Indian Linguistics 82(1-2) 2021: 1-25 ISSN: 0378-0759

USING INTERLINEAR GLOSS TEXTS TO IMPROVE LANGUAGE DESCRIPTION

SHOBHANA CHELLIAH, MARY BURKE and MARTY HEATON
University of North Texas
shobhana.chelliah@unt.edu

ABSTRACT

Interlinear-glossed text (IGT) is a method of representing semantic, morphological and phonological information about lexemes along with phrase and clause level translations of connected text. While the Leipzig Glossing Rules (LGR) provide general standards and principles for IGT, we argue here that language-family specific guidelines are necessary to facilitate rapid creation of new interpretable IGT that can be used for language description, typological discovery, and cross-language comparison. Using selected examples of Tibeto-Burman IGTs, we demonstrate how linguists create their own terminology and conventions for representing linguistic phenomena which fall outside the scope of the LGR. To date, there are few, at least within the Sino-Tibetan linguistics community, that have discussed language-family specific IGT conventions, so new annotators lack guidance on IGT creation. This paper examines how typical Tibeto-Burman constructions (e.g., reduplication, verb stem alternation, directionals) are represented in IGT from several South Central Tibeto-Burman languages. We offer some remarks on the purposes of IGT and some principles for new IGT creators.

Keywords: Interlinear-glossed text (IGT), language description, Tibeto-Burman, South Central Tibeto-Burman

1. Introduction

Interlinear-glossed text (IGT) is a method of representing semantic, morphological and phonological information about lexemes along with phrase and clause level translations of connected text. IGT representations often include three lines of analysis, as seen in (1): the top line is the target language transcribed in a practical orthography or IPA, the next line includes a lexeme-by-lexeme gloss, and the final line in single quotes provides a free translation.

(1) Hakha Lai (Van Bik, 2010, p. 139)[1]

nihîn	*tsŭu*	*kàymaʔ*	*tivăa*	*ka-kâl*	*làay*
today	DEM	1SG.PRON	river	1SG.S-go.I	FUT

'Today, I will go to the river.'

These representations are created iteratively, that is, new versions of IGT for the same connected text are generated by the annotator (the IGT creator) as that annotator's understanding of constructions improves. The process of creating IGT is useful for improving analyses because the annotator sees constructions in cultural and pragmatic context and can use that context to understand lexical and morphological semantics. This is needed because constructions accessible through natural speech are often difficult to elicit through more traditional questionnaires and word or sentence elicitation (Chelliah, 2001; Rice, 2001; Davis et al., 2014). Also, the annotator can benefit from the use of software which stores analyses and representations and allows for comparison across constructions via concordancing features. Automating this process also allows

for easy revisions and export of examples for further analysis and incorporation of connected text examples in grammar writing.

Although IGT is a widely used form of representing and analyzing data, there are almost no guidelines for the novice annotator on how to move from transcribed text to IGT. There are some general principles for annotation in the form of the Leipzig Glossing Rules (LGR) developed at the Max Planck Institute for Evolutionary Anthropology (2015). LGR provide a set of morpheme labels for describing common linguistic features. One of the stated goals of the LGR is to offer an avenue for comparison of analyses by different linguists, which could enable data-driven typological discoveries and computational efforts.

The LGR offer guidelines in terms of 'syntax' (e.g., using an equal sign to indicate a morpheme is an enclitic) and 'semantics' (e.g., assigning a label to each morpheme, such as NEG for 'negation, negative'). Since the LGR-suggested annotation labels are for typologically universal or common linguistic phenomena, the labels are not intended to capture all features possible in human languages. Thus, LGR states that "most authors will feel the need to add (or modify) certain conventions (especially category labels)" (Max Planck Institute for Evolutionary Anthropology Department of Linguistics, 2015, p. 1). Where language-specific features fall outside the scope of the LGR, linguists create their own specialized way of representing phenomena. To date, there are few, at least within the Sino-Tibetan linguistics community, that have discussed language-family specific IGT conventions (see discussion in Chelliah, 2020; Lahaussois, 2021). As we discuss in this paper, more robust IGT guidelines could increase production of IGT and improve descriptive outputs.

Our paper addresses this gap in annotation conventions at the language-family level. As a first step in exploring this issue, we describe IGT variation in selected published grammars and descriptions of South-Central Tibeto-Burman languages.[2] We divide our discussion of IGT into two components:

(1) IGT syntax, i.e., how IGT reflects researchers' analyses of:

- Words and word boundaries (e.g., spaces between forms)
- Morpheme categories (e.g., bound, free, inflectional, derivational, enclitic) and how morpheme categories are indicated (e.g., hyphens, spaces, equal sign, uppercase or lowercase for glosses)
- Constructions such as compounds, verb serialization, clauses, or phrases and how these are represented (e.g., using white space, line breaks, or bracketing to separate constituents).

(2) IGT semantics, or how the following considerations are addressed in IGT:

- Grammaticalized, lexicalized or polysemous forms: e.g., where a case maker such as the 'dative' develops a second use such as clausal subordinate 'after.that' should both the case maker and clausal subordinator be glossed 'dative' or glossed differently to match the function.
- Morphemes without suitable abbreviations in the LGR category labels list. LGR broadly aims for 1:1 relationship between form and gloss (p. 2), but leaves open researchers' decisions for trickier cases. How should the researcher represent morphemes or constructions for which they have only a partial definition? Or, where they can identify the function, but cannot yet describe the semantics? For example, it may be clear that a

morpheme is a clausal subordinator, but the exact meaning of the subordinator may not be known yet.

In this paper, we provide a brief introduction to the South Central Tibeto-Burman languages under discussion. Section 2 presents examples of IGT syntax and semantics with discussion on why and how these align or do not align between writers. In Section 3, we offer discussion of potential applications of IGT, including some principles to consider when creating IGT. Section 4 concludes.

2. Variation in IGT

2.1 Introduction to South Central Tibeto-Burman IGT

We consider the representation of constructions typical for South Central Tibeto-Burman languages. We focus on constructions commonly seen in South Central, such as:

- Derivational and inflectional morphology (Chhangte, 1993; VanBik, 2010; Bedell et al., 2013)
- Reduplication and expressive collocations (Chhangte, 1993; Peterson, 2008; Peterson, 2010; Davis, 2017; Zakaria, 2017; Chelliah et al., 2021)
- Case marking (So-Hartmann, 2009; Zakaria, 2017; Davis, 2017)
- Verb stem alternation (Chhangte, 1993; VanBik, 2010; Haokip, 2012)
- Information structure (Peterson, 2011)
- Complex participant marking systems (Chelliah & Utt, 2017; Ozerov, 2019; Konnerth & DeLancey; 2019).

We focus on variability in the IGT representation of categories within these types of constructions.

2.2 IGT syntax

In this section, we consider punctuation and spacing used to represent morphological structure in IGT. By convention, a space (also called 'white space') is used to indicate division between lexemes. One of the most challenging aspects of creating IGT for South Central Tibeto-Burman languages is deciding where to place these spaces. Morphemes, whether bound or free, may be separated from the main verb with white space, as in the work of Bedell et al. (2013), Davis (2017), or VanBik (2010). Alternatively, the verb head may be written together with derivational and inflectional morphology separated by hyphens (as is done for Mizo in Chhangte (1993)). The conventions on how to represent IGT at this level are often based on nascent practical orthographies and are subject to change as these orthographies develop. Indeed, as the vast majority of these orthographies are still in development, there is at least some variation between writers with respect to what is written together and what is written separately. This variation spills over to considerations of the placement of white space in IGT.

2.2.1 Wordhood and Clitics

Consider, for example as seen in (2), (3) and (4), how these languages are represented as more isolating or agglutinative with similar future tense constructions. In the Sizang example, the temporal is analyzed as an enclitic; in Mizo, as a bound (presumably) derivational suffix; while, for Hakha Lai, the temporal is written separately, and whether the temporal is a bound or free morpheme is not represented in the syntax.

(2) Sizang (Davis, 2017, p. 39)
*ziŋ=**tɛ̌:*** *pɔ̂:ai* *ŏm* *tû:* *hî:*
morning **=TEMP** festival exist.I IRR be
'There will be a festival in the morning'

(3) Hakha Lai (VanBik, 2010, p. 138)
nihîn *tsŭu* *nàŋmaʔ* *tivâa* *na-kâl* ***làay***
today DEM 2SG.PRON river 2SG.S-go.I FUT
'Today, you will go to the river.'

(4) Mizo (Chhangte, 1993, p. 111)
naktùùk-aʔ *kán-hóóŋ-**áŋ***
tomorrow-LOC 1S.PL-return.home-**FUT**
'We will return home tomorrow.'

A comparison of these examples illustrates the issue we are highlighting in this paper. In the descriptions in which these examples occur, there is no overt guide as to how we should read the IGT. Therefore, we assume that, per LGR, hyphens "-" are used for bound morphemes, white space " " for free morphemes, and the equal sign "=" for clitics. Thus, if we go by the IGT, in Sizang, the temporal is indicated by an enclitic, in Hakha Lai by an independent lexeme, and in Mizo, by a bound morpheme. In the case of Sizang, the representation commits the annotator to see these as enclitics. For the Hakha Lai and Mizo, however, we cannot be sure if it is orthography or analysis determining the IGT syntax.

2.2.2 Wordhood and Reduplication

South Central languages have very similar lexical constructions as well as similarities in the verbal and nominal phrases. One common feature of lexica are copied or duplicated segments which function as verbal modifiers (Chhangte, 1993; So-Hartmann, 2009; Zakaria, 2017; Chelliah et al., 2021), in some instances called verbal classifiers (Peterson, 2008; Peterson, 2011). Copied or duplicated forms can be seen as following a verb stem but preceding inflection and, in this sense, the duplicated forms can be considered verbal bound morphology. They are represented as such in example (5) for Mizo and (6) for Khumi. In the Lamkang Naga example (7) or Daai Chin example (8), the copied forms are written as independent lexemes. For Lamkang Naga, it is not clear from the IGT that these copied forms are in fact part of the verbal complex. We created the IGT for Lamkang Naga over the course of an 8-year language documentation project. When looking for the most accurate and efficient way to represent this IGT, we certainly wanted to be true to the genius of this particular language but, as annotators often do, we turned to the literature to see what others had done. There is no model at the level of World Atlas of Language Structures Online (WALS) (Dryer & Haspelmath, 2013) or in LGR because, although common for South Central, these forms are not common for all the languages of the world. There is also no one model for South Central. So, we were on our own to make decisions on representation which we based first on orthography, then morphological structure (once this was discovered), and then back to orthography to make use of software more consistent with the transcription provided by community transcribers. In short, a language family-specific discussion of the relevant issues would have proven useful in speeding up our process and in offering context for making this decision, even if we decided in the end to employ a language-specific convention.

(5) Mizo (Chhangte, 1993, p. 183)
*tha-**èèm-èèm**-în* *a-lów-kúáy-ta-**nhèèp-nhèè*** *â*
good-**very-very**-ADV 3S-come-sprout-PRPF-**INT-INT** FP
'It sprouted luxuriously (it sprouts bending softly).'

(6) Khumi (Peterson, 2008, p. 121)

puykhawng	*p-kung-**phuphuu***	*khaá*	*oeyngkeéwng*	*ngaang*	*hawplay*
(name)	CAUS-enter-**AUGVCL**	when	tree.ALL	leaf	well

leng-ceng	*khay-yo=pray=loee*	*naang=poee*
stuff.into-tightly	leave-IMPERF=INTENS=TOP	2S=FOC

'When Puykhawng made him [a bear] go inside, he packed the tree tightly with leaves, and left, you see.'

(7) Lamkang Naga (Chelliah et al., 2021, p. 178)

a-ktxek	***sek–sek***	*ráh*
2A-tear	**ideo–ideo**	3A.FUT

'you will slice it into pieces'

(8) Daai Chin (So-Hartmann, 2009, p. 121)

Nghia-theih	*ta*	*ang'aai*	***jak-jak**=a*	*hmin=kti.*
Mango-fruit	FOC	yellowishly	**INTENSF:very**=CF	ripe=NON.FUT

'As for the mango fruits they are very yellowishly ripe.'

Compare these to the Hakha Lai forms in (9) where the verbal modifiers are written separately. These examples illustrate a common source of variation in how these forms are written. They may be written as separate words because, prosodically, there is break after the stem even though, morphologically, they form a unit with the stem. This prosodic break gets represented in the IGT, obscuring the morphological status of the reduplicated form. The Lamkang Naga example similarly shows a break which is prosodically determined, reflected in the spelling, and then again in the IGT.

(9) Hakha Lai (Patent, 1998, p. 177)

a. hŋaak-tshia ʔa-vaak **ʔuaʔ-maʔ**
baby 3SS-crawl **IDEO**
'The [big, fat] baby crawls around.'

b. hŋaak-tshia ʔa-vaak **ʔiaʔ-maʔ**
baby 3SS-crawl **IDEO**
'The [small, thin] baby crawls around.'

Comment [u1]: SS is not there in the Abbreviation list of this language.

Let us also consider reduplicated forms that act as nominal modifiers. The reduplicants are written as separate words and, sometimes, with no space between the copied forms, as in (6) and (10). For many languages, however, the reduplicants are joined by hyphens.

(10) Hyow (Zakaria, 2017, p. 159)

yówyówâ kɔ́n hngát hmúʔhɔ̂ní bɔ́kphóngphóngâ útsúní úpúm khêní hɛ́n dzídzíâ thónéy dɔ̂k hángháng.

yówyówâ	*kɔ́n*	*hngát*	*hmúʔ-hɔ̂-ní*	*bɔ́k**phóngphón**gâ*	*ú-tsú-ní*
glitteringly	pond	one	see.II-PM-TEMP	in.**utterly.white**	3S-jump-TEMP

ú-púm	*khê=ní*	*hɛ́n*	*dzídzíâ*	*(thón-éy-dɔ̂k)*	*hángháng*
3SG.POSS-body	all=FOC	silver	just	(happen-MID)-ANT	all

'Since he saw a pond glittering, when he jumped in utterly white, all his body became just silver, all.'

Oppose the nominal modifier 'glitteringly' in (10) above with the adverbial with clausal scope of 'like that' in (11) below, which is written separately. Finally, note that fully duplicated verb phrases are usually written separately, as seen in the Hyow and Mizo examples (11) and (12), respectively.

(11) Hyow (Zakaria 2017, p. 459)
èybɔ́, èybɔ́, khònpé, khònpé únûlá ání prèâ èyméyhyɔ̂. prètsàʔání ááphɔ̀âl.

èybɔ́	***èybɔ́***	***khón-pé***	***khón-pé***	*ú-nû*	*lâ*
like.that	**like.that**	**strip.off.I-SIM**	**strip.off.I-SIM**	3SG.POSS-mother	CONJ

ání	*pré=â*	*èy*	*mêy-hyɔ̂*	*pré-tsáʔ=â=ní*
3SG	country=LOC	ANAPH.DEM	exist.II-PM	country-border=LOC=FOC

á-á-phɔ́-âl
3S-DIR-reach.II-DEP
'Like that, like that, stripping off, stripping off the banana tree stem, his mother and he stayed. He reached the border of the country.'

(12) Mizo (Chhangte, 1993, p. 92)

ruaʔ	*hî*	***a-sûûr***	***a-sûûr***	*â*
rain	DEM	**3S-pour**	**3S-pour**	FP

'It rains every time.'

One aspect we will touch on in section 2.3 is how morphemes are glossed. Here, we see the differences in which level of structure is being captured by the glossing. The semantics of the form are captured in the glosses as in Mizo (5) and Daai Chin (8). But, annotators differ on whether they specify the manner or extent of intensity (e.g., 'very' or 'somewhat'), or just label the morpheme as 'intensity' as in (6). Some specify the grammatical function, as in the Khumi (6) example, which specifies the use of the reduplicated form as an 'augmentative verbal classifier.' Similarly an abstract category label such as 'ideophone' in the Lamkang Naga example (7) or 'reduplicant' as in the Anal Naga (13) example.

(13) Anal Naga (Ozerov, 2019, p. 32)

va-sìŋ-kʰín!	*sàːn*	*íː-**íː**-má-kʰín!*
IMP.INTR-clever-IMP.PL	night	sleep-**RDP**-NEG-IMP.PL

'Be clever! Do not sleep in the night!'

2.2.3 Serial verbs

When two verbs occur in sequence, the IGT creator needs to decide whether to represent the verbs as separate lexical items, or, where relevant, to indicate they are part of a serialized verb construction. In verb serialization, often, one of the verbs acts as the main verb for the clause, while the other has undergone some semantic bleaching and, just like the pre-verbal directionals discussed in section 2.3.2, the IGT representation of this bleached verb can vary between annotators. That is, it can be represented as an independent verb (15-16), or as a stem in a compound (14).

(14) Mizo (Chhangte, 1993, p. 143)

keel-in	*pâl*	*a-**sû-chia***
goat-ERG	fence	3S-**butt-bad**

'A/the goat butted the fence and broke it.'

In (14), the verbs *sû* 'butt' and *chia* '(to be) bad' occur in a sequence. The verbs share a subject ('goat') and are both marked with a single participant marker (*a-*'3s'). The first verb, *sû* 'butt', is considered the main verb due to the ergative marking on the NP (*keel* 'goat'). The verbs sequences are represented with a hyphen in (14) but as seen in (15) and (16), similar serial verb constructions can be represented as two unbound lexical items one after another, as in Ranglong and Hakha Lai.

(15) Ranglong (Haokip, 2021, p. 124)

há	*nâai*	*há*	*a-*	***hông***	***ànzìr***	*achú*
PROX.DET	child	PROX.DET	3.S-	**come**	**birth**	DIST.DET

a-	*râmîng*	*Ralngam*	*a-*	*phûa*	*-u*	*-ná*
3.S-	name	Ralngam	3.S-	call	-PL1	-SUB:then

'When that baby was born, they named him Ralngam.'

(16) Hakha Lai (Bedell, Mang, Nawl & Suantak, 2013, p. 2)

Ramvai	*pawl*	*ramlak*	*in*	*an*	***hung***	***tlung***.
hunter	group	jungle	from	3PL	**come**	**arrive**

'The hunting party returned from the jungle.'

In these examples, the main verbs (*ànzìr* 'to birth', *tlung* 'to arrive') share a subject (the child, the hunting group, respectively). Although the first verb 'to come' may have undergone some semantic bleaching, it can function as an independent lexical item in other constructions. As such, the IGT creators chose to represent them as two unbound forms. These examples show two strategies for representing sequences of two verbs in IGT.

2.2.4 Case marking

Another example of differences in IGT syntax is how case makers are represented. They are mostly represented as clitics and set off by the equal sign as in Hyow (17), but may also be written as independent words, as in Hakha Lai (18). The mismatch between representing the case marker as an independent lexeme versus an enclitic implies a difference in analysis, but most likely is a result of differing orthographies influencing the IGT (as discussed and illustrated in Section 2.2.2).

Given this common mismatch between prosodic, orthographic, and morphological wordhood, it would be useful for IGT creators of South Central Tibeto-Burman languages to, by agreed upon convention, represent orthographic or prosodically defined word breaks on one line of IGT and word breaks determined by morphological considerations on a second line. This would greatly speed up the creation of IGT, as the researchers would not need to wait or change the wordhood analysis based on changes in orthographic choice, nor do they need to compromise or hedge on the distributional value of different morphemes based on prosody alone. See (17) and (18) for how this might be done.

(17) Hyow (Zakaria, 2017, p. 528)

krûng ínítɛ̀ʔhyɔ̂ yɔ́ lá díŝng.

krûng	*í-ní-taéʔ-hyɔ̂*	*[yɔ́*	*lá*	*dí=ŝng]*$_{\textbf{OBL}}$
roof	3A-PL-weave.roof.II-PM	[bamboo	and	reed=INST]

'They weave the roof with bamboo and reed.'

(18) Hakha Lai (VanBik & Tluangneh, 2017, p. 148)
khuabawipa khuahlun ah a hung

/khùa-bòoy-pǎa khua-hlŭun ʔaʔ ʔa-hûŋ/
village-chief-male village-old LOC 3SG.S-go up.I
'The chief goes up to the old village.'

These examples of similar reduplicated forms, serial verb constructions, and case marking show us where a better understanding of the possible syntactic and semantic annotation strategies available could guide new annotators to make principled decisions, keeping in mind how readers might evaluate annotation due to implicit or explicit comparison across languages. We turn in more detail to glossing conventions in Section 2.3.

2.2.5 Subordination

South Central Tibeto-Burman languages employ two major strategies to form subordinate clauses: affixes on the verb often derived from postpositions, and post-verbal lexical subordinators. (19) illustrates by contrasting a subordinate clause with a simple clause.

(19) Thadou (Haokip, 2012, p. 2)

Main clause				Subordinate clause			
sâa	*kâ*	*nêe*	*êe*	*sâa*	*kâ*	*nèq*	***lèq***…
meat	1	eat.1	DECL	meat	1	eat.2	**if**
'I eat/eat meat.'				'If I eat meat…'			

As shown in (19), the verb is followed by the lexical subordinator *lèq* 'if' which appears unbound. Similar subordinators may be represented as concatenated to the verb stem as seen for the subordinator *boeloee* in (20) for Khumi. The use of small caps along with the hyphen shows that this *boeloee* is considered a bound morpheme.

(20) Khumi (Peterson, 2010, p. 94)

vaáwy	*vaáwy*	*vaáwy-**boeloee***	*láwyáa*	*kni-khóeleewng*	*náy-hay-noe*
return	return	return-**WHEN**	poor.thing	sky-ELAB	rain-APP-NZ

suy-ple-ngaw=khue=coee…
wet-DIMVCL-ACCID=just=AFFIRM
'He returned and returned and when he returned, the poor thing, it rained on him, he was maybe completely wet…'

2.2.6 Clause constituents

Another useful--though not widely followed--convention is to mark the boundaries of phrasal and clausal constituents within a larger clause using square brackets as seen in (21). The reader will notice the small caps subscripts after the clauses to identify the level of the clause as subordinate (here, 'dependent') or main. It is useful to include this type of notation to support reader recovery of constituent boundaries. Because the architecture of the Tibeto-Burman clause can be complex (Genetti, 1991), there is room for misinterpretation of boundaries.

(21) Hyow (Zakaria, 2017, p. 644)
zúldə̂ krɔ́hìtsə̂, shɔ̀thnɑ́nghnɔ̀ʔtî.
[zúl=də̂ krɔ̂-hí =tsə̂**]$_{DC}$** [shɔ̀t-hnɑ́ng-hnɔ́ʔ-tî**]$_{MC}$**
[even=EMPH fall-COND =TOP] [butcher.I-PH.CAP-DEL.NEG-2SG.NEG]
'If the leaves fall in even number, you will not be able to butcher [the goat].'

(22) Lamkang (Thounaojam & Chelliah, 2007, p. 102)

məwpá	*[thrálikəmə*	*sunú]*	*khət*	*əmdə*
məw-pá	thrá-lin+kV-máŋ	sá-nú	khət	əm-də
elder-male	good-big+NMLZ-not	body-female	one	be-PFV
elder one	wicked	woman	one	was

'The elder brother's wife was wicked.'

It is worth noting that one reason why this is not done is because common software used for creating IGT such as SIL's FLEx do not provide the ability to indicate clause and phrase level constituency. Thus, many published IGTs may skip this level of detail.

2.3 IGT semantics (glossing conventions)

In addition to IGT syntax, an additional layer for representing grammatical and semantic information about a linguistic construct (constructions, clauses, phrases, or lexical and sublexical categories), is glossing. LGR provide two overarching principles for glossing: there should be a 1:1 correspondence between the number of morphemes in the top line and glosses in the middle line, and those glosses may come from a set lexicon, what they refer to as a "grammatical category labels." The vocabulary of glosses can be expanded as needed.

LGR recommendations for glossing include various structural levels above the morpheme, e.g., word, phrase, and clause, and systems such as case alignment systems. For example, LGR abbreviations include case labels such as ablative, dative, ergative or labels associated with alignments systems, i.e., A for 'agent-like argument of canonical transitive verb', S for 'single argument of canonical intransitive verb', and P for 'patient-like argument of canonical transitive verb'. So, one type of labeling for nominals is semantically based and the other syntactic. The semantically-based analysis does not require the analyst to have discovered the alignment system although the A/P/S glossing may, in fact, imply discovery and representation of such a system.

LGR abbreviations also include glossing of clause level functions of morphemes such as subordination (COMP 'complementizer' or CVB 'converb') and clause type (IND 'indicative' or IMP 'imperative'). Here, it is the function that is represented in the gloss, and not the semantics of the morpheme. A subordinator may be adverbial, or indicate purpose or conditionality. The purpose of the LGR category labels is not to differentiate these levels of structure or complex features of morphemes, but, as we will see in section 2.4, for South Central Tibeto-Burman languages, it may be beneficial to indicate both the clausal level function (e.g., subordination) and the specific type of subordination (e.g., time adverbial, manner adverbial).

The LGR grammatical category labels include expected categories used in typological description such as those for indicating gender (M 'masculine'), number (D 'dual'), and person (1 'first person), tense (PST 'past'), aspect (PROG 'progressive'), mood (SBJV 'subjunctive'), valence affecting morphology (ANTIP 'antipassive', CAUS 'causative'), and information structure (DEF 'definite', FOC 'focus', TOP 'topic'). Other labels straddle part of speech and function, (e.g., ADV 'adverbial', DET 'determiner').

We review the LGR abbreviations here to illustrate that LGR grammatical category labels reference various levels of grammar, from the less to more specific. Importantly, the glossing itself reveals analysis and is not meant as scaffolding for ongoing analysis. As we argue in this paper, IGT annotation in language documentation and description is an iterative process. The analyst is more confident of higher structural labeling at an early stage of annotation and must circle back for finer-grained semantic annotation as understanding of the grammar increases. Therefore, it is useful to differentiate glossing of different levels of structure. We build on this concept in Section 2.4 below.

First, let us look specifically at glossing conventions in South Central languages' IGT with respect to the first overarching principle: there should be a unique gloss associated with each morpheme. We do this by considering how the polysemy and semantic change of morphemes challenges the annotator in deciding how to adhere to the LGR principle of 1:1 glossing. For Tibeto-Burman languages in general, Genetti (1991) points to widespread use of case markers for clausal subordination and Saxena (1988) discusses the use of the verb 'say' for a quotative subordinator. As is common for this type of semantic change (Traugott, 1989), the grammaticalized morpheme persists alongside the originating lexical item. These facts present a challenge to the annotator as two morphemes with the same historical origin may have differing functions in the synchronic grammar but retain the same form. For example, should the verb 'say' be glossed as 'say' everywhere the form occurs or where the form functions as a quotative, should it be glossed as such. In this section, we consider how morphological polysemy and grammaticalization are represented in IGT by looking at case marking and directionals. Through this review, we illustrate terminological variation in glossing and the causes of this variation.

2.3.1 Case and information structure

With relatively free ordering of noun phrases, South Central Tibeto-Burman languages rely heavily on case marking, including the locative, genitive, instrumental, and dative for local cases. An example of more than one semantic value to a form is the polysemy between local cases such as dative and locative such as in Daai Chin where the *=üng* clitic encodes both cases. It is common for the dative and locative cases to overlap (LaPolla, 2006), with the dative case marker appearing on locations, recipients, or indirect objects. Knowing this, the IGT creator chose to gloss the *=üng* clitic according to the meaning in its given context. See (23) and (24).

(23) Daai Chin (So-Hartmann, 2009, p. 165)

Msi:-mna	*nakün*	*sun*	*uum**=üng***	*nih*	*thaan.*
seed-grain	all	DEM	container**=DAT**	S.AGR:1DU/PL.INCL	put.in

'We put all grains into containers.'

(24) Daai Chin (So-Hartmann, 2009, p. 166)

La:m	*kdo*	*nu:**=üng***	*ah*	*seh*	*püi.*
road	good	very**=LOC**	S.AGR:3S	take.along	APP:COM

'He took him along a very good road.'

Next, let us consider core arguments where we find differential marking of agents and patients (LaPolla, 2006). Ongoing conversations among many South Central Tibeto-Burman scholars (DeLancey, 2011; Chelliah, 2017) call into question the use of 'ergative' to describe NP with A-like role in these languages because unlike canonical ergative languages, the single argument of a one-place predicate can be marked with 'ergative' and the agent of a two-place predicate can under some circumstances be unmarked. There is a great deal of variation in how the A and P arguments are labelled because of these common patterns. The labels range from 'ergative', 'pragmatic-ergative', and 'agent', to 'focus' or 'foregrounding' for A and 'object', or 'patient' for P. This variation

obscures the fact that there are some bona fide ergative systems such as Mizo (Chhangte, 1993) and others with A and/or P differentially marked (Chelliah & Hyslop, 2011). As well, there is a common grammaticalization of case to information structure so that a morpheme may indicate agent and polysemous morpheme indicate focus or related meaning. Compare for example (25) and (26) in Sizang.

(25) Sizang (Davis, 2017, p. 27)

pă:taŋ	*=pă:*	***=nă:***	*bɔ.lŭŋ*	*=∅*	*kʰât*	*nŭ:mê:i*	*kû:ŋ*	*=a:*	*lɔ̂:n*	*hî:*
boy	=MASC	**=ERG**	all		one	female	place	=at	throw.I	be

Lit. 'boy ball one female place at throw'
'The boy threw a ball to the girl.' (Elicited)

(26) Sizang (Davis, 2017, p. 41)

á	*= ní:*	***-in***	***-nă:***	*tu:a*	*mûn*	*=a:*	*tʰi:*	*kʰɔ̂:m*	*hî:*
3	= two-	**ERG**	**-ERG**	that	place	=at	die.I	together	be

Lit. 'They two that place at die together'
'The two of them died together in that place.'

We can see the tensions between the traditions of the 'ergative' terminology in examples like (26). Here, two morphemes labelled 'ergative' occur in succession, suggesting that these markers encode two related but different meanings and thus, should be glossed differently.

2.3.2 Directionals

Another common occurrence in SCTB is for verbs of motion to be grammaticalized to function as a directional affix. The IGT creator must decide how to differentiate the original verb of motion from the affix in the glossing. See, for example, the Lamkang sentences where *yung* is glossed as 'down' in (27) where it functions as the main verb, but, in (28), the gloss appears in small caps to indicate its function as a directional prefix on the main verb 'roll', rather than an independent lexical item.

(27) Lamkang Naga (Chelliah & Utt, 2017, p. 30)

ar-	***yung***	*-da*
VEN-	**down**	-3PFV

'He came down'

(28) Lamkang Naga (Chelliah & Utt, 2017, p. 37)

m-rthlii	*ar-*	***yung***	*-chaai-da*
3POS-tear	VEN-	**DOWN**	-roll-3PFV

'His tears rolled down.'

Again, we see this pattern in examples (29-30), where the verb *hang* functions as an independent verb stem in (29), but the partially reduced, grammaticalized form *han* appears in (30). As a full verb, *hang* means 'to climb', but when functioning as a directional prefix, we (Chelliah & Utt) chose the gloss 'UP' to most accurately reflect the meaning.

(29) Lamkang Naga (Chelliah & Utt, 2017, p. 30)

hang	*-da*
climb	-3PFV

'He went up'

(30) Lamkang Naga (Chelliah & Utt, 2017, p. 36)
*ar-**han**-loon=nu*
VEN-**UP**-climb=IMP
'climb up towards me!'

In Hakha Lai, some directionals can still function as independent verbs. Other semantically similar forms, (e.g., *va*in 31) are not able to function as main verbs of an independent clause. These are considered 'pre-verbal particles' and are thusly glossed differently.[3]

(31) Hakha Lai (VanBik & Tluangneh, 2017, p. 147)
Aw! a va nuam ee!
/oo! ʔa-**va**-nŭam ʔêe!/
Excl! 3SG.S-**DIR**-be pleasant.I Excl!
'Oh my God! It's so pleasant!'

(32) Hakha Lai (VanBik & Tluangneh, 2017, p. 149)
khuahlun i a run tikah aa hawile nih an don
/khua-hlŭun ʔĭi ʔa-**rŭn** tikʔaʔ ʔa-hôoy-lĕe niʔ ʔân-dŏn/
village-old LOC 3SG.S-**go down.II** when 3SG.POS-friend.PLU ERG 3PL.S-(SG.O)-meet.II
'When he goes down to the old village, his friends welcome him.'

(33) Hakha Lai (VanBik & Tluangneh, 2017, p. 149)
Nan run i daw lai
/nân-**rŭn**-i-dŏolâay/
3PL.S-**DIR**-RFL-love.IFUT
'You will love one another.' (Implication: "after I am gone.")

From these examples, we can see that IGT creators gloss a given form based on its meaning in the target sentence (32-33), but a change in the function or morphemic status of the morpheme may affect the gloss. Other points of interest, as discussed in Lahaussois (2021), would be features of the free translation (literal or idiomatic) and preserving source language word order in phrase and word-level glosses.

2.4 Establishing vocabulary for glossing

There are similar constructions but differing traditions and training of individual analysts lead to differing methods of glossing. In this section, we provide examples of this phenomenon.

2.4.1 Verb stem alternation

Verb stem alternation is a defining characteristic of South-Central Tibeto-Burman languages. This is the process by which the clause type conditions the shape of the verb stem--most commonly, a subordinate clause and main clause will exhibit different verb stems (VanBik, 2009), as demonstrated in (34).

(34) Thadou (Haokip, 2012, p. 2)

Main clause				Subordinate clause			
sâa	*kâ*	***nêe***	*êe*	*sâa*	*kâ*	***nèq***	*lèq…*
meat	1	**eat.1**	DECL	meat	1	**eat.2**	if
'I eat/eat meat.'				'If I eat meat…'			

Differences between verb stems alternates include tone, final consonant, or length of the vowel, (So-Hartmann, 2009; Zakaria, 2017). Typically, two variants are identified and glossed as ‘stem 1’ and ‘stem 2’, though some languages may exhibit more than two stems (Chhangte, 1993; VanBik, 2009; VanBik, 2010).

There do not exists uniform conventions for this shared feature. When glossing verb stem alternation, IGT creators may gloss both verb stems identically (35), or indicate which category the stem belongs to with Roman numerals (36), Arabic numerals (34) and (37), or a small caps letter, as in (38).

(35) Mizo (Chhangte, 1993, p. 84)
*ka-**thûû*** *kâ-**thut**-nâ*
1S-**sit** 1POSS-**sit**-NOM
‘I sit’ ‘my seat’

(36) Hyow (Zakaria, 2017, p. 273)
èy-thôn *ú-**hmú?**-thôn* ***hmú**-á?*
ANAPH.DEM-CONCESS 3A-**see.II**-CONCESS **see.I**-3SG.NEG
‘Even if it was that, even if he saw, he did not see.’

(37) Anal Naga (Ozerov, 2019, p. 44)
tɛàmàŋpá *và-**tɛ́ː**-vàl*
giant.cannibal 3-**eat$_2$**-PERF
‘A giant cannibal has eaten him.’
(Stem 1 tɛà ‘eat’)

(38) Daai Chin (So-Hartmann, 2009, p. 339)
sha-ui: *ta* *hnampo* ***mpyu*** *vaai* *kkhai=a* *sit* *betü=kti*
fox FOC banana **steal.B** DIR:go FUT=CF go ASP=NON.FUT
‘The fox went to steal again bananas.’

Note in example (39) it is indicated if the verb stem alternates 1 or 2. This is consistently done when verb stem alternation is under discussion. However, the same IGT creator may omit the verb stem category from the gloss line, as in (39), when that is not the focus of the discussion.

(39) Anal Naga (Ozerov, 2019, p. 47)
***tɛ́**-máŋ-nì*
eat-DUB-3
‘He probably ate.’

While (37) was intended specifically to illustrate verb stem alternation, (39) is featured in a discussion of person indexation and TAM. Other languages have preserved verb stem alternation only in a subset of verbs. In those cases, IGT creators indicate the category of verb stems only for verbs that exhibit verb stem alternation, and other verbs are left unmarked. While the correspondence of Arabic numerals to Roman numbers is explicit, it is not necessarily clear whether stem A is analogous to stem 1/I. Additionally, in cases where stems are not marked, it will be difficult for future readers to directly compare these verb stems.

To be sure, further discussion of how to represent verb stem alternation and consensus among the community of South-Central Tibeto-Burman researchers would be a useful process to make

linguistic descriptions more comparable. However, such a discussion would not be just of terminology, but also of analysis and representation. For example, with respect to Stem 1 and 2, the question arises if Stem 2 is derived from Stem 1, or if Stem 2 occurs with subordinate clauses because it is historically derived from a nominalization, and so on. Enroute to creating shared conventions for representing stem variants in IGT, one would also be committing to grammatical analyses.

2.4.2 Subordination

We have discussed challenges in annotation for subordinators with respect to IGT syntax in section 2.2.4 where we noted that similar types of morphemes are noted as bound or free. We note here the variable glossing of subordinators based on function or meaning. By function, we mean use of terms like 'subordinator', 'nominalizer', or 'relativizer'. By meaning, we mean either generalized grammatical description such as 'conditional' as in the Hyow example (40), or specific semantic value such as 'when', 'upon.doing.X', or 'even.though.X' as in example (41). While the forms are similar across languages, the glossing conventions across IGTs vary across these parameters.

(40) Hyow (Zakaria, 2017, p. 644)
zúldô krɔ́hìtsɛ̂, shɔ̀thn~~ɨ~~nghnɔ̀ʔtî.
[zúl =dô krɔ̂-**hí** = tsɛ̂]$_{DC}$ [shɔ̀t-hn~~ɨ~~ng-hnɔ́ʔ-tî]$_{MC}$
[even=EMPH fall-COND=TOP] [butcher.I-PH.CAP-DEL.NEG-2SG.NEG]
'If the leaves fall in even numbers, you will not be able to butcher [the goat].'

(41) Hyow (Zakaria, 2017, p. 658)
kárbárí khɛ̂w khìnàʔthɔ̀nàtsɛ̂ ɔ́pɔ̂y.
[kárbárí khɛ̂w khín-áʔ-**thɔ́ná** =tsɛ̂]$_{DC}$ [ɔ́-pɔ̂y]$_{MC}$
[village.chief word listen.I-3SG.NEG-SC.**CONCESS**=TOP] [3S-be.good]
'Even though he did not listen to the village chief's words, it was good.'

The glossing may indicate function, as in this Anal Naga example (42).[4]

(42) Anal Naga (Ozerov, 2019, p. 48)
háŋ-kàl-(Ø)-**so**
UP.TEMP-climb-(3)-3.IRR.**SUB**
'It will climb up and…'

The glossing may indicate both the function and meaning. In the Daai example, we have SUBO 'subordinator' followed by a colon and then the semantics of that subordinator.

(43) Daai Chin (So-Hartmann, 2009, p. 334)
*Nah jah mtheh hü=a athoon**=üng***
S.AGR:2S IO.AGR:1/3DU/PL tell DIR:around=CF happen**=SUBO:if**
kah ni:ng man-ei yai ni.
S.AGR:1S O.AGR:2S catch-AO SUBJ EMPH
'If it happens that you tell them, I would catch you.'

We return in Section 3 to further discussion of this hierarchical glossing going from more abstract category to specific meaning. IGT creation is an iterative process because the annotator's understanding of the language grows with each annotation. The best way to understand this is to

imagine the annotator tackling glossing with no knowledge of the specifics of the grammar of the text being annotated. Rather, the annotator has a general knowledge of how grammar works in related languages. Using that information, the annotator starts labeling the most obvious features. Oftentimes, many iterations of annotation are needed to arrive at the details of morpheme semantics. Until then, it is likely that there will be several morphemes that fall under similar labeling. For example, it may be clear that a morpheme is acting as subordinator, but the specific function or meaning of that subordinator may not become clear until the end of the annotation process for this text or even the end of annotating a number of texts. If this is the case, then breaking the gloss for a subordinator up into two parts as seen in (43) would allow the annotator to represent what is known (that this morpheme is a subordinator) and use the colon convention to show that the specifics are unknown.

The type of hierarchical annotation convention would allow annotators to represent their analysis as it develops. As it happens, annotators will want to indicate what they know of a morpheme, but reserve the opportunity to define a morpheme at a later stage. In the following Thadou example for instance, the annotator recognizes three distinct morphemes as indicating various shades of negation. As illustrated in Haokip (2012), these three forms (*poo*, *hiq*, *low)* indicate nuanced differences in meanings based on the expectation of the speaker. Glossing all three as NEG does capture their meaning accurately, but, as is often the case with ongoing analysis and slowly growing understanding of semantics, it is not always possible at the time of IGT creation to accurately gloss meanings so as to reflect those nuanced differences. The IGT does not reflect this awareness that the annotator intends to return to these morphemes to flesh out the semantic differences. As we suggest below, there could be a useful mechanism to do so in the future. The hierarchical annotation method would allow the annotator to do this, but still follow LGR principles by adding another layer of glossing, i.e., "NEG:not.expected", "NEG:expected", and "NEG:unknown". Notice that the wording used on the right side of the colon is non-technical and allows the annotator to accurately capture knowledge at a moment in time. While the left hand side of the colon will likely remain constant, the right hand side may change as analysis improves.

(44) Thadou (Haokip, 2012, p. 5)
gòo â zùu ***pòo*** *êe*
rain 3 fall.1 **NEG** DECL
'It is not raining' (Not expected)

(45) Thadou (Haokip, 2012, p. 5)
gòo â zùu ***hiq*** *êe*
rain 3 fall.1 **NEG** DECL
'It is not raining' (Expected)
(It was supposed to rain but did not rain)

(46) Thadou (Haokip, 2012, p. 5)
gòo ø zùu ***lòw*** *dîŋ â-hîi*
rain 3 fall.1 **NEG** FUT 3be
'It is not going to rain'

2.4.3 Information structure and discourse markers

Similarly, the semantics of information packaging morphology can be notoriously difficult to pin down. In Tibeto-Burman languages, many information packaging morphemes have multiple

uses (Chelliah, 2009; Ozerov, 2020). The glossing for such morphemes varies across individual annotators, depending on the use or convention followed by individuals. In addition, the same oppositions are not uniformly noted across languages. For example, what is glossed as foregrounded in Khumi (Peterson, 2011) might be glossed as agent, ergative, focus, or topic in another language. In other words, as a community of analysts, we are still working out the semantics of these morphemes, and this is reflected in the variation in the vocabulary used in IGT annotations. Comparing information structure across these languages based on IGT can be misleading.

Another area where there is general agreement on the scope and function of a morpheme but difficulty of pinning down the semantics of the morpheme is with clause final affective markers. These are often glossed as 'non-final' to indicate that another clause is following. See (47-49).

(47) Daai Chin (So-Hartmann, 2009, p. 52)

Aai	*kthi=e*	*sun*	*jah*	*ng'yet-ei*	*püi=u*	***lü…***
chicken	dead=PL	DEM	IO.AGR:1/3DU/PL	share	APPL:COM=PL	NF

'They shared the dead chicken among each other and...'

(48) Sizang Chin (Davis, 2017, p. 15)

t^{h}âu	*tăm*	*mámâ:*	*kă:p*	***a:***
gun	be.many	INTENSE	shoot.I	**NF**

'[They] shot a lot of guns and…'

(49) Thadou (Haokip, 2021, p. 5)

â-tûi	*â-dòon-sàh-û*	*lêh*	*â-mûh*	*jôl*	*chŏm-hîh-**în***
3-water	3-drink$_2$-CAUS-PL3	SUB:and	3-lip	oily	change-NEG:proh-**NFP**

'They made her drink her water, but her lips were not oily.'

More information about the discourse-pragmatic functions of these forms would be needed in order to understand their function or to compare across languages. Furthermore, there may be more than one per language. Hierarchical glossing would thus help the annotator confidently label these morphemes by their function but leave the specific semantics open until determined, e.g., NF:X.

2.5 Summary

In this section, we reviewed practices of expert IGT creators for South Central Tibeto-Burman languages. We considered these practices in light of 2 LGR principles:

- 1:1 correspondence between number of constituents and glosses
- Use of an established vocabulary for annotations

We illustrated that there are common features in South Central Tibeto-Burman languages where it becomes difficult to apply these principles in an obvious fashion because of polysemy, as in the case of case marking and directionals. We also noted that, for those hard to define morphemes, it is a practice to provide a gloss that is closest to the meaning that might be accurate. This practice could obscure true correspondences between languages and what is known about the language (such as the more abstract category) and what is yet to be discovered (semantics). In the next section, we argue for the importance of moving forward on developing IGT conventions for improved representations of these languages.

3. Discussion

At the 53rd meeting of the International Conference on Sino-Tibetan Languages and Linguistics we heard from IGT creators.[5,6] There were three main strands of discussion regarding the creation of IGT standards. First, that it is neither possible nor reasonable to harmonize IGT across languages. Each language works slightly differently, and insisting on similar glossing would artificially restrict the most appropriate representation for a language as deemed by the analyst.

Second, it was noted that harmonizing IGT could happen post-analysis by mapping individual annotations to some universal annotation vocabulary; therefore, there was no need to undertake the exercise of community discussion and agreement on IGT annotation beforehand. But given our discussion above, we can see that misalignment is not just a question of morpheme labelling. It is also a question of syntax or representation, e.g., spaces between forms or use of hyphens. Farrar and Lewis (2007) addressed this issue of shifting terminology in linguistics, noting: "If it were only a question of terminology, then many-to-many, or even simple term mappings could be constructed...But linguists not only employ different surface terminologies, they actually conceptualize the discipline in divergent, and often, incompatible ways" (p. 53). Farrar and Lewis speak to both the ambiguity created by changes in terminology and also the set of issues that arise when two linguists attribute the same term to vastly different phenomena. Finally, it was noted that IGT in itself is a representation of language structure that deals in morphemes rather than larger constituents or constructions and, therefore, IGT glossing of morphemes may not be the right focus for representation of connected text in the future.

Indeed, if IGT is meant simply as representation for the reader of linguistic examples in a typological paper, then, as stated by LGR, "depending on the author's purposes and the readers' assumed background knowledge, different degrees of detail will be chosen" (Max Planck Institute for Evolutionary Anthropology Department of Linguistics, 2015, p. 1). More recently, Haspelmath (2016) writes that interlinear glosses are "not abbreviations of deep analyses but reading aids to the reader" (p. 301). Therefore, he too intimates that IGT is representation and that representation should fit the needs of the audience. For example, he says that if the goal is typological comparison, then more general labels may be used, and if the goal is language description, then another set could be used. See also Lehmann (2004, p. 1837) on this point. In Section 2.4.1, we saw two examples of IGT by the same linguist representing verb stem alternation differently according to the feature in focus in the target language. IGT creators may emphasize certain aspects of an example, for instance, by bolding or underlining morphemes (9), adding brackets around constituents (17), or including explanatory notes (37) on verb stem alternation. In a reference grammar, however, IGT tends to remain constant and maximally represented throughout the description. So, when IGT includes constituent analysis or indicates morpheme ordering then this does seem to be an 'abbreviation of a deep analysis'. For example, in (50), the ordering of morphemes is shown through subscript numbers which correspond to the categories of elements in the verb phrase.

(50) Hyow (Zakaria, 2017, p. 315)
nàngá ínítʰɔ́wpèkálɛ̀ʔyhnɨ́ngùngánú tîng.
náng=á kí-ní-hɔ́w-êy$_{1}$-pék$_{2}$-âl$_{3}$-ɛ́ʔy$_{4}$-hnɨ̂ng$_{5}$-ùngâ$_{6}$=nú$_{7}$
2SG=DAT 1A-INV-say.I-MID$_{1}$-BEN$_{2}$-DEP$_{3}$-IRR$_{4}$-PH.CAP$_{5}$-1PL.EXC.NEG$_{6}$=SS.EVID$_{7}$
'After that, he said, "So, I will not be able to ask (that) for you."'

We must also recognize that in the world of language description and documentation, IGT creation is an important tool for language discovery. It is a process with several stages and resulting

in several products. The products are well known: word lists and dictionaries based on texts, discourse, linguistic and cultural examples for scientific discovery (Epps, Webster, & Woodbury, 2017), and language and culture revitalization (Hinton, Russ & Roche, 2018). What is less discussed is that the process of creating IGT is central to supporting grammatical analysis. While some work on morpheme semantics is arrived at via elicitation of targeted structures, it is generally agreed that connected utterances used in human-to-human interactions yield the language samples with none of the typical errors introduced through the elicitation process (Schütze, 2008; Chelliah, 2016). As Dixon (2007, p. 22), puts it, "texts are the lifeblood of linguistic fieldwork...The only way to understand the grammatical structure of a language is to analyze recorded texts in that language (not by asking how to translate sentences from the lingua franca)." In addition, a collection of analyzed texts can "show the language as it really is, and among other things provides a corpus against which the grammar's claims can be tested, and which subsequent linguists may scrutinize for generalization overlooked by the original grammarian" (Evans & Dench, 2006, p. 10).

Since the analysis of connected speech is central to linguistic discovery, we look for methodologies for creating a corpora of analysed connected text. Currently, documentary and descriptive methodologies include this workflow: record natural interactions, transcribe in a practical orthography or phonetically, use software to analyze, keep record of and disseminate lexical and grammatical discoveries, publish on individual grammatical topics or publish a linguistic grammar using textual examples. As discussed above, the analyst starts with connected text typically transcribed from source audio or video recordings. The linguist must then perform several analytic tasks on this transcribed material. The most basic translations, whether by construction, intonational phrases, or words can be reconstructed along with native speaker input. Further linguistic analysis including morpheme semantics and determining phrase and clause boundaries happen in a gradual fashion as described in Section 2.2.6. This is because text analysis is interwoven with additional translation of the texts and supportive elicitation tasks. For example, suppose I am creating IGT for a text on how to build a house. I may translate many of the constructions, but have trouble with the exact meaning for the modals (e.g., 'you *should* wet the bamboo' or 'you *might* use a substitute'). In this case, I could create an elicitation schedule on modals to clarify some doubts, and this might include translation from a contact language (e.g., English or other lingua franca) or elicitation with prompts (Michael, 2015; Burton & Matthewson, 2015; Chelliah, 2001). Another method would be to examine another text with modals and use comparison of grammar and pragmatics to arrive at modal semantics. This is what we mean by analysis being interwoven with text translation and analysis.

What aids do we provide novice annotators to undertake this workflow? Transformational to this process has been IGT creation software such as SIL's FLEx.[7] What we are suggesting is the need for further scaffolding for novice annotators. This scaffolding does not need to be strict guidelines or efforts at standardization. Rather, this scaffolding can be in the form of overt discussion of annotation practices and experiences both in the published IGT as well as through discussion via asynchronous discussion boards for those working on related languages. Such discussion by seasoned annotators would give novice annotators a starting point and a place to vet glossing decisions. It is with a view to creating this type of support that we posit the following preferred practices.

- Read existing descriptions of related languages.
- Create a list comparing annotation conventions by individual authors. The areas that show the most variation in annotation will probably also be most difficult for the annotator - consider using hierarchical glossing for these.
- Decide on a baseline representation: Early decisions on the transcription that will be used in IGT annotation will save time in having to make revisions later. This is especially true

for word breaks See Bedell (2001) and references in Chelliah & Garton (in prep) for discussion on orthographies for South Central Tibeto-Burman languages.

- Use prosodic cues to divide connected texts into phrases and clauses. Look for repeated morphology at prosodic edges which could signal clausal subordinators or discourse markers to support decisions on constituent edges (Woodbury, 1985).
- Create IGT with differing levels of granularity in analysis.
    - A connected text with clause level division and clause translations can be useful for creating bilingual story books.
    - A connected text with word and clause level translation can be used to populate a dictionary or word list.
    - A connected text with word, clause and some morpheme level glossing will be useful for further grammatical investigation.
- Be consistent. As the annotator's understanding grows, annotations will change. See, for example, Peterson (2011)'s discussion of the *=mö*[8] clitic in Khumi (p. 74). It is very difficult to keep track of necessary changes over the course annotation which could take months to years. Therefore, programs like FLEx, which keep a record of glosses and allow for global and local edits and are useful for maintaining consistency, even as analyses develop.
- Create a hierarchical annotation strategy: The strategy seen in (43) is extremely useful in that it encodes the meaning of the morpheme ('if') in addition to the grammatical function ('SUBO'). Use hierarchical annotation when:
    - The function is clear, but the semantics are not; for instance, the annotator can tell that a morpheme is functioning as a clausal subordinator, but the semantics are unclear.
    - There are forms related through semantic change: For example, conjunctions are frequently grammaticalized to indicate subordination. In this case, it is useful to indicate how a particular form is being used. Thus glossing might look as follows SUB:and or CONJ:and. See also examples from Daai Chin where the same form may act as a lexical subordinator (43) or a case marker (23-24).
    - Different morphemes have very similar semantics: For example, in Thadou, we see three morphemes which could be glossed 'negative'. To distinguish these forms, we avoid naming all three identically as 'negative'. Rather, we provide NEG as the overarching semantics and a variable after the colon, i.e., NEG:X or NEG:y. See Haokip (2021, pp. 122-123) for examples. The variables are filled if and when the semantics become clearer.
- Be transparent**:** The IGT creator may also write a guide to the IGT conventions to provide explanations of the IGT syntax and glossing. For example, the guide may:
    - Explain how clause breaks are determined, or how hyphens are used for reduplication.
    - Provide guidance on how to cite the IGT by explaining how lines are numbered, how the texts correspond to any available audio recordings, and where such audio recordings are available.
    - Explain translation practices followed, or any information about the orthography that is not obvious from the text.
    - Provide a list with definitions of each functional and semantic category label. Typically, abbreviation lists do not include definitions. Often the use of Latinate terms like 'nominative' only partially describes the use of a morpheme.

In sum, for language description to reach its potential, we suggest that the IGT annotation process more readily support analysis. For this process to work effectively, IGT creators can

Comment [u2]: Can you check the endnote 8? Does that sync here?

follow the general principles of LGR. Based on the variation seen for syntax and semantics for South Central languages, we feel discussion of language-family specific IGT practices would be useful especially to new annotators.[9] Such discussion could take place via wikis, in overt explanation of IGT conventions, and in annotators following glossing conventions that recognize polysemy, grammaticalization, homophony, morpheme senses, and the incremental gains an analyst makes in understanding grammar. In particular, we champion hierarchical glossing and annotation guides.

4. Conclusion

This paper has examined how certain features (e.g., reduplication, verb stem alternation, directionals) are represented in IGT from several South-Central Tibeto-Burman languages. We offer some remarks on the recognized purposes of IGT: (1) connected texts provide source data for novel grammatical constructions; (2) language samples form connected natural texts are not influenced by the elicitation method; (3) connected annotated texts are useful for checking hypotheses; (4) annotating connected texts helps add to lexical and cultural information for various purposes, including language revitalization. Further, we foreground three often overlooked functions of IGT: (1) IGT creation is an analytic process that supports the annotators growing understanding of a language by providing a record of previous and possible analyses; (2) comparison of existing IGT from related languages can provide scaffolding for a new annotator; (3) IGT creation encourages the analysis of systems rather than individual constructions, as the rules and predictions for one text must hold throughout the corpus. Therefore, IGT creation is an analytic method, not just a representation of existing analysis.

In the Indian context, IGT creation will encourage linguists-in-training to reflect on aspects of grammar which they might otherwise have no occasion to analyze. For example, if I decide to investigate question formation in a language, I might collect translations of various types of questions and limit my analysis to those constructions. This is a common descriptive method for Indian linguists, and has been helpful in setting a baseline description for many languages. However, this method has also potentially kept amazing grammatical information locked in unexplored language samples of connected text. Past technological restrictions no longer exist. Free software for data recording, transcription, and analysis is available, and the training to use this software is readily available (Computational Resource for South Asian Languages, 2021; Endangered Language Documentation Programme, 2021). Coupled with language-family specific discussions of IGT syntax and semantics, IGT can greatly improve the final products of language description.

NOTES

[1] All IGT examples retain their original formatting and glosses, with emphasis (bolding) added to the segment relevant to the present discussion. Abbreviations and glosses used in IGT examples are reproduced in Appendix A. An '*' indicates an inferred gloss for an abbreviation which was not stated in the source.

[2] This branch has formerly been known as Kuki-Chin, but this term is considered unacceptable by many speakers of these languages. Therefore, we follow the newer naming practice with "South-Central."

[3] See VanBik and Tluangneh (2017) for a full discussion of this topic.

[4] In (41), the subordinate clauses are demarcated with brackets and labeled with subscript DC to indicate 'dependent clause.'

[5] (https://icstll.ci.unt.edu/)

[6] (https://www.youtube.com/watch?v=jDwYl75FRUE)

[7] To support text collection and annotation, there exist transcription and data management tools such as SayMore: https://software.sil.org/saymore/ and ELAN :https://archive.mpi.nl/tla/elan. See also FLEx:

FieldWorks Language Explorer https://software.sil.org/fieldworks/. Earlier versions of FLEx, such as Toolbox: https://software.sil.org/toolbox/ are also in use. Finally, expex permits multiple interlinear glossing lines. https://www.ctan.org/pkg/expex.

[8] See DeLancey (2011) for a thorough discussion of the history of scholarship on optional ergativity in Tibeto-Burman languages.

[9] The first International Conference on Linguistic Terminology, Glossing and Phonemicization (LiTGaP 2020), hosted by the Japanese Linguistics Internationalization Committee at Yonezawa City, Yamagata, Japan, February 22–24, 2020, is headed in the right direction. LiTGaP focused on the history of linguistic terminology and interlinear glossing with a focus on Japanese grammar description.

ABBREVIATIONS AND GLOSSES USED IN EXAMPLES

Anal Naga (Ozerov, 2019): 3 'third person'; DUB 'dubitative'; IMP 'imperative'; INTR 'intransitive'; IRR 'irrealis'; NEG 'negation'; PERF* 'perfective'; PL 'plural'; RDP* 'reduplicant'; SUB 'subordinated'; TEMP 'permanent/temporary movement'; UP 'UP directional'.

Daai Chin (So-Hartmann, 2009): 1DU '1st person dual'; 1s '1st person singular'; 2s '2nd person singular'; 3DU '3rd person dual'; 3s '3rd person singular'; AO 'agent orientating'; APP* 'applicative'; APPL 'applicative'; ASP 'aspect'; CF 'constituent final'; COM 'comitative'; DAT 'dative'; DEM 'demonstrative'; DIR 'directional'; EMPH 'emphasis'; FOC 'focus'; FUT 'future'; INCL 'inclusive'; INTENSF 'intensifier'; IO.AGR* 'indirect object agreement'; LOC 'locative'; NF 'non-final'; NON.FUT 'non-future'; O.AGR 'object agreement'; PL 'plural'; S.AGR 'subject agreement'; SUBJ 'subjunctive'; SUBO:if 'subordinator'.

Hakha Lai (VanBik & Tluangneh, 2017): DIR 'directional'; ERG 'ergative'; Excl!*'Exclamatory'; FUT 'Future'; LOC 'Locative'; O 'Object'; PLU* 'plural'; PL 'plural'; POS 'possessive'; RFL 'reflexive'; S 'subject'; SG 'Singular'; 3 'third person'; PL 'plural suffix or particle'; 3s* 'Third person singular'; S* 'subject'; IDEO* 'ideophone'.

Hakha Lai (VanBik, 2010): 1 'first person'; 2 'second person'; 3 'third person'; DEM 'demonstrative'; FUT 'Future Marker'; PRON 'pronoun'; S 'subject'; sg 'singular'.

Hakha Lai (Bedell, Mang, Nawl & Suantak, 2013): 3 'third person'; PL 'plural'.

Hakha Lai (Patent, 1998): 3s* 'third person singular'; S* 'subject'; IDEO* 'ideophone'.

Hyow (Zakaria, 2017): 1 'First person'; 2 'Second person'; 3 'Third person'; A* 'Agent'; ANAPH 'Anaphoric'; ANT* 'Anterior'; BEN* 'Benefactive'; CAP* 'Capability'; CONCESS 'Concessive'; COND* 'Conditional'; CONJ* 'Conjunction'; DAT* 'Dative'; DEL 'Delayed'; DEM 'Demonstrative'; DEP* 'Dependent'; DIR 'Directional prefix'; EMPH 'Emphatic suffix'; EVID 'Evidential'; EXC* 'Exclusive'; FOC 'Focus clitic'; INST 'Instrumental case'; INV* 'Inverse'; IRR* 'Irrealis'; LOC 'Locative case'; MID 'Middle'; NEG 'Negative suffix'; PH 'Physical'; PL 'Plural number'; PM 'Predicate Marker'; POSS* 'Possessive'; S* 'Subject'; SC* 'Simple clause'; SG 'Singular number'; SIM 'Simultaneous'; SS 'Sensory evidential'; TEMP* 'Temporal'; TOP 'Topic clitic'.

Khumi (Peterson, 2008): 2 'second person'; 3 'third person'; ALL 'allative'; AUGVCL 'augmentative verbal classifier'; CAUS 'causative'; FOC 'focus'; IMPERF 'imperfect'; INTENS 'intensifier'; TOP 'topic'.

Khumi: (Peterson, 2010): ACCID 'accidental'; AFFIRM 'affirmative'; APP 'applicative'; DIMVCL 'diminutive verbal classifier'; ELAB 'elaboration (in an elaborate expression)'; NZ 'nominalizer';

Lamkang Naga (Thounaojam & Chelliah, 2007): NMLZ 'nominaliser'; PFV 'perfective'.

Lamkang Naga (Chelliah & Utt, 2017): 3 'third person'; 3 'third person'; FUT 'future'; VEN 'venitive'; 3PFV '3rd person subject perfective aspect'; 3POS '3rd possessive'; DOWN 'downward movement'; UP 'upward movement'; IMP 'imperative'.

Lamkang Naga (Chelliah et al. 2021): 2 'second person'; A* 'agent'; AGT 'agent'; IDEO 'ideophone'.

Mizo (Chhangte, 1993): 1 'first person'; 3 'third person'; Adv 'adverb'; DEM 'demonstrative'; ERG 'ergative'; FP 'final particle'; FUT 'future'; INT 'intensifier'; LOC 'locative'; NOM 'nominalizer'; pl 'plural'; Poss 'possessive'; PRPF 'present perfect'; s 'subject'.

Sizang (Davis, 2017): 3 'third person'; ERG 'ergative'; INTENSE 'intensifier'; IRR 'irrealis'; MASC 'masculine'; NF 'non-final coordinating marker'; TEMP 'temporal'.

Thadou (Haokip, 2012): 1 'first person'; 3 'third person'; DECL 'declarative'; FUT 'future'; NEG 'negative'.

Thadou (Haokip, 2021): 3 'third person participant marker'; CAUS 'causative'; NEG:proh 'negative: prohibitive'; NFP 'non-final particle'; PL3 'plural 3'; SUB:and 'subordinator: and'.

Ranglong (Haokip, 2021): 3.s 'third person singular participant marker'; DIST.DET 'distal determiner'; PL1 'plural for verbs'; PROX.DET 'proximate determiner'; SUB:then 'subordinator: then'.

REFERENCES

Bedell, G., Mang, K.S., Nawl, R.S., & Suantak, K. (2013). *The morphosyntax of verb stem alternation*. Paper presented at 46th International Conference on Sino-Tibetan Languages and Linguistics, Hanover, New Hampshire, USA.

Burton, S., & Matthewson, L. (2015). Targeted construction storyboards in semantic fieldwork. In R. Bochnak & L. Matthewson (Eds.), *Methodologies in semantic fieldwork* (pp. 135-156). Oxford University Press.

Chelliah, S. (2001). The role of text collection and elicitation in linguistic fieldwork. In P. Newman & M. Ratliff (Eds.), *Linguistic fieldwork* (pp. 152-165). Cambridge University Press.

Chelliah, S. (2009). Semantic role to new information in Meithei. In J. Barðdal & S. Chelliah (Eds.), *The role of semantic, pragmatic, and discourse factors in the development of case* (pp. 337-400). John Benjamins.

Chelliah, S., & Hyslop, G. (2011). Optional case marking in Tibeto-Burman. Special Edition of *Linguistics of the Tibeto-Burman Area, 34*(2), 1-7.

Chelliah, S., & Utt, T., P. (2017). The syntax and semantics of spatial reference in Lamkang verb. In C. Genetti & K. Hildebrandt (Eds.), Special Issue on the Grammatical Encoding of Space, *Himalayan Linguistics*, *16*(1), 28-40.

Chelliah, S., Blair, E., Robinson, M., Khullar, R.R., & Khular, S. (2021). Reduplication in Lamkang: Form, function, feeling. In J. Williams (Ed.), *Expressive morphology in the languages of South Asia* (pp. 165-186). Routledge.

Chelliah, S., & Garton, R. (2021). *Orthography development for languages of the South-Central branch of Tibeto-Burman: Lessons from Lamkang* (Manuscript in preparation).

Chhangte, L. (1993). *Mizo syntax* (Unpublished dissertation). University of Oregon.

Davis, H., Gillion, C., & Matthewson, L. (2014). How to investigate linguistic diversity: Lessons from the Pacific Northwest. *Language, 90*(4), e180—e226. doi: 10.1353/lan.2014.0076.

Davis, T. D. (2017). *Verb stem alternation in Sizang Ching narrative discourse* (Unpublished Master's thesis). Payap University.

DeLancey, S. (2011). "Optional" ergativity in Tibeto-Burman languages. *Linguistics of the Tibeto-Burman Area, 34*(2), 9-20.

Dixon, R. M. W. (2007). Field linguistics: A minor manual. *Sprachtypol. Univ. Forsch. (STUF), 60*(1), 12-31.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology.

Epps, P., Webster, A., & Woodbury, A. (2017). A humanities of speaking: Franz Boas and the continuing centrality of texts. *International Journal of American Linguistics, 83*(1), 41-78.

Evans, N., & Dench, A. (2006). Introduction: Catching language. In F. K. Ameka, A. Dench & N. Evans (Eds.), *Catching language: The standing challenge of grammar writing* (pp. 1-39). Mouton de Gruyter.

Farrar, S, & Lewis, W.D. (2007). The GOLD community of practice: An infrastructure for linguistic data on the web. *Language Resources and Evaluation*, *41*(1), 45-60.

Genetti, C. (1991). From postposition to subordinator in Newari. In E. C. Traugott & B. Heine (Eds.), *Approaches to grammaticalization 2*, (pp. 227-255). John Benjamins.

Haokip, P. (2012). Negation in Thadou. *Himalayan Linguistics, 11*(2), 1-20.

Haspelmath, M. (2016). The challenge of making language description and comparison mutually beneficial. *Linguistic Typology, 20*(2), 299-303.

Hinton, L., Huss, L., & Roche, G. (Eds.). (2018). *The Routledge handbook of language revitalization.* Routledge.

Ozerov, P. (2019). Person indexation in Anal verbal paradigms. *Himalayan Linguistics, 18*(1), 26-53.

Ozerov, P. (2020). Information structure and intonational accent in Burmese. *Linguistics of the Tibeto-Burman Area, 43*(2), 191-224.

Patent, J. (1998). A willy-nilly look at Lai ideophones. *Linguistics of the Tibeto-Burman Area, 21*(1), 155-206.

Peterson, D. (2008). Bangladesh Khumi verbal classifiers and Kuki-Chin 'chiming'. *Linguistics of the Tibeto-Burman Area, 31*(1), 109-138.

Peterson, D. (2010). Khumi elaborate expressions. *Himalayan Linguistics, 19*(1), 81-100.

Peterson, D. (2011). Core participant marking in Khumi. *Linguistics of the Tibeto-Burman Area, 34*(2), 73-100.

Rice, K. (2001). Learning as one goes. In P. Newman & M. Ratliff (Eds.), *Linguistic fieldwork* (pp. 230-249). Cambridge University Press. doi: 10.1017/CBO9780511810206.012.

Saxena, A. (1988). On syntactic convergence: The case of the verb 'say' in Tibeto-Burman. *Proceedings of the 14th annual meeting of the Berkeley linguistics society*, pp. 375-388.

So-Hartmann, H. (2009). *A descriptive grammar of Daai Chin.* University of California Press.

Traugott, E. (1989). On the rise of epistemic meaning in English: An example of subjectification in semantic change. *Language, 65*, 31-55.

VanBik, K. (2009). *Proto-Kuki-Chin: A reconstructed ancestor of the Kuki-Chin languages.* University of California Press.

VanBik, K. (2010). The syntax of psycho-collocation in Hakha Lai. *Linguistics of the Tibeto-Burman Area, 33*(2), 137-150.

VanBik, K., & Tluangneh, T. (2017). Directional pre-verbal particles in Hakha Lai. *Himalayan Linguistics, 16*(1), 141-150.

Woodbury, A. (1985). The functions of rhetorical structure: A study of Central Alaskan Yupik Eskimo Discourse. *Language in Society, 14*(2), 153-190.

Zakaria, M. (2017). *A grammar of Hyow* (Unpublished dissertation). Nanyang Technological University.

INTERNET SOURCES

Bedell, G. (2001). *'Word Combination' in Lai* (Unpublished manuscript). Retrieved 6/22/2021 from http://sealang.net/sala/archives/pdf8/bedell2001word.pdf.

Chelliah, S. (2020). Standards for interlinear-glossed texts in related languages [Conference session]. *53rd International conference on Sino-Tibetan languages and linguistics*, Denton, Texas, United States. https://drive.google.com/file/d/11W9xqgvSSspDE37LKjOmrxrRsovtO9re/view?usp=sharing

Chelliah, S. (2017). Ergativity in Tibeto-Burman. In J. Coon, D. Massam, & L. D. Travis (Eds.), *The Oxford handbook of ergativity*. Oxford University Press. doi: 10.1093/oxfordhb/9780198739371.013.38

Chelliah, S. (2016). Language documentation improved through rhetorical structure analysis. In M. W. Post, S. Morey & S. Delancey (Eds.), *Language and culture in Northeast India and beyond: In honor of Robbins Burling* (vol. A-PL 023, pp. 293-330). Asia-Pacific Linguistics. http://hdl.handle.net/1885/38458.

Computational Resource for South Asian Languages. (2021). Collaborative language archiving curriculum. https://corsal.unt.edu/curriculum.

Endangered Languages Documentation Programme. (2021). About ELDP trainings. https://www.eldp.net/en/our+trainings/about/.

Haokip, P. (2021). *Annotated texts of languages of the Barak Valley: Thadou, Saihriem, Hrangkhol, Ranglong*. Aquiline Books. doi: https://doi.org/10.12794/sps.corsal-060-1

Konnerth, L., & DeLancey, S. (2019). Verb agreement in languages of the Eastern Himalayan region. *Himalayan Linguistics, 18*(1). doi: http://dx.doi.org/10.5070/H918144455.

Lahaussois, A. (2021). Glossing in the linguistic survey of India. Some insights into early 20th century glossing practices. *Historiographia Linguistica.* doi: https://doi.org/10.1075/hl.00081.lah.

LaPolla, R. J. (2006). Overview of Sino-Tibetan morphosyntax. In R.J. LaPolla & G. Thurgood (Eds.), *The Sino-Tibetan languages*. Routledge. doi: https://doi.org/10.4324/9780203221051.

Lehmann, C. (2004). Interlinear morphemic glossing. In G. Booij, C. Lehmann, J. Mugdan & S. Skopeteas (Eds.), *Morphologie. Ein internationals Handbuch zur Flexion und Wortbildung.* (Handbücher Der Sprach- Und Kommunikationswissenschaft 17/2), (pp.1834-1857). W. de Gruyter. doi: https://doi.org/10.1515/9783110172782.2.20.1834

Max Planck Institute for Evolutionary Anthropology Department of Linguistics. (2015). Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Michael, L. (2015). Master class: Elicitation and documentation of evidentiality. Paper presented at 4th international conference on language documentation and conservation. Manoa, Hawaii. http://hdl.handle.net/10125/25394

Schütze, C. (2008). Thinking about what we are asking speakers to do. In S. Kepser & M. Reis (Eds.), *Linguistic evidence* (pp. 457-484). De Gruyter Mouton. doi: https://doi.org/10.1515/9783110197549.457