Annotation Guidelines for Ṛgvedic Similes and Related Constructions

This document contains the annotation guidelines for Ṛgvedic similes and related constructions, and constitutes an appendix to the book *Rigvedic Similes: a corpus-based analysis of thheir forms and functions*.

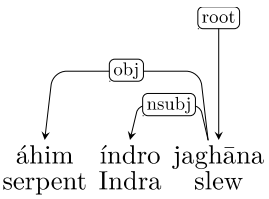
The ṚV contains about 2400 similes. For this study, all similes have been syntactically annotated, together with other constructions employing one of the three comparative particles; these include the approximative use of *ná*, and *iva*, amounting to 61 occurrences, and 87 comparative clauses introduced by *yáthā*. The annotation has been carried out according to the Universal Dependency scheme, with some language-specific modification introduced during the creation of the Vedic Treebank.

The Vedic Treebank

The Vedic Treebank (VTB; Hellwig et al. 2020, Biagetti et al. 2021, Hellwig, Nehrdich, and Sellmer 2023) contains selected passages from texts of Vedic literature, syntactically annotated according to the Universal Dependencies standard. The treebank is hosted within the *Digital Corpus of Sanskrit*[1] (Hellwig 2010-2021), which provides *sandhi* splits and morphosyntactic annotations alongside the raw source texts. The annotation is performed directly in the web interface of the *Digital Corpus of Sanskrit* which features a supportive, trainable machine learning classifier (see Hellwig et al. 2020 for details).[2] In the VTB, about a quarter of the extant Vedic corpus has been manually annotated. This highly accurate annotation, referred to as gold annotation, was used to train a parser, which was then fed with the remaining portions of Vedic texts (Hellwig, Nehrdich, and Sellmer 2023).

Universal Dependencies

Universal Dependencies (UD; Nivre et al. 2016) is a project that is developing cross-linguistically consistent treebank annotation for many languages.[3] Syntactic annotation in the UD scheme consists of typed dependency relations (*deprel*) between words. The basic representation forms a tree, where exactly one word is the head of the sentence depending on a conventional root and all the other words depend on exactly one word. A simple dependency tree for the sentence *áhim índro jaghāna* 'Indra slew the serpent' (ṚV 2.15.1d) is given in Figure 1.



'Indra slew the serpent.'

Figure 1. Dependency tree for ṚV 2.15.1d.

[1] http://www.sanskrit-linguistics.org/dcs/.

[2] http://www.sanskrit-linguistics.org/dcs/index.php?contents=texte.

[3] The latest version (2.13, released November 15, 2023) includes 259 treebanks of 148 languages.

The inventory of relations is given in Table 1:

Table 1. Universal Dependency relations.[4]

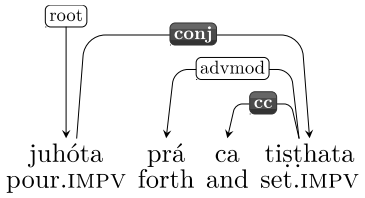
	Nominals	Clauses	Modifier words	Function words
Core Arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl* dislocated	advcl	advmod discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf* case
Coordination	MWE[5]	Loose	Special	Other
conj cc	fixed flat compound	list* parataxis	orphan goeswith* reparandum	punct* root dep

The following principles are observed in the annotation to maximize parallelism while accounting for differences between languages. Dependency relations hold primarily between content words, rather than being mediated by function words (*primacy of content words*).[6] Thus, case-marking elements like prepositions, postpositions, and clitic case markers are treated as dependents of the nouns they attach to or introduce (case). Coordination follows a similar treatment, with the leftmost conjunct as the head, and other conjuncts as well as the coordinating conjunction depending on it via conj and cc respectively (Figure 2). Finally, auxiliary verbs (aux) and copulas (cop) are not the head of the clause but depend on the lexical predicate (Figure 3).

[4] *Deprels* marked with an asterisk are not employed in the annotation of the VTB. Abbreviations: acl 'adjectival clause' (clause modifier of noun), advcl 'adverbial clause modifier', advmod 'adverbial modifier', amod 'adjectival modifier', appos 'apposition', aux 'auxiliary', case 'case marking', cc 'coordinating conjunction', ccomp 'clausal complement', clf 'classifier', compound, conj 'conjunct', cop 'copula', csubj 'clausal subject', dep 'unspecified dependency', det 'determiner', discourse 'discourse element', dislocated 'dislocated element', expl 'expletive', fixed 'fixed multiword expression', flat 'flat multiword expression', goeswith 'goes with', iobj 'indirect object', list, mark 'marker', nmod 'nominal modifier', nsubj 'nominal subject', nummod 'numeral modifier', obj 'object', obl 'oblique', orphan 'orphaned dependent', parataxis, punct 'punctuation', reparandum 'overridden disfluency', root, vocative, xcomp 'open clausal complement'.

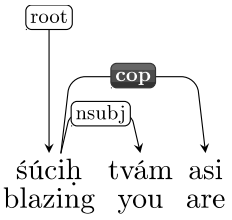
[5] Multi-word expression: sequences of words which, in different degrees, behave as a single lexical unit. Three kinds of them are distinguished and given three different relations: fixed links together grammaticalized sequences of function words, flat is used to analyze exocentric expression like names, titles, and honorifics, whereas compound is used for (mostly endocentric) compounded words. Considered the peculiarities of compounding in Vedic, the compound relation is used differently in the VTB (see Hellwig et al. 2020; cf. also Biagetti 2018).

[6] This decision is motivated by the observation that marking relations between content words maximizes parallelism. Indeed, the same grammatical relation can be expressed by morphology in some languages or constructions and by function words in other languages or constructions, while some languages may not mark the information at all.



‘Pour and set it forth.’ (ṚV 1.15.9b)

Figure 2. Coordination.



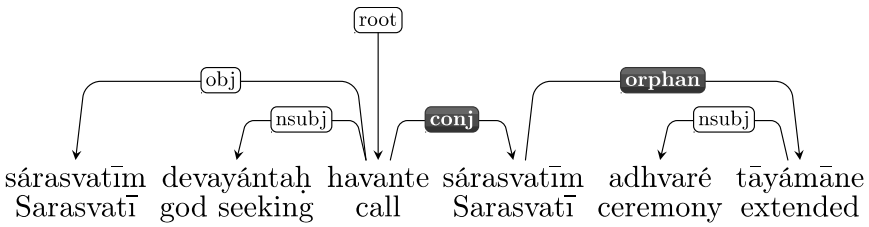
‘You are blazing pure’ (ṚV 9.88.8)

Figure 3. Copula annotation.

In UD, the treatment of central dependency relations between content words is based on the distinction between core arguments (subjects, objects, clausal complements) and obliques. Even if the major role of syntactic analysis is to represent function, the scheme also provides for some structural analysis, distinguishing between a) nominal phrases, b) clauses headed by a predicate, and c) different kinds of modifier words. This distinction is clearly encoded in dependency labels. For example, if a verb is taking an adverbial modifier, it may bear one of three relations a) `obl`, b) `advcl`, or c) `advmod` depending on which of the three categories above it belongs to. In the same way, the core grammatical relations differentiate core arguments that are clauses, such as `csubj` or `ccomp`, from those that are nominal phrases, such as `nsubj` and `obj`.

Within clausal dependents, UD does not distinguish between finite and non-finite clauses. Rather, a distinction is made between clausal dependents that feature obligatory control (`xcomp`) and those that do not (`ccomp`) as well as between clausal subjects and adverbs (`csubj`, `advcl`), which have verbal attachment, and clausal modifiers of nouns (`acl`, `acl:rel`).

The principle of the primacy of content words has consequences on the annotation of ellipsis. Differently from other formalisms based on dependency grammar, such as the PROIEL[7] scheme (Haug and Jøhndal 2008), UD does not make use of empty nodes to represent ellipsis or gapping, but marks all kinds of ellipsis by promoting a member of the elliptical clause to the head position on the base of a “coreness” hierarchy.[8] The promoted member takes the syntactic relation that the elided element would otherwise bear; to signal that the dependency structure is incomplete, all non-promoted dependents of the elided element receive the relation `orphan`. Take for instance Figure 4, which represents the treatment of ellipsis in coordination: as a consequence of the elision of the verb *havante* ‘they call’ in the second conjunct, the object *sarasvatīm* ‘Sarasvatī’ is promoted to the head position of the coordinate clause (`conj`), whereas the adjunct *tāyamāne* depends on it via the relation `orphan`.



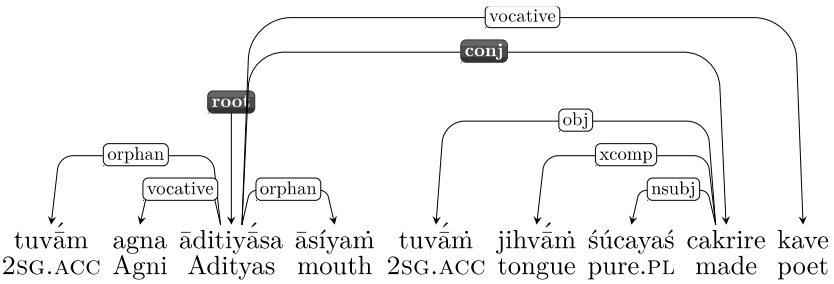
‘Sarasvatī do those seeking the gods invoke, Sarasvatī when the ceremony is being extended.’ (ṚV 10.17.7)

Figure 4. Annotation scheme for verb ellipsis.

[7] http://dev.syntacticus.org/proiel.html#downloads.

[8] Orphaned dependents are considered for promotion in the following order: `nsubj` > `obj` > `iobj` > `obl` > `advmod` > `csubj` > `xcomp` > `ccomp` > `advcl` > `dislocated` > `vocative`.

In the case of leftward gapping, the dependent which in the first conjunct has the highest rank is promoted to the "new-head" position, while the second conjunct, i.e. the one bearing the verb, is connected to the new head via `conj` and does not require any `orphan` relation. This is shown by Figure 5, where the subject *ādityā́saḥ* 'the Adityas' is promoted to the root position, the object *tuvā́m* 'you' and the predicative *āśíyam* 'mouth' take the relation `orphan`, whereas the verb *cakrire* 'they made' is linked to the root via `conj`.



'The Adityas made you their mouth, o Agni, the pure ones made you their tongue, o poet!' (ṚV 2.1.13ab)

Figure 5. Annotation scheme for leftward gapping in coordination.

Annotation scheme for similes and related constructions

UD guidelines provide annotation schemes for basic equatives and for clausal ones. In the former, the standard is linked to the parameter via the relation `obl`, while the standard marker depends on the standard via the relation `case` (Figure 6). In clausal comparison, the verb of the comparative clause is attached to the main verb through the relation `advcl`, the standard marker depending on it via `mark` (Figure 7).

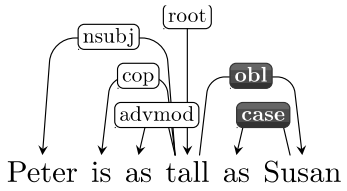


Figure 6. Annotation scheme for basic comparatives.

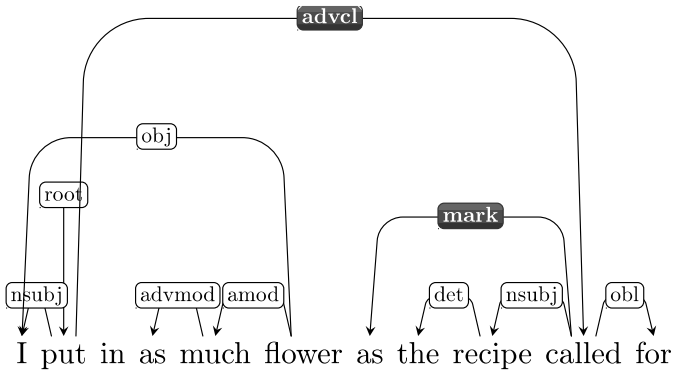


Figure 7. Annotation scheme for clausal comparatives.

The annotation of gapping structure in comparative clauses is mentioned in the report of a working group dedicated to comparative constructions.[9] The report provides the sentence in Figure 8 as an

[9] https://universaldependencies.org/workgroups/comparatives.html.

example of gapping in comparative clauses and suggests analyzing such comparative gapping using the `orphan` relation, much like the more widespread gapping in coordination.

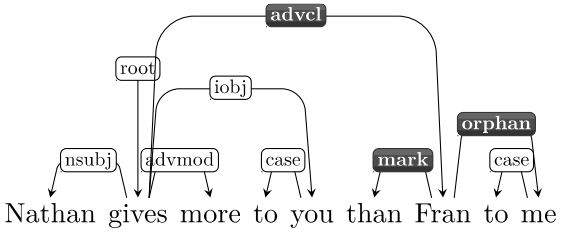


Figure 8. Annotation scheme for gapping in comparison.

Ṛgvedic similes introduced by the particles *ná*, *iva*, and *yáthā*/*yathā* systematically lack a verb in the standard clause. From a descriptive point of view (i.e. for the purposes of annotation) it is useful to analyze simple similes as cases of verb ellipses in which the promoted element has no dependents, and double and or triple similes as cases of gapping, in which the second remnant is attached to the promoted one with the relation `orphan`.

In UD, there are no relations designed specifically to mark equative and similative constructions. First of all, UD employs the same scheme for equality and inequality comparison, as shown by the annotation of the two comparative constructions in Figure 9. Furthermore, phrasal comparatives are simply assimilated to other obliques (`obl`), whereas comparative clauses are treated in the same way as other adverbial clauses (`advcl`). Similarly, standard markers take the same *deprel* as other function words such as adpositions (`case`) and subordinating conjunctions (`mark`). Take for instance the two trees in Figure 10, where the clausal comparative contained in the first sentence takes the same labels as the temporal clause contained in the second; in fact, the same annotation scheme is employed for adverbial clause modifiers of all types.

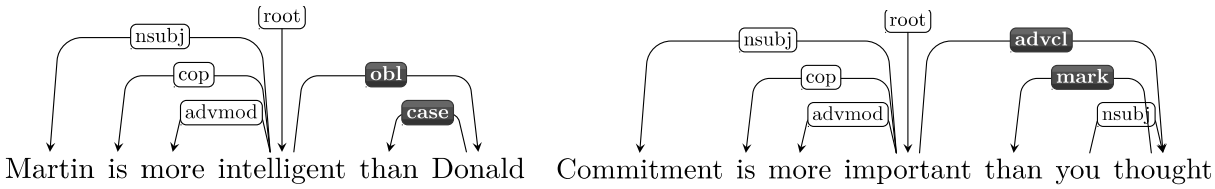


Figure 9. Annotation scheme for basic and clausal comparison of inequality.

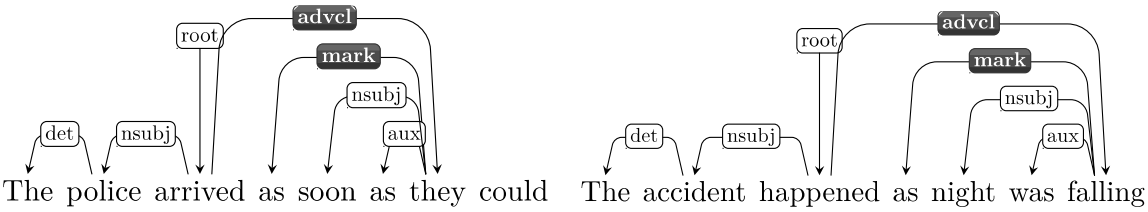


Figure 10. Annotation scheme for adverbial clause modifiers. Left: comparison; right: temporal clause.

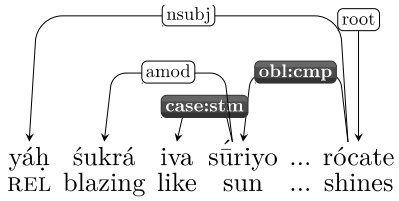
Since the particles *ná*, *iva*, and *yáthā*/*yathā* have other functions beside that of standard marker, and since comparison is also expressed by other strategies, it was necessary to increase the informativeness of the annotation to be able to make granular and targeted queries on different types of constructions.

In order to represent the syntax of similes in detail, the VTB distinguishes the following subtypes of comparative constructions:

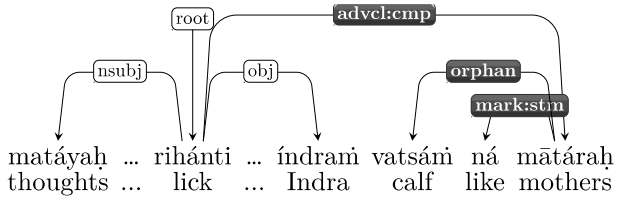
Table 2. Comparative constructions of equality with their respective annotation.

CONSTRUCTION	EXAMPLE	ANNOTATION (dependent → `relation` → head)
PREDICATIVE SIM.	'Agni is like the sun.'	*like* → `case:stm` → *sun* *Agni* → `nsubj` → *sun*
SIM. WITH ELLIPSIS	'Agni shines like the sun.'	*like* → `case:stm` → *sun* → `obl:cmp` → *shines*
SIM. WITH GAPPING	'Thoughts lick Indra like mothers a calf.'	*like* → `mark:stm` → *mothers* → `advcl:cmp` → *lick*; *calf* → `orphan` → *mothers*
CLAUSAL SIM.	'Just as you drank the previous soma drinks, so take a drink today.'	*as* → `mark` → *drank* → `advcl:cmp` → *drink*; *previous drinks* → `obj` → *drank*; *so* → `advmod` → *drink*

As shown by Table 2, the VTB formally distinguishes similes with ellipsis (annotated with `obl` and `case`) from similes with gapping (annotated with `advcl` and `mark`).



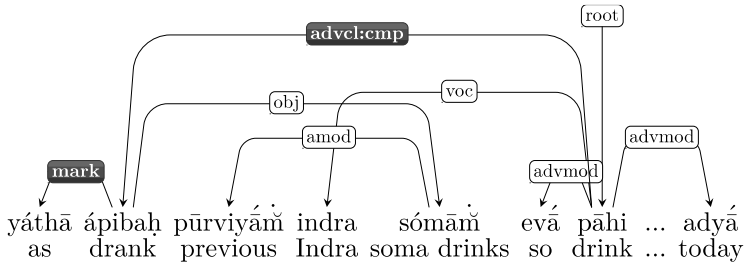
'He (Agni) who shines like the blazing sun […].' (ṚV 1.43.5ab)
Figure 11. Annotation scheme for similes with ellipsis.



'Thoughts lick Indra […] like mothers a calf.' (ṚV 3.41.5)
Figure 12. Annotation scheme for similes with gapping.

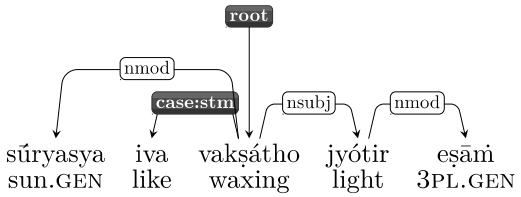
In addition to the universal dependency taxonomy, UD allows the employment of language-specific extensions that capture peculiar constructions found in a given language or in a group of languages. These extensions are regarded as subtypes of existing UD relations and have the format `universal:extension`: for instance, `obl:manner` stands for the language-specific `manner` extension of the UD relation `obl`. In the VTB, the sublabel `:cmp` added to the relations `obl` and `advcl` allows distinguishing standards of comparison from other kinds of adverbial modifiers. Furthermore, the sublabel `:stm` ('standard marker') attached to the relations `case` and `mark` allows the user to easily retrieve all particles that introduce basic similes and to distinguish them from those

that introduce clausal similes (which take `mark` alone). Compare for instance the annotation of basic similes like those in Figure 11 and Figure 12 with that of a clausal simile like the one in Figure 13:

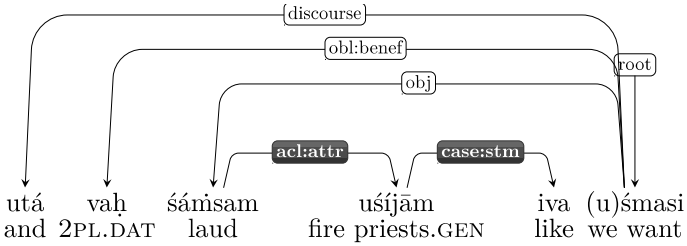


'Just as you drank the previous soma drinks, Indra, so take a drink today […].' (ṚV 3.36.3cd)
Figure 13. Annotation scheme for clausal similes.

Constructions characterized by ellipsis and gapping that were presented above do not exhaust the forms that Ṛgvedic similes can take. Thanks to the annotation scheme presented above, predicative similes are easily retrieved from the corpus by looking for all those constructions whose head does *not* take any of the `obl:cmp` or `advcl:cmp` relations, but nevertheless governs one of the three particles via the relation `case:stm` or `mark:stm`. As Figure 14 and Figure 15 show, this query returns both similes constituting the sentence main predication (in which case their head is the `root`), and those that function as a secondary predicate (in which case the relation of the head is variable).



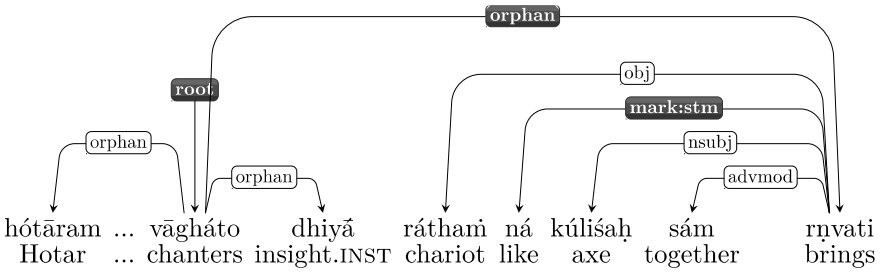
'Their light is like the waxing of the sun.' (ṚV 7.33.8a)
Figure 14. Predicative simile, head = `root`.



'And we want a laud for you (that is) like that of the fire-priests.' (ṚV 2.31.6a)
Figure 15. Predicative simile, head = secondary predicate (`acl:attr`).

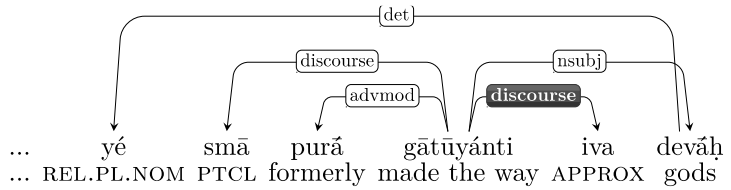
Sometimes, word order and especially verbal agreement suggest that the verb is exceptionally constructed with the standard of a similes rather than with the comparee. As shown by Figure 16, these cases are also captured by the annotation scheme. In this example, we would expect a plural verb **sám ṛṇvanti* in agreement with the plural nominative *vāghátas* 'chanters'; on the contrary, the verb *sám ṛṇvati* 'bring'.PRS.3SG agrees with the singular nominative *kúliśaḥ* 'axe' which constitutes the standard of the simile. As a whole, the sentence is treated similarly to a case of leftward gapping in coordination (see Figure 5): as a consequence of the lack of the verb in the comparee, the subject

vāghátas is promoted to the `root` position, while the object *hótāram*, the adjunct *dhiyā́*, as well as the verb contained in the standard depend on it via the relation `orphan`.



'As an axe brings together a chariot, the chanters (bring together) with their insight the Hotar.' (ṚV 3.2.1cd)
Figure 16. Annotation of similes whose verb is constructed with STAND.

Finally, when *iva* e *ná* function as approximation markers, they depend on their head via the relation `discourse`, which in UD is reserved to discourse markers. For instance, in Figure 17 *iva* modifies the verb and depends on it via the relation `discourse`.



'… the gods who up till now have provided the way, as it were.' (ṚV 1.169.5d)
Figure 17. Annotation of *iva* as approximation marker (`discourse`).

References

Biagetti, Erica. 2018. “A Dependency Treebank of Classical Sanskrit.” MA thesis, University of Pavia.

Haug, Dag T. T. and Marius L. Jøhndal. 2008. “Creating a Parallel Treebank of the Old Indo-European Bible Translations.” In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data* (LaTeCH 2008), edited by Caroline Sporleder and Kiril Ribarov, 27–34.

Hellwig, Oliver. 2010-2021. *The Digital Corpus of Sanskrit*. http://www.sanskrit-linguistics.org/dcs/index.php.

Hellwig, Oliver, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The Treebank of Vedic Sanskrit. In Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al. (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference* (LREC 2020), 5137–5146.

Hellwig, Oliver and Sven Sellmer. 2021. The Vedic Treebank. In Erica Biagetti, Chiara Zanchi, and Silvia Luraghi, *Building New Resources for Historical Linguistics*. Pavia: Pavia University Press.

Hellwig, Oliver, Sebastian Nehrdich, and Sven Sellmer. “Data-driven dependency parsing of Vedic Sanskrit.” *Language Resources and Evaluation* (2023): 1–34.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, et al. 2016. “Universal Dependencies V1: A Multilingual Treebank Collection.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC’16), edited by Johann-Mattis List, Michael Cysouw, Robert Forkel, and Nicoletta Calzolari, 1659–1666. European Language Resources Association (ELRA).