Multi-CAST

Sanzhi Dargwa

annotation notes

Diana Forker
University of Jena

Nils Norman Schiborr
University of Bamberg

May 2019
v1.0

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Australian Government
Australian Research Council

University of Bamberg

DFG

Multi-CAST

Multilingual Corpus of
Annotated Spoken Texts

Contents

1   Notes on the GRAID annotations

The following comprises selected notes on the GRAID (Haig & Schnell 2014) and RefIND (Schiborr et al. 2018) annotations of Sanzhi Dargwa. A grammar of Sanzhi (Forker Under revision), compiled on the basis of the material from which the texts in this corpus were taken, is currently in preparation, and will be freely available from the website of the publisher in the future.

This document corresponds to version 1905 of the Sanzhi annotations, published in May 2019. Unless a more recent version of this document exists, it also applies to any later versions of the annotations.

1.1   Ergativity and the distinction between S and A

Sanzhi is a language with ergative alignment, that is, the Patient-like argument (P) of transitive clauses and the single argument (S) of intransitive clauses receive the same marking in the morphology, distinct from that of the Agent-like argument (A) of transitive clauses. In Sanzhi, the A argument is marked with the ergative case, while S and P arguments are in the unmarked absolutive case. S and P control gender agreement – where instantiated – while A may trigger person agreement under certain conditions.

As a general rule for the GRAID annotations of A, S, and P in Sanzhi, we follow the language-specific morphology in glossing the ergative-marked argument as ⟨:a⟩, an absolutive-marked object as ⟨:p⟩, and an absolutive-marked subject as ⟨:s⟩. Examples (1) and (2) illustrate this practice:

(1)   *dul abiʔuˤnda q:arq:a, čak'al dul akaxubda.*

	du-l	*a-b-iʔ-uˤn-da*	*q:arq:a*
	1SG-ERG	NEG-N-steal.PFV-PRET-1	stone
##ds.neg	pro.1:a	v:pred	np:p

	ča-k'al	*du-l*	*a-kax-ub-da*
	who-INDEF	1SG-ERG	NEG-kill.PFV-PRET-1
##ds.neg	indef_other.h:p	pro.1:a	v:pred

'[He said,] "I did not steal the stone, I did not kill anyone."'

[mc_sanzhi_barkalla_0059]

(2)   *hinna š:uleb kabis:uncab šajt'an.*

	hin-na	*š:ule-b*	*ka-b-is:-un*	*ca-b*	*šajt'an*
	water-GEN	at.side-N	down-N-sleep.PFV-PRET	be-N	devil
##	ln_np:poss	np:l	v:pred	rv_aux	np.h:s

'The devil was sleeping at the water's edge.'   [mc_sanzhi_devils_0019]

For the vast majority of predicates in the corpus, adherence to the language-specific morphology yields results that are compatible with our general approach to A and P, as outlined in Andrews (2007: 137). Our adaptation has the advantage of providing clearly defined criteria for the purposes of annotation, and we have maintained it as the default throughout the corpus.

Under certain circumstances, however, the morphology does not align with our cross-linguistic definition of S, A, and P, which necessitated adaptations to our annotation conventions. These cases are discussed in Sections 1.1.1, 1.1.2, 1.1.3, and 1.2 below.

1.1.1 Experiencer verbs

Sanzhi has a number of predicates that assign quirky subject case. The most notable of these are experiencer verbs ("affective verbs" in Forker Under revision) such as 'be cold', 'want', 'see', and certain other verbs such as 'find', whose Experiencer-argument takes dative case. With bivalent experiencer verbs, the Stimulus is in the absolutive, and with certain verb forms the Experiencer may alternatively be marked with ergative case (see Forker Under revision: 395).[1]

There are numerous ways of interpreting the argument structure of these verbs. As a compromise solution, we have treated the Stimulus as ⟨:p⟩, since it is coded as absolutive, while the Experiencer receives the gloss ⟨:ncs⟩ 'non-canonical subject' if it is in dative case and ⟨:a⟩ if it is ergative. If the subject of a bivalent experiencer verb is unexpressed, the glossing defaults to ⟨:ncs⟩. Examples of experiencer verbs with dative and ergative Experiencers are illustrated in (3) and (4), respectively.

(3) a. *rucbaj qixbe dik:ul, …*

	ruc-b-a-j	*qix-be*	*d-ik:-ul*
	sister-PL-OBL-DAT	nut-PL	NPL-want.IPFV-ICVB
#ds_cv	np.h:ncs_cv	np:p	v:pred

'[She said,] "My sisters want nuts, …"' [mc_sanzhi_patima_0012]

b. *dam ʔaˤħ mus:a balχad.*

	dam	*ʔaˤħ*	*mus:a*	*b-alχ-ad*
	1SG.DAT	good	place	N-know.IPFV-PRS.1
##ds	pro.1:ncs	ln_adj	np:p	v:pred

'[The fox said,] "I know a good place."' [mc_sanzhi_patima_0027]

(4) a. *it:alaʁuna amzudex dul nalla čibažibil ak:u.*

	it:a-la	*=ʁuna*	*amzu-dex*	*du-l*	*nalla*
	3PL.OBL-GEN	=EQ	clean-NMLZ	1SG-ERG	until.then
#pc.neg	ln_poss_pro.h	=ln	np:p	pro.1:a_pc	other

	či-b-až-ib-il	*ak:u*
	SPR-N-see.PFV-PRET-PTCP	NEG
	v:pred	rv

'A cleanliness like theirs I have seen nowhere else.' [mc_sanzhi_ramazan_0067]

b. *hana ik'ulde dul lidil sahadarč:ibda!*

	hana		*ik'-ul-de*
	now	0_Ismail	say.IPFV-ICVB-2SG
#ds_cv	other	0.2:s_ds_cv	v:pred

	du-l	*lidil*	*sa-ha-d-arč:-ib-da*
	1SG-ERG	all<PL>	ANTE-upwards-NPL-find.PFV-PRET-1
##ds	pro.1:a	np:p	v:pred

'[He said,] "And now (you) say I made it all up!"' [mc_sanzhi_barkalla_0069]

---

[1] The conditions that trigger this alternation are as yet improperly understood and bear further investigation.

c.   *dul urč'emc'anu urč'emra juldaš barč:ibda.*

du-l	*urč'em-c'a-nu*	*urč'em-ra*	*juldaš*	*b-arč:-ib-da*
1SG-ERG	nine-ten-tenth	nine-NUM	friend	HPL-find.PFV-PRET-1
##ds pro.1:a	ln	ln	np.h:p	v:pred

'[He said,] "I have found ninety-nine friends."'   [mc_sanzhi_happy_0013]

1.1.2   Antipassive verbs

Sanzhi has an antipassive construction in which the case marking of the A and P arguments of a canonical transitive verb is reversed (Forker Under revision: 398). Notably, there is no formal marking on the verb for this type of construction. It most frequently occurs with the consumption verbs 'eat' and 'drink', and with the verb *b-irq'-* 'do, make' as in (5).

   In line with the morphology, antipassive constructions have been annotated as special types of intransitive clauses, that is, the absolute-marked Agent is ⟨:s⟩, and the ergative-marked Patient ⟨:obl⟩.

(5)   *nuš:a χalq' ʡaˤčil dirq'an χalq'deq'al!*

nuš:a	*χalq'*
1PL	people
##ds pro.1:s	np.h:appos

	ʡaˤči-l	*d-irq'-an*	
0_people	work-ERG	1/2PL-do.IPFV-PTCP	
#rc_pc rc_f0.1:s	np:obl	v:pred	%

χalq'-de	*=q'al*
people-PST	=MOD
np.h:pred	=other

'[They said,] "We people who are working are people too!"'   [mc_sanzhi_mill_0014]

1.1.3   Verbs of speech

Verbs of speech in Sanzhi can be morphologically transitive or intransitive, and hence assign ergative or absolutive case to the NP expressing the speaker. In either case, in the narratives included in this corpus, they are more often than not used to introduce direct speech, and even transitive verbs of speech only infrequently occur with nominal objects. As is standard practice in GRAID, direct speech complements are not formally analyzed as a ⟨:p⟩ arguments of the clauses that introduce them, but merely receive the tag ⟨#ds⟩ 'direct speech' (as per Haig & Schnell 2014: 26). As a consequence, in some Multi-CAST corpora (notably Northern Kurdish, among others), the subjects of morphologically transitive verbs of speech have been annotated as cases of S.

   For Sanzhi, we have decided to maintain the morphological distinction between A and S for verbs of speech, even when the former is not paired with a P argument in the annotations. To denote their special status, the subjects of these verbs have been glossed either ⟨:a_ds⟩ 'subject of a transitive verb of speech' as in (6a) and (6b), or ⟨:s_ds⟩ 'subject of an intransitive verb of speech' as in (7).

(6)   a.   *dul helic:e haʔibda, helit:e bursa!, daˁʔaˁn-ne.*

        du-l   *hel-i-c:e*   *ha-ʔ-ib-da*
        1SG-ERG   that-OBL-IN upwards-say.PFV-PRET-1
    ## pro.1:a_ds pro.h:g   v:pred

                                      hel-it:e   *b-urs-a*   *daˁʔaˁn-ne*
       0_Ramazan   0_that that-ADVZ   N-tell-IMP   secret-ADVZ
    #ds im_f0.2:a_ds 0:p   dem_pro:other vother:pred % other

    'I told him secretly, "Say (it) like this!"'   [mc_sanzhi_kurban_0009]

  b.   *i Pat'ima rursibcar helʁuna χabar, …*

        i   *Pat'ima*   *r-urs-ib*   *ca-r*   *hel*   *=ʁuna χabar*
        and   Patima   F-tell-PRET be-F that   =EQ   story
    ## other pn_np.h:a_ds v:pred   rv_aux ln_dem_pro =ln np:p

    'And Patima told that kind of story, …'   [mc_sanzhi_patima_0021]

(7)   *ʔaˁlišiχ ik'ʷi, cellij ʔaˁʁunile darš:al eč:a?*

        ʔaˁlišiχ   *ik'ʷ-i*   *cellij* *ʔaˁʁuni-le*   *darš:al*   *eč:a*
        Alishikh   say.IPFV-HAB.PST   why   needed-ADVZ one.hundred nanny.goat
    ## pn_np:s_ds v:pred   ##ds other other   ln_num   np:s

    'Alishikh said, "Why would you need a hundred goats?"'   [mc_sanzhi_happy_0005]

Where verbs of speech frame non-clausal quotations of the kind in (8) that cannot unambiguously be analyzed as objects, the quoted material is glossed ⟨:other⟩ rather than ⟨:p⟩:

(8)   *i dul waˁwherʔanda, k:aza k'at'a qix:unaj!*

        i   *du-l*   *waˁw*   *h-erʔ-an-da*
        and   1SG-ERG   shout   upwards-say.IPFV-PTCP-1
    #ds_pc other pro.1:a_ds_cps_pc other:lvc v:pred

    k:aza   *k'at'a* *qix:unaj*
    fork   spade qixunaj
    np:other rn_np rn_np

    '[The fox says,] "And I will shout 'fork, spade, qixunaj'!"'   [mc_sanzhi_patima_0024]

1.2   Complex predicates

Complex predicates (CPs) combine a semantically weak light verb (or "vector verb") such as *do*, *take*, or *be*, with a non-verbal element of some kind, the latter of which supplies most of the semantic content to the expression. Crucially, the non-verbal element fails to exhibit many of the properties of regular objects (see, e.g., Berlage 2010), and is hence not identified as such in the annotation. Instead, it is glossed ⟨:lvc⟩ 'light verb complement', which marks it out as a special type of expression. Because the lexical category of the complement cannot always be ascertained, it invariably receives the form gloss ⟨other⟩.

The complement and light verb contribute jointly to the argument structure of the entire expression (cf. Butt 2010). In most cases, the case marking of the subject is determined by the light verb: if the light verb is agentive (e.g. 'do', 'take') or experiential (e.g. 'see', 'hear'; see Section 1.1.1), the subject of the complex predicate is, respectively, ergative or dative; if not (e.g. with

'be'), it is absolutive. At the same time, the presence or absence of a direct object in the complex predicate is not strictly controlled by the light verb's valency. For this reason, it is not uncommon to find complex predicates with two absolutive arguments, as in (9b), or ones with an ergative subject but no discernible (or even implied) object, as in (10a).

A system of extended annotation has been introduced to help capture the complexities of these predicates. The base GRAID function gloss of the subject of a complex predicate is determined by its case marking, usually assigned by the light verb. A subject in the ergative is glossed ⟨:a⟩, a subject in the absolutive ⟨:s⟩, and a subject in the dative ⟨:ncs⟩, as discussed in the previous sections. To this gloss is then added an additional specifier ⟨_cps⟩ or ⟨_cpa⟩, which marks the overall transitivity of the complex predicate, that is, the absence or (at least implied, i.e. ⟨0⟩) presence of a P argument. As such, there are a total of six possible configurations:

1. a. subject is absolutive + CP is intransitive ➙ ⟨:s_cps⟩ e.g. (9a), (13)
   b. subject is absolutive + CP is transitive ➙ ⟨:s_cpa⟩ e.g. (9b)

2. a. subject is ergative + CP is intransitive ➙ ⟨:a_cps⟩ e.g. (10a), (8)
   b. subject is ergative + CP is transitive ➙ ⟨:a_cpa⟩ e.g. (10b)

3. a. subject is dative + CP is intransitive ➙ ⟨:ncs_cps⟩ e.g. (11a)
   b. subject is dative + CP is transitive ➙ ⟨:ncs_cpa⟩ e.g. (11b)

The following examples illustrate each case.

(9) a. *urk' uqun il miskin, …*

	urk'	*uq-un*	*il*	*miskin*
	fright	go.PFV-PRET	that	poor
##	other:lvc	v:pred	ln_dem	np.h:s_cps

'The pauper was frightened, …' [mc_sanzhi_barkalla_0015]

b. *il šakričibcar bec'li berk:unce.*

	il	*šak*	*r-ič-ib*	*ca-r*
	that	feel	F-occur.PFV-PRET	be-F
##	dem_pro.h:s_cpa	other:lvc	v:pred	rv_aux

	bec'-li	*b-erk:-un-ce*	
	wolf-ERG	HPL-eat.PFV-PRET-ATTR.SG	0_sisters
#cc:p	np.d:a	v:pred	0.h:p

'She suspected that the wolf had eaten (her sisters).' [mc_sanzhi_patima_0018]

(10) a. *dul at kumek birq'anda.*

	du-l	*at*	*kumek*	*b-irq'-an-da*
	1SG-ERG	2SG.DAT	help	N-do.IPFV-PTCP-1
##ds_pc	pro.1:a_cps_pc	pro.2:obl	other:lvc	v:pred

'[The fox said,] "I will help you."' [mc_sanzhi_patima_0022]

b. *ca eč:a dulira at peškeš birq'id.*

	ca	*eč:a*	*du-li*	*=ra*	*at*	*peškeš*	*b-irq'-id*
	one	nanny.goat	1SG-ERG	=and	2SG.DAT	gift	N-do.IPFV-1.PRS
##ds	ln	np:p	pro.1:a_cpa	=other	pro.2:g	other:lvc	v:pred

'[He said,] "I will give you one goat as a gift."' [mc_sanzhi_happy_0016]

(11) a. *dam q:ihin dičible cad.*

	dam	*q:ihin*	*d-ič-ib-le*	*ca-d*
	1SG.DAT	difficult	NPL-occur.PFV-PRET-CVB	be-NPL
##ds_cc_cv:p	pro.1:ncs_cps_cv	other:lvc	v:pred	rv_aux

'[She said,] "[Tell the friends that] I have met with hardship."'

[mc_sanzhi_happy_0030]

b. *ilij han bičib, t:ura herʁanne žaˤndar, …*

	il-i-j	*han*	*b-ič-ib*
	that-OBL-DAT	seem	N-occur.PFV-PRET
##	dem_pro.h:ncs_cpa	other:lvc	v:pred

	t:ura	*h-erʁ-an-ne*	*žaˤndar*
	outside	upwards-come-PTCP-PRS.3	Kumyk
#cc:p	other:g	v:pred	np.h:s

'He thought that a Kumyk man would come out, …'

[mc_sanzhi_barkalla_0013]

In some cases, the same light verb occurs with different case assignment in different complex predicates. The light verb 'occur', for instance, itself governs the absolutive as in (12a), but in (12b), the complex predicate 'remember occur' is coded as an experiencer verb with a dative-marked Experiencer, which we have decided to gloss analogously to other bivalent experiencer verbs such as 'know' or 'want' (see Section 1.1.1 above):

(12) a. *waˤwdexlij guči bičibcab χalq'.*

	waˤw-dex-li-j	*guči*	*b-ič-ib*	*ca-b*	*χalq'*
	shout-NMLZ-OBL-DAT	gather	HPL-OCCUR.PFV-PRET	be-HPL	people
##	np:other	other:lvc	v:pred	rv_aux	np.h:s_cps

'Because of the shouting, people gathered.'   [mc_sanzhi_barkalla_0054]

b. *dam hanbičib na libil, …*

	dam	*han*	*b-ič-ib*	*na*	*libil*
	1SG.DAT	remember	N-occur.PFV-PRET	now	all\<N>
##	pro.1:ncs_cpa	other:lvc	v:pred	other	np:p

'I remember it all now, …'   [mc_sanzhi_ramazan_0034]

1.3   Structurally and pragmatically suppressed arguments

In GRAID, unexpressed clausal referents ⟨0⟩ are annotated only where they are

1. licensed by the predicate,
2. specific and retrievable from the discourse context, and
3. not in an argument slot that is systematically suppressed by the predicate.

The third criterion assumes that it is possible to distinguish two types of referentiall null argument: those that are structurally licensed, but remain empty due to context-specific pragmatic factors, and those that are either systematically suppressed or not licensed due to purely structural factors. As a general rule, GRAID glosses only the former kind of argument with ⟨0⟩, because only in this case do speakers exercise any choice of expression; the latter remains unannotated.

This distinction has proven difficult to maintain in the annotation of a number of languages, including Sanzhi, where it is particularly contentious in the context of certain non-finite verb

forms such as converbs and participles, and with imperatives and certain typs of complement clauses. Furthermore, not capturing arguments that fail to meet the third criterion, such as the gapped constituents in relative clauses, leads to conceptual issues regarding the implicitness of discourse.

It is for this reason that we have decided to introduce a form gloss ⟨f0⟩ 'forced zero' to capture categorically suppressed referents, as a counterpart to contrastively suppressed zero ⟨0⟩. It should be noted that ⟨f0⟩ is not a type of ⟨0⟩, so the two categories should never be conflated during analysis. In the current version of Multi-CAST, the ⟨f0⟩ symbol is used only in the Sanzhi corpus; it is planned to become an optional gloss in the standard GRAID specification in the future.

In Sanzhi, the ⟨f0⟩ symbol is primarily applied to gaps in relative clauses, which are discussed in Section 1.3.1, and to the subjects of certain types of complement clause, discussed in Section 1.3.2. Issues with non-finite constructions and related issues are addressed in Section 1.3.3, imperatives in Section 1.3.4.

1.3.1   Gapped constituents in relative clauses

The gapped constituents in relative clauses cannot be overtly expressed. Where in standard GRAID they would remain unannotated, in the Sanzhi corpus they receive the form gloss ⟨f0⟩ with an additional specifier ⟨rc_⟩ to mark out the context in which they occur.

(13)   *umxu birχʷi ruc:ic:eb wac'ac:eb errirχʷan.*

	umxu	*b-irχ-i*	*ruc:i-c:e-b*
	key	N-be.IPFV-HAB.PST	sister-IN-N
##	np:s	cop	np.h:pred_l

		wac'a-c:e-b	*er*	*r-irχʷ-an*
	0_sister	forest-IN-N	life	F-be.IPFV-PTCP
#rc_pc	rc_f0.h:s_cps_pc	np:l	other:lvc	v:pred

'The key was with the sister who lived in the forest.'   [mc_sanzhi_devils_0027]

(14)   *wat', na ilt:i il c'ikuri rarč:ibcari, hana xadi karižibil.*

	wat'	*na*	*ilt:i*	*il*	*c'ikuri*	*r-arč:-ib*	*cari*
	well	now	3PL	that	bride	F-find.PFV-PRET	be.F
##	other	other	dem_pro.h:s	ln_dem	np.h:p	v:pred	rv_aux

	hana		*xadi*	*ka-r-iž-ib-il*
	now	0_bride	married	down-F-BE.PFV-PRET-PTCP
#rc_pc	other	rc_f0.h:s_cps_pc	other:lvc	v:pred

'Well, they found this bride who had only just married.'   [mc_sanzhi_mill_0027]

1.3.2   Infinitival complement clauses

In infinitival complement clauses in which the subject is co-referential with the subject of the matrix clause, the former is structurally suppressed (i.e. complement control, see Forker Under revision: §24.2.6; §24.5), and is hence annotated ⟨f0⟩. In the Sanzhi corpus, infinitival complement clauses most commonly occur with verbs such as 'want', 'decide', and 'forget', and with phasal verbs such as 'begin' and 'end':

(15) *na ilijq'ar bik:ul caw direktur agaraj.*

	na	*il-i-j*	*=q'ar*	*b-ik:-ul*	*ca-w*
	now	that-OBL-DAT	=MOD	N-want.IPFV-ICVB	be-M
#cv	other	dem_pro.h:ncs_cv	=other	v:pred	rv_aux
		0001			

		direktur	*ag-ar-aj*
	0_Kurban	director	go.PFV-PRS-INF2.3
#cc_in:p	f0.h:s_in	np.h:obl	v:pred
	0001		

'He wants to become director.'   [mc_sanzhi_kurban_0016]

(16) *waˤw ik'ʷij waʔiš:ibcaw.*

		waˤw	*ik'ʷ-ij*	
	0_Ismail	shout	say.IPFV-INF1	
## #cc_in:p	f0.h:s_ds_in	other:lvc	v:pred	%
	0001			

	w-aʔ-iš:-ib	*ca-w*
0_Ismail	M-begin-become.PFV-PRET	be-M
0.h:a	v:pred	rv_aux
0001		

'(He) started to shout.'   [mc_sanzhi_barkalla_0006]

1.3.3 Subordinate verb forms

The concept of finiteness is difficult to apply to Dargwa languages, including to Sanzhi (see, e.g. Kalinina & Sumbatova 2007; see also Forker 2011 and 2013 for a discussions of finiteness in Hinuq, a related Nakh-Daghestanian language). Lack of finiteness, however defined, is often accompanied by an inability to express certain kinds of argument, most notably subjects. Sanzhi has a number of verb forms that lack some of the features exhibited by normal finite verbs, such as the full range of TAM marking and agreement morphology. They may also possess certain nominal properties like case marking. These verb forms include converbs, participles, infinitives, and the masdar. Yet although they may appear morphologically deficient in some ways, distributionally they are often very similar to finite verbs, and appear to govern arguments in an identical manner.

For the GRAID annotations, this means that when one of these "less finite" verb forms governs a referential argument, but that argument is not overtly present, it is difficult to decide whether its absence is caused by the structural inability of the verb to license the argument, or by contextual factors. It is for this reason that we have decided to assume a somewhat non-committal stance in the treatment of these verbs.

Firstly, we annotate any of their omitted arguments as ⟨0⟩ 'contrastively suppressed' rather than ⟨f0⟩ 'structurally suppressed', and the head of the verbal complex as regular ⟨v⟩ rather than ⟨vother⟩ 'non-canonical verb form'. Secondly, a series of specifiers are applied to the GRAID function glosses of the subject, allowing these contentious forms to be readily distinguished: ⟨_cv⟩ for converb clauses, ⟨_pc⟩ for participial clauses, and ⟨_in⟩ for infinitival clauses. The masdar is exceedingly rare in the annotated texts, and so is not labelled; the two variants of the infinitive (non-inflecting INF1 and inflecting INF2, called "infinitive" and "subjunctive" in Forker Under revision) are functionally equivalent (Forker Under revision: 351–354), and hence do not receive

separate symbols. Lastly, the same three symbols ⟨cv⟩, ⟨pc⟩, and ⟨in⟩ are added to the clause boundary marker ⟨#⟩. While to a degree redundant with the function specifiers, these tags allow for easier identification of clauses of particular types.

The following examples illustrate the annotation patterns, (17) for converb clauses, (18) for participle clauses, and (19) for infinitive and subjunctive clauses. See also Section 1.6 for how these extra specifiers are ordered relative to the base GRAID symbols.

(17) *nuš:al hinna urχ:ab birq'ulda.*

	nuš:a-l	*hin-na*	*urχ:ab*	*b-irq'-ul-da*
	1PL-ERG	water-GEN	mill	N-do.IPFV-ICVB-1
#ds_cv	pro.1:a_cv	ln_np:poss	np:p	v:pred

'[They said,] "We are building a watermill."'   [mc_sanzhi_mill_0019]

(18) *du hel zamana ala salab arganda.*

	du	*hel*	*zamana*	*ala*	*sala-b*	*arg-an-da*
	1SG	that	time	2SG.GEN	front-N	go.IPFV-PTCP-1
#ds_pc	pro.1:s_pc	ln_dem	np:other	pro.2:l	adp	v:pred

'[It said,] "And I will appear before you right away."'   [mc_sanzhi_devils_0040]

(19) *t:ura buqij, calli ečne as:ij, calli juldaš:e barč:ij.*

	t:ura	*b-uq-ij*
	0_they	outside	N-go.PFV-INF1
#ac_in	0.h:s_in	other:g	v:pred

	ca-l-li	*eč-ne*	*as:-ij*
	one-OBL-ERG	nanny.goat-PL	buy.PFV-INF1
#ac_in	np.h:a_in	np:p	v:pred

	ca-l-li	*juldaš:-e*	*b-arč:-ij*
	one-OBL-ERG	friend-PL	N-find.PFV-INF1
#ac_in	np.h:a_in	np.h:p	v:pred

'(They) left, one to buy nanny goats, one to find friends.'   [mc_sanzhi_happy_0008]

A final point of contention concerns syntactic hierarchization. Especially with converb clauses, which are highly frequent and often appear in long chains, it can be difficult to determine exactly which independent clause, if any, they are subordinated to. The Sanzhi annotations thus implement a slight relaxation of the definition of the two left-edge clause boundary markers in GRAID: while ⟨##⟩ is still defined as the beginning of a fully independent syntactic unit, ⟨#⟩ is not used specifically for identifiably subordinated units, but for all clauses that do not meet the criteria for being glossed ⟨##⟩. In the Sanzhi corpus, then, ⟨#⟩-clauses can freely occur outside of the boundaries of a matrix ⟨##⟩-clause. While this change could cause issues for analyses that rely on the precise syntactic hierarchization of clause units (for which GRAID was not designed in the first place, it should be noted), we believe the benefits of this approach outweigh any conceivable disadvantages.

1.3.4   Imperatives

While the subjects of imperatives and prohibitives (i.e. negated imperatives) are almost always left unexpressed, they are in fact not categorically suppressed (Forker Under revision: 460). For this

reason, like the subjects of subordinate verb forms discussed in Section 1.3.3, they are annotated ⟨0⟩ rather than ⟨f0⟩, and the verbal complex ⟨v⟩ rather than ⟨vother⟩:

(20) *dam kumek barq'aja!*

	dam	*kumek*	*b-arq'-aja*
0	1SG.DAT	help	N-DO.PFV-IMP.PL
##ds 0.2:a_cps	pro.1:obl	other:lvc	v:pred

'[He cried,] "Help me!"'   [mc_sanzhi_barkalla_0006]

(21) *u abdaldew? maˤq'aˤt:a!*

	u	*abdal-de*	*=w*		*maˤ-q'-aˤt:a*
	2SG	fool-2SG	=Q	0_brother	PROH-go-PROH.SG
##ds	pro.2:s	np.h:pred	=other	##ds.neg 0.2:s	v:pred

'[They said,] "Are you a fool? Don't go!"'   [mc_sanzhi_dragon_0006]

Imperatives and prohibitives with non-zero subjects are illustrated in (22a) and (22b).

(22) a. *u sala kabiže*

	u	*sala*	*ka-b-iž-e*
	2SG	front	down-N-be.PFV-IMP
#ds	pro.2:s	np:l	v:pred

'[The fox said,] "You sit down in front."'   [mc_sanzhi_patima_0031]

b. *warilla wari u iχt:aj erčimahark'ut:a!*

	warilla	*wari*	*u*	*iχ-t:-a-j*	*er*
	no	way	2SG	DEM.down-PL-OBL-DAT	look
##ds.neg	other	other	pro.2:s_cps	dem_pro:obl	other:lvc

či-ma-hark'-ut:a
SPR-PROH-look.IPFV-PROH.SG
v:pred

'[He said,] "Whatever happens, do not look at them!"'

[mc_sanzhi_bazhuk_0004]

1.4   Predicative expressions of possession

In Sanzhi, predicative expressions of possession involve locational copulae (Forker Under revision: 641; 467–470), with the possessor in the genitive and the possessee in the absolutive. We annotate these clauses analogously to existential constructions, that is, the copula is glossed ⟨other:predex⟩, the possessor ⟨:obl⟩, and the possessee ⟨:s⟩:

(23) *dila darš:al juldaš biχʷar, …*

	di-la	*darš:al*	*juldaš*	*b-iχʷ-ar*
	1SG-GEN	hundred	friend	N-be.PFV-PRS
#ds_ac	pro.1:obl	ln	np.h:s	other:predex

'[He said,] "If I had a hundred friends, …"'   [mc_sanzhi_happy_0006]

1.5   Third person pronominal forms

Sanzhi does not have a separate paradigm of third person personal pronouns. Their role is filled by an extensive set of demonstrative pronouns expressing proximity, elevation, and cardinal direction, in addition to case and number. See Forker (Under revision: Ch. 4) for a full listing of pronominal forms in Sanzhi.

Only first and second person pronouns receive the form gloss ⟨pro⟩; third person pronouns and other demonstrative forms are glossed ⟨dem_pro⟩ instead.

(24)   *c'ili ilt:ali kaq:ible q'uˤq', …*

	c'ili	*ilt:a-li*	*k-aq:-ib-le*	*q'uˤq'*
	then	3PL-ERG	down-carry-PRET-CVB	hammer
#cv	other	dem_pro.h:a_cv	v:pred	np:p

'Then they brought a hammer, …'   [mc_sanzhi_mill_0031]

(25)   *il mar χabar cab, ac:e dul bursulda.*

	il		*mar*	*xabar*	*ca-b*
	that		truth	story	be-N
##	dem_pro:s		ln	np:pred	cop

	a-c:e	*du-l*	*b-urs-ul-da*
	2SG-IN	1SG-ERG	N-tell-ICVB-1
#cv	pro.2:g	pro.1:a_ds	v:pred

'This is a true story, I tell you.'   [mc_sanzhi_kurban_0053]

1.6   On the relative order of additional symbols

The GRAID annotations for Sanzhi Dargwa make use of a number of additional symbols for complex predicates (Section 1.2) and subordinate verb forms (Section 1.3.3) that attach to function glosses and clause boundary markers. While they are not by themselves particularly numerous, some complexity arises from their combination with other symbols that occupy the same space.

In the Sanzhi texts, the subject of a complex predicate of speech in converb form, for instance, might receive the function gloss ⟨:s_ds_cps_cv⟩, and in the direct speech that might follow, a negated complement clause with an infinitival predicate in P role would have ⟨#ds_cc_in.neg:p⟩ as its clause boundary marker. While these are, thankfully, the worst case scenarios, they are far from uncommon occurrences.

In order to avoid confusion, these symbols combine in accordance with a strictly defined pattern. As a general rule, symbols that are not part of the base GRAID inventory always attach after (or outside of) those that are; in the Sanzhi Dargwa annotations, additional specifiers on function glosses (of subjects, mostly) are always added in the following order:

1.   base function symbol: e.g. ⟨:s⟩, ⟨:a⟩, ⟨:ncs⟩
2.   subject of direct speech: ⟨_ds⟩
3.   complex predicate transitivity: ⟨_cps⟩, ⟨_cpa⟩
4.   clause type (converb, participle, infinitive): ⟨_cv⟩, ⟨_pc⟩, ⟨_in⟩

The various clause boundary tags likewise combine as follows (an extension of Haig & Schnell 2014: 25, Tab. 6); the first element to follow after the boundary marker ⟨#⟩ has no delimiter (⟨_⟩ or ⟨.⟩):

1. boundary marker: ⟨#⟩ or ⟨##⟩
2. direct speech clause: ⟨ds⟩
3. clause type (complement, adverbial, relative): ⟨_cc⟩, ⟨_ac⟩, ⟨_rc⟩
4. clause type (converb, participle, infinitive): ⟨_cv⟩, ⟨_pc⟩, ⟨_in⟩
5. negated: ⟨.neg⟩
6. function: e.g. ⟨:s⟩, ⟨:a⟩, ⟨:p⟩, etc.

2 Notes on the RefIND annotations

2.1 Cataphoric introductions

Occassionally, the first 'mention' of a new referent will be in a subordinate clause that precedes the formal introduction in the matrix clause. Mentions are always annotated as they occur in linear order, with the RefLex gloss being applied to the first occurrence of a referent index, irrespective of the form (or lack thereof) of the referring expression. These cases are readily identifiable in the analysis, which allows cataphoric distance and other aspects of the referential relationship to be determined.

(26) *helka sat:i argan k:urt:aj čiražible hel.*

hel-ka	*sat:i*			*arg-an*	
that-down	as.soon.as		0_fox	go.IPFV-PTCP	
#cv dem_pro:other	adp	#rc_pc	rc_f0.d:s_pc	v:pred	%
0000			0015		
			new		

k:urt:a-j	*či-r-až-ib-le*	*hel*
fox-DAT	SPR-F-see.PFV-PRET-CVB	that
np.d:ncs_cv	v:pred	dem_pro.h:p
0015		0000

'As soon as she (sat) down, a fox that was passing by saw her.'

[mc_sanzhi_patima_0019]

2.2 Referents in clauses otherwise not considered

Segments that have not been annotated for whatever reason, be that because they are incomplete or not syntactically well-formed, or because they are taken out of the normal flow of narration (e.g. because they address the listener, directly replay to the interviewer's questions, or are not produced by the primary speaker), are marked as ⟨#nc⟩ 'not considered', and all of the elements they contain are glossed ⟨nc⟩.

However, these segments may still contain identifiable discourse references, which are presumably registered by the listener even in cases where the clause in question is abandoned partway through. So as to preserve the genuine sequence of references in the annotations, mentions in ⟨#nc⟩ segments are indexed with RefIND, even though they do not receive meaningful GRAID annotations. This is true of all Multi-CAST corpora with RefIND.

For the Sanzhi corpus, however, we have attempted to go one step further by adding form and person/animacy glosses back onto those ⟨nc⟩'d elements that have referent indices. The glosses

are added as specifiers to the righthand side of the ⟨nc⟩ symbol, yielding, for instance, ⟨nc_np⟩ or ⟨nc_pro.h⟩. Grammatical functions are not glossed.

(27) a. *caw q:umuqlandew?*

	ca-w	*q:umuqlan-de*	*=w*
	REFL-M	Kumyk-PST	=Q
#nc	nc_refl.h	nc	nc
	0000		

[Someone in the audience asks,] 'Was he himself Kumyk?'

b. *aʔa, dark:ʷande.*

	aʔa		*dark:ʷan-de*
	no	0_journalist	Dargwa-PST
#nc	nc	nc_0.h	nc
		0000	

'No, he was Dargi.'   [mc_sanzhi_ramazan_0012-0013]

Why not simply gloss these elements normally? Doing it in this (admittedly) roundabout way makes it clear that while some information can be gleaned from these elements, one cannot (and should not) rely on retrieving any information from the rest of the ⟨#nc⟩ segment. For most types of analysis, the ⟨nc_⟩ glosses should not be conflated with related GRAID symbols.

References

Andrews, Avery. 2007. The major functions of the noun phrase. In Shopen, Timothy (ed.), *Language typology and syntactic description*, vol. 1: Clause structure, 132–223. Cambridge: Cambridge University Press.

Berlage, Eva. 2010. Lexicalisation of predicative complements in English. *Transactions of the Philological Society* 108(1). 53–67.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Amberber, Mengistu & Harvey, Mark & Baker, Brett (eds.), *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge: Cambridge University Press.

Forker, Diana. Under revision. *An grammar of Sanzhi Dargwa*. Berlin: Language Science Press.

Forker, Diana. 2011. Finiteness in Hinuq. *Linguistic Discovery* 9. 3–29.

Forker, Diana. 2013. Hinuq verb forms and finiteness. *Acta Orientalia Academiae Scientarium Hungaricae* 66. 69–93.

Forker, Diana & Schiborr, Nils N. 2019. Multi-CAST Sanzhi Dargwa. In Haig, Geoffrey & Schnell, Stefan (eds.), *Multi-CAST: Multilingual corpus of annotated spoken texts*. (http://multicast.aspra.uni-bamberg.de/#sanzhi) (Accessed 2019-04-05).

Haig, Geoffrey & Schnell, Stefan. 2014. *Annotations using GRAID (Grammatical Relations and Animacy in Discourse): Introduction and guidelines for annotators (version 7.0)*. (http://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).

Haig, Geoffrey & Schnell, Stefan (eds.). 2016. *Multi-CAST: Multilingual Corpus of Annotated Spoken Texts*. (http://multicast.aspra.uni-bamberg.de/) (Accessed 2019-03-08).

Kalinina, Elena & Sumbatova, Nina. 2007. Clause structure and verbal forms in Nakh-Daghestanian. In Nikolaeva, Irina (ed.), *Finiteness: Theoretical and empirical foundations*, 183–249. Oxford: Oxford University Press.

Schiborr, Nils N. & Schnell, Stefan & Thiele, Hanna. 2018. *RefIND — Referent Indexing in Natural-language Discourse: Annotation guidelines (v1.1)*. Bamberg / Melbourne: University of Bamberg / University of Melbourne. (http://multicast.aspra.uni-bamberg.de/#annotations) (Accessed 2019-03-08).

Appendices

A   List of corpus-specific GRAID symbols

The following is a list of the non-standard GRAID symbols used in the annotation of the Multi-CAST Sanzhi Dargwa corpus. Please refer to the *GRAID manual* (Haig & Schnell 2014: 54–55) for an inventory of basic GRAID symbols.

Form symbols and specifiers

⟨f0⟩	structurally suppressed argument slot of a predicate
⟨dem_pro⟩	demonstrative pronoun
⟨poss_pro⟩	possessive pronoun
⟨pn_np⟩	proper name
⟨intrg_other⟩	interrogative pronoun
⟨indef_other⟩	indefinite pronoun
⟨rc_⟩	*specifier:* gapped argument of a relative clause; attaches to ⟨f0⟩

Function symbols and specifiers

⟨:lvc⟩	light verb complement
⟨_ds⟩	*specifier:* subject of a verb of speech; attaches to ⟨:s⟩, ⟨:a⟩, and ⟨:ncs⟩
⟨_cps⟩	*specifier:* subject of an intransitive complex predicate
⟨_cpa⟩	*specifier:* subject of a transitive complex predicate
⟨_cv⟩	*specifier:* subject of a converb clause
⟨_pc⟩	*specifier:* subject of a participial clause
⟨_in⟩	*specifier:* subject of an infinitival clause

Clause boundary symbols

⟨ds⟩, ⟨_ds⟩	*tag:* direct speech clause
⟨cv⟩, ⟨_cv⟩	*tag:* converb clause
⟨pc⟩, ⟨_pc⟩	*tag:* participial clause
⟨in⟩, ⟨_in⟩	*tag:* infinitival clause

Subconstituent symbols

⟨_adj⟩	adjectival modifier; attaches to ⟨ln⟩ and ⟨rn⟩
⟨_dem⟩	demonstrative determiner; attaches to ⟨ln⟩ and ⟨rn⟩
⟨_deti⟩	indefinite determiner; attaches to ⟨ln⟩ and ⟨rn⟩
⟨_num⟩	numeral modifier; attaches to ⟨ln⟩ and ⟨rn⟩
⟨_aux⟩	auxiliary; attaches to ⟨lv⟩ and ⟨rv⟩

Other symbols

⟨nc_⟩	*specifier:* marks form glosses with RefIND indices in segments otherwise not considered (i.e. those marked ⟨#nc⟩)

B   List of abbreviated morphological glosses

1	first person	LOC	locative
1/2	first/second person	M	human masculine singular
2	second person	MOD	modal
3	third person	MODQ	modal interrogative
ABL	ablative	MSD	masdar
ADD	additive	N	neuter singular
ADVZ	adverbializer	NEG	negative
ANTE	spatial case 'before'	NMLZ	nominalizer
ATTR	attributive	NPL	neuter plural
CAUS	causative	NUM	numeral
COMIT	comitative	OBL	oblique stem
COMP	comparative	OPT	optative
CONC	concessive	ORD	ordinal
COND	conditional	POST	spatial case 'behind'
CVB	converb	PRET	preterite
DAT	dative	PRS	present
DEM	demonstrative	PROH	prohibitive
DIR	directional	PTCP	participle
ERG	ergative	PL	plural
EQ	equative enclitic	PRT	particle
F	human feminine singular	PST	past
GEN	genitive	PFV	perfective
HAB	habitual	Q	question
HPL	human plural	SG	singular
ICVB	imperfective converb	SPR	spatial case 'on'
IMP	imperative	SUB	spatial case 'under'
IN	spatial case 'in'		
INF1	non-inflecting infinitive	NC	not classified
INF2	inflecting infinitive	[I]	code switching to Icari Dargwa
INDEF	indefinite		
INDQ	embedded question	[R]	code switching to Russian
IPFV	imperfective		

Multi-CAST

Multilingual Corpus of Annotated Spoken Texts

multicast.aspra.uni-bamberg.de/