The *RigVeda* goes “universal”: annotation and analysis of equative constructions in Vedic and beyond

**Erica Biagetti**  
University of Pavia  
[erica.biagetti01@universitadipavia.it](mailto:erica.biagetti01@universitadipavia.it)

**Abstract**

By presenting a case study on Rigvedic equative and simulative constructions, this paper demonstrate[s] that treebanks constitute an important support for research in historical linguistics for two main reasons. First, by providing quantitative evidence on linguistic phenomena, they can confirm or dismiss hypotheses formulated on the base of qualitative data. Second, by capturing correlations among linguistic phenomena which could hardly be grasped by linguists’ naked eye, treebank-based analyses allow scholars to formulate new hypotheses. Since an analysis of Rigvedic equative constructions calls for a granular and informative annotation scheme, the Vedic Treebank implements the UD scheme for equative constructions with sub-relations; while some such extensions were specifically designed for a study on Rigvedic similes, others might be adopted by every treebank developer interested in representing equative strategies.

1 **Introduction**

Historical linguistics has always relied on collections of written texts, i.e., corpora, which constitute the only source of evidence available for ancient languages. Annotated corpora revolutionized historical linguistics because they allow scholars to automatically retrieve large quantitative evidence on linguistic phenomena whose account has been previously based on qualitative evidence and to capture correlations among them which could hardly be grasped by linguists’ naked eye (Eckhoff et al., 2018: 303; Biber, 2009; Anthony, 2013). Furthermore, morphosyntactically annotated corpora require automatic data selection through explicit query expressions, crucially making historical linguistic research replicable (Haug, 2015).

By presenting a case study on Rigvedic equative and simulative constructions, in this paper I provide further evidence for the relevance of treebanks for the study of ancient languages. The *Rigveda* (RV) is a collection of 1028 hymns, dating back to the second half of the second millennium BCE (Witzel, 1995), which constitutes the oldest layer of Vedic literature and whose language is strongly conditioned by the poetic and ritual character of the text. The division of the collection into ten books reflects the internal chronology of the work. The core of the collection and its oldest part are books II to VII (the so-called “Family Books”), whereas book X is the most recent. Books I, VIII, and IX are generally younger than the Family Books.

The Rigvedic treebank was created as part of the larger Vedic Treebank (VTB; Hellwig et al., 2020; Biagetti et al., 2021), a corpus of selected passages from Vedic Sanskrit literature syntactically annotated according to the Universal Dependency (UD) standard. The VTB is maintained within the Digital Corpus of Sanskrit, which provides a web-based interface for collaborative dependency annotation. A first version of the treebank was published in occasion of the release of UD version 2.6

---

1 Although the UD standard covers most of the syntactic phenomena found in Vedic texts, some constructions require special attention during annotation and their annotation scheme within the VTB may deviate slightly from the official UD scheme (see the annotation guidelines available at: https://github.com/OliverHellwig/sanskrit/tree/master/papers/2020lrec/paper). While some such deviations were removed in occasion of the treebank release within the UD platform, others remain and are fully documented in Hellwig et al. (2020).

2 [http://www.sanskrit-linguistics.org/dcs/index.php?contents=texte](http://www.sanskrit-linguistics.org/dcs/index.php?contents=texte)
(15 May 2020); a new version, revised and considerably expanded, is currently under development (Hellwig and Sellmer, forthc.).

The case study presented in this paper is part of a project devoted to the study of Rigvedic similes. Similes, which are the most frequent trope found in the RV, are explicit comparative constructions that owe their figurative meaning to the fact that the compared entities are felt as being fundamentally unlike each other, and therefore unlikely to be compared (Israel et al., 2004). While the language of the RV disposes of different strategies for the encoding of comparison, equative and similative constructions introduced by the particles ná ‘as, like’, iva ‘as, like’ and yáthā (lyáthā) ‘as, like’ have specialized for the encoding of figurative comparison. The aim of this paper is to demonstrate that a treebank-backed study on the syntax of these constructions allows us not only to understand their synchronic distribution, but also to confirm previous hypotheses on their origin and development, as well as to formulate new ones. Such a study calls for a granular and informative annotation scheme, which is able to capture the different strategies employed in the RV for the expression of comparison of equality; therefore, a second, major purpose of this paper is to present a new annotation scheme based on the UD standard for comparative constructions implemented with sub-relations.

The paper is organized as follows: Section 2 introduces the main strategies employed in the RV for the encoding of comparison of equality, among which we find similes introduced by ná, iva, and yáthā. After summarizing UD guidelines for the annotation of equative constructions (3.1), Section 3.2 introduces the implemented annotation scheme adopted by the VTB for the analysis of such constructions. Section 4.1 shows that quantitative data extracted from the treebank can provide interesting insights about the syntax and origin of Rigvedic similes. Section 5 suggests extending part of the enhanced scheme to other languages and constructions. Section 6 contains the conclusions.

2 Comparison of equality in the RV

Equative and similative constructions encode similarity between a comparee (CPREE) and a standard (STAND) with respect to some action or property, called parameter (PAR), and by means of a standard marker (STM; Haspelmath and Buchholz, 1998; Treis, 2017). While equative constructions encode quantitative comparison of equality (e.g. Peter is as tall as Susan), similative constructions encode qualitative comparison, or comparison of manner (e.g. Peter runs like a hare.)

In the RV, constructions introduced by the STMs ná, iva, and yáthā constitute the main strategy for the encoding of comparison of equality. They are characterized by systematic ellipsis of the verb in the STAND and by case transparency (Haspelmath and Buchholz, 1998: 307), i.e., identity of case and function between CPREE and STAND (Bergaigne 1887; Jamison 1982; Pinault 1997). Quantitative and qualitative comparison are encoded by the same constructions and are therefore nearly impossible to distinguish (henceforth: equatives). Rigvedic equatives occur in three main configurations of CPREE(s) and STAND(s). Single equatives can take an adjectival predicate as PAR or a verbal one, as in (1).³

(1) ví śloka etu pathyā iva sūrēḥ  
LP signal_call.NOM go.IMPV.3SG pathway.NOM like patron.GEN
PAR= CPREE= PAR STAND STM -CPREE

‘Let the signal-call of the patron go forth afar like a pathway.’⁴ (RV 10.13.1)

Double equatives are characterized by the presence of two parallel elements in the CPREE and in the STAND, and thus have a gapping structure (2). Less often, equatives may be triple, with CPREE and STAND consisting of three elements each.

(2) matāyāḥ rihánti … īndraṁ iva vatsāṁ ná mātāraḥ  
thought.NOM.PL lick.PRS.3PL Indra.ACC calf.ACC like mother.NOM.PL
CPREE₁= PAR -CPREE₂= STAND₁= STM STAND₂=

³ In glosses, the nominal number is specified only if it is plural or dual while gender is specified only if it is feminine or neuter (singular and masculine are not indicated). Among verbal categories, indicative mood and active voice are not indicated.
⁴ Translations of Rigvedic passages are taken from Jamison and Brereton (2014).
‘Thoughts lick … Indra like mothers a calf.’ (RV 3.41.5)

Besides being employed in syntagmatic comparison, the accented particle yāthā also introduces comparative clauses, whose main clause often contains a correlative adverb such as evā ‘so, in this way’ in (3). Note that the difference between clausal and syntagmatic comparison is not limited to the presence vs. absence of a verb: while in the former yāthā functions as a subordinator and occurs in clause-initial position, in the latter yāthā (with its unaccented variant yathā), ná, and iva have a clitic behavior and follow the STAND.

(3) yāthā jaghāntha dhṛṣatā purā cid

like smite.PF.2SG boldly before PTCL

evā jahi śātrum asmākam indra

so smite.IMPV.2SG rival.ACC 1PL.GEN Indra.VOC

‘Just as you also smote boldly before, so smite our rival, o Indra.’ (RV 2.30.4cd)

Finally, comparison of equality can be expressed in the RV by a number of other constructions, including comparative compounds as in (4), adjectives meaning ‘same’ (samá-), or less grammaticalized strategies involving a verb whose meaning is ‘reach’ (“reach equatives” in Haspelmath et al., 2017), as in (5). For comparison and gradation in Vedic, see Kulikov (2021).

(4) agnī-bhrājaso vidyūto gābhastiyo

fire-flash.NOM.PL lightning_bolt.NOM.PL fist(M/F).LOC.DU

STAND-PAR CPREE

‘Lightning bolts flashing like fire (are) in your fists.’ (RV 5.54.11c)

(5) nāki ṭāṁ kārmanā naśan

no_one 3SG.ACC ritual_work.INST reach.SUBJ.AOR.3SG

CPREE STAND-STM PAR PM

‘No one can equal [lit. reach] him (Agni) in his ritual work.’ (RV 8.31.17)

3 Annotating Rigvedic similes

3.1 UD annotation scheme for equative constructions

UD guidelines provide annotation schemes for both basic and clausal equatives. In the former, the standard is linked to the parameter via the relation obl, while the standard marker depends on the standard via case (Figure 1). In clausal equatives, the verb of the comparative clause is attached to the main verb through advcl, the standard marker depending on it via mark (Figure 2).

![Figure 1. Basic equatives.](https://universaldependencies.org/u/overview/specific-syntax.html#comparatives)

![Figure 2. Clausal equatives.](https://universaldependencies.org/u/overview/specific-syntax.html#comparatives)

Gapping occurring in comparative constructions is treated in the same way as coordinate gapping. Thus, in the Swedish equative in (6), the promoted element Joakim takes the relation that the elided verb would otherwise bear (advcl), tennis takes the orphan relation, and the standard marker än, being a functional element, retains its relation mark (Figure 3).

(6) Dan spelar badminton bättre än Joakim tennis

Dan play.PRS badminton better than Joakim tennis

5 https://universaldependencies.org/u/overview/specific-syntax.html#comparatives
‘Dan plays badminton better than Joakim (does) tennis.’

![Diagram](https://universaldependencies.org/workgroups/comparatives.html)

**Figure 3.** Annotation scheme for gapping in comparison.

### 3.2 Extending the scheme: language-specific relations

In UD, there are no relations designed specifically to mark equative constructions. First, UD adopts the same scheme for equality and inequality comparison. Furthermore, basic comparatives are simply assimilated to other obliques (obl), whereas clausal equatives 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 4, where the clausal comparative contained in the first sentence takes the same labels as the temporal clause contained in the second.

![Diagram](https://universaldependencies.org/docs/u/dep/advcl.html)

**Figure 4.** UD scheme for adverbial clause modifiers. L: comparative clause; R: temporal clause.

In Early Vedic, the particles ná, iva, and yáṭhā/yathā have other functions beside that of standard marker of equative constructions: for instance, when employed as a subordinator, Vedic yáṭhā also introduces temporal, final, causal, and content clauses with verbs of knowing and saying (Delbrück 1888: 592-596). Furthermore, as we have seen in Section 2, Vedic has at its disposal several strategies for the encoding of comparison of equality.

Following the UD scheme, it would be possible to extract, e.g., all basic equatives featuring a gapping structure by retrieving all nodes a) that are not a finite verb, b) whose deprel is advcl, c) that have a child whose deprel is mark and d) that have at least another child whose deprel is orphan. In order to exclude other types of subordinate clauses characterized by gapping structure, it would also be necessary to specify e) the lemma of the former child. Even so, one would obtain all basic equatives introduced by ná, iva, and yáṭhā (and not subordinates introduced, e.g., by yáḍ ‘that’), but also other subordinates introduced by yáṭhā that present an elided verb. Cf. Figure 5:

```python
    cat rv.conllu | udapy -TM util.Mark node='a) node.feats("VerbForm") == "" and b) node.deprel == "advcl" and c) len([x for x in node.children if x.deprel == "orphan"]) == 1 and d) len([x for x in node.children if x.deprel == "mark" and e) x.lemma in ("nā", "iva", "yathā")]) == 1' | less -R
```

**Figure 5.** Udapi³ query: ‘display all basic equatives with gapping structure’.

---

6 https://universaldependencies.org/workgroups/comparatives.html
7 https://universaldependencies.org/docs/u/dep/advcl.html
8 https://udapi.github.io
Such query would also prevent one from detecting and isolating hybrid constructions such as the one in (7), whose standard has no verb, as in syntagmatic comparison, but in which yāthā precedes the standard, as in clausal comparison.

(7) yāthā nah pitāraḥ párāsaḥ pratnāso...
    like 1PL.GEN father.NOM.PL further.NOM.PL ancient.NOM.PL
    súcād ayan didhitim ukthaśāh
    blazing.ACC.N come.SUBJ.3PL vision(1).ACC reciting_praise.NOM.PL

‘Like our further forefathers of old […], those reciting solemn speech (now) will come to the blazing (udder of sacrifice [=Vala]), to visionary power.’ (RV 4.2.16)

In order to represent the syntax of equatives in detail and to be able to make granular and targeted queries on different types of constructions, the VTB makes use of language-specific extensions that enrich the universal dependency taxonomy. Like language-specific extensions found in UD, extensions employed within the VTB are regarded as subtypes of existing UD relations and have the format universal:extension: for instance, obl:manner stands for manner extension of the UD relation obl. As in UD, extensions employed within the VTB are neither recursive nor multidimensional, which means that one node can instantiate at most one subtype of a universal relation. However, the VTB allows the user to employ a considerably high number of sub-relations for research-related purposes, provided that such extensions are fully documented in the guidelines.

Table 1 summarizes the scheme employed by the VTB for equative constructions.

| CONSTRUCTION            | EXAMPLE                                                                 | ANNOTATION (dependent → relation → head) |
|-------------------------|------------------------------------------------------------------------|------------------------------------------|
| PREDICATIVE SIMILE      | ‘Agni is like the sun.’                                                | sun → root                               |
|                         |                                                                        | sun → nsubj → Agni                       |
|                         |                                                                        | sun → case:sim → like                    |
| SIMILE WITH ELLIPSIS    | ‘Agni shines like the sun.’                                            | shines → obl:grad → sun → case:sim → like|
|                         | ‘The lightning bellows like a cow.’                                    | bellow → obl:manner → cow → case:sim → like|
| SIMILE WITH GAPPING     | ‘Thoughts lick Indra like mothers a calf.’                              | lick → advcl:manner → mothers → mark:sim → like; mothers → orphan → calf |
| CLAUSAL SIMILE          | ‘Just as you drank the previous soma drinks, so take a drink today.’   | drink → advcl:manner → drank → mark → as; drank → obj → previous drinks; drink → advmod → so |
‘He (Agni) who shines like the blazing sun.’ (RV 1.43.5)  
**Figure 6.** Extended scheme for simple equatives.

‘Thoughts lick Indra like mothers a calf.’ (RV 3.41.5)  
**Figure 7.** Extended scheme for equatives with gapping structure.

The sublabel :sim\(^9\) attached to the relations case and mark allows the user to easily retrieve all particles that introduce basic equatives and to distinguish them from those that introduce clausal similes (which take mark alone). Compare for instance the annotation of basic equatives like those in Figure 6 and Figure 7 with that of a clausal equative like the one in Figure 8:

‘Just as you drank the previous soma drinks, Indra, so take a drink today.’ (RV 3.36.3cd)  
**Figure 8.** Extended scheme for clausal equatives.

In some cases, the verb is exceptionally constructed with the standard rather than with the comparee. As shown by Figure 9, such cases are also captured by the annotation scheme.\(^{10}\)

‘As an axe brings together a chariot, the chanters Ø the Hotar with their insight.’ \(^{11}\) (RV 3.2.1)  
**Figure 9.** Annotation of equatives whose verb is constructed with STAND.

---

\(^9\): sim stands for “simile”.

\(^{10}\) In this example, we would expect a plural verb sám rṝvati in agreement with the comparee vāghātás.NOM.PL ‘chanters’; the verb sám rṝvati.PRS.3SG ‘brings’ agrees instead with the nominative singular kūḷāḥ ‘axe’ which constitutes the standard of the simile. As a whole, the sentence is treated similarly to a case of leftward gapping in coordination.
4  Treebank-based analysis of Rigvedic similes

Despite employing different standard markers, Rigvedic comparisons introduced by nā, iva, and yāthā, constitute a coherent construction from the point of view of both syntax and semantics. Syntactically, they have a syntagmatic nature and present clitic standard markers; semantically, they are specialized for figurative comparison and can be defined as similes in all respects.

With the support of extant literature on the origin of Rigvedic similes, quantitative evidence provided by the treebank can help understanding how different particles came to be employed in the kind of constructions attested in the RV. In particular, four groups of queries run on a corpus of 857 similes yielded interesting results in this regard. Queries employed in this study are reported in Appendix A. Before presenting the results, two premises are in order. First, due to his complex internal chronology, the RV constitutes a diachronic corpus, thus lending itself to the study of language change. Second, in presenting word-order patterns attested in similes, I will only take similes introduced by nā and iva into account: basic equatives introduced by yāthā occur only 76 times in the RV and thus do not lend themselves to quantitative studies on word order (Levshina et al., forthcoming).

Besides a language word-order preferences, heaviness is also responsible for the relative order of standard and parameter. As shown by Table 2, similes with gapping, whose standard consists of at least two arguments of the verb, have PAR - STAND order more frequently than simple similes (62% vs. 52%). In turn, Table 3 shows that the percentage of STAND - PAR order is especially high (68%) in those similes whose standard consists of a single element (e.g., pitā iva ‘like a father’, putram nā ‘like a son’), and it decreases to 57% in those similes whose standard has adjectival, participial, or genitive modifiers (e.g., nityam nā sūnīm ‘like a dear son’).

Table 2. Order of STAND and PAR in similes a) with ellipsis and b) with gapping.

| ORDER     | SIMILES WITH ELLIPSIS | SIMILES WITH GAPPING |
|-----------|-----------------------|----------------------|
| STAND-PAR | 360                   | 151                  |
| PAR-STAND | 212                   | 134                  |
| TOTAL     | 572                   | 285                  |
| p-value (χ² test) | 0.0064         |

Table 3. Order of STAND and PAR in a) similes with ellipsis and simple STAND, and b) similes with ellipsis and complex STAND.

| ORDER     | ELLIPSIS AND SIMPLE STAND | ELLIPSIS AND COMPLEX STAND |
|-----------|----------------------------|-----------------------------|
| STAND-PAR | 197                        | 163                         |
| PAR-STAND | 91                         | 121                         |
| TOTAL     | 288                        | 284                         |
| p-value (χ² test) | 0.0083          |

Finally, the percentage of PAR - STAND order is increased by the high frequency of thetic sentences (e.g., *The telephone’s ringing*), which in Vedic have verb-initial order (Lambrecht, 1994: 143; Viti, 2008). Cf. example (8):

---

12 The annotated portion of the RV is available at: https://github.com/EricaBiagetti/VTB_Rigveda.
13 Differently from nā and iva similes, whose origin is disputed, we do not need quantitative evidence in order to confirm the emergence of yāthā similes from comparative clauses and the consequent cliticization of the subordinator.
14 In the annotated portion of the RV in the VTB (24109 tokens in 3092 sentences) OV occurs in 63% of cases. However, Ryan and Gunkel (2015) have shown that, in metrically neutral contexts, non-imperative finite verbs display OV order in 78% of cases (37 in total) and imperative forms in 77% of cases (22 in total).
(8) próthad áśvo ná yávase avisyán
snort.INJ.PRS.3SG horse.NOM like pasture(N).LOC eager.NOM
‘He has snorted like a hungry horse in his pasture.’ (RV 7.3.2a)

Knowing which factors determine the order of standard and parameter helps envisaging diachronic tendencies in the development of equative constructions as attested in the RV, presented in points 2 to 4 below.

2. Query: Frequency of STAND - PAR and PAR - STAND orders in iva e ná similes
Two main hypotheses have been proposed in the literature on the development of ná similes: a) according to Vine (1978), they derive from coordinate negative constructions with ellipsis of the verb in the second conjunct, (9); b) according to Pinault (1985), they stem from the so-called negative parallelism, i.e., a rhetorical device typical of Baltic and Slavic folk literature, consisting of two sentences, the first of which presents a negation and optional ellipsis of the negated verb (10).

Thus while, according to Vine, similes introduced by ná originate from constructions in which the PAR (verb) preceded the STAND, according to Pinault they stem from constructions with the opposite order of STAND and PAR:

(9) Coordinate negative constructions: PAR - STAND
ná ta indra sumatáyo ná ráyah
NEG 3PL.NOM.N Indra.VOC favor(F).NOM.PL NEG rich.NOM.PL
SAIČAKSE pùrvā uśāsār ná nūmāh
enumerate.DAT earlier.NOM.PL.F dawn(F).NOM.PL NEG/like recent.NOM.PL.F

‘Neither your favors nor your riches, O Indra, can be entirely surveyed, through the previous dawns, nor through the current ones.’ > ‘Neither your favors nor your riches, O Indra, can be entirely surveyed, just like the previous and the current dawns (cannot be entirely surveyed).’ (RV 7.18.20)

(10) Negative parallelism: STAND - PAR
vérv ná ādusāc
bird.NOM NEG/like wood_sitting.NOM
camuvoir ā asadad dhāriḥ
cup(F).LOC.DU LP seat.AOR.3SG tawny.NOM

‘It is not a bird sitting in the wood, the tawny one (Soma) has taken his seat in the two cups.’ > ‘Like a bird sitting in the wood the tawny one has taken his seat in the two cups.’ (RV 9.72.5)

Observing the relative order of standard and parameter separately for iva and ná similes, we gain some important insights on the origin of these constructions. Table 4 shows that simple similes introduced by ná have STAND - PAR order more frequently than those introduced by iva (68% vs. 60%). While this difference is statistically only weakly significant (χ² test, p-value 0.06), the picture changes if we focus on similes whose standard is composed of one single element, with no modifiers: here, the percentage of STAND - PAR order reaches 78% with standards marked by ná, against 63% of standards marked by iva (p-value 0.013). On the contrary, no significant difference can be observed in word-order patterns of similes with gapping, since ná and iva similes of this type show STAND - PAR order in 54% and 52% of cases respectively.

Table 4. N. of STAND - PAR and PAR - STAND orders in simple similes and in simple similes whose standard consists of only one element.

| SIMILE TYPE | ALL SIMPLE SIMILES | STANDARD = ONE ELEMENT |
|-------------|--------------------|-----------------------|
| STM ORDER   | iva similes | ná similes | iva similes | ná similes |
| STAND - PAR | 114 | 60% | 234 | 68% | 65 | 63% | 121 | 78% |
| PAR - STAND | 76 | 40% | 108 | 32% | 37 | 37% | 33 | 22% |
| TOTAL      | 190 | 342 | 102 | 154 |
| p-value (χ² test) | 0.06 | 0.013 |
If we assume that, in the absence of other syntactic and pragmatic factors presented under point 1, similes tend to retain the original relative position of standard and parameter, the fact that simple ná similes have a more marked preference for the STAND - PAR pattern than iva similes may constitute an important clue in favor of their origin from the negative parallelism (Pinault 1985), where the standard always precedes the verb. The fact that the preference for the STAND - PAR order is less marked for iva similes, on the other hand, may support the hypothesis of its origin as a marker for syntagmatic comparison, which does not tie the standard to any position with respect to the parameter (see points 2 and 3). Finally, the fact that ná and iva similes behave in the same way in the presence of gapping would be due to the heaviness of the standard in such constructions.

Turning to semantics, the origin of ná equatives from negative parallelism provides some interesting insights on their specialization for figurative comparison: in negative parallelism, the subject of the first clause usually represents a prototype participant of the action or quality expressed by the verb and thus lends itself to figurative readings.\(^\text{15}\)

3. Query: equatives whose verb (PAR) is construed with the STAND, and not with CPREE
Query number 2 returns five cases in which the verb is constructed with a standard introduced by ná (as in Figure 9) and three cases in which yáthá occurs in a hybrid construction, as the one presented in (7). In contrast, the query does not return any case in which a standard marked by iva is clearly constructed with the verb. If we interpret such cases as remnants of a stage in which both the comparee and the standard clause could contain a verb, the presence of such evidence in ná similes confirms point 2 on the clausal origin of the latter; accordingly, the lack of such evidence in iva similes may suggest that iva has always introduced syntagmatic comparison.

4. Query: frequency of equatives with gapping structure
If, as suggested by point 3, iva similes were always syntagmatic, we can assume that they originally had simpler standards and that only later allowed gapping structure on the model of ná similes (which, as suggested by point 2 and 3, originally contained a verb). By dividing the corpus into the ten books that make up the RV, we can check whether similes with gapping became more frequent in younger books (I, VIII-X) than they were in older ones (II-VII). Table 5 reports the frequencies of simple similes and similes with gapping introduced by iva and ná throughout the ten books; note that, if the whole RV is considered (last raw), the ratio of simple and gapped standards is virtually the same for iva and ná similes.

### Table 5. Percentage of simple similes and of similes with gapping in each book.

| Book | iva similes | Simple similes | ná similes | With gapping | Simple similes | With gapping |
|------|-------------|----------------|-------------|--------------|----------------|--------------|
|      | With gapping|                |             |              |                |              |
| I    | 22          | 56%            | 17          | 44%          | 63             | 58%          | 45           | 42%          |
| II   | 31          | 76%            | 10          | 24%          | 17             | 65%          | 9            | 35%          |
| III  | 12          | 75%            | 4           | 25%          | 14             | 67%          | 7            | 33%          |
| IV   | 7           | 78%            | 2           | 22%          | 16             | 73%          | 6            | 27%          |
| V    | 19          | 90%            | 2           | 10%          | 13             | 72%          | 5            | 28%          |
| VI   | 10          | 67%            | 5           | 33%          | 32             | 70%          | 14           | 30%          |
| VII  | 10          | 67%            | 5           | 33%          | 27             | 64%          | 15           | 36%          |
| VIII | 25          | 62.5%          | 15          | 37.5%        | 30             | 68%          | 14           | 32%          |
| IX   | 19          | 59%            | 13          | 41%          | 71             | 74%          | 25           | 26%          |
| X    | 35          | 55%            | 29          | 45%          | 59             | 71%          | 24           | 29%          |
| Total| 190         | 65%            | 102         | 35%          | 342            | 67%          | 164          | 32%          |

| p-value | 0.01 | 0.024 |

\(^{15}\) Furthermore, Pinault (1985: 138-143) suggests that the comparative reading of ná must have spread thanks to the existence of comparative compounds (e.g. viśvajīta- lit. ‘wind-swift’) and comparisons with an ablative STAND (e.g. manāso. ABL. jāvīyas ‘swifter that thought’), which shared the STAND - PAR order with the negative parallelism. Comparative compounds are known cross-linguistically for their preference for generic comparisons (Haspelmath and Buchholz, 1998) and, at least within the IE domain, idiomatic ablative comparatives are also often employed in this function (cf. the type Latin melle dulcior ‘sweeter than honey’).
Table 5 suggests that gapping structure did indeed become more common for iva similes in younger books: a significant difference can be observed between, e.g., 9% of similes with gapping in book V and 43% in book I, or 45% in book X. Similes introduced by ná present a different picture: while book I has indeed the higher percentage of similes with gapping (41%), these were already frequent in old books such as II, III, and VII. In fact, Kruskal-Wallis tests suggest that older and younger books differ from each other in the frequency of iva similes with gapping (p-value 0.01) as well as in the frequency of ná similes with gapping (p-value 0.02). Due to the low absolute counts reported in Table 5, the tests do not point to clear diachronic differences in the structure of ná and iva similes and suggest that the issue should be investigated further on a larger data set.

To sum up, with the partial exception of point 4, results obtained from the four queries suggest that equative constructions introduced by ná and iva probably influenced each other: by systematic ellipsis of the negated verb in the negative parallelism, ná similes became syntagmatic and the standard marker ná developed a clitic behavior; iva similes, on the other hand, specialized for figurative comparison and started to feature gapping structure.

5 Thinking big: cross-linguistic extensions

As anticipated above, the annotation scheme presented in Section 3.2 was developed within a project devoted to the study of Rigvedic similes. As showed in Section 4, the introduction of language-specific extensions made it possible to perform precise, quantitative analyses on the syntax of Rigvedic similes; however, some language-specific extensions would be superfluous if employed in analyses of more general interest or for languages other than Early Vedic.

This suggests that, in view of the next UD release, some extensions might be discarded whereas other might be considered for employment in other treebanks. For instance, the distinction between standard markers of clausal and phrasal equatives, which in the VTB are annotated as mark and mark:sim respectively, should be discarded as the difference between such constructions results in the presence vs. absence of a verb in the standard. Furthermore, the information stored in the :manner and :grad extensions should be moved to the MISC field of the CoNLL-U format and assigned on a lexical basis to the parameter, depending on whether it encodes a gradable or non-gradable quality.

More interesting is the possibility of extending the relation subtype :sim to standard markers of equative and similitive constructions in other languages and construction types. In many languages, standard markers of equative constructions can be identical with conjunctive particles and subordinators (Haspelmath et al., 2017): remaining within the Indo-European domain, cf. Latin ut ‘as, how’, which introduces several other kinds of subordinate clauses. Beside particles and conjunctions, standards of equatives and similitives can be marked by adpositions or by case markers. When the parameter marker is expressed by an adjective or verb, the standard is marked by a case selected by the governing adjective or verb: cf. the Latin adjective consimilis in (11) and the Ancient Greek participle eidómenon in (12), both governing a dative standard. Figure 10 shows the suggested annotation scheme for example (12).

(11) | harum | est | consimilis | capris | figura |
    | this GEN.PL | be.PRS.3SG | similar.NOM | goat.DAT.PL | shape.NOM |
    | ‘their shape (scil. of elks) is similar to [that of] goats’ (Caes. Gall. 6.27.1; Ittzés 2021: 479)

(12) | élithé | moi | phásma | eidómenon | Aristōni |
    | come.AOR.3SG | 1SG.DAT | phantom.NOM | resemble.PTCP.PRS.NOM | Ariston.DAT |
    | ‘A phantom came to me that resembled Ariston.’ (Herodotus 6.69.1; de Kreij 2021: 350)

---

16 Note that negative ná, from which comparative ná derives (cf. Pinault, 1985), stands either in clause-initial position or before the predicate.

17 Note that the CoNLL-U format adopted by the DCS does not include a MISC field. This determined the choice of extending the syntactic relations obl and advcl of the STAND with semantic information pertaining the whole construction such as :manner and :grad.
Extending the relation subtype :sim would allow accounting for equative and similitative constructions that are otherwise not covered by the UD taxonomy. This is the case, for instance, of reach equatives such as (5), which are tagged like usual transitive clauses in the UD scheme (Figure 11). While in Early Vedic such constructions are sporadic and scarcely grammaticalized (Biagetti 2021), in some languages they constitute a major comparison strategy and extending their annotation scheme would enhance the possibility of studying equative constructions cross-linguistically. The extended annotation for reach equatives is illustrated by Figure 12 from Malgwa (Chadic; Löhr, 2002: 107).

The reason for adding relation subtypes to standard markers and not to parameter markers of equative constructions is suggested by Haspelmath et al. (2017: 25) Generalization 1, according to which “[n]o language has only a degree-marker, leaving the standard unmarked”. In other words, while constructions such as “Kim is Ø tall like Pat” are cross-linguistically common, constructions such as “Kim is equally/as tall Ø Pat” are not attested; thus, marking only standard markers with relation subtypes would allow capturing all types of equatives while avoiding redundancy. Finally, assigning the label :sim to elements of equative constructions would allow distinguishing them from elements of comparative constructions proper, which encode comparison of inequality (Treis 2017) and are marked by the extension :cmpr in some treebanks.20

6 Conclusion

By presenting a case study on Rigvedic equative constructions, in this paper I argued that treebanks constitute an important support to research in historical linguistics because they allow to confirm or dismiss previously formulated hypotheses (see especially query 2) and to observe correlations between language phenomena that could hardly be grasped by the naked eye (queries 1, 3, and 4). However, the need to account for formal variations or hybrid constructions that may play a role in language change sometimes calls for more granular and informative annotation schemes. In the case of Rigvedic similes, I suggested implementing the UD scheme for equative and similitative constructions with sub-relations; crucially, such extensions are not meant to be language specific and some of them might be adopted by every treebank developer interested in representing equative constructions.

---

18 With Dixon (2012), we might say that they constitute comparative strategies rather than constructions proper.
19 See for instance the examples from Malgwa (Chadic), Malian Tamashek (Berber), or Zay (Semitic) in Haspelmath et al. (2017: 21-22).
20 Treebanks of Latin, Polish, and Tamil employ obl: cmp for comparative oblique arguments and advcl: cmp for comparative clauses. While the former is limited to comparison of inequality, the latter is instantiated with examples of clausal equatives. In order to increase consistency, I suggest limiting advcl: cmp to proper comparative clauses and adding a new relation subtype (such as :sim) for clausal equatives. Note, in passing, that Telugu employs obl: cmp and Moksha obl:comp with the same purpose of obl: cmp. Finally, Erzya employs advmod: comp for adverbs functioning as standard markers in comparatives proper. Cf. https://universaldependencies.org/ext-dep-index.html.
Acknowledgements

I wish to thank the three anonymous reviewers who, with their comments and suggestions, have provided insightful feedback for a substantial improvement of this work. I am grateful to Oliver Hellwig for his constant support, from the first annotation tests to the final stages of this article, and to Salvatore Scarlata for the lengthy discussions on what was the best way to annotate Rigvedic similes. Final responsibility remains my own.

Reference

Andersen, Paul Kent. 1983. *Word Order Typology and Comparative Constructions*. John Benjamins, Amsterdam.

Anthony, Laurence. 2013. A critical look at software tools in corpus linguistics. *Linguistic research* 30(2): 141–161.

Bergaigne, Abel. 1887. *La syntaxe des comparaisons védiques*. Mélanges Renier, 75-101. Vieweg, Paris.

Biagetti, Erica. 2021. *Rigvedic similes: a corpus-based analysis of their forms and functions*. PhD thesis. University of Pavia.

Biagetti, Erica, Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2021. Evaluating Syntactic Annotation of Ancient Languages: Lessons from the Vedic Treebank. *Old World: Journal of Ancient Africa and Eurasia* 1, no. 1: 1–32.

Biber, Douglas. 2009. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*, Bernd Heine & Heiko Narrog (eds), 159–191. Oxford University Press, Oxford.

Delbrück, Berthold. 1888. *Altindische syntax*. Verlag der Buchhandlung des Waisenhauses.

Dixon, Robert M.W. 2012. *Basic Linguistic Theory*. Volume 3. Further Grammatical Topics. Oxford University Press, Oxford.

Eckhoff, Hanne, Bech, Kristin, Bouma, Gerlof, Eide, Kristine, Haug, Dag, Haugen, Odd Einar & Jøhndal, Marius. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1), 29–65.

Gunkel, Dieter, and Kevin Ryan. 2015. Investigating Rigvedic word order in metrically neutral contexts. Handout. Vienna.

Haspelmath, Martin & Buchholz, Oda. 1998. Equative and similative constructions in the languages of Europe. In *Adverbial Constructions in the Languages of Europe*, Van der Auwera, Johan (ed.), 277–334. Mouton de Gruyter, Berlin.

Haspelmath, Martin & the Leipzig Equative Constructions Team 2017. Equative constructions in world-wide perspective. In *Similative and Equative Constructions: A Cross-linguistic Perspective*, Yvonne Treis & Martine Vanhove (eds.) 9–32. John Benjamins, Amsterdam.

Haug, Dag T. T. 2015. Treebanks in historical linguistics research. In *Perspectives on Historical Syntax*, Carlotta Viti (ed), 187–202. John Benjamins, Amsterdam.

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

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

Israel, Michael, Jennifer Riddle Harding, and Vera Tobin. 2004. On simile. *Language, culture, and mind* 100.

Jamison, Stephanie W. 1982. Case disharmony in Rigvedic similes. *Indo-Iranian Journal* 24, no. 4: 251–271.

Jamison, Stephanie W. & Brereton, Joel P. 2014. *The Rigveda: the Earliest Religious Poetry of India*. Oxford University Press, New York.
Appendix A: Queries

This Appendix contains all the queries employed for the case study presented Section 4. All queries were written in Udapi query language (https://udapi.github.io).

Query 1:

a. N. of STAND - PAR and PAR - STAND orders in all similes
   cat RV.conllu | udapy util.See node='node.deprel in ("advcl:manner", "obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā")]) == 1'

b. N. of STAND - PAR and PAR - STAND orders in all similes with ellipsis
   cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case:sim"]) == 1'

c. N. of STAND - PAR and PAR - STAND orders in all similes with gapping
   cat RV.conllu | udapy util.See node='node.deprel in ("advcl:manner") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "mark:sim"]) == 1'

d. N. of STAND - PAR and PAR - STAND orders in similes with ellipsis and simple STAND
   cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case:sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
e. N. of STAND - PAR and PAR - STAND orders in similes with ellipsis and complex STAND

```python
cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case: sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]}) == 1'
```

**Query 2:**

a. N. of STAND - PAR and PAR - STAND orders in similes introduced by ná:

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel in ("case: sim", "mark: sim")]) == 1'
```

b. N. of STAND - PAR and PAR - STAND orders in similes introduced by iva:

```python
cat RV.conllu | udapy util.See node='node.deprel in (len([x for x in node.children if x.lemma == "iva" and x.deprel in ("case: sim", "mark: sim")]) == 1'
```

c. N. of STAND - PAR and PAR - STAND orders in ná-similes with ellipsis

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel == "case: sim"]}) == 1'
```

d. N. of STAND - PAR and PAR - STAND orders in iva-similes with ellipsis

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "iva" and x.deprel == "case: sim"]}) == 1'
```

e. N. of STAND - PAR and PAR - STAND orders in ná similes with ellipsis and simple STAND

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma in ("na") and x.deprel == "case: sim"]}) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
```

f. N. of STAND - PAR and PAR - STAND orders in iva similes with ellipsis and simple STAND

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma in ("iva") and x.deprel == "case: sim"]}) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
```

g. N. of STAND - PAR and PAR - STAND orders in ná-similes with gapping

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel == "mark: sim"]}) == 1'
```

h. N. of STAND - PAR and PAR - STAND orders in iva-similes with gapping

```python
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "iva" and x.deprel == "mark: sim"]}) == 1'
```

**Query 3:**

a. STAND constructed with a finite verb

```python
cat RV.conllu | udapy -TM util.Mark node='node.lemma in ("na", "iva", "yathā") and node.deprel in ("case: sim", "mark: sim") and node.parent.upos == "VERB" and node.parent.feats["VerbForm"] == ""' | less -R
```

**Query 4:**

a. N. of iva similes with ellipsis in each book

```python
cat rv1.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma == "iva" and x.deprel == "case: sim"]}) == 1'
```

b. N. of ná similes with ellipsis in each book

```python
cat rv1.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma == "na" and x.deprel == "case: sim"]}) == 1'
```

c. N. of iva similes with gapping in each book
d. N. of *ná* similes with gapping in each book

```
cat rv1.conllu | udapy util.See node='node.deprel == "advcl:manner" and len([x for x in node.children if x.lemma == "iva" and x.deprel == "mark:sim"]) == 1'
cat rv10.conllu | udapy util.See node='node.deprel == "advcl:manner" and len([x for x in node.children if x.lemma == "na" and x.deprel == "mark:sim"]) == 1'
```