial correlation. reverse sign so that positive pbi gets
           associated with "higher value -> more DAT" (confusing default is "
           higher value -> more NOM")
119    cat("\nPoint-biserial correlation for", var, "and NOM (0) ~ DAT (1): "
           )
120    print(-1 * biserial.cor(combi[,1], combi[,2]))
121    cat("\n")
122
123    # logistic regression for single variable
124    if(var == "distance_to_last" | var == "competitors"){ tmp.lrm = lrm(
           dom$DOM ~ dom[,var]) }
125    # most quantitative variables are exponentially related to DAT
126    else{ tmp.lrm = lrm(dom$DOM ~ dom[,var]^2) }
127    print(anova(tmp.lrm))
128    print(tmp.lrm)
129
130    # plot distribution of random variable compared to random variable +
           DAT
```

```
131     pdf(paste("/Users/robertschikowski/ /The core of all things/graphics/"
            ,var,".pdf", sep=""))
132     # par(mar=c(2,4,0.1,0.1), cex=2, family="Linux Libertine O") # LLO not
            available with pdf device, but does work on the console
133     par(mar=c(2,4,0.2,0.2), cex=2)
134     xvar <- table(dom[,var])
135     xdat <- as.table(xtabs(~ dom[,var] + dom$DOM)[,"DAT"])
136     xnom <- as.table(xtabs(~ dom[,var] + dom$DOM)[,"NOM"])
137     interpol=50
138     linetype="l"
139     if(var == "competitors"){
140        interpol=max(dom[,var])
141        linetype="o"
142     }
143     # make sure that the ranges of x and y are the original ones
144     plot(approx(rownames(xvar), xvar, n=interpol), type=linetype, lwd="1",
            xlab="", ylab="frequency")
145     points(approx(rownames(xdat), xdat, n=interpol), col="red", type=
            linetype, lwd="1")
146     points(approx(rownames(xnom), xnom, n=interpol), col="green", type=
            linetype, lwd="1")
147     legend("topright",c("all","NOM","DAT"),col=c("black","green","red"),
            pch=15)
148     dev.off()
149 }
```

C.7 dom-regression.R

```
1  # Does multivariate logistic regression of DOM annotations in the NNC and
       compares the results with a rule-based model of DOM.
2  # Usage: source("path/to/dom-regression.R")
3
4
5  ######################
6  ### PRELIMINARIES ###
7  ######################
8
9  source("dom-regression-preprocessing.R")
10 library(rms)
11 datadist(dom) -> domdata
12 options(datadist = "domdata")
13 options(contrasts=c("contr.treatment","contr.treatment")) # needed to
       deal with ordered factors
14
15 all_variables <- c("human", "humposs", "process", "abstract",
16    "quantifiability", "situation", "pronoun", "demonstrative",
17    "givenness", "ranked_frequency_sofar", "ranked_frequency_total", "
          competitors", "distance_to_last", "focus",
18    "role", "co_argument_case", "relative_position", "distance_from_
          predicate", "diathesis", "genre")
19 evaluators <- c("R2", "C", "Dxy", "P", "Model L.R.")
20
21
22 ####################################
23 ### OVERVIEW OF SINGLE VARIABLES ###
24 ####################################
25
```

```
26 good_variables <- numeric(0)
27 var_eval = data.frame("variable" = numeric(100), "R2" = numeric(100), "C"
       = numeric(100), "Dxy" = numeric(100), "P" = numeric(100), "Model L.R.
      " = numeric(100))
28 counter = 1
29 # go through all variables
30 for(var in all_variables){
31    cat("\n---", var, "---\n\n")
32    if(var %in% c("distance_from_predicate", "ranked_frequency_sofar", "
          ranked_frequency_total")){ tmp.lrm = lrm(dom$DOM ~ dom[,var]^2) }
33    else{ tmp.lrm = lrm(dom$DOM ~ dom[,var]) }
34    print(anova(tmp.lrm))
35    print(tmp.lrm)
36    if(tmp.lrm$stats["P"] < 0.05){ good_variables <- c(good_variables, var
          ) }
37    var_eval[counter,"variable"] = var
38    for(stat in evaluators){
39       var_eval[counter,stat] = tmp.lrm$stats[stat]
40    }
41    counter = counter+1
42 }
43 cat("variables with significant effect taken alone:", paste(good_
      variables, collapse=", "), "\n\n")
44 var_eval <- var_eval[var_eval$R2 > 0,]
45 print(var_eval[order(var_eval$R2, decreasing=T),])
46 cat("\n")
47
48
49 ########################################
50 ### MULTIVARIATE LOGISTIC REGRESSION ###
51 ########################################
52
53 # check maximal model
54 # reg_variables <- good_variables
55 reg_variables <- setdiff(all_variables, c("givenness","ranked_frequency_
      sofar", "distance_to_last"))
56 varforreg <- paste(reg_variables, collapse="+")
57 varforreg <- gsub("distance_from_predicate", "distance_from_predicate^2",
       varforreg)
58 varforreg <- gsub("ranked_frequency_total", "ranked_frequency_total^2",
      varforreg)
59 regr_formula <- as.formula(paste("DOM ~", varforreg))
60 dom.lrm <- lrm(regr_formula, data=dom, x=T, y=T)
61 cat("\nMaximal model including all variables:\n")
62 print(dom.lrm)
63
64 # get good variables using fast backwards elimination
65 cat("\nDoing validation with fast backwards elimination...\n")
66 runs <- 10000
67 validation <- validate(dom.lrm, bw=T, B=runs) # default
68 # validation <- validate(dom.lrm, bw=T, B=runs, aics=10000) # high
      threshold for keeping variables drops all in order of importance
69 final_reg_variables <- setdiff(reg_variables, c("role", "situation", "
      relative_position", "pronoun"))
70 varforreg <- paste(final_reg_variables, collapse="+")
71 regr_formula <- as.formula(paste("DOM ~", varforreg))
72 dom.lrm <- lrm(regr_formula, data=dom, x=T, y=T)
73 cat("\nModel with variables retained in fast backwards elimination:\n")
```

```
74  print(dom.lrm)
75
76  # find out best penalty for avoiding overfitting
77  cat("\nTrying to find best penalty for model with retained variables...\n
        ")
78  dom.penalty <- pentrace(dom.lrm, seq(0,1,by=0.05))$penalty
79  # refit model with penalty
80  dom.lrm <- lrm(regr_formula, data=dom, x=T, y=T, penalty=dom.penalty)
81  cat("\nFinal model with penalty:\n")
82  print(dom.lrm)
83
84  # plot relations between variables and probability for DAT
85  pdf("/Users/robertschikowski/ /The core of all things/graphics/final-
        variables.pdf")
86  p <- Predict(dom.lrm, fun=plogis)
87  p$.predictor. <- factor(p$.predictor, final_reg_variables)
88  print(plot(p))
89  dev.off()
90
91  # build mixed-effects model
92  library(lme4)
93  regr_formula_mixed <- as.formula(paste("DOM ~", varforreg, "+ (1|speaker_
        name)"))
94  mixed.eff <- lmer(regr_formula_mixed, data=dom, family="binomial")
95  cat("\nMixed model:\n\n")
96  print(mixed.eff)
97  random.effects <- ranef(mixed.eff)$speaker_name
98  fixed.effects <- fixef(mixed.eff)
99
100
101 ########################################
102 ### PREDICT CASE WITH VARIOUS MODELS ###
103 ########################################
104
105 cat("\nPredicting case based on probabilistic, rule-based and hybrid
        models...\n")
106
107 # probabilistic model: insert probability based on logit(DAT) function
108 domf <- Function(dom.lrm)
109 dom$pred_prob <- plogis(sapply(1:nrow(dom), function(i) do.call(domf, dom
        [i,final_reg_variables])))
110
111 # insert probability based on mixed-effects model. lme4 doesn't have a
        predict() function, so a custom function has to be built with the
        values from mixed.eff
112 mixpred <- function(speaker_name, human, humposs, process, abstract,
        quantifiability, demonstrative, ranked_frequency_total, competitors,
        focus, co_argument_case, distance_from_predicate, diathesis, genre){
113     # start with intercept
114     p = -12.96956
115     # add random effect of speaker identity
116     p = p + random.effects[speaker_name,]
117     # add fixed effects
118     p = p + (human=="human")*4.41893 + (humposs=="humposs")*1.08233 + (
          process=="non-process")*2.04366 + (abstract=="abstract")*0.74935 +
          (quantifiability=="qnt")*1.96179 + (demonstrative=="dem")*1.75519 +
           ranked_frequency_total*1.96743 + competitors*-0.12717 + (focus=="
          fragile")*1.52449 + (co_argument_case=="ordinary 0")*3.26479 +
```

```
        distance_from_predicate*-0.22072 + (diathesis=="active")*2.00735 +
        (genre=="written")*1.09919
    # probability from logit
    p = plogis(p)
    return(p)
}

random_and_fixed <- c("speaker_name",final_reg_variables)
dom$pred_mix_eff <- sapply(1:nrow(dom), function(i) do.call(mixpred, dom[
    i,random_and_fixed]))

# drop rows where probability is NA (< distance_from_predicate = NA),
    otherwise they will be counted as incorrect predictions later
dom <- dom[!is.na(dom$pred_prob) & !is.na(dom$pred_mix_eff),]

# rule-based and hybrid model: prediction based on salient cases
for(pred_type in c("pred_rule", "pred_hybrid")){
    dom[dom$animacy == "human" & dom$quantifiability == "qnt", pred_type]
        = 1
    dom[dom$ctag == "pro", pred_type] = 1
    dom[dom$demonstrative == "dem" & dom$quantifiability == "qnt", pred_
        type] = 1
    dom[dom$ranked_frequency_total > 0.26, pred_type] = 1
    dom[dom$animacy == "process", pred_type] = 0
    dom[dom$quantifiability == "nonq", pred_type] = 0
    dom[dom$co_argument_case == "T with G-DAT", pred_type] = 0
}
# rule-based model: cases not covered yet are all NOM
dom$pred_rule[is.na(dom$pred_rule)] = 0
# hybrid model: cases not covered yet are taken from mixed-effects model
dom$pred_hybrid[is.na(dom$pred_hybrid)] = dom$pred_mix_eff[is.na(dom$pred
    _hybrid)]

# build summary table where one line = one unique combination of values
    of final variables + its relative frequency and case predictions
library(plyr)
dom$numdom[dom$DOM != "DAT"] <- 0
dom$numdom[dom$DOM == "DAT"] <- 1
combivar <- c(final_reg_variables, "pred_prob", "pred_mix_eff", "pred_
    rule", "pred_hybrid")
summary <- ddply(dom, combivar, summarise, freq = length(DOM), dat_prop =
     mean(numdom))


###################################
### CALCULATE PREDICTIVE POWER ###
###################################

# simple accuracy
cat("\nSimple accuracy:\n")
cat("\tNull model:", (nrow(dom[dom$DOM=="NOM",])/nrow(dom)), "(all), 0.0
    (DAT), 1.0 (NOM)\n")
for(pred_type in c("pred_rule", "pred_prob", "pred_mix_eff", "pred_hybrid
    ")){
    correct_all <- nrow(dom[(dom$DOM == "DAT" & dom[,pred_type] > 0.5)
                      | (dom$DOM == "NOM" & dom[,pred_type] <= 0.5),]) /
                          nrow(dom)
```

```
163     correct_DAT <- nrow(dom[dom$DOM == "DAT" & dom[,pred_type] > 0.5,]) /
            nrow(dom[dom$DOM=="DAT",])
164     correct_NOM <- nrow(dom[dom$DOM == "NOM" & dom[,pred_type] <= 0.5,]) /
             nrow(dom[dom$DOM=="NOM",])
165     cat("\t", pred_type, ": ", correct_all, " (all), ", correct_DAT, " (
            DAT), ", correct_NOM, " (NOM); ", sep="")
166
167     # accuracy when p() is taken as predictor of DAT proportions in
            summary table
168     cat("mean digression of probability/proportion: ",
169     sum((abs(summary$dat_prop - summary[,pred_type]) * summary$freq) / sum
            (summary$freq)), "\n", sep="")
170
171 }
172
173 # accuracy after resampling, optionally with simulation
174 runs <- 10000
175 counter = 1
176 cat("\nMean accuracies after", runs, "runs of resampling - ")
177 # simulation function
178 simulate_case <- function(prob){
179    return(sample(c("DAT","NOM"), 1, prob=c(prob, 1-prob)))
180 }
181
182 # do resampling
183 resampling_eval = array(0, dim=c(runs,4,7), dimnames=list(c(), c("all","
        DAT","NOM","mean_digr"), c("pred_rule","pred_prob","pred_mix_eff","
        pred_hybrid","pred_sim","pred_sim_mix_eff","pred_null")))
184 cat("run ")
185 for(i in 1:runs){
186    cat(i, "... ", sep="")
187    # determine number of observations to be drawn, minimum is one where
            1000 observations are thrown away
188    obs_number <- sample((nrow(dom)-1000):nrow(dom)-100,1)
189    # get subset of DOM
190    subdom <- dom[sample(nrow(dom),obs_number),]
191    # some rows where distance_from_predicate is NA (= 0 referents) also
            have p=NA
192    subdom <- na.omit(subdom)
193
194    # build summary table
195    summary <- ddply(subdom, combivar, summarise, freq = length(DOM), dat_
            prop = mean(numdom))
196
197    # accuracy of various prediction methods
198    for(pred_type in c("pred_rule","pred_prob","pred_mix_eff","pred_hybrid
            ")){
199       resampling_eval[counter,"all",pred_type] <- nrow(subdom[(subdom$DOM
              == "DAT" & subdom[,pred_type] > 0.5)
200                                          | (subdom$DOM == "NOM" & subdom
                                               [,pred_type] <= 0.5),]) /
                                              nrow(subdom)
201       resampling_eval[counter,"DAT",pred_type] <- nrow(subdom[subdom$DOM
              == "DAT" & subdom[,pred_type] > 0.5,]) / nrow(subdom[subdom$DOM
              =="DAT",])
202       resampling_eval[counter,"NOM",pred_type] <- nrow(subdom[subdom$DOM
              == "NOM" & subdom[,pred_type] <= 0.5,]) / nrow(subdom[subdom$DOM
              =="NOM",])
```

```
203         resampling_eval[counter,"mean_digr",pred_type] <- sum((abs(summary$
                dat_prop - summary[,pred_type]) * summary$freq) / sum(summary$
                freq))
204     }
205
206     # concordance with simulation
207     subdom$sim_case = sapply(subdom$pred_prob, simulate_case)
208     resampling_eval[counter,"all","pred_sim"] <- nrow(subdom[subdom$DOM ==
            subdom$sim_case,]) / nrow(subdom)
209     resampling_eval[counter,"DAT","pred_sim"] <- nrow(subdom[subdom$DOM ==
            "DAT" & subdom$sim_case == "DAT",]) / nrow(subdom[subdom$DOM=="DAT
            ",])
210     resampling_eval[counter,"NOM","pred_sim"] <- nrow(subdom[subdom$DOM ==
            "NOM" & subdom$sim_case == "NOM",]) / nrow(subdom[subdom$DOM=="NOM
            ",])
211     resampling_eval[counter,"mean_digr","pred_sim"] <- NA
212     subdom$sim_case_mix_eff = sapply(subdom$pred_mix_eff, simulate_case)
213     resampling_eval[counter,"all","pred_sim_mix_eff"] <- nrow(subdom[
            subdom$DOM == subdom$sim_case_mix_eff,]) / nrow(subdom)
214     resampling_eval[counter,"DAT","pred_sim_mix_eff"] <- nrow(subdom[
            subdom$DOM == "DAT" & subdom$sim_case_mix_eff == "DAT",]) / nrow(
            subdom[subdom$DOM=="DAT",])
215     resampling_eval[counter,"NOM","pred_sim_mix_eff"] <- nrow(subdom[
            subdom$DOM == "NOM" & subdom$sim_case_mix_eff == "NOM",]) / nrow(
            subdom[subdom$DOM=="NOM",])
216     resampling_eval[counter,"mean_digr","pred_sim_mix_eff"] <- NA
217
218     # accuracy of null model
219     resampling_eval[counter,"all","pred_null"] <- nrow(subdom[subdom$DOM
            == "NOM",]) / nrow(subdom)
220     resampling_eval[counter,"DAT","pred_null"] <- 0
221     resampling_eval[counter,"NOM","pred_null"] <- 1
222     resampling_eval[counter,"mean_digr","pred_null"] <- nrow(subdom[subdom
            $DOM == "DAT",]) / nrow(subdom)
223
224     # digression of prediction from proportion within unique value
            combinations
225     # subdom$numdom[subdom$DOM != "DAT"] <- 0
226     # subdom$numdom[subdom$DOM == "DAT"] <- 1
227     # summary <- ddply(subdom, final_reg_variables, summarise, freq =
            length(DOM), dat_prop = mean(numdom))
228     # summary$pred_prob <- plogis(sapply(1:nrow(summary), function(i) do.
            call(domf, summary[i,final_reg_variables])))
229     # summary$diff <- abs(summary$dat_prop - summary$pred_prob)
230     # resampling_eval[counter,"all","mean_digr"] <- sum(summary$diff *
            summary$freq) / sum(summary$freq)
231
232     counter=counter+1
233 }
234
235 # print/plot means and densities of predictive accuracy and of models in
        resampling
236 pdf("/Users/robertschikowski/ /The core of all things/graphics/predictive
        -accuracy.pdf")
237 plot(x=0, y=0, xlim=c(0.4,1), ylim=c(0,160), xlab="proportion of
        concordant predictions", ylab="density", main="")
238 params <- data.frame(output=c("brown",    "black",    "red",      "green"
        , "violet",   "blue",        "grey",     "solid","dashed","dotted"),
```

```
239                 row.names=c("pred_prob","pred_mix_eff","pred_rule","pred_
                        hybrid","pred_sim","pred_sim_mix_eff","pred_null","all"
                        ,"DAT","NOM"))
240 cat("\n")
241 for(pred_type in c("pred_rule","pred_prob","pred_mix_eff","pred_sim","
        pred_sim_mix_eff","pred_hybrid","pred_null")){
242    cat("\t", pred_type, ":\n", sep="")
243    for(aspect in c("all","DAT","NOM","mean_digr")){
244       cat("\t\t", aspect, ": ", mean(resampling_eval[,aspect,pred_type]),
              "\n", sep="")
245       if(pred_type %in% c("pred_rule","pred_mix_eff","pred_sim_mix_eff")
              & aspect %in% c("all","DAT","NOM")){
246          lines(density(resampling_eval[,aspect,pred_type]), lty=as.vector
                 (params[aspect,"output"]), col=as.vector(params[pred_type,"
                 output"]))
247       }
248    }
249 }
250 legend(x=0.4, y=160, ncol=3, legend=c("prob. pred. all", "prob. pred. DAT
        ", "prob. pred. NOM", "prob. sim. all", "prob. sim. DAT", "prob. sim.
        NOM", "rule-based all", "rule-based DAT", "rule-based NOM"), cex=0.8,
        lty=c("solid","dashed","dotted"), col=c("black","black","black","blue"
        ,"blue","blue","red","red","red"))
251 dev.off()
```

Appendix D

Verb paradigms

D.1 Chintang

This section contains full paradigms for all finite and non-finite forms of Chintang, both as found in isolation and in verb compounding. Most cells show bipersonal agreement with the row indicating A and the column indicating P. Monopersonal agreement is shown in the last column with the row indicating S. Each paradigm cell contains two forms, the upper one being the non-past form, the lower one past. Negation is so regular that it is not necessary to list negated forms separately. The paradigms are the joint result of research carried out during the CPDP and my own field work.

D.1.1 Indicative

<table>
<tr><th>A/P</th><th>1s</th><th>1de</th><th>1pe</th><th>1di</th><th>1pi</th><th>2s</th><th>2d</th><th>2p</th><th>3s</th><th>3ns</th><th>3ref</th><th>0 (itr)</th></tr>
<tr><td>1s</td><td colspan="5">Σ-ŋa-ʔã-ci-ŋ
Σ-ŋ-a-ŋ-cĩ-h-ẽ</td><td>Σ-na-ʔã
Σ-n-e-hẽ</td><td>Σ-na-ʔã-ce
Σ-n-a-c-e</td><td>Σ-na-ʔã-ni
Σ-n-a-ni-hẽ</td><td>Σ-u-ku-ŋ
Σ-u-h-ẽ</td><td>Σ-u-ku-ŋ-ci-ŋ
Σ-u-ŋ-cĩ-h-ẽ</td><td rowspan="8">(impossible)</td><td>Σ-ŋa-ʔã
Σ-e-h-ẽ</td></tr>
<tr><td>1de</td><td colspan="8" rowspan="2">Σ-na-ʔa-ncĩ-ya
Σ-n-a-ncĩ-e-h-ẽ</td><td>Σ-c-o-ko-ŋa
Σ-a-ce-h-ẽ</td><td rowspan="2">Σ-u-ku-m-ci-m-ma
Σ-u-m-ci-m-m-e</td><td>Σ-ce-ke-ŋa
Σ-a-ce-h-ẽ</td></tr>
<tr><td>1pe</td><td>Σ-u-ku-m-ma
Σ-u-m-m-e</td><td>Σ-i-ki-ŋa
Σ-i-e-hẽ</td></tr>
<tr><td>1di</td><td colspan="8" rowspan="2">Σ-na-ʔa-ncĩ
Σ-n-a-ncĩ-hẽ</td><td>Σ-c-o-ko
Σ-a-c-e</td><td rowspan="2">Σ-u-ku-m-ci-m
Σ-u-m-ci-mh-e</td><td>Σ-ce-ke
Σ-a-c-e</td></tr>
<tr><td>1pi</td><td>Σ-u-ku-m
Σ-u-mh-e</td><td>Σ-i-ki
Σ-i-hẽ</td></tr>
<tr><td>2s</td><td>a-Σ-ŋa-ʔã
a-Σ-e-h-ẽ</td><td rowspan="3">{a-ma}Σ-ce-ke
{a-ma}Σ-a-c-e</td><td rowspan="3">{a-ma}Σ-no
{a-ma}Σ-e</td><td colspan="5" rowspan="2">a-Σ-na-ʔa-ce
a-Σ-n-a-c-e</td><td>a-Σ-u-ku-ce
a-Σ-u-c-e</td><td>a-Σ-no
a-Σ-e</td></tr>
<tr><td>2d</td><td>a-Σ-ŋa-ʔa-ŋ-ci-ŋ
a-Σ-a-ŋ-cĩ-h-ẽ</td><td>a-Σ-c-o-ko
a-Σ-a-c-e</td><td rowspan="2">a-Σ-u-ku-m-ci-m
a-Σ-u-m-ci-mh-e</td><td>a-Σ-ce-ke
a-Σ-a-c-e</td></tr>
<tr><td>2p</td><td>a-Σ-ŋa-ʔa-ŋ-ni-ŋ
a-Σ-a-ŋ-nĩ-h-ẽ</td><td colspan="5">a-Σ-na-ʔa-ncĩ
a-Σ-n-a-ncĩ-hẽ</td><td>a-Σ-u-ku-m
a-Σ-u-mh-e</td><td>a-Σ-i-ki
a-Σ-i-hẽ</td></tr>
<tr><td>3s</td><td>u-Σ-ŋa-ʔã
u-Σ-e-h-ẽ</td><td rowspan="3">ma-Σ-ce-ke
ma-Σ-a-c-e</td><td rowspan="3">ma-Σ-no
ma-Σ-e</td><td rowspan="3">mai-Σ-ce-ke
mai-Σ-a-c-e</td><td rowspan="3">mai-Σ-no
mai-Σ-e</td><td rowspan="3">na-Σ-no
na-Σ-e</td><td rowspan="3">na-Σ-ce-ke
na-Σ-a-c-e</td><td rowspan="3">na-Σ-i-ki
na-Σ-i-hẽ</td><td>Σ-o-ko
Σ-e</td><td>Σ-u-ku-ce
Σ-u-c-e</td><td>Σ-na-ʔa-ce
Σ-n-a-c-e</td><td>Σ-no
Σ-e</td></tr>
<tr><td>3d</td><td>u-Σ-ŋa-ʔa-ŋ-ci-ŋ
u-Σ-a-ŋ-cĩ-h-ẽ</td><td>u-Σ-c-o-ko
u-Σ-a-c-e</td><td rowspan="2">u-Σ-u-ku-ce
u-Σ-u-c-e</td><td rowspan="2">u-Σ-na-ʔa-ncĩ
u-Σ-n-a-ncĩ-hẽ</td><td>u-Σ-ce-ke
u-Σ-a-c-e</td></tr>
<tr><td>3p</td><td>u-Σ-ŋa-ʔa-ŋ-ni-ŋ
u-Σ-a-ŋ-nĩ-h-ẽ</td><td>u-Σ-o-ko
u-Σ-e</td><td>u-Σ-no
u-Σ-e</td></tr>
</table>

D.1.2 Complex indicative

A/P	1s	1de	1pe	1di	1pi
1s	Σ_1-ŋa-Σ_2-ŋa-ʔã-ci-ŋ Σ_1-ŋa-Σ_2-ŋ-a-ŋ-ci-ŋ-hẽ				
1de 1pe	Σ_1-na-Σ_2-na-ʔa-ncĩ-ya Σ_1-n-a-Σ_2-n-a-ncĩ-ẽ-hẽ				
1di 1pi	Σ_1-na-Σ_2-na-ʔa-ncĩ Σ_1-n-a-Σ_2-n-a-ncĩ-hẽ				
2s	a-Σ_1-ŋa-Σ_2-ŋa-ʔã a-Σ_1-a-ŋ-Σ_2-e-h-ẽ	{a-ma}Σ_1-ci-Σ_2-ce-ke {a-ma}Σ_1-a-Σ_2-a-c-e	{a-ma}Σ_1-na-Σ_2-no {a-ma}Σ_1-a-Σ_2-e	a-Σ_1-na-Σ_2-na-ʔa-ce a-Σ_1-n-a-Σ_2-n-a-c-e	
2d	a-Σ_1-ŋa-Σ_2-ŋa-ʔa-ŋ-ci-ŋ a-Σ_1-a-ŋ-Σ_2-a-ŋ-cĩ-h-ẽ				
2p	a-Σ_1-ŋa-Σ_2-ŋa-ʔa-ŋ-ni-ŋ a-Σ_1-a-ŋ-Σ_2-a-ŋ-nĩ-h-ẽ				
3s	u-Σ_1-ŋa-Σ_2-ŋa-ʔã u-Σ_1-a-ŋ-Σ_2-e-h-ẽ	ma-Σ_1-ci-Σ_2-ce-ke ma-Σ_1-a-Σ_2-a-c-e	ma-Σ_1-na-Σ_2-no ma-Σ_1-a-Σ_2-e	mai-Σ_1-ci-Σ_2-ce-ke mai-Σ_1-a-Σ_2-a-c-e	mai-Σ_1-na-Σ_2-no mai-Σ_1-a-Σ_2-e
3d	u-Σ_1-ŋa-Σ_2-ŋa-ʔã-ŋ-ci-ŋ u-Σ_1-a-ŋ-Σ_2-a-ŋ-cĩ-h-ẽ				
3p	u-Σ_1-ŋa-Σ_2-ŋa-ʔã-ŋ-ni-ŋ u-Σ_1-a-ŋ-Σ_2-a-ŋ-nĩ-h-ẽ				

<table>
<tr><th>A/P</th><th>2s</th><th>2d</th><th>2p</th><th>3s</th><th>3ns</th><th>3ref</th><th>0 (itr)</th></tr>
<tr><td>1s</td><td>Σ1-na-Σ2-na-ʔã
Σ1-n-a-Σ2-n-e-h-ẽ</td><td>Σ1-na-Σ2-na-ʔã-ce
Σ1-n-a-Σ2-n-a-c-e</td><td>Σ1-na-Σ2-na-ʔã-ni
Σ1-n-a-Σ2-n-a-ni-hẽ</td><td>Σ1-u-ŋ-Σ2-u-ku-ŋ
Σ1-u-ŋ-Σ2-u-h-ẽ</td><td>Σ1-u-ŋ-Σ2-u-ku-ŋ-ci-ŋ
Σ1-u-ŋ-Σ2-u-ŋ-cĭ-h-ẽ</td><td rowspan="10">(impossible)</td><td>Σ1-ŋa-Σ2-ŋa-ʔã
Σ1-a-ŋ-Σ2-e-h-ẽ</td></tr>
<tr><td>1de</td><td colspan="3" rowspan="2">Σ1-na-Σ2-na-ʔa-ncĭ-ya
Σ1-n-a-Σ2-n-a-ncĭ-e-h-ẽ</td><td>Σ1-c-u-Σ2-c-o-ko-ŋa
Σ1-u-Σ2-a-ce-h-ẽ</td><td rowspan="2">Σ1-u-m-Σ2-u-ku-m-ci-m-ma
Σ1-u-m-Σ2-u-m-ci-m-m-e</td><td>Σ1-ci-Σ2-ce-ke-ŋa
Σ1-a-Σ2-a-ce-h-ẽ</td></tr>
<tr><td>1pe</td><td>Σ1-u-m-Σ2-u-ku-m-ma
Σ1-u-m-Σ2-u-m-m-e</td><td>Σ1-i-Σ2-i-ki-ŋa
Σ1-i-Σ2-i-e-h-ẽ</td></tr>
<tr><td>1di</td><td colspan="3" rowspan="2">Σ1-na-Σ2-na-ʔa-ncĭ
Σ1-n-a-Σ2-n-a-ncĭ-hẽ</td><td>Σ1-c-u-Σ2-c-o-ko
Σ1-a-Σ2-a-c-e</td><td rowspan="2">Σ1-u-m-Σ2-u-ku-m-ci-m
Σ1-u-m-Σ2-u-m-ci-mh-e</td><td>Σ1-ci-Σ2-ce-ke
Σ1-a-Σ2-a-c-e</td></tr>
<tr><td>1pi</td><td>Σ1-u-m-Σ2-u-ku-m
Σ1-u-m-Σ2-u-mh-e</td><td>Σ1-i-Σ2-i-ki
Σ1-i-Σ2-i-hẽ</td></tr>
<tr><td>2s</td><td colspan="3">a-Σ1-na-Σ2-na-ʔa-ce
a-Σ1-n-a-Σ2-n-a-c-e</td><td>a-Σ1-u-Σ2-o-ko
a-Σ1-u-Σ2-e</td><td>a-Σ1-u-Σ2-u-ku-ce
a-Σ1-u-Σ2-u-c-e</td><td>a-Σ1-na-Σ2-no
a-Σ1-a-Σ2-e</td></tr>
<tr><td>2d</td><td colspan="3" rowspan="2">a-Σ1-na-Σ2-na-ʔa-ncĭ
a-Σ1-n-a-Σ2-n-a-ncĭ-hẽ</td><td>a-Σ1-c-u-Σ2-c-o-ko
a-Σ1-a-Σ2-a-c-e</td><td rowspan="2">a-Σ1-u-m-Σ2-u-ku-m-ci-m
a-Σ1-u-m-Σ2-u-m-ci-mh-e</td><td>a-Σ1-ci-Σ2-ce-ke
a-Σ1-a-Σ2-a-c-e</td></tr>
<tr><td>2p</td><td>a-Σ1-u-m-Σ2-u-ku-m
a-Σ1-u-m-Σ2-u-mh-e</td><td>a-Σ1-i-Σ2-i-ki
a-Σ1-i-Σ2-i-hẽ</td></tr>
<tr><td>3s</td><td rowspan="3">na-Σ1-na-Σ2-no
na-Σ1-a-Σ2-e</td><td rowspan="3">na-Σ1-ci-Σ2-ce-ke
na-Σ1-a-Σ2-a-c-e</td><td rowspan="3">na-Σ1-i-Σ2-i-ki
na-Σ1-i-Σ2-i-hẽ</td><td>Σ1-u-Σ2-o-ko
Σ1-u-Σ2-e</td><td>Σ1-u-Σ2-u-ku-ce
Σ1-u-Σ2-u-c-e</td><td>Σ1-na-Σ2-na-ʔa-ce
Σ1-n-a-Σ2-n-a-c-e</td><td>Σ1-na-Σ2-no
Σ1-a-Σ2-e</td></tr>
<tr><td>3d</td><td>u-Σ1-c-u-Σ2-c-o-ko
u-Σ1-a-Σ2-a-c-e</td><td rowspan="2">u-Σ1-u-Σ2-u-ku-ce
u-Σ1-u-Σ2-u-c-e</td><td rowspan="2">u-Σ1-na-Σ2-na-ʔa-ncĭ
u-Σ1-n-a-Σ2-n-a-ncĭ-hẽ</td><td>u-Σ1-ci-Σ2-ce-ke
u-Σ1-a-Σ2-a-c-e</td></tr>
<tr><td>3p</td><td>u-Σ1-u-Σ2-o-ko
u-Σ1-u-Σ2-e</td><td>u-Σ1-na-Σ2-no
u-Σ1-a-Σ2-e</td></tr>
</table>

D.1.3 Subjunctive

<table>
<tr><th>A/P</th><th>1s</th><th>1de</th><th>1pe</th><th>1di</th><th>1pi</th><th>2s</th><th>2d</th><th>2p</th><th>3s</th><th>3ns</th><th>3ref</th><th>0 (itr)</th></tr>
<tr><td>1s</td><td colspan="5">Σ-ŋa-ci-ŋ
Σ-ŋ-a-ŋ-ci-ŋ</td><td>Σ-na
Σ-n-a</td><td>Σ-na-ce
Σ-n-a-ce</td><td>Σ-na-ni
Σ-n-a-ni</td><td>Σ-u-ŋ
Σ-u-ŋ</td><td>Σ-u-ŋ-ci-ŋ
Σ-u-ŋ-ci-ŋ</td><td rowspan="8">(impossible)</td><td>Σ-ŋa
Σ-a-ŋ</td></tr>
<tr><td>1de</td><td colspan="8" rowspan="2">Σ-na-ncĩ-ya
Σ-n-a-ncĩ-ya</td><td>Σ-c-o-ŋa
Σ-a-c-o-ŋa</td><td rowspan="2">Σ-u-m-ci-m-ma</td><td>Σ-ce-ŋa
Σ-a-ce-ŋa</td></tr>
<tr><td>1pe</td><td>Σ-u-m-ma
Σ-u-m-ma</td><td>Σ-i-ŋa
Σ-i-ŋa</td></tr>
<tr><td>1di</td><td colspan="8" rowspan="2">Σ-na-ncĩ
Σ-n-a-ncĩ</td><td>Σ-c-o
Σ-a-c-o</td><td rowspan="2">Σ-u-m-ci-m
Σ-u-m-ci-m</td><td>Σ-ce
Σ-a-ce</td></tr>
<tr><td>1pi</td><td>Σ-u-m
Σ-u-m</td><td>Σ-i
Σ-i</td></tr>
<tr><td>2s</td><td>a-Σ-ŋa
a-Σ-a-ŋ</td><td rowspan="3">{a-ma}Σ-ce
{a-ma}Σ-a-ce</td><td rowspan="3">{a-ma}Σ
{a-ma}Σ-a</td><td colspan="5">a-Σ-na-ce
a-Σ-n-a-ce</td><td>a-Σ-o
a-Σ-o</td><td>a-Σ-u-ce
a-Σ-u-ce</td><td>a-Σ
a-Σ-a</td></tr>
<tr><td>2d</td><td>a-Σ-ŋa-ci-ŋ
a-Σ-a-ŋ-ci-ŋ</td><td colspan="5" rowspan="2">Σ-na-ncĩ
Σ-n-a-ncĩ</td><td>a-Σ-c-o
a-Σ-a-c-o</td><td rowspan="2">a-Σ-u-m-ci-m
a-Σ-u-m-ci-m</td><td>a-Σ-ce
a-Σ-a-c-o</td></tr>
<tr><td>2p</td><td>a-Σ-ŋa-ni-ŋ
a-Σ-a-ŋ-ni-ŋ</td><td>a-Σ-u-m
a-Σ-u-m</td><td>a-Σ-i
a-Σ-i</td></tr>
<tr><td>3s</td><td>u-Σ-ŋa
u-Σ-a-ŋ</td><td rowspan="3">ma-Σ-ce
ma-Σ-a-ce</td><td rowspan="3">ma-Σ
ma-Σ-a</td><td rowspan="3">mai-Σ-ce
mai-Σ-a-ce</td><td rowspan="3">mai-Σ
mai-Σ-a</td><td rowspan="3">na-Σ
na-Σ-a</td><td rowspan="3">na-Σ-ce
na-Σ-a-ce</td><td rowspan="3">na-Σ-i
na-Σ-i</td><td>Σ-o
Σ-o</td><td>Σ-u-ce
Σ-u-ce</td><td>Σ-na-ce
Σ-n-a-ce</td><td>Σ
Σ-a</td></tr>
<tr><td>3d</td><td>u-Σ-ŋa-ci-ŋ
u-Σ-a-ŋ-ci-ŋ</td><td>u-Σ-c-o
u-Σ-a-c-o</td><td rowspan="2">u-Σ-u-ce
u-Σ-u-ce</td><td rowspan="2">u-Σ-na-ncĩ
u-Σ-n-a-ncĩ</td><td>u-Σ-ce
u-Σ-a-ce</td></tr>
<tr><td>3p</td><td>u-Σ-ŋa-ni-ŋ
u-Σ-a-ŋ-ni-ŋ</td><td>u-Σ-o
u-Σ-o</td><td>u-Σ
u-Σ-a</td></tr>
</table>

D.1.4 Complex subjunctive

<table>
<tr><th>A/P</th><th>1s</th><th>1de</th><th>1pe</th><th>1di</th><th>1pi</th></tr>
<tr><td>1s</td><td colspan="5">Σ₁-ŋa-Σ₂-ŋa-ci-ŋ
Σ₁-ŋ-a-Σ₂-ŋ-a-ŋ-ci-ŋ</td></tr>
<tr><td>1de</td><td colspan="5" rowspan="2">Σ₁-na-Σ₂-na-ncĩ-ya
Σ₁-n-a-Σ₂-n-a-ncĩ-ya</td></tr>
<tr><td>1pe</td></tr>
<tr><td>1di</td><td colspan="5" rowspan="2">Σ₁-na-Σ₂-na-ncĩ
Σ₁-n-a-Σ₂-n-a-ncĩ</td></tr>
<tr><td>1pi</td></tr>
<tr><td>2s</td><td>a-Σ₁-ŋa-Σ₂-ŋa
a-Σ₁-a-ŋ-Σ₂-a-ŋ</td><td rowspan="3">{a-ma}Σ₁-ci-Σ₂-ce
{a-ma}Σ₁-a-Σ₂-a-ce</td><td rowspan="3">{a-ma}Σ₁-na-Σ₂
{a-ma}Σ₁-a-Σ₂-a</td><td colspan="2" rowspan="3">a-Σ₁-na-Σ₂-na-ce
a-Σ₁-n-a-Σ₂-n-a-ce</td></tr>
<tr><td>2d</td><td>a-Σ₁-ŋa-Σ₂-ŋa-ci-ŋ
a-Σ₁-a-ŋ-Σ₂-a-ŋ-ci-ŋ</td></tr>
<tr><td>2p</td><td>a-Σ₁-ŋa-Σ₂-ŋa-ni-ŋ
a-Σ₁-a-ŋ-Σ₂-a-ŋ-ni-ŋ</td></tr>
<tr><td>3s</td><td>u-Σ₁-ŋa-Σ₂-ŋa
u-Σ₁-a-ŋ-Σ₂-a-ŋ</td><td rowspan="3">ma-Σ₁-ci-Σ₂-ce
ma-Σ₁-a-Σ₂-a-ce</td><td rowspan="3">ma-Σ₁-na-Σ₂
ma-Σ₁-a-Σ₂-a</td><td rowspan="3">mai-Σ₁-ci-Σ₂-ce
mai-Σ₁-a-Σ₂-a-ce</td><td rowspan="3">mai-Σ₁-na-Σ₂
mai-Σ₁-a-Σ₂-a</td></tr>
<tr><td>3d</td><td>u-Σ₁-ŋa-Σ₂-ŋa-ci-ŋ
u-Σ₁-a-ŋ-Σ₂-a-ŋ-ci-ŋ</td></tr>
<tr><td>3p</td><td>u-Σ₁-ŋa-Σ₂-ŋa-ni-ŋ
u-Σ₁-a-ŋ-Σ₂-a-ŋ-ni-ŋ</td></tr>
</table>

A/P	2s	2d	2p	3s	3ns	3ref	0 (itr)
1s	Σ_1-na-Σ_2-na Σ_1-n-a-Σ_2-n-a	Σ_1-na-Σ_2-na-ce Σ_1-n-a-Σ_2-n-a-ce	Σ_1-na-Σ_2-na-ni Σ_1-n-a-Σ_2-n-a-ni	Σ_1-u-ŋ-Σ_2-u-ŋ Σ_1-u-ŋ-Σ_2-u-ŋ	Σ_1-u-ŋ-Σ_2-u-ŋ-ci-ŋ Σ_1-u-ŋ-Σ_2-u-ŋ-ci-ŋ		Σ_1-ŋa-Σ_2-ŋa Σ_1-a-ŋ-Σ_2-a-ŋ
1de				Σ_1-c-u-Σ_2-c-o-ŋa			Σ_1-ci-Σ_2-ce-ŋa
1pe	Σ_1-na-Σ_2-na-ncĩ-ya Σ_1-n-a-Σ_2-n-a-ncĩ-ya			Σ_1-a-Σ_2-a-c-o Σ_1-u-m-Σ_2-u-m-ma	Σ_1-u-m-Σ_2-u-m-ci-m-ma Σ_1-u-m-Σ_2-u-m-ci-m-ma		Σ_1-a-Σ_2-a-ce-ŋa Σ_1-i-Σ_2-i-ŋa Σ_1-i-Σ_2-i-ŋa
1di				Σ_1-c-u-Σ_2-c-o			Σ_1-ci-Σ_2-ce
1pi	Σ_1-na-Σ_2-na-ncĩ Σ_1-n-a-Σ_2-n-a-ncĩ			Σ_1-a-Σ_2-a-c-o Σ_1-u-m-Σ_2-u-m	Σ_1-u-m-Σ_2-u-m-ci-m Σ_1-u-m-Σ_2-u-m-ci-m	(impossible)	Σ_1-a-Σ_2-a-ce Σ_1-i-Σ_2-i Σ_1-i-Σ_2-i
2s	a-Σ_1-na-Σ_2-na-ce a-Σ_1-n-a-Σ_2-n-a-ce			a-Σ_1-u-Σ_2-o a-Σ_1-u-Σ_2-o	a-Σ_1-u-Σ_2-u-ce a-Σ_1-u-Σ_2-u-ce		a-Σ_1-na-Σ_2 a-Σ_1-a-Σ_2-a
2d	a-Σ_1-na-Σ_2-na-ncĩ a-Σ_1-n-a-Σ_2-n-a-ncĩ			a-Σ_1-c-u-Σ_2-c-o a-Σ_1-a-Σ_2-a-c-o a-Σ_1-u-m-Σ_2-u-m a-Σ_1-u-m-Σ_2-u-m	a-Σ_1-u-m-Σ_2-u-m-ci-m a-Σ_1-u-m-Σ_2-u-m-ci-m		a-Σ_1-ci-Σ_2-ce a-Σ_1-a-Σ_2-a-ce a-Σ_1-i-Σ_2-i a-Σ_1-i-Σ_2-i
2p							
3s	na-Σ_1-na-Σ_2 na-Σ_1-a-Σ_2-a	na-Σ_1-ci-Σ_2-ce na-Σ_1-a-Σ_2-a-ce	na-Σ_1-i-Σ_2-i na-Σ_1-iΣ$_2$-i	Σ_1-u-Σ_2-o Σ_1-u-Σ_2-o	Σ_1-u-Σ_2-u-ce Σ_1-u-Σ_2-u-ce	Σ_1-na-Σ_2-na-ce Σ_1-n-a-Σ_2-n-a-ce	Σ_1-na-Σ_2 Σ_1-a-Σ_2-a
3d				u-Σ_1-c-u-Σ_2-c-o u-Σ_1-a-Σ_2-a-c-o	u-Σ_1-u-Σ_2-u-ce	u-Σ_1-na-Σ_2-na-ncĩ	u-Σ_1-ci-Σ_2-ce u-Σ_1-a-Σ_2-a-ce
3p				u-Σ_1-u-Σ_2-o u-Σ_1-u-Σ_2-e	u-Σ_1-u-Σ_2-u-ce	u-Σ_1-n-a-Σ_2-n-a-ncĩ	u-Σ_1-na-Σ_2 u-Σ_1-a-Σ_2-e

D.1.5 Imperative

<table>
<tr><td>A/P</td><td>1s</td><td>1de</td><td>1pe</td><td>1di</td><td>1pi</td><td>2s</td><td>2d</td><td>2p</td><td>3s</td><td>3ns</td><td>0 (itr)</td></tr>
<tr><td>2s</td><td>Σ-a-h-ã</td><td rowspan="3">ma-Σ-a-c-a</td><td>ma-Σ-a</td><td colspan="5" rowspan="2">Σ-n-a-c-a</td><td>Σ-a</td><td>Σ-u-c-a</td><td>Σ-a</td></tr>
<tr><td>2d</td><td>Σ-a-ŋ-cĩ-h-ã</td><td>ma-Σ-a-c-a</td><td>Σ-a-c-a</td><td rowspan="2">Σ-a-n-u-m-ci-mh-a</td><td>Σ-a-c-a</td></tr>
<tr><td>2p</td><td>Σ-a-ŋ-nĩ-h-ã</td><td>ma-Σ-a-ni-hã</td><td colspan="5">Σ-n-a-ncĩ-hã</td><td>Σ-a-n-u-mh-a</td><td>Σ-a-ni-hã</td></tr>
</table>

D.1.6 Complex imperative

<table>
<tr><td>A/P</td><td>1s</td><td>1de</td><td>1pe</td><td>1di</td><td>1pi</td><td>2s</td><td>2d</td><td>2p</td><td>3s</td><td>3ns</td><td>0 (itr)</td></tr>
<tr><td>2s</td><td>Σ_1-a-ŋ-Σ_2-ã-h-ã</td><td rowspan="3">ma-Σ_1-a-Σ_2-a-c-a</td><td>ma-Σ_1-a-Σ_2-a</td><td colspan="5">Σ_1-n-a-Σ_2-n-a-c-a</td><td>Σ_1-u-Σ_2-a</td><td>Σ_1-u-Σ_2-u-c-a</td><td>Σ_1-a-Σ_2-a</td></tr>
<tr><td>2d</td><td>Σ_1-a-ŋ-Σ_2-a-ŋ-cĩ-h-ã</td><td>ma-Σ_1-a-Σ_2-a-c-a</td><td colspan="5" rowspan="2">Σ_1-n-a-Σ_2-n-a-ncĩ-hã</td><td>Σ_1-a-Σ_2-a-c-a</td><td rowspan="2">Σ_1-a-Σ_2-a-n-u-mh-a</td><td>Σ_1-a-Σ_2-a-c-a</td></tr>
<tr><td>2p</td><td>Σ_1-a-ŋ-Σ_2-a-ŋ-nĩ-h-ã</td><td>ma-Σ_1-a-Σ_2-a-ni-hã</td><td>Σ_1-a-Σ_2-a-n-u-mh-a</td><td>Σ_1-a-Σ_2-a-ni-hã</td></tr>
</table>

D.1.7 Imperative with *-khag* [CON]

<table>
<tr><td>A/P</td><td>1s</td><td>1de</td><td>1pe</td><td>1di</td><td>1pi</td><td>2s</td><td>2d</td><td>2p</td><td>3s</td><td>3ns</td><td>0 (itr)</td></tr>
<tr><td>2s</td><td>Σ-a-ŋ-kh-a-ŋ</td><td rowspan="3">ma-Σ-a-kh-a-ce</td><td>ma-Σ-a-kha</td><td colspan="5">Σ-n-a-kh-a(ŋ-na)-ce</td><td>Σ-o-kh-o</td><td>Σ-o-kh-o-ce</td><td>Σ-a-kha</td></tr>
<tr><td>2d</td><td>Σ-a-ŋ-kh-a-ŋ-ci-ŋ</td><td>ma-Σ-a-kh-a-ce</td><td colspan="5" rowspan="2">Σ-n-a-kh-a(ŋ-na)-ncĩ</td><td>Σ-a-kh-a-c-o</td><td rowspan="2">Σ-a-kh-a-n-u-m-ci-m</td><td>Σ-a-kh-a-ce</td></tr>
<tr><td>2p</td><td>Σ-a-ŋ-kh-a-ŋ-ni-ŋ</td><td>ma-Σ-a-kh-a-nĩ</td><td>Σ-a-kh-a-n-u-m</td><td>Σ-a-kh-a-nĩ</td></tr>
</table>

D.1.8 Non-finite forms

	default	reflexive
INF	Σ-ma	Σ-ma-ncĩ
PURP	Σ-si	Σ-cĩ-si
CVB.FGR	Σ-saŋa	Σ-cĩ-saŋa
RECP	Σ-ka-Σ (lus-)	
ACT.PTCP	ka-Σ(-pa)	-
PASS.PTCP	Σ-mayaŋ	-

D.1.9 Complex non-finite forms

	complex	complex reflexive
INF	Σ_1-ma-Σ_2-ma	Σ_1-ma-Σ_2-ma-ncĩ
PURP	Σ_1-ø-Σ_2-si	Σ_1-Σ_2-cĩ-si
CVB.FGR	Σ_1-ø-Σ_2-saŋa	Σ_1-Σ_2-cĩ-saŋa
RECP	Σ_1-ka-Σ_1 (lus-Σ_2-)	
ACT.PTCP	ka-Σ-pa-Σ-pa	-
PASS.PTCP	Σ-ma-Σ-mayaŋ	-

D.2 Nepali

This section shows full paradigms for Nepali verbal inflection. Similar tenses and the corresponding negative forms are shown next to each other. The base for the paradigms were Genetti (1994) and Hutt and Subedi (1999). Forms that are not listed there were elicited by me. I also carried out the analysis.

D.2.1 Overview of TMA and agreement markers

TMA, polarity and agreement markers follow the verb stem and show a high degree of fusion. All suffixes and allomorphies are shown in Table D.1. Many agreement markers (especially feminine ones) trigger umlaut in preceding negation markers (*-dʌin* [NPST.NEG] > *-din*, *-en* [PST.NEG] > *-in*). This is indicated by ^{i} in the table. The high-honorific second/third person does not have dedicated agreement suffixes but is marked by the combination of *-nu* [INF_1] with tensed 3rd person singular forms of *hu-* 'happen', e.g. *gʌr-nubhʌyo* [do-PST.2/3HH] 'he did' < *gʌr-nu bhʌ-y-o* [do-INF_1 become-PST-3s] 'doing happened'. It is therefore not included in the table.

	NPST -ch	**NPST.NEG** -dʌin	**PST(.HAB)** -y, -thy	**PST.NEG** -en	**OPT** -Ø	**FUT** -la
1s	-u	-^{i}ʌ̃	-ẽ	-^{i}ʌ̃	-ũ	-ũ
1p	-ʌũ	-ʌũ	-ʌũ	-ʌũ	-ʌũ	-ʌũ
2sLH	-ʌs	-ʌs	-is	-^{i}ʌs	-es	-s
2sfLH	-es	-^{i}ʌs	-is	-^{i}ʌs	-es	-is
2MH	-ʌu	-ʌu	-ʌu	-ʌu	-ʌu	-ʌu
2fMH	-eu	-^{i}ʌu	-ʌu	-^{i}ʌu	-ʌu	-eu
3s	-ʌ	-ʌ	-o	-ʌ	-os	-Ø
3sf	-e	-^{i}Ø	-i	-^{i}ʌ	-os	-i
3p/3MH	-ʌn	-ʌn	-e	-ʌn	-un	-n
3pf/3fMH	-in	-^{i}ʌn	-in	-^{i}ʌn	-un	-in

Table D.1: Nepali TMA, polarity, and agreement

D.2.2 Simple nonpast

	NPST	**NPST.NEG I**	**NPST.NEG II**
1s	Σ-ch-u	Σ-din-ʌ̃	Σ-nn-ʌ̃
1p	Σ-ch-ʌũ	Σ-dʌin-ʌũ	Σ-nn-ʌũ
2sLH	Σ-ch-ʌs	Σ-dʌin-ʌs	Σ-nn-ʌs
2sfLH	Σ-ch-es	Σ-din-ʌs	Σ-nn-ʌs
2MH	Σ-ch-ʌu	Σ-dʌin-ʌu	Σ-nn-ʌu
2fMH	Σ-ch-eu	Σ-din-ʌu	Σ-nn-ʌu
3s	Σ-ch-ʌ	Σ-dʌin-ʌ	Σ-nn-ʌ
3sf	Σ-ch-e	Σ-din-ʌ	Σ-nn-ʌ
3p/3MH	Σ-ch-ʌn	Σ-dʌin-ʌn	Σ-nn-ʌn
3pf/3fMH	Σ-ch-in	Σ-din-ʌn	Σ-nn-ʌn
2/3HH	Σ-nuhunchʌ	Σ-nuhunnʌ	Σ-nuhunnʌ

Table D.2: Nepali nonpast

D.2.3 Simple and habitual past

	PST	PST.NEG	PST.HAB	PST.HAB.NEG I	PST.HAB.NEG II
1s	Σ-ẽ	Σ-in-ʌ̃	Σ-th-ẽ	Σ-dinʌ-th-ẽ	Σ-th-in-ʌ̃
1p	Σ-y-ʌũ	Σ-en-ʌũ	Σ-thy-ʌũ	Σ-dʌinʌ-thy-ʌũ	Σ-th-en-ʌũ
2sLH	Σ-i-s	Σ-in-ʌs	Σ-thi-s	Σ-dinʌ-thi-s	Σ-th-in-ʌs
2MH	Σ-y-ʌu	Σ-en-ʌu	Σ-thy-ʌu	Σ-dʌinʌ-thy-ʌu	Σ-th-en-ʌu
2fMH	Σ-y-ʌu	Σ-in-ʌu	Σ-thy-ʌu	Σ-dinʌ-thy-ʌu	Σ-th-in-ʌu
3s	Σ-y-o	Σ-en-ʌ	Σ-thy-o	Σ-dʌinʌ-thy-o	Σ-th-en-ʌ
3sf	Σ-i	Σ-in-ʌ	Σ-th-i	Σ-dina-th-i	Σ-th-in-ʌ
3p/3MH	Σ-e	Σ-en-ʌn	Σ-th-e	Σ-dʌinʌ-th-e	Σ-th-en-ʌn
3pf/3fMH	Σ-in	Σ-in-ʌn	Σ-th-in	Σ-dinʌ-th-in	Σ-th-in-ʌn
2/3HH	Σ-nubhʌyo	Σ-nubhʌenʌ	Σ-nuhunthyo	Σ-nuhunnʌthyo	Σ-nuhunthenʌ

Table D.3: Nepali past

D.2.4 Other screeves

	OPT	PROB.FUT	IMP
1s	Σ-ũ	Σ-ũ-la	-
1p	Σ-ʌũ	Σ-ʌũ-la	-
2sLH	Σ-es	Σ-la-s	Σ
2MH	Σ-ʌu	Σ-l-au, Σ-ʌu-la	Σ-ʌ
2fMH	Σ-ʌu	Σ-l-eu, Σ-ʌu-la	Σ-ʌ
3s	Σ-os	Σ-la-Ø	-
3sf	Σ-os	Σ-l-i	-
3p/3MH	Σ-un	Σ-la-n	-
3pf/3fMH	Σ-un	Σ-l-in	-
2/3HH	Σ-nu(ho)s	Σ-nuhola	Σ-nu(ho)s

Table D.4: Nepali other screeves

For negation the prefix *nʌ-* is added in all forms.

D.2.5 Non-finite forms

INF_1	Σ-nu	**CVB_1**	Σ-erʌ
INF_2	Σ-nʌ	**CVB_2**	Σ-i
IPFV.PTCP	Σ-ne	**CVB_3**	Σ-ikʌnʌ
PRFV.PTCP	Σ-eko	**CVB_4**	Σ-da
COND/NMLZ	Σ-e	**CVB_5**	Σ-dakheri
CHAR.PTCP	Σ-do	**CVB_6**	Σ-unjel
PROG	Σ-dʌi	**LNK**	Σ-i-

Table D.5: Nepali infinite forms

D.2.6 Inflection of *hu-*

The verb *hu-* has three senses, 'be', 'become, happen, be okay', and 'be there' (also 'be' with adjectives). These three senses are expressed by the same stem in some forms (e.g. in the infinitive *hu-nu*) but by different stems in others. In tenses without a distinction the gloss [COP] is used instead of the senses. All stems are shown in Table D.6.

	'be'	'become'	'be there'
NPST	h:	hu-	ch:
NPST.HAB	hu-	hu-	hu-
PST	thi-	bhʌ-	thi-
PST.HAB	hu-	hu-	hu-
OPT	hu-	hu-	hu-/chʌ-
PROB.FUT	hu-	hu-	hu-
IMP	hu-	hu-	hu-
NONF *-n*	hu-	hu-	hu-
NONF *-e*	bhʌ-	bhʌ-	bhʌ-
NONF *-d*	hu-	hu-	hu-/chʌ-
NONF *-i*	hu-	bhʌ-	hu-
NONF *-u*	hu-	bhʌ-	hu-/chʌ-

Table D.6: Stems of *hu-*

In the forms marked with a colon, suffixes are joined to stems without an intervening *-ch* [NPST]. The result are the irregular forms in Table D.7. When *hu-* is combined with *-ch* [NPST] and suffixes like a regular verb (cf. section D.2.2) the form becomes habitual. The habitual nonpast exists only for this verb and is formally identical to NPST for all other verbs.

	'be'	'be' (NEG)	'be there'	'be there' (NEG)
1s	h-ũ	h-oin-ʌ̃	ch-u	ch-ʌin-ʌ̃
1p	h-ʌũ	h-oin-ʌũ	ch-ʌũ	ch-ʌin-ʌũ
2sLH	h-os	h-oin-ʌs	ch-ʌs	ch-ʌinʌ-s
2sfLH	h-os	h-oin-ʌs	ch-es	ch-ʌinʌ-s
2mMH	h-ʌu	h-oin-ʌu	ch-ʌu	ch-ʌin-ʌu
2fMH	h-ʌu	h-oin-ʌu	ch-eu	ch-ʌin-ʌu
3s	h-o	h-oin-ʌ	ch-ʌ	ch-ʌin-ʌ
3sf	h-o	h-oin-ʌ	ch-e	ch-ʌin-ʌ
3pm/3mMH	h-un	h-oin-ʌn	ch-ʌn	ch-ʌin-ʌn
3pf/3fMH	h-un	h-oin-ʌn	ch-in	ch-ʌin-ʌn
2/3HH	hu-nuhunchʌ	hu-nuhunnʌ	hu-nuhunchʌ	hu-nuhunnʌ

Table D.7: Irregular nonpast forms of *hu-*

D.2.7 Composite tenses

The two participles and the progressive can be used in connection with *hu-* to form composite tenses. The tense on *hu-* marks a reference point relative to speaking time. The participle then locates an event before (*-eko* [PST.PTCP]), at (*-dʌi* [PROG]), or after (*-ne* [NPST.PTCP]) that reference point.

The same functional distinctions that are marked by different stems in various tenses of *hu-* can also be made in the corresponding composite tenses. Thus, there is, for instance, a difference between *gʌreko chu* 'I have done' (auxiliary *ch:* 'be there' in NPST) and *gʌreko hũ* 'I am the one who has done' (auxiliary *h:* 'be' in NPST) and between *gʌreko thiẽ* 'I had done' (auxiliary *thi-* 'be,

be there' in PST) and *gʌreko bhʌẽ* 'I got into a state where I had done' (auxiliary *bhʌ-* 'become' in PST). In addition, both habitual tenses of *hu-* can be used to create habitual composite tenses, e.g. *gʌreko hunchu* 'I usually have done', *gʌreko hunthẽ* 'I usually had done'.

Table D.8 below lists all possible combinations in the 1st person singular of *Σ-* 'do' with an approximate translation.

	PST.PTCP	PROG	NPST.PTCP
'be' (NPST)	*Σ-eko ho* 'he's the one who has done'	*Σ-dʌi ho* 'he's the one who is doing'	*Σ-ne ho* 'he's the one who will do'
'be there' (NPST)	*Σ-eko chʌ* 'he's done'	*Σ-dʌi chʌ* 'he's doing'	*Σ-ne chʌ* 'he will do'
'become' (NPST), 'be, be there, become' (NPST.HAB)	*Σ-eko hunchʌ* 'he usually has done'	*Σ-dʌi hunchʌ* 'he's usually doing'	*Σ-ne hunchʌ* 'he usually will do'
'be, be there' (PST)	*Σ-eko thiyo* 'he had done'	*Σ-dʌi thiyo* 'he was doing'	*Σ-ne thiyo* 'he was going to do'
'become' (PST)	*Σ-eko bhʌyo* 'he happened to have done'	*Σ-dʌi bhʌyo* 'he happened to be doing'	*Σ-ne bhʌyo* 'he happened to do'
'be, be there, become' (PST.HAB)	*Σ-eko hunthyo* 'he usually had done'	*Σ-dʌi hunthyo* 'he was usually doing'	*Σ-ne hunthyo* 'he was usually going to do'
'be, be there, become' (OPT)	*Σ-eko hos* 'may he have done'	*Σ-dʌi hos* 'may he be doing'	*Σ-ne hos* 'may he be going to do'
'be, be there, become' (PROB.FUT)	*Σ-eko hola* 'probably he will have done'	*Σ-dʌi hola* 'probably he will be doing'	*Σ-ne hola* 'probably he will be going to do'
'be, be there, become' (IMP)	*Σ-eko hou* 'have done!'	*Σ-dʌi hou* 'be doing!'	*Σ-ne hou* 'do (in the future)!'

Table D.8: Nepali composite tenses

Bibliography

Abadie, Peggy. 1974. Nepali as an ergative language. *Linguistics of the Tibeto-Burman Area* 1:156--177.

Abbott, Barbara. 2006. Definiteness and Indefiniteness. In *The Handbook of Pragmatics*, ed. Laurence R. Horn and Gregory Ward, chapter 6, 122--151. Oxford: Blackwell.

Abbott, Barbara. 2010. *Reference*. Oxford: Oxford University Press.

Abney, Steven. 1996. Statistical Methods and Linguistics. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, ed. Judith Klavans and Philip Resnik. Cambridge, Massachusetts: MIT Press.

Acharya, Jayaraj. 1991. *A descriptive grammar of Nepali and an analyzed corpus*. Washington: Georgetown University Press.

Ackerman, Farrell, and John Moore. 2001. *Proto-properties and grammatical encoding*. Stanford Monographs in Linguistics. Stanford: CSLI.

Adhikārī, Hemāṇarāja. 2052 V.S. *Nepālī kāraka vyākaraṅa*. Kāṭhmāḍauṁ: Royal Nepal Academy. 1995/1996 AD.

Aissen, Judith. 2003. Differential object marking: iconicity vs. economy. *Natural Language & Linguistic Theory* 21:435--483.

Allen, N.J. 1975. *Sketch of Thulung Grammar*. Ithaca (New York): Cornell University.

Angdembe, Tej Man. 1998. Antipassive via noun incorporation: future of the Limbu object agreement. *Nepalese Studies* 2:17--24.

Apte, Mahadeo L., and D.P. Pattanayak. 1967. *An outline of Kumauni grammar*. Durham, North Carolina: Program in Comparative Studies on Southern Asia, Duke University.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24:65--87.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London: Croom Helm.

Baayen, Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Bailey, T. Grahame, and T. Cummings. 1912 [1994]. *Panjabi manual and grammar*. New Delhi: Asian educational services.

Barāla, Īśvara, ed. 2046 V.S. *Sayapatrī*. Kāṭhmāḍauṁ: Sājhā Prakāśana. 1989/1990 AD.

Bates, Douglas, Martin Maechler, and Ben Bolker. 2012. lme4: Linear mixed-effects models using S4 classes. R package, http://www.R-project.org/.

Beames, John. 1872-79 [1966]a. *A comparative grammar of the modern Aryan languages of India: to wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oriya and Bangali*, volume 1. Delhi: Munshiram Manoharlal.

Beames, John. 1872-79 [1966]b. *A comparative grammar of the modern Aryan languages of India: to wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oriya and Bangali*, volume 2. Delhi: Munshiram Manoharlal.

Bhatia, Tej K. 1993. *Punjabi*. London/New York: Routledge.

Bickel, Balthasar. 1999. Face vs. empathy: The social foundation of Maithili verb agreement. *Linguistics* 37:481--518.

Bickel, Balthasar. 2003a. Belhare. In *The Sino-Tibetan languages*, ed. Graham Thurgood and Randy LaPolla, 546--570. London/New York: Routledge.

Bickel, Balthasar. 2003b. Referential density in discourse and syntactic typology. *Language* 79:708--736.

Bickel, Balthasar. 2004a. Hidden syntax in Belhare. In *Himalayan languages: past and present*, ed. Anju Saxena, 141--190. Berlin: Mouton de Gruyter.

Bickel, Balthasar. 2004b. The syntax of experiencers in the Himalayas. In *Non-nominative Subjects*, ed. Peri Bhaskararao and Karumuri Venkata Subbarao, 77--112. Amsterdam: John Benjamins.

Bickel, Balthasar. 2006. Referential Density in Typological Perspective. In *Plenary lecture at the Leipzig Spring School on Linguistic Diversity*. Leipzig.

Bickel, Balthasar. 2007. Alignment typology revisited: ditransitives in general and in Southeastern Kiranti. In *Conference on Ditransitive Constructions*.

Bickel, Balthasar. 2008a. Aspects of Kiranti syntax: grammatical relations. Paper presented at the Central Department of Linguistics Tribhuvan University, Kirtipur, August 14, 2008.

Bickel, Balthasar. 2008b. Grammatical relations in Chintang. Paper presented at the Max Planck Institute for Evolutionary Anthropology, Leipzig, October 7, 2008, 10 2008.

Bickel, Balthasar. 2008c. On the scope of the referential hierarchy in the typology of grammatical relations. In *Case and grammatical relations: studies in honor of Bernard Comrie*, ed. Greville Corbett and Michael Noonan, 191--210. Amsterdam: John Benjamins.

Bickel, Balthasar. 2011. Grammatical relations typology. In *The Oxford Handbook of Linguistic Typology*, ed. Jae Jung Song, 399--444. Oxford: Oxford University Press.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha Rai, Manoj Rai, Novel Kishor Rai, and Sabine Stoll. 2007a. Free prefix ordering in Chintang. *Language* 83:43--73.

Bickel, Balthasar, Martin Gaenszle, Arjun Rai, Prem Dhoj Rai, Shree Kumar Rai, Vishnu S. Rai, and Narayan P. Sharma. 2007b. Two ways of suspending object agreement in Puma: between incorporation, antipassivization, and optional agreement. *Himalayan Linguistics* 7:1--19.

Bickel, Balthasar, and Johanna Nichols. 2009. Case marking and alignment. In *The Handbook of Case*, ed. Andrej Malchukov and Andrew Spencer, 304--321. Oxford University Press.

Bickel, Balthasar, Manoj Rai, Netra Paudyal, Goma Banjade, Toya Bhatta, Martin Gaenszle, Elena Lieven, Ichchha Rai, Novel Kishor Rai, and Sabine Stoll. 2010. The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti). In *Studies in ditransitive constructions*, ed. Andrej Malchukov, Martin Haspelmath, and Bernard Comrie. Berlin: Mouton de Gruyter.

Bickel, Balthasar, and Yogendra Yadava. 2000. A fresh look at grammatical relations in Indo-Aryan. *Lingua* 110:343--373.

Birner, Betty, and Gregory Ward. 1994. Uniqueness, familiarity, and the definite article in English. In *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society*, 93--102.

Bittner, Maria. 1987. On the Semantics of the Greenlandic Antipassive and Related Constructions. *International Journal of American Linguistics* 53:194--231.

Bjørnum, Stig. 2003. *Grønlandsk Grammatik*. Nuuk: Atagkuat.

Blake, Barry J. 1979. *A Kalkatungu Grammar*. Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.

Bloch, Jules. 1914 [1970]. *The formation of the Marāṭhī language*. Delhi/Patna/Vanarasi: Motilal Banarsidas.

Bloomfield, Leonard. 1946. Algonquian. In *Linguistic structures of Native America*, ed. Harry Hoijer, volume 6, 85--129. New York: Viking Fund Publications in Anthropology.

Bloomfield, Leonard. 1957. *Eastern Ojibwa*. Ann Arbor: University of Michigan Press.

Bok-Bennema, Renate. 1991. *Case and agreement in Inuit*. Berlin/New York: Foris.

Borchers, Dörte. 2008. *A grammar of Sunwar*. Leiden/Boston: Brill.

Bossong, Georg. 1982. Der präpositionale Akkusativ im Sardischen. In *Festschrift für Johannes Hubschmid zum 65. Geburtstag: Beiträge zur allgemeinen, indogermanischen, und romanischen Sprachwissenschaft*, ed. Otto Winkelmann and Maria Braisch, 579--599. Bern: Francke.

Bossong, Georg. 1985. *Empirische Universalienforschung: differentielle Objektmarkierung in den neuiranischen Sprachen*. Tübingen: Narr.

Bossong, Georg. 1998. Le marquage differential de l'objet dans les langues de l'Europe. In *Actance et valence dans les langues de l'Europe*, ed. Jack Feuillet, 193--257. Mouton de Gruyter.

Bossong, Georg. 2001. Ausdrucksmöglichkeiten für grammatische Relationen. In *Language Typology and Language Universals. An International Handbook*, ed. Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, volume 1, chapter VIII, 657--669. Berlin/New York: Walter de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, ed. Irene Krämer and Jost Zwarts, 69--94. Koniklijke Nederlandse Akademie van Wetenschapen.

Brown, Goold. 1851. *The Grammar of English Grammars*. New York: Samuel S. and William Wood.

Bunt, Harry C. 1979. Ensembles and the Formal Semantic Properties of Mass Terms. In *Mass terms: some philosophical problems.*, ed. Francis Jeffrey Pelletier, 249--277. Dordrecht: Reidel.

Bunt, Harry C. 1985. *Mass terms and model-theoretic semantics*. Cambridge: Cambridge University Press.

Butt, Miriam. 1993. Object Specificity and Agreement in Hindi/Urdu. In *Papers from the 29th Regional Meeting of the Chicago Linguistic Society*, ed. Katharine Beals, Gina Cooke, David Kathman, Sotaro Kita, Karl-Erik Cullough, and David Testen, 89--103. Chicago: Chicago Linguistic Society.

Butt, Miriam, and Tikaram Poudel. 2007. Distribution of the ergative in Nepali. Unpublished manuscript.

Bykova, E.M. 1981. *The Bengali language.* Moscow: Nauk.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22:249--254.

Caṭṭopādhyāya, Śrīsunītikumāra. 1966. *Sarala bhāṣā prakāśa bāṅgālā vyākaraṇa.* Calcutta.

Central Bureau of Statistics. 2001. *Population Census.* Kathmandu: National Planning Commission.

Chafe, Wallace, ed. 1980. *The Pear Stories. Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Norwood: Ablex.

Chandralal, Dileep. 2010. *Sinhala.* Amsterdam/Philadelphia: John Benjamins.

Chesterman, Andrew. 1991. *On definiteness.* Cambridge: Cambridge University Press.

Childers, Robert Caesar. 1875 [2005]. *A Dictionary of the Pali Language.* New Delhi: Munshiram Manoharlal.

Christophersen, Paul. 1939. *The articles - a study of their theory and use in English.* Copenhagen: Munksgaard.

Clark, T. W. 1963. *Introduction to Nepali.* London: School of Oriental and African Studies.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37--46.

Collins, Steven. 2005. *A Pali grammar for students.* Chiang Mai: Silkworm.

Comrie, Bernard. 1978. Ergativity. In *Syntactic Typology: Studies in the Phenomenology of Language*, ed. Winfred Lehmann, 329--394. Austin: University of Texas Press.

Cooreman, Ann. 1994. A functional typology of antipassives. In *Voice: form and function*, ed. Barbara Fox and Paul Hopper, 50--87. Amsterdam: John Benjamins.

Cox, D.R., and E.J. Snell. 1968. A General Definition of Residuals. *Journal of the Royal Statistical Society* 30:248--275.

Croft, William. 1990. *Typology and Universals.* Cambridge: Cambridge University Press.

Croft, William. 1991. *Syntactic Categories and Grammatical Relations.* Chicago/London: University of Chicago Press.

Davis, Alice Irene. 1984. *Basic Colloquial Maithili.* Delhi: Motilal Banarsidas.

Deo, Ashwini, and Devyani Sharma. 2006. Typological Variation in the Ergative Morphology of Indo-Aryan Languages. *Linguistic Typology* 10:369--418.

Ḍhakāl, Santoś. 2008. *Yatīko khojīmā.* Kathmandu: Visual Production Center.

Dhongde, Ramesh Vaman, and Kashi Wali. 2009. *Marathi.* Amsterdam/Philadelphia: John Benjamins.

Dik, Simon C. 1989. *The theory of functional grammar. Part I: The structure of the clause.* Dordrecht: Foris.

Dixon, Robert. 1979. Ergativity. *Language* 55:59--138.

Dixon, Robert. 1994. *Ergativity.* Cambridge University Press.

Dixon, Robert. 2010. *Basic Linguistic Theory*, volume 1. Oxford: Oxford University Press.

Doctor, Raimond. 2004. *A grammar of Gujarati.* München: Lincom.

Doornenbal, Marius. 2009. *A Grammar of Bantawa.* Utrecht: LOT.

Dowty, David. 1979. *Word meaning and Montague Grammar*. Dordrecht: Reidel.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67:547--619.

Driem, George van. 1987. *A grammar of Limbu.* Berlin: Mouton de Gruyter.

Driem, George van. 1993. The Proto-Tibeto-Burman verbal agreement system. *Bulletin of the School of Oriental and African Studies* 56:292--334.

Dryer, Matthew. 1986. Primary objects, secondary objects, and antidative. *Language* 62:808--845.

Dryer, Matthew S. 2006. Descriptive theories, explanatory theories, and basic linguistic theory. In *Catching Language: Issues in Grammar Writing*, ed. Felix Ameka, Alan Dench, and Nicholas Evans, 207--234. Berlin: Mouton de Gruyter.

Ebert, Karen. 1997a. *Camling*. München: Lincom.

Ebert, Karen. 1997b. *A grammar of Athpare.* München: Lincom.

Ebert, Karen. 2003. Kiranti languages: an overview. In *The Sino-Tibetan languages*, ed. Graham Thurgood and Randy LaPolla, chapter 31, 505--517. London/New York: Routledge.

Enç, Mürvet. 1991. The Semantics of Specificity. *Linguistic inquiry* 22:1--25.

Epstein, Richard. 1999. Roles, frames and definiteness. In *Discourse Studies in Cognitive Linguistics*, ed. Karen van Hoek, Andrej Kibrik, and Leo Noordman, 53--74. Amsterdam: John Benjamins.

Epstein, Richard. 2002. The definite article, accessibility, and the construction of discourse referents. *Cognitive Linguistics* 12:333--378.

Fahs, Achim. 1989. *Grammatik des Pali.* Leipzig: VEB.

Farkas, Donka. 1994. Specificity and scope. In *Actes du Premier Colloque Langues & Grammaire*, ed. Léa Nash and George Tsoulas, 119--137. Paris.

Farkas, Donka. 2002. Specificity Distinctions. *Journal of Semantics* 19:1--31.

Farrell, Patrick. 2005. *Grammatical Relations.* Oxford: Oxford University Press.

Fauconnier, Gilles. 1984. *Espaces mentaux. Aspects de la construction du sens dans les langues naturelles.* Paris: Les éditions de Minuit.

Fauconnier, Gilles. 1994. *Mental spaces. Aspects of meaning construction in natural languages.* Cambridge: Cambridge University Press.

Fauconnier, Stefanie. 2011. Differential Agent Marking and animacy. *Lingua* 121:533--547.

Féry, C., G. Fanselow, and M. Krifka, ed. 2007. *The notions of information structure*, volume 6 of *Working Papers of the SFB632: Interdisciplinary Studies on Information Structure (ISIS).* Potsdam: Universitätsverlag Potsdam.

Fortescue, Michael. 1984. *West Greenlandic.* London: Croom Helm.

Frantz, Donald G. 1991. *Blackfoot Grammar*. Toronto: University of Toronto Press.

Fraurud, Kari. 1996. Cognitive Ontology and NP Form. In *Reference and Referent Accessibility*, 65--88. Amsterdam: John Benjamins.

Frawley, William, and Roberta Michnick Golinkoff. 1995. Linguistic explanation. In *Handbook of Pragmatics*, ed. Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen, 608--615. Philadelphia: John Benjamins.

Gaenszle, Martin. 2011. Binomials and the Noun-to-Verb Ratio in Puma Rai Ritual Speech. *Anthropological Linguistics* 53:365--382.

Gair, James, and Kashi Wali. 1989. Hindi agreement as anaphor. *Linguistics* 27:45--70.

Gair, James W., and John C. Paolillo. 1997. *Sinhala*. München: Lincom.

Genetti, Carol. 1988. Notes on the structure of the Sunwar transitive verb. *Linguistics of the Tibeto-Burman Area* 11:62--92.

Genetti, Carol. 1994. Introduction (with a Sketch of Nepali Grammar). In *Aspects of Nepali grammar*, ed. Carol Genetti, 1--40. Santa Barbara: Department of Linguistics, University of Berkeley at Santa Barbara.

Genetti, Carol. 1999. Variation in Agreement in the Nepali Finite Verb. In *Topics in Nepalese Linguistics*, ed. Yogendra Yadava and Warren Glover, 542--55. Kathmandu: Royal Nepal Academy.

Genetti, Carol. 2011. Nominalization in Tibeto-Burman languages of the Himalayan area: A typological perspective. In *Nominalization in Asian Languages: Diachronic and typological perspectives*, ed. Yap Foong Ha and Janick Wrona, 163--193. John Benjamins.

Ghimire, L. 2002. Dative Subject Construction in Nepali. Master's thesis, Tribhuvan University, Kathmandu.

Givón, Talmy, ed. 1983. *Topic continuity in discourse*. Amsterdam/Philadelphia: John Benjamins.

Givón, Talmy. 1997. Grammatical Relations: An Introduction. In *Grammatical relations*, ed. Talmy Givón, 1--84. Amsterdam/Philadelphia: John Benjamins.

Gompel, Roger P. G. van, and Martin J. Pickering. 2007. Syntactic parsing. In *The Oxford Handbook of Psycholinguistics*, ed. M. Gareth Gaskell, chapter 17, 289--307. Oxford University Press.

Götze, Michael, Cornelia Endriss, Stefan Hinterwimmer, Ines Fiedler, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, Ruben Stoel, and Thomas Weskott. 2007. Information structure. In *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, ed. Stefanie Dipper, Michael Götze, and Stavros Skopeteas, volume 7 of *Working Papers of the SFB632: Interdisciplinary Studies on Information Structure (ISIS)*, 90--113. Potsdam: Universitätsverlag Potsdam.

Greaves, Edwin. 1921 [1983]. *Hindi grammar*. New Delhi: Asian educational services.

Gundel, Jeanette. 1994. On Different Kinds of Focus. In *Focus. Linguistic, Cognitive, and Computational Perspectives*, ed. Peter Bosch and Rob van der Sandt, Studies in Natural Language Processing, 293 -- 305. Cambridge: Cambridge University Press.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69:274--307.

Gupta, Bidhu Bhudan Das, and Madhav Lal Karmacharya. 1981. *Nepali self-taught*. Calcutta: Das Gupta Prakashan.

Gupta, Bidhu Bhusan Das. 1976. *Gujarati Self-Taught*. Calcutta: Das Gupta Prakashan.

Gupta, Sagar, and Jyoti Tuladhar. 1979. Dative-subject constructions in Hindi, Nepali and Marathi and relational grammar. *Contributions to Nepalese Studies 7* 7:119--153.

Hardie, Andrew. 2005. Categorisation for automated morphosyntactic analysis of Nepali: introducing the Nelralec Tagset (NT-01). Technical report, Bhasa Sanchar, Kathmandu.

Harrell, Frank E. 2001. *Regression Modeling Stragies.* New York: Springer.

Harrell, Frank E. 2011. rms: Regression Modeling Strategies. R package, http://www.R-project.org/.

Hartmann, Iren, Martin Haspelmath, and Bradley Taylor (ed.). 2013. Valency Patterns Leipzig. http://www.valpal.info/.

Haspelmath, Martin. 2004. Does linguistic explanation presuppose linguistic description? *Studies in Language* 28:554--579.

Haspelmath, Martin. 2005. Argument marking in ditransitive alignment types. *Linguistic Discovery* 3:1--21.

Haspelmath, Martin. 2011. On S, A, P, T, and R as comparative concepts for alignment typology. *Linguistic Typology* 15:535--689.

Hawkins, John. 1978. *Definiteness and Indefiniteness.* London: Croom Helm.

Heath, Jeffrey. 1976. Antipassivization: a functional typology. In *Proceedings of the Second Annual Meeting of the Berkeley Linguistic Society*, 202--211.

Heim, Irene. 1983. File change semantics and the familiarity theory of definiteness. In *Meaning, use and interpreation of language*, ed. Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow, 223--248. Mouton de Gruyter.

Heine, Bernd, and Tania Kuteva. 2002. *World Lexicon of Grammaticalization.* Cambridge: Cambridge University Press.

Heusinger, Klaus von. 1997. Salience and Definiteness. *The Prague Bulletin of Mathematical Linguistics* 67:5--23.

Heusinger, Klaus von. 2007. Accessibility and definite noun phrases. In *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference*, ed. Monika Schwarz-Friesel, Manfred Consten, and Mareile Knees, volume 86 of *Studies in Language Companion Series*, 123--144. Amsterdam: John Benjamins.

Hewson, John. 1972. *Article and noun in English.* The Hague: Mouton.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161--195.

Hoernle, A.F. Rudolf. 1880 [1975]. *A comparative grammar of the Gauḍian (Indo-Aryan) languages.* Amsterdam: Philo.

Hoop, Helen de, and Peter de Swart, ed. 2008. *Differential Subject Marking*. Dordrecht: Springer.

Hopper, Paul, and Sandra Thompson. 1980. Transitivity in Grammar and Discourse. *Language* 56:251--299.

Hughes, Morland. 1947. *A grammar of the Nepali language.* London: Luzac.

Hume, David. 1739 [2003]. *A Treatise of Human Nature.* Adelaide: The University of Adelaide Library. URL http://ebooks.adelaide.edu.au/h/hume/david/h92t/.

Hutt, Michael. 1988. *Nepali: A National Language and its Literature.* New Delhi: Sterling.

Hutt, Michael, and Abhi Subedi. 1999. *Teach Yourself Nepali.* London: Hodder Headline.

Ichihashi-Nakayama, Kumiko. 1994. On dative `subject' constructions in Nepali. In *Aspects of Nepali grammar*, ed. Carol Genetti, Papers in Linguistics 6, 41--76. Santa Barbara: Department of Linguistics, University of California at Santa Barbara.

Iemmolo, Giorgio. 2011. Towards a typological study of Differential Object Marking and Differential Object Indexing. Doctoral Dissertation, Università degli studi di Pavia, Pavia.

Ioup, Georgette. 1977. Specificity and the interpretation of quantifiers. *Linguistics and Philosophy* 1:233--245.

Isaak, André. 1999. The Antipassive, Split Ergativity, and Transitivity. *CLS* 35:175--186.

Jackendoff, Ray. 1983. *Semantics and Cognition.* Cambridge, Massachusetts: MIT Press.

Jain, Usha R. 1995. *Introduction to Hindi grammar*. Berkeley, California: Centers for South and Southeast Asia Studies, University of California at Berkeley.

Johnson, Marion. 1980. *Ergativity in Inuktitut in Montague grammar and in relational Ergativity in Inuktitut in Montague grammar and in relational Ergativity in Inuktitut in Montague grammar and in relational Ergativity in Inuktitut in Montague grammar and in relational grammar*. Bloomington: Indiana University Linguistic Club.

Joshi, Kabindra. 2012. Nepal Maps. http://www.digitalhimalaya.com/collections/maps/nepalmaps/.

Kachru, Yamuna. 2006. *Hindi.* Amsterdam/Philadelphia: John Benjamins.

Kakati, Banikanta. 1941. *Assamese, its formation and development*. Gauhati: Narayani Handiqui Historical Institute.

Kalmár, Ivan. 1979a. The antipassive and grammatical relations in Eskimo. In *Ergativity: towards a theory of grammatical relations*, ed. Frans Plank, 117--143. London/New York: Academic Press.

Kalmár, Ivan. 1979b. *Case and Context in Inuktitut*. Ottawa: National Museum of Canada.

Kambarov, Zaur. 2008. *The concept of definiteness and its application to automated reference resolution.* New York: Peter Lang.

Kamp, Hans. 1981. A Theory of Truth and Semantic Representation. In *Formal Methods in the Study of Language*, ed. J. Ballweg and H. Glinz, 103--114. Amsterdam: Mathematisch Centrum.

Kärkkäinen, Elise. 1994. On the *i(n)* Construction in Nepali. In *Aspects of Nepali grammar*, ed. Carol Genetti, 77--115. Santa Barbara: Department of Linguistics, University of Berkeley at Santa Barbara.

Kellogg, S.H. 1875 [1972]. *A grammar of the Hindi language.* New Delhi: Munshiram Manoharlal.

Khaḍkā, Śundara Sāna. 2055 V.S. *Prayogika Nepālī vyākaraṇa.* Kāṭhmāḍauṁ: Dhuba and Narendra Khaḍkā. 1998/1999 AD.

Kibrik, Andrej. 2011. *Reference in Discourse.* Oxford: Oxford University Press.

Kiparsky, Paul. 1998. Partitive Case and Aspect. In *The projection of arguments*, ed. Miriam Butt and Wilhelm Geuder, 265--308. Stanford: CSLI.

Kittilä, Seppo. 2002. Remarks on the basic transitive sentence. *Language Sciences* 24:107--130.

Kittilä, Seppo. 2008. Animacy effects on differential goal marking. *Linguistic Typology* 12:245--268.

Klages-Kubitzki, Monika. 1995. *Article Usage in English.* Frankfurt: Peter Lang.

Kleinschmidt, Samuel. 1851 [1968]. *Grammatik der grønlændischen Sprache : mit teilweisem Einschluss des Labradordialekts.* Hildesheim: Olms.

König, Ekkehard. 1991. *The Meaning of Focus Particles.* London/New York: Routledge.

Korolev, I. 1965. *Jazyk Nepali.* Moskva: Nauka.

Krifka, Manfred. 2007. Basic notions of information structure. In *Working Papers of the SFB632: Interdisciplinary Studies on Information Structure (ISIS)*, volume 6, 13--56. Potsdam: Universitätsverlag Potsdam.

Kulikov, Leonid. 2011. Voice typology. In *The Oxford Handbook of Linguistic Typology*, ed. Jae Jung Song, chapter 18, 368--398. Oxford: Oxford University Press.

Lahaussois, Aimée. 2002. Aspects of the grammar of Thulung Rai. Doctoral Dissertation, University of California, Berkeley, California.

Lahaussois, Aimée. 2009. Koyi Rai: An initial grammatical sketch. *Himalayan Linguistics Archive* 4:1--33.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things.* Chicago: University of Chicago Press.

Lambrecht, Knud. 1994. *Information structure and sentence form.* Cambridge: Cambridge University Press.

Lamsāla, Rāmcandra. 2062 V.S. *Nepālī bhāṣā ra vyākaraṇa.* Kirtipur: Sunlight publication. 2005/2006 AD.

Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159--174.

Lawless, J.F., and K. Singhal. 1978. Efficient screening of nonnormal regression models. *Biometrics* 34:318--327.

Lazard, Gilbert. 2001. Le marquage differéntiel de l'objet. In *Language typology and language universals: an international handbook*, ed. Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, volume 2, 873--885. Berlin: Mouton de Gruyter.

Leiss, Elisabeth. 2000. *Artikel und Aspect. Die grammatischen Muster von Definitheit*. Berlin/New York: Walter de Gruyter.

Li, Chao. 2007a. Split ergativity and split intransitivity in Nepali. *Lingua* 117:1462--1482.

Li, Chao. 2007b. Split ergativity in Nepali and its typological significance. *University of Pennsylvania Working Papers in Linguistics* 13:169--182.

Lichtenberk, Frantisek. 1982. Individuation hierarchies in Manam. In *Studies in transitivity*, ed. Paul Hopper and Sandra Thompson, 261--276. New York: Academic Press.

Link, Godehard. 1983. The Logical Analysis of Plurals and Mass Terms: A Lattice-Theoretic Approach. In *Meaning, use and interpreation of language*, ed. Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow, 302--324. Berlin: Blackwell.

Lyons, Christopher. 1999. *Definiteness.* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Macdonell, Arthur Anthony. 1916 [1941]. *A Vedic Grammar for Students.* Delhi: Oxford University Press.

Mahapatra, Bijay Prasad. 2007. *A synchronic grammar of Oriya.* Mysore: Central Institute of Indian Languages, Mysore.

Malchukov, Andrej. 2007. Animacy and asymmetries in differential case marking. *Lingua* 118:203--221.

Malchukov, Andrej, Martin Haspelmath, and Bernard Comrie. 2010. Ditransitive constructions: a typological overview. In *Studies in Ditransitive Constructions: A Comparative Handbook*, ed. Andrej Malchukov, Martin Haspelmath, and Bernard Comrie, 1--64. Berlin/New York: Mouton de Gruyter.

Margetts, Anna. 2008. Transitivity Discord in some Oceanic Languages. *Oceanic Linguistics* 47:30--44.

Margetts, Anna. 2011. Transitivity in Saliba-Logea. *Studies in Language* 35:650--675.

Masica, Colin. 1982. Identified Object Marking in Hindi and other Languages. In *Topics in Hindi Linguistics*, ed. Omkar N. Koul, volume 2, 16--51. Chandigarh: Bahri.

Masica, Colin. 1991. *The Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Massam, Diane. 2001. Pseudo Noun Incorporation in Niuean. *Natural Language & Linguistic Theory* 19:153--197.

Matisoff, James. 1972. Lahu nominalization, relativization, and genitivization. In *Syntax and semantics*, ed. John Kimball, volume 1, 237--257. New York: Seminar Press.

Matras, Yaron. 2002. *Romani*. Cambridge: Cambridge University Press.

Matthews, David. 1984. *A course in Nepali*. New Delhi: Heritage.

Matuszewski, Jeanine M. 2011. Properties of an R^2 Statistic for Fixed Effects in the Linear Mixed Model for Longitudinal Data. Doctoral Dissertation, University of North Carolina, Department of Biostatistics, Chapel Hill.

Meerendonk, M. 1949. *Basic Gurkhali Grammar*.

Menard, Scott. 2002. *Applied logistic regression analysis*. Quantitative Applications in the Social Sciences. Thousand Oaks: Sage.

Mistry, P. J. 1997. Objecthood and specificity in Gujarati. In *The Life of Language*, ed. Jane H. Hill, P. J. Mistry, and Lyle Campbell, 425--442. Berlin/New York: Mouton de Gruyter.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. Stanford: CSLI.

Monier-Williams, Monier. 1899 [1974]. *A Sanskrit-English dictionary*. Oxford: Clarendon.

Montaut, Annie. 2004. *A Grammar of Hindi*. München: Lincom.

Mukherjee, Shantanu. 1985. *Kasus und Diathese im Bengalischen*. Heidelberg: Groos.

Müller-Gotama, Franz. 1994. *Grammatical Relations. A Cross-Linguistic Perspective on their Syntax and Semantics*. Berlin/New York: Mouton de Gruyter.

Næss, Åshild. 2007. *Prototypical transitivity*. Amsterdam/Philadelphia: John Benjamins.

Nagelkerke, N.J.D. 1991. A Note on a General Definition of the Coefficient of Determination. *Biometrika* 78:691--692.

Neukom, Lukas, and Manideepa Patnaik. 2003. *A grammar of Oriya*. Arbeiten des Seminars für Allgemeine Sprachwissenschaft (ASAS). Zürich: Universität Zürich.

Nichols, Joanna. 1986. Head-marking and Dependent-marking Grammar. *Language* 62:56--119.

Opgenort, Jean Robert. 2004. *A grammar of Wambule*. Leiden/Boston: Brill.

Opgenort, Jean Robert. 2005. *A grammar of Jero*. Leiden/Boston: Brill.

Palmer, Frank R. 1994. *Grammatical roles and relations.* Cambridge: Cambridge University Press.

Pandharipande, Rajeshwari V. 1990. Experiencer (Dative) NPs in Marathi. In *Experiencer Subjects in South Asian Languages*, ed. K. P. Mohanan Manindra K. Verma, 161--180. Stanford: Center for the Study of Language and Information, Stanford University.

Pandharipande, Rajeshwari V. 1997. *Marathi.* London/New York: Routledge.

Paudyal, Netra. 2009. The syntax of three-argument verbs in Nepali. Seminar paper written at the University of Leipzig.

Paudyāl, Netra Prasād, Balthasar Bickel, Robert Schikowski, Sabine Stoll, Elena Lieven, Gomā Banjade, Iccha Pūrṇa Rāī, Manoj Rāī, Martin Gaenszle, Noval Kiśor Rāī, and Toyā Nāth Bhaṭṭa. 2010. Non-finite adverbial subordination in Chintang. *Nepalese Linguistics* 25:121--132.

Peirce, Charles. 1906. Prolegomena to an Apology for Pragmaticism. *The Monist* 16:492--546.

Peterson, John. 1998. *Grammatical Relations in Pāli and the Emergence of Ergativity in Indo-Aryan.* München: Lincom.

Plank, Frans. 1984. *Objects. Towards A Theory Of Grammatical Relations.* London: Academic Press.

Pokharela, Bālakṛṣṇa, ed. 2031 V.S. *Pānc Say Varṣa.* Lālitapura: Jagadambā. 1974/1975 AD.

Pokharela, Mādhavaprasāda. 2054 V.S. *Nepālī Vākya-Vyākaraṇa.* Kāṭhamāḍauṁ: Nepāla Rājakīya Prajñā-Pratiṣṭhāna (Royal Nepal Academy). 1997/1998 AD.

Pradhāna, Paraśu. 1997. Ṭēbalamāthiko tyasa ākāśavāṇī. In *Modern literary Nepali*, ed. Michael Hutt, 75--79. New Delhi: Oxford University Press.

Prasain, Balaram. 2011. A computational analysis of Nepali morphology: a model for natural language processing. Doctoral Dissertation, Tribhuvan University, Kathmandu.

Primus, Beatrice. 1999. *Cases and Thematic Roles.* Tübingen: Niemeyer.

R Development Core Team. 2012. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.r-project.org.

Rāī, Noval Kiśor. 1984. A descriptive study of Bantawa. Doctoral Dissertation, Deccan college, Pune.

Rāī, Noval Kiśor, Manoj Rāī, Netra Prasād Paudyāl, Robert Schikowski, Balthasar Bickel, Sabine Stoll, Martin Gaenszle, Gomā Banjade, Iccha Pūrṇa Rāī, Toyā Nāth Bhaṭṭa, Sebastian Sauppe, Rikhī Māyā Rāī, Janak Kumārī Rāī, Lās Kumārī Rāī, Durga Bahādur Rāī, Gaṇeś Rāī, Dayārām Rāī, Durga Kumārī Rāī, Anitā Rāī, Candra Kumārī Rāī, Śanti Māyā Rāī, Ravindra Kumār Rāī, Judy Pettigrew, and Tyko Dirksmeyer. 2011. *Chintāṅa-Nepālī-Ãgrejī śabdakośa tathā vyākaraṇa.* Kāṭhmāḍauṁ: Chintang Language Research Programme.

Raible, Wolfgang. 1992. *Junktion. Eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration.* Heidelberg: Winter.

Ray, Punya Sloka, Muhammad Abdul Rai, and Lila Ray. 1966. *Bengali language handbook.* Washington: Center for Applied Linguistics.

Riccardi, Theodore. 2003. Nepali. In *The Indo-Aryan Languages*, ed. George Cardona and Dhanesh Jain, Routledge Language Family Series, chapter 15, 538--580. London: Routledge.

Rijkhoff, Jan. 2002. *The Noun Phrase.* Oxford: Oxford University Press.

Ritter, Elizabeth, and Sara Thomas Rosen. 2010. Animacy in Blackfoot: Implications for Event Structure and Clause Structure. In *Lexical Semantics, Syntax, and Event Structure*, ed. Malka Rappaport Hovav, Edit Doron, and Ivy Sichel, chapter 7, 124--152. Oxford: Oxford University Press.

Rizopoulos, Dimitris. 2011. ltm: Latent Trait Models under IRT. R package, http://www.R-project.org/.

Roberts, Craige. 2003. Uniqueness in definite noun phrases. *Linguistics and Philosophy* 26:287--350.

Russell, Bertrand. 1905. On denoting. *Mind* 479--493.

Rutgers, Roland. 1998. *Yamphu*. Leiden: Research School CNWS, School of Asian, African, and Amerindian Studies.

Sadock, Jerrold. 2003. *A Grammar of Kalaallisut*. München: Lincom.

Sandahl, Stella. 2000. *A Hindi Reference Grammar*. Leuven: Peeters.

Saussure, Ferdinand de. 1915 [1975]. *Cours de linguistique générale*. Paris: Payot.

Schackow, Diana. In preparation. Aspects of Yakkha Grammar. Doctoral Dissertation, University of Leipzig.

Schikowski, Robert. 2011. Chintang morphology. Unpublished manuscript.

Schikowski, Robert, Balthasar Bickel, and Netra Paudyal. forthcoming. Flexible valency in Chintang. In *Valency Classes: a comparative Handbook*, ed. Bernard Comrie and Andrej Malchukov.

Schmidt, Bodil Kappel. 2003. West Greenlandic antipassive. In *Proceedings of the 19th Scandinavian Conference on Linguistics*, ed. Peter Svenonius Anne Dahl, Kristine Bentzen, volume 31.2, 385--399. Tromsø.

Schmidt, Ruth Laila. 1993. *A Practical Dictionary of Modern Nepali*. Delhi: Ratna Sagar.

Schmidt, Ruth Laila. 1999. *Urdu. An essential grammar*. London/New York: Routledge.

Seiter, William. 1980. *Studies in Niuean Syntax*. New York: Garland Press.

Shukla, Shaligram. 1981. *Bhojpuri grammar*. Washington: Georgetown University Press.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In *Grammatical Categories in Australien Languages*, ed. Robert Dixon, 112--171. Canberra: Australien Institute of Aboriginal Studies.

Sim, Julius, and Chris C. Wright. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy* 85:257--268.

Sinnemäki, Kaius. forthcoming. A typological perspective on differential object marking. In *Special issue of Linguistics*, ed. Giorgio Iemmolo and Gerson Klumpp.

Smith, W.L. 1997. *Bengali Reference Grammar*. Stockholm: Association of Oriental Studies.

Sommer, Anton F.W. 1993. *Einführung in das Nepali*. Wien: Self-published.

Srivastava, Dayanand. 1962. *Nepali language: its history and development*. Calcutta: Calcutta University.

Stoll, Sabine, and Balthasar Bickel. 2009. How deep are differences in referential density? In *Crosslinguistic Approaches to the Psychology of Language*, ed. Jiansheng Guo, Elena Lieven, Nancy Budwig, Susan Ervin-Tripp, Keiko Nakamura, and Şeyda Özçalışkan, 543--555. New York: Psychology Press.

Stoll, Sabine, and Balthasar Bickel. 2012. How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications. In *Potentials of language documentation: methods, analyses, utilization*, ed. Frank Seifart, Geoffrey Haig, Nikolaus Himmelmann, Dagmar Jung, Anna Margetts, Paul Trilsbeek, and Peter Wittenburg, 84--90. Manoa: University of Hawai'i Press.

Stoll, Sabine, and Balthasar Bickel. in press. The acquisition of ergative case in Chintang. In *The acquisition of ergativity*, ed. Edith Bavin and Sabine Stoll. Amsterdam: John Benjamins.

Swart, Henriette de. 2006. Aspectual implications of the semantics of plural indefinites. In *Non-definiteness and plurality*, ed. Svetlana Vogeleer and Liliane Tasmowski, 161--189. Amsterdam/Philadelphia: John Benjamins.

Swart, Peter de. 2007. *Cross-linguistic Variation in Object Marking*. Utrecht: LOT.

Tauberschmidt, Gerhard, and Alfred Bala. 1992. Transitivity and Ergativity in Sinaugoro. *Language and Linguistics in Melanesia* 23:179--191.

Taylor, Geo. P. 1908. *The student's Gujarātī grammar*. Bombay: Thacker.

Tolsma, Gerard Jacobus. 2006. *A grammar of Kulung*. Leiden/Boston: Brill.

Tolstaya, N.I. 1960. *The Panjabi language*. London/Boston/Henley: Routledge Kegan Paul.

Turnbull, A. 1923 [1992]. *Nepali Grammar Vocabulary*. New Delhi: Asian Education Services, 3 edition.

Turner, Ralph Lilley. 1931 [1990]. *A comparative and etymological dictionary of the Nepali language*. New Delhi: Allied Publishers.

UNESCO Ad Hoc Expert Group on Endangered Languages. 2003. Language Vitality and Endangerment. Technical report, UNESCO, Paris.

United Nations Cartographic Section. 2007. General Maps. http://www.un.org/Depts/Cartographic/english/htmain.htm.

Valin, Robert Van, and Randy LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

Verkuyl, H. J. 1972. *On the compositional nature of the aspects*. Dordrecht: Reidel.

Verkuyl, Henk. 1993. *A theory of aspectuality*. Cambridge: Cambridge University Press.

Verma, Manindra. 1992. Nepali. In *International Encyclopedia of Linguistics*, ed. William Bright, volume 3, 76--79. Oxford: Oxford University Press.

Verma, Manindra K. 2003. Bhojpuri. In *The Indo-Aryan Languages*, ed. George Cardona and Dhanesh Jain, 515--537. London/New York: Routledge.

Wallace, William David. 1981. Object-marking in the history of Nepali. *Studies in The Linguistics Sciences* 11:107--129.

Wallace, William David. 1985. *Subjects and subjecthood in Nepali: an analysis of Nepali clause structure and its challenges to relational grammar and government and binding*. Michigan: UMI.

Weekley, Ernest. 1921. *An Etymological Dictionary of Modern English*. London: John Murray.

Weidert, Alfons, and Bikram Subba. 1985. *Concise Limbu Grammar and Dictionary*. Amsterdam: Lobster.

Whelpton, John. 2011. *A History of Nepal*. Cambridge: Cambridge University Press.

Whitney, William Dwight. 1889 [1974]. *Sanskrit grammar*. Cambridge, Massachusetts: Harvard University Press.

Williams, Robert S. 1994. A Statistical Analysis of English Double Object Alternation. *Issues in Applied Linguistics* 5:37--58.

Witzlack-Makarevich, Alena. 2011. Typological variation in grammatical relations. Doctoral Dissertation, University of Leipzig.

Woodbury, Tony. 2003. Defining documentary linguistics. *Language documentation and description* 1:35--51.

Wulff, Stefanie. 2003. A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics* 245--282.

Yadav, Ramawatar. 1996. *A Reference Grammar of Maithili.* Berlin/New York: Mouton de Gruyter.

Yadava, Yogendra. 2003. Language. In *Population Monograph of Nepal, 2001*, 137--171. Kathmandu: Central Bureau of Statistics.

Yadava, Yogendra P. 1996. Verb agreement in Maithili. *Journal of Nepalese Studies* 1:109--121.

Lebenslauf

Robert Schikowski
Schwandenholzstrasse 212
8046 Zürich

Telefon: +41 (0)44 63 40236
Email: robert.schikowski@uzh.ch
Homepage: www.spw.uzh.ch/schikowski

Persönliche Daten

Geburtsdatum:	27.12.1982
Geburtsort:	München
Nationalität:	Deutsch

Arbeitserfahrung und Lehre

02 2013 - jetzt:	Administrativer Koordinator im ZüKL (Zürcher Kompetenzzentrum Linguistik)
08 2012 - jetzt:	Workflow-Manager des CLRP (Chintang Language Research Program)
09 2012 - 12 2012:	Unterricht Seminar Areallinguistik (Allgemeine Sprachwissenschaft, UZH)
09 2009 - 08 2012:	wissenschaftlicher Mitarbeiter im EuroBABEL-Projekt RHIM (Gruppe Leipzig/Zürich, „Differential agreement vs differential case“)
02 2012 - 06 2012:	Unterricht Übung Phonologie (Allgemeine Sprachwissenschaft, UZH)
04 2009 - 07 2009:	Unterricht Proseminar Phonologie (Allgemeine Sprachwissenschaft, LMU München)
10 2008 - 01 2009:	Tutorium zum Kurs Dokumentationslinguistik (LMU München)
01 2009 - 03 2009:	wissenschaftliche Hilfskraft im IMPACT-Projekt (Gruppe CIS München, „Verbesserung der OCR historischer deutscher Texte“)
10 2007 - 01 2008:	Tutorium zum Proseminar Syntax (Allgemeine Sprachwissenschaft, LMU München)
10 2007 - 01 2008:	Unterricht Einführung ins Japanische (Allgemeine Sprachwissenschaft, LMU München)
11 2006 - 07 2007:	privater Deutschlehrer in Tokyo
11 2006 - 07 2007:	Übersetzer Japanisch > Deutsch bei goga Tokyo
04 2006 - 07 2006:	Tutorium zum Proseminar Typologie (Allgemeine Sprachwissenschaft, LMU München)
04 2006 - 07 2006:	Unterricht Einführung ins Westgrönländische (Allgemeine Sprachwissenschaft, LMU München)
10 2005 - 01 2006:	Tutorium zum Proseminar Computerlinguistische Morphologie und Lexikographie (Computerlinguistik, LMU München)

10 2005 - 01 2006:	Tutorium zur Einführung in die Allgemeine Sprachwissenschaft (Allgemeine Sprachwissenschaft, LMU München)

Ausbildung

2011 - 2013:	Doktorat an der UZH (Fach Allgemeine Sprachwissenschaft), Abschluss *summa cum laude*
2009 - 2011:	Doktorat an der Universität Leipzig (Fach Linguistik)
2004 - 2009:	Magisterstudium an der LMU München (Hauptfach Allgemeine Sprachwissenschaft, Nebenfächer Philosophie und Computerlinguistik), Abschluss mit 1,3
2006 - 2007:	Auslandsstudium an der University of Tokyo
1993 - 2002:	Besuch des Dom-Gymnasiums Freising, Abiturnote 1,7
1989 - 1993:	Besuch der Grundschule Haimhausen

Ehrenamtliche Tätigkeit

01 2011 - 10 2012:	Sprecher des Doktoratsprogramms Linguistik (UZH)
10 2004 - 10 2008:	Fachschaftssprecher der Allgemeinen Sprachwissenschaft (LMU München)
2004 - 2007:	freiwilliger Betreuer auf Behindertenfreizeiten der OBA München
07 2003 - 10 2003:	Anderer Dienst im Ausland in einer Behindertenwerkstätte bei Tokyo