Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 26–30
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

Annotating "Absolute" Preverbs in the Homeric and Vedic Treebanks

Luca Brigada Villa,[1] Erica Biagetti,[2] Chiara Zanchi[2]
[1]University of Bergamo/University of Pavia, [2]University of Pavia
luca.brigadavilla@unibg.it, {erica.biagetti, chiara.zanchi01}@unipv.it

Abstract

Indo-European preverbs are uninflected morphemes attaching to verbs and modifying their meaning. In Early Vedic and Homeric Greek, these morphemes held ambiguous morphosyntactic status raising issues for syntactic annotation. This paper focuses on the annotation of preverbs in so-called "absolute" position in two Universal Dependencies treebanks. This issue is related to the broader topic of how to annotate ellipsis in Universal Dependencies. After discussing some of the current annotations, we propose a new scheme that better accounts for the variety of absolute constructions.

Keywords: Universal Dependencies, preverbs, Ancient Greek and Sanskrit linguistics

1. Introduction

In this paper, we discuss the current annotation scheme of Early Vedic and Homeric Greek preverbs (PVs) and propose a new one. Our data is extracted from the Ṛgvedic portion of the Vedic TreeBank (VTB, Hellwig et al., 2020)[1] and from a rule-based Universal Dependencies (UD) conversion of the Iliad and Odyssey (Homeric TreeBank; HTB)[2] treebanked at the *Perseus Project*.[3] This paper documents the first step toward the larger goal of systematizing the annotation of PVs occupying other syntactic positions through semiautomatic methods. In particular, we deal with ancient Indo-European PVs in the so-called "absolute position", virtually replacing a verbal form. This issue relates to the larger question of how elliptical structure should be annotated in UD.

Section 2 familiarizes the readers with Indo-European PVs and their UD annotation. Section 3 focuses on absolute usages. Section 4 describes data extraction. Section 5 discusses their current annotation in the VTB and HTB. Finally, section 6 contains the annotation proposal.

2. PVs in Early Vedic and Homeric Greek

Indo-European PVs are uninflected morphemes attaching to verbs and modifying verbal meaning (e.g., Homeric Greek *baínō* 'walk' vs. *ana-baínō* 'upward-walk'). In Proto-Indo-European, PVs were free positioning spatial adverbs, which later underwent functional bifurcation into univerbated prefixes proper and adpositions.

In both Early Vedic and Homeric Greek, this diachronic development was still ongoing: the same uninflected morphemes held an ambiguous morphosyntactic status, functioning as adverbs, nominal or verbal modifiers, adpositions, and PVs proper (for discussion and examples, see Zanchi, 2019: 65-116, 173-183 with references). In early Indo-European languages, PVs could semantically modify verbs without morphological univerbation. Take for instance examples (1) and (2) from Early Vedic and Homeric Greek, in which the preverbs *prá* 'forward' and *en* 'in' are separated from the verbs *vocam* 'say' and *ebḗsamen* 'stepped' that they modify (Zanchi 2019: 98; 181).

(1) ṚV 1.59.6a

prá	*nū́*	*mahitvā́m*
forward	now	greatness(F).ACC
vṛṣabhā́sya	***vocaṃ***	
bull.GEN	say.INJ.AOR.1SG	

'Now I proclaim the greatness of the bull [=Indra].'

(2) *Od*. 11.4

en	*dè*	*tà*	*mêla*
in	PTC	DEM.ACC.PL.N	sheep(N).ACC.PL
labóntes		***ebḗsamen***	
take.PTCP.AOR.NOM.PL		walk.AOR.3PL	

'As we have taken the sheep, we stepped into (the ships).'

The original syntactic freedom of PVs, shown by examples (1) and (2), emerges even more clearly when they occur in the so-called "absolute" position, virtually "substituting" a verbal form (cf. section 3).

2.1 PVs Current Annotation in UD

PVs' ambiguous categorial status raises issues for annotation: disambiguating PVs' function is a non-trivial task, even for human annotators. Ideally, the *deprel advmod* should be used for adverbs, *compound:prt* for PVs (even if "detached" from verbs), and *case* for adpositions. Practically, it is often very difficult to distinguish adverbial from preverbal usages in both languages (in the VTB, *advmod* is exclusively used, whereas the HTB employs *compound:prt*). Furthermore, especially in Early Vedic, adpositional usages can be difficult to sort out, and the *case* label is sometimes used in unclear cases. The syntactic annotation of absolute PVs is thoroughly discussed in section 5.

Ambiguities encountered at the syntactic level are mirrored in the assignation of part-of-speech tags. In the HTB, the part-of-speech tag ADV is given to adverbs and preverbs, whereas adpositions are tagged as ADP; the occurrences in which PVs hold an ambiguous status are not annotated in a consistent way. For example, in (3) the PV *ek* '(lit.) out of' occurring in tmesis initial position is annotated as ADP taking the genitive plural *astragálōn*.

[1] https://github.com/OliverHellwig/sanskrit/tree/master/papers/2020lrec/treebank.

[2] https://github.com/francescomambrini/katholou/tree/main/ud_treebanks/agdt/data.

[3] https://perseusdl.github.io/treebank_data/.

(3) *Od.* 11.64-65
ek *dé* *moi* *aukhèn*
out_of PTC 3SG.DAT neck.NOM
astragálōn *eágē*
neck_vertebra.GEN.PL break.AOR.PASS.3SG
'My neck broke from the vertebrae.'

In the quasi-identical passage in *Od*. 10.559-560, the same preverb is annotated as ADV.[4]
In the VTB, ADV is employed in all cases.

3. Preverbs in Absolute Position

PVs in absolute position seem to function as "proxies" for the verb (Chantraine, 1953: 82). The "omitted" verbal form can be recovered either from the previous linguistic material or from the extralinguistic context. For example, in (4) the verb *bhare* 'I bring' can be recovered from the hymn's opening verse *prá vaḥ ... suṣṭutím ... bhare* 'I bring forth to you my good praise'. In (5), instead, a motion verb such as √*gam*- 'go' or *sám* √*car*- 'converge' can be recovered based on similar comparative constructions involving cows moving towards (*abhí*) their calves that occur elsewhere in the RV.

(4) ṚV 2.16.7
prá *te* *nā́vaṁ* *ná*
forward 2SG.DAT boat.ACC like
sámane *vacasyúvam*
assembly.LOC eloquent.ACC
'Within the assembly, (I bring) forth to you my eloquent (formulation), like a boat.'

(5) ṚV 9.86.2
dhenúr *ná* *vatsám* *páyasā*
milk-cow.NOM like calf.ACC milk.INST
abhí *vajríṇam* *índram*
towards with_mace.ACC Indra.ACC
índavo
drop.NOM.PL
'As a milk-cow (goes) to her calf with milk, the drops […] (go) to Indra, possessor of the mace.'

Absolute uses of PVs are found both in independent clauses, such as (4) and (5), and in dependent ones, such as (6); in the latter case, the preverb *epí* "substitutes" for the compound verb *épeimi* 'be there'. Absolute PVs are also frequently found in coordinated clauses: in (7), the first conjunct contains the compound verb *pàr ... etíthei* '(he) placed beside', whereas in the second conjunct the PV alone is repeated (on PV repetition see Dunkel, 1979; Klein, 2007).

(6) *Il.* 1.514
nēmertès *mèn* *dḗ* *moi*
infallible.ACC PTC PTC 1SG.DAT
hypóscheo ... *epeì*
give.IMPV.AOR.2SG.MID CONJ
oú *toi* ***épi*** *déos.*
NEG 2SG.DAT upon fear.NOM
'Give me your infallible promise […] for there (is) nothing to make you afraid.'

(7) *Od*. 8.70
pàr *d'* ***etíthei*** *káneon*
beside PTC place.IMPF.3SG basket.ACC
kalḗn *te* *trápezan,*
beautiful.ACC PTC table.ACC
pàr *dè* *dépas* *oínoio*.
beside PTC cup.ACC wine.GEN
'And beside him he placed a basket and a beautiful table, and a cup of wine.'

4. Data Extraction

In order to analyze the annotations for PVs occurring in absolute position in Homeric Greek and Early Vedic (see Section 5), we implemented two queries to extract from the HTB and the VTB all sentences in which such elements occur. First, we identified the patterns that may involve such lemmas (as discussed in section 3), then, we wrote a Python script[5] to get all the matching sentences from the treebanks. All the functions used to design the queries rely on the Python conllu module.[6]
The target elements of the queries were tokens whose lemmas are included in the list of PVs (see Appendix) and another token, from which the former depended. In the HTB, we looked for tokens whose part-of-speech is NOUN or PRON and that have *advcl* or *conj* as dependency relation. Furthermore, the PV's *deprel* has to be *compound:prt*. In the VTB, we looked for tokens that govern a PV via the relation *orphan*. The token can have any part-of-speech but cannot be a finite verb. As finite verbs lack VerbForm in the conllu *feats* field, if the head of the PV was a verb, we restricted the selection to those for which the VerbForm feature was specified. If it had any other part-of-speech, we included the pattern in the results without checking anything else.

5. Current Annotation of Absolute PVs

UD marks all kinds of ellipses by promoting a member of the elliptical clause to the head position on the base of a "coreness" hierarchy:

(8) *nsubj* > *obj* > *iobj* > *obl* > *advmod* > *csubj* > *xcomp* > *ccomp* > *advcl* > *dislocated* > *vocative*

The promoted member takes the syntactic relation that the elided element would bear; to signal that the dependency structure is incomplete, all non-promoted dependents of the elided elements receive the relation *orphan* (Figure 1; see also Schuster, Lamm, and Manning, 2017 with references).[7]

[4] The passage in *Od*. 10.559-560 is quasi-identical to that in *Od.* 11.64-65 in that it includes the dative *hoi* of the third person pronoun instead of *moi*. This difference is not relevant for the purposes of our analysis.

[5] The analyzed data and the Python (Van Rossum and Drake, 2009) script employed for this study are available at: https://github.com/unipv-larl/preverbs. The scripts were only used to extract the patterns and analyze their annotation. They can be used to extract data from other portions of the VTB and can easily implemented to fix the annotation automatically with the correction proposed in this paper.

[6] https://pypi.org/project/conllu/#description

[7] https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis

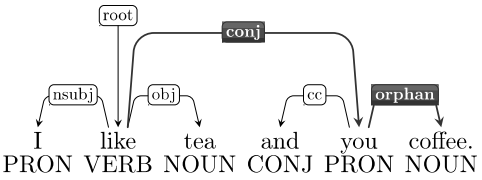


Figure 1: Annotation scheme for verb ellipsis.

In the VTB and HTB, PVs' absolute uses are annotated following the basic ellipsis representation: one argument of the "omitted" verb is promoted to the head position and governs the PV. However, the two treebanks differ as to the relation holding between the PV and the promoted element: in the VTB, the PV takes the *orphan* relation, whereas in the HTB it depends on the promoted noun via the *compound:prt* relation. Compare Figures 2 and 3:

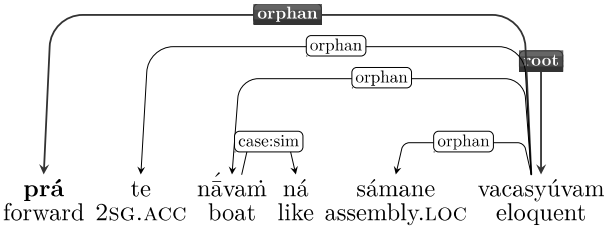


Figure 2: Dependency tree for (1) in the VTB.

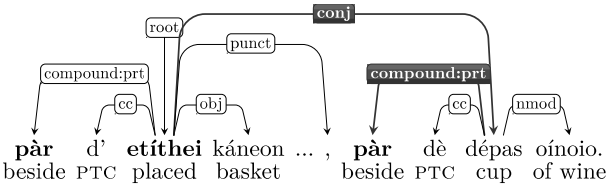


Figure 3: Dependency tree for (4) in the HTB.

In the case of Early Vedic, this annotation scheme derives from the fact that verbal arguments and adjuncts precede adverbial modifiers in the UD promotion hierarchy in (8): since PVs are always annotated as *advmod*, they cannot be promoted to the head position if other verbal arguments or adjuncts are available, given that the latter occupy a higher position in the hierarchy. However, this results in a linguistically unrealistic annotation, since in such constructions it is the PV, and not a verbal argument/adjunct, that "substitutes" for the verb. Furthermore, the variety of constructions in which absolute PVs can occur is lost in the annotation, as the scheme always treats PVs as *orphans*.

Issues related to the promotion hierarchy especially arise when the clause contains an explicit subject: in (9), the coordinated subjects *díphros* 'chariot' and *híppoi* 'horses' are promoted to head position, instead of the PV *pára*, which stands for the compound verb *páreimi* 'be present, or the adjunct *toi* 'for you'.

(9) *Od*. 3.325
ei *d'* *ethéleis* *pezós,* ***pára***
CONJ PTC will.2SG by_land.NOM beside
toi *díphros* *te* *kaì* *híppoi*
2SG.DAT chariot.NOM PTC CONJ horse.NOM.PL
'If you will go by land, here (are) a chariot and horses at hand for you.'

Besides the same problematic head promotion, the HTB shows the additional issue of employing *compound:prt* to tag the relation between the PV and the promoted element: this relation should exclusively be used for idiomatic syntagmatic verbs and not for PVs depending on nouns.[8]

Differently from the VTB, the HTB is enriched with enhanced dependencies which allow for adding empty nodes to represent verb ellipsis. In the case of absolute PVs, the enhanced graph contains an empty node for the "omitted" verb, on which the PV depends via the *compound:prt* relation, as in Figure 4.

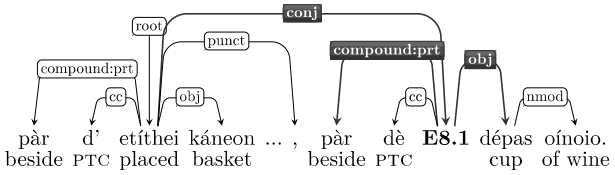


Figure 4: Enhanced graph for (4) in the HTB.

At first sight, the enhanced representation is satisfying, for it allows for representing the different constructions in which absolute PVs can occur (the empty node may be *root*, *advcl*, *acl*, *conj*, among others). However, while in the case of ellipsis in coordination there is general agreement on the need to recover a verbal form, it is not clear that the same holds for non-coordinative contexts. In some languages such as spoken Russian, the overt expression of motion via a motion verb is unnecessary in many contexts, as shown in (10) (Zanchi, 2019: 106). Similarly, in German, the combination of a modal verb such as *müssen* and a prepositional phrase such *as in die Stadt* 'into town', shown in (11), can express motion without any motion verb; example (11) is taken from the UD version of the Hamburg Dependency Treebank (HDT),[9] where the modal verb is promoted to head position (Figure 5).

(10) a. *Ty* ***kuda****?*[10]
'Where (are) you (going)?'
b. *V metro*[11]
'To the subway.'

(11) *Was tut die Kleinfamilie, wenn sie … nicht mehr* ***in die Stadt muss****?*
'What does the small family do when […] they no longer have to (go) into town?'

[8] https://universaldependencies.org/u/dep/compound-prt.html
[9] https://github.com/UniversalDependencies/UD_German-HDT/blob/master/README.md
[10] Vasin, Ivan Švedov, Muž, 37, 1969, http://www.ruscorpora.ru/en/
[11] Sergej, Sergej Puskepalis, Muž, 40, 1966, http://www.ruscorpora.ru/en/

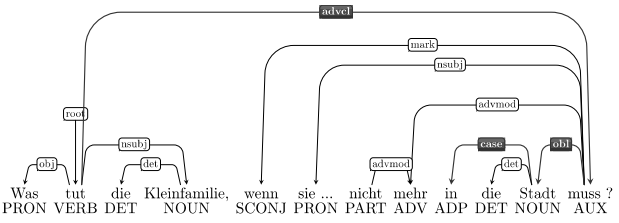


Figure 5: Dependency tree for (11) in the HDT.

The same argument can be used for cases such as (5), in which the Vedic PV *abhí* alone might express directed motion without necessarily assuming a verbal form. Finally, in copular sentences, such as (6) and (9), the need to recover a verbal form is even less straightforward. While overt copulas are optional in Homeric Greek and Early Vedic, in many languages the relation between a subject and a nominal predicate systematically lacks overt marking (Stassen, 2013).

6. Annotation Proposal

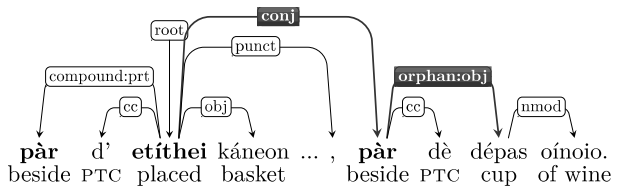
Considering the above discussion, we suggest the following annotation scheme for absolute uses of PVs in the HTB and VTB:

a) Verb ellipsis in coordinative contexts:

If the treebank is enriched with enhanced dependencies, empty nodes should be substituted for the elided verb. The PV should depend on the empty node via *compound:prt* (cf. the annotation shown in Figure 4).

If enhanced dependencies are not employed, as in the VTB, PVs should be promoted to the role of head, instead of verbal arguments. Note that *compound(:prt)* is not included in the coreness hierarchy presented in Section 5 and thus preverbs taking this relation are not considered possible candidates for promotion to the head position. However, since absolute PVs convey most information on the kind of motion event expressed in the sentence, they would be better substitutes for elided verbs than subjects or other dependents in the hierarchy.

Accordingly, all verb arguments should depend on the promoted PV via the *orphan* relation. In order to retain syntactic information on the type of argument of each orphaned element, we suggest adding sub-relations such as *:subj*, *:obj*, or *:obl*.[12] See Figure 6, based on example (7):



[12] This proposal resembles the one developed by Joakim Nivre and Daniel Zeman as part of the discussion of the second version of the UD guidelines (see Schuster, Lamm, and Manning, 2017: 130-131). To retain syntactic information on each remnant, Nivre and Zeman suggested employing composite relations of the type `conj>subj`, `conj>obj`, etc. Our proposal is moved by the same intention but exploits UD extensions to ordinary dependency relations. This allows users to decide whether to include sub-relations or not when querying the treebanks. Note that sub-relations would be especially useful for Homeric Greek where the same case form can fulfill different syntactic relations.

Figure 6: Suggested basic graph for ellipsis in coordination.

b) Verb ellipsis in non-coordinative contexts:

Regardless of whether the treebank contains enhanced dependencies or not, we suggest promoting the PV to the head position (*root*, *advcl*, *xcomp*, etc.), without the mediation of empty nodes. Other elements should depend on the promoted PV via *orphan* specified by the relevant sub-relation. See Figure 7, based on example (4):

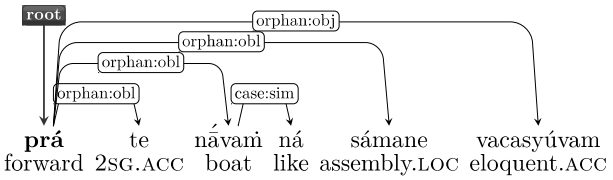


Figure 7: Suggested graph for ellipsis outside of coordination.

c) Zero copula for predicate nominals:

Finally, PVs in zero copula constructions should also be treated as the heads of the construction. In this case, however, the subject depends on the PV via *subj*, as in ordinary copula constructions. Compare Figure 8, where the PV *pàra* functions as *root* (see ex. (9)), with Figure 9, where the adjective *hatró(os)* is the *root*.

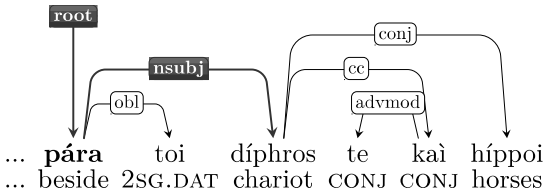
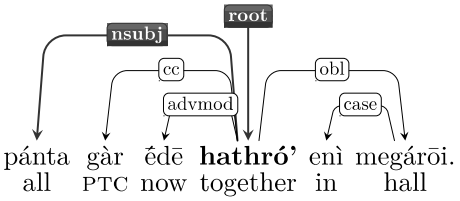


Figure 8: Suggested annotation for zero copula constructions with PV = *root*.



'For all are now gathered together in the hall.' (*Od*. 2.410)

Figure 9: UD annotation scheme for zero copula constructions with Adj = *root*.

7. Conclusion and future perspectives

After discussing the current syntactic annotation of Early Vedic and Homeric Greek PVs in absolute position, in this paper we proposed a new annotation scheme for these constructions.

The proposed annotation has multiple advantages:

- it does not level out the variety of constructions in which absolute PVs occur;
- it uses annotation practices that are already part of UD guidelines;

- it does not add empty nodes where unnecessary, thus making the subsequent evaluation of the data easier;
- it keeps all the syntactic information in the annotation thanks to the addition of sub-relations;
- it can easily be extended to similar constructions of other ancient and modern languages.

This paper documents the first step toward the larger goal of systematizing the annotation of PVs occupying other syntactic positions through semiautomatic methods. Such systematization would represent a major improvement for the UD treebanks of ancient Indo-European languages: it would coherently account for the variety of constructions in which these uninflected items occur, thus facilitating (cross-linguistic) research upon them. The interest in PVs goes beyond Indo-European studies, crosscutting grammaticalization studies and lexical typological studies in the expression of motion events.

Acknowledgments

We wish to thank Oliver Hellwig and the three anonymous reviewers who provided insightful feedback for a substantial improvement of the final version of this paper. Final responsibility remains our own.

8. Bibliographical References

Chantraine, P. 1953. *Grammaire homérique*. Tome 2: Syntaxe. Paris: Klincksieck.

Dunkel, G.E. 1979. Preverb repetition. *Münchener Studien zur Sprachwissenschaft* 38: pages 41–82.

Hellwig, O., Scarlata, S., Ackermann, E., and Widmer, P. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of The 12th Language Resources and Evaluation Conference* (LREC 2020), pages 5137–146. http://www.lrec-conf.org/proceedings/lrec2020/index.html.

Klein, J.S. 2007. On the Nature and Function of Preverb Repetition in the Rigveda. *Studien zur Indologie und Iranistik* 24: pages 91-103.

Schuster, S., Lamm, M., and Manning, C.D. 2017. Gapping constructions in universal dependencies v2. *In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies* (UDW 2017), pages 123–132.

Stassen, L. 2013. Zero Copula for Predicate Nominals. In *The World Atlas of Language Structures Online*, Dryer, Matthew S. and Haspelmath, Martin (eds.), Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/120, Accessed on 2021-11-12.

Van Rossum, G. and Drake, F.L. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Zanchi, C. 2019. *Multiple preverbs in ancient Indo-European languages*. Tübingen: Narr.

Appendix: List of PV lemmas

Early Vedic: *apa*, *ava*, *ā*, *ud*, *ni*, *nis*, *parā*, *puras*, *pra*, *sam*, *vi*, *achā*, *ati*, *adhi*, *anu*, *antar*, *api*, *abhi*, *upa*, *tiras*, *paras*, *pari*, *puras*, *purā*, *prati*.[13]

Homeric Greek: *amphí*, *aná*, *antí*, *apó*, *diá*, *en*, *eis*, *ek*, *epí*, *hypér*, *hypó*, *katá*, *metá*, *pará*, *perí*, *pró*, *prós*, *sýn*.

[13] Note that, differently from later Vedic and Classical Sanskrit, Early Vedic texts contain word accents (*ápa*, *áva*, *ā́*, etc.). However, in order to lemmatize words attested in Early Vedic text together with those attested in later Vedic or Classical Sanskrit, the digitized text of the *Ṛgveda* employed by the VTB does not contain word accents.