Detecting Diachronic Syntactic Developments in Presence of Bias Terms

Oliver Hellwig, Sven Sellmer
Heinrich Heine Universität Düsseldorf
Institute for Language and Information
{Oliver.Hellwig, sellmer}@uni-duesseldorf.de

Abstract

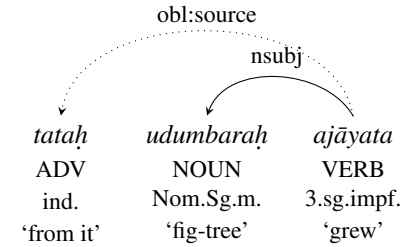
Corpus-based studies of diachronic syntactic changes are typically guided by the results of previous qualitative research. When such results are missing or, as is the case for Vedic Sanskrit, are restricted to small parts of a transmitted corpus, an exploratory framework that detects such changes in a data-driven fashion can support the research process. In this paper, we introduce an infinite relational model (Kemp et al., 2006) that groups syntactic constituents based on their structural similarities and their diachronic distributions. We propose a simple way to control for register and intellectual affiliation, and discuss our findings for four syntactic structures in Vedic texts.

Keywords: Historical syntax, Vedic Sanskrit, infinite relational model

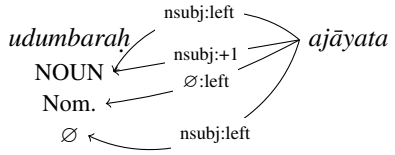
1. Introduction

When studying diachronic linguistic change from a corpus-based perspective, it is – often silently – assumed that the linguistic phenomena of interest are known from previous qualitative studies (Hilpert and Gries, 2016, 44ff.). Such an assumption may not hold for premodern languages with a limited history of research. Here, it can be useful to have an exploratory framework that draws attention to phenomena that change in time, while simultaneously controlling for other influence variables such as intellectual affiliation or register. We propose such a framework for detecting diachronic trends in the syntax of Vedic Sanskrit (or Vedic), a premodern Indo-Aryan language with a large corpus of religious and ritualistic texts composed in the second and first millenia BCE (Renou, 1956). Most previous research on Vedic syntax has concentrated on the oldest Vedic texts and a limited number of research questions (see Sec. 2 of this paper), and only few studies have tried to quantify the phenomena they describe, especially from a diachronic perspective. Given that a syntactic treebank of Vedic is now available (see Sec. 3), this research situation leaves ample space for quantitative approaches.

Studying diachronic syntactic changes is challenging because the number of interacting units grows nonlinearly with the size of the dependency tree and interesting phenomena may not be directly read off the joint surface representation. In this paper, we therefore focus on individual syntactic constituents which are composed of the morpho-syntactic representation of a word and its relation to its syntactic head. Consider, as an example, the solid dependency arc in Fig. 1a. This left-branching arc of length one marks the nominative singular noun *udumbaraḥ* 'fig-tree' as the subject of the finite verb *ajāyata* 'grew'. While the joint representation of this arc (*udumbara-* acting as a subject and placed directly to the left of its head) is rare and therefore offers only limited linguistic insights, combinations of its



(a) A sample sentence with morpho-syntax: 'A fig-tree grew from it.'



(b) Constructing abstract representations of the constituent *udumbaraḥ* in Fig. 1a

Figure 1: A sample sentence (Kaṭha-Saṃhitā 6.1.6) and some constituents generated from its subject *udumbaraḥ*

features are better suited to highlight linguistically interesting and especially interpretable phenomena (see Fig. 1b; details in Sec. 4.1). We therefore extract the features of the noun *udumbaraḥ*, i.e. part-of-speech, morpho-syntax and details about the syntactic relation, and form their Cartesian product. This step generates new, more abstract representations of the constituent such as NOUN/∅/nsubj/+1 (a nominal subject in unspecified case [∅: wildcard] is found directly to the left of its head), or NOUN/∅/∅/left (an unspecified nominal dependent is found anywhere to the left of its head). In this way we abstract from the syntactic surface as, for instance, ∅/Nom/nsubj/∅ (a word in nominative case serves as subject) may apply to a verbal clause as in Fig. 1b or to a nominal identity statement

such as *yajñaḥ prajāpatiḥ* '[the god] Prajāpati [is the] sacrifice' where the subject *prajāpatiḥ* is found to the right of the predicate *yajñaḥ*. We expect that such abstract constituents ('constituents' in the following) reveal syntactic patterns that are easier to interpret and more frequent than the joint surface representations and therefore more useful for exploratory linguistic studies.

We now want to determine which constituents show diachronic variation while controlling for the Vedic school of each text (see Sec. 3) and for register, two variables that influence the linguistic form of Vedic texts (Witzel, 1989; Hock, 1997; Cohen, 2008). While, for instance, Cochran–Mantel–Haenszel tests (Agresti, 2007) could be applied here, the large number of factor combinations results in sparse count tensors which do not allow for a meaningful statistical interpretation. Moreover, many constituents differ from each other only in minor aspects (e.g. a noun placed to the left of its head vs. a noun placed directly to the left of its head), and it is not clear a priori which of such variants should be studied in greater detail. We hypothesize, however, that similar constituents have similar chronological distributions. Aggregating similar constituents may therefore produce more stable results. Starting from these ideas, we interpret the constituents as nodes in a similarity graph G. G has an edge between two constituents if they occur in the analysis of the same surface form. From the sentence in Fig. 1 we can, for instance, deduce that the constituents NOUN/Nom/∅/+1 and ∅/∅/nsub/left are connected in G because they are both analyses of the same surface form *udumbaraḥ*. The graph-based approach has the advantage that abstract representations of a surface form have good chances to occur in diverse syntactic constructions and thereby connect constructions that differ strongly at first view.

As the resulting graph is large and therefore difficult to interpret, we partition its nodes (i.e. the constituents) using a variant of the infinite relation model (IRM; Kemp et al. (2006)). The IRM is a Bayesian model that groups objects from multiple domains based on their n-ary relations. In our case, all objects (constituents) come from the same domain, and a binary relation between two objects is present if both represent the same surface form at least twice in the treebank. We extend this model by constituent specific normal distributions that record how much the chronological distribution of a constituent deviates from the maximum likelihood estimate of the corpus distribution. This chronological information is represented in the form of unary attributes attached to each node of the IRM. The aim of the IRM is therefore to find partitions containing constituents that are both similar with regard to their morpho-syntactic information and their chronological distributions. Constituents in the partitions produced by the IRM are finally ranked using an information-theoretic criterion that takes the bias terms (school, register) into account.

Section 2 of this paper gives a short overview of related work. The data is described in Sec. 3, and the model is defined in Sec. 4. Section 5 reviews four syntactic phenomena detected by the model and discusses possible limitations of the approach. Section 6 summarizes the paper. Code and data are available at `https://github.com/OliverHellwig/sanskrit/tree/master/papers/2022-lt4hala-syntax`.

2. Related research

Much of the work on Sanskrit syntax has been devoted to Vedic (Hock, 2015b; Hock, 2015a). Issues of morpho-syntax such as case syntax, verbal nouns and converbs form the bulk of this research, whereas only few studies deal with word and constituency order. Apart from overviews like Gonda (1971), diachronic approaches are rarely found. Changes in word order are sometimes investigated in the context of the Indo-European background and the influence of substrate languages (Lehmann, 1974; Hock, 1984). There are also studies that trace the development of certain syntactic phenomena over time (Renou, 1937), or study temporal stratification (Wüst, 1928), but all in all the amount of research helpful for our task is limited. One major problem is that the youngest stratum of Vedic has largely been neglected in linguistic research (Wezler, 2001).

While diachronic semantics have recently received much attention in linguistics and NLP (see e.g. Haase et al. (2021) and Frermann and Lapata (2016)), the question how to detect syntactic changes is only rarely addressed. Closest to what we aim at in this paper is the exploratory tool described by Schätzle et al. (2019) which visualizes the relationship between multiple linguistic features and can thus be used for detecting previously unnoticed diachronic syntactic changes. Further data-driven approaches to historical syntax are discussed in Hilpert and Gries (2016).

3. Data

The syntactic data are taken from the Vedic Treebank (VTB).[1] Compared to previous versions of the VTB (Hellwig et al., 2020; Biagetti et al., 2021; Hellwig and Sellmer, 2021) its current version has been extended substantially. It now contains 18,061 sentences from 37 texts, including texts from the White Yajurveda as well as extracts from the Śrauta Sūtras, the manuals of the solemn ritual. Table 1 describes the composition of the VTB in terms of the influence variables considered in this paper. As the treebank is biased towards late prose texts and the schools of the Rig- and Yajurveda, controlling these variables is even more important.

Most Vedic texts contain a large number of mantras, i.e. verbatim citations from the old metrical Saṃhitās.

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

Layer				
RV	MA	PO	PL	SU
9,817	28,989	25,211	40,618	31,636
School				
AV	Black YV	RV	SV	White YV
16,053	31,536	51,732	19,910	17,040
Register				
metrical	prose			
38,806	97,465			

Table 1: Number of words in the Vedic Treebank grouped by the influence variables considered in this paper. The layers (first compartment) are sorted in ascending chronological order (*Rigveda* proper, Mantra language, old prose, late prose, Sūtra language).

Mantras are cited at virtually every step of a sacrifice in order to guarantee its success (Patton, 2006) and therefore account for about 6% of all words in the VTB. Because mantras contain archaic linguistic material they impede the chronological analysis of Vedic syntax. We therefore completely remove mantras from the data.

Each Vedic text can be assigned to one of five Vedic schools. These schools differ in which role their main priests assume in the solemn sacrifice, and their texts therefore focus on different aspects of this sacrifice (Renou, 1947). We also know the register of each text.

The most problematic part is chronological information. There have been numerous attempts to date the Vedic corpus as a whole or parts thereof, none of which has found unanimous support in the scholarly community (Hellwig, 2020). As a consequence, all we have is a vague relative order of Vedic texts which is disputed in many details. Given this state of research, each text is assigned to one of five consecutive diachronic layers whose arrangement is based on ideas proposed by Witzel (1989) and Kümmel (2000):

1. Early Vedic [= RV]: *Rigveda* 2-7, 9
2. Old Vedic [= MA]: *Rigveda* 1, 8, 10; metrical portions of the *Atharvaveda-* and *Yajurveda-Saṃhitā*s ('Mantra language')
3. Middle Vedic [= PO]: prose portions of the Saṃhitās, the older parts of Brāhmaṇas, Āraṇyakas, and Upaniṣads
4. Young Vedic [= PL]: younger parts of Brāhmaṇas, Āraṇyakas (both prose), and Upaniṣads (partly prose, partly verse)
5. Late Vedic [= SU]: ancillary texts (Sūtras), mostly prose

4. Model

4.1. Creating the constituents and their distributions

Starting from the intuition that part-of-speech, morpho-syntax, the Universal Dependencies (UD; Nivre et al. (2016)) label and the placement of a word with regard to its syntactic head are relevant features when studying syntactic change, we create constituents by forming all possible combinations of these four features. We include a wildcard option (∅) for each of them; this means that the respective feature can take any value in a given constituent. In the following, a semicolon separates options, and a number in square brackets indicates the number of options for each feature:

POS: the actual POS tag; ∅ [2]

Morpho-syntax: the case for words with nominal inflection, ind(eclinable), fin(ite) or inf(inite) for verbal forms; ∅[2]

UD label: the actual label; ∅ [2]

Placement: the signed distance between the position of the head and the dependent, cut off at a distance of 4; the absolute value of this distance; dependent to the left or to the right of its head; ∅ [4]

There are $2 \cdot 2 \cdot 2 \cdot 4 = 32$ constituent representations of each surface form. We denote a constituent by the quadruple of its values; NOUN/Gen/nmod/-1, for example, is a noun in genitive case that acts as a nominal modifier and stands directly to the left of its head.

We calculate the empirical distribution $\mathbf{o}_i^f \in \mathbb{R}_+^5, \sum_j o_{ij}^f = 1$ of one of the three influence variables f (time, school, register) for a given constituent i by counting the frequency of the constituent in each factor level of the variable and normalizing these counts. The respective expected distribution $\mathbf{e}_i^f \in \mathbb{R}_+^5, \sum_j e_{ij}^f = 1$ is obtained by subtracting the counts for i from the respective corpus counts and normalizing. The differences $\mathbf{d}_i^t \in \mathbb{R}^5$ between the expected and observed distributions for the chronological variable describe to which degree the distribution of constituent i deviates from the global estimate calculated without knowledge about i. These differences are used as input for the unary relations in our model ($\mathbf{n} \in \mathbb{Z}_{\geq 0}^5$: vector of global corpus counts for the five chronological slots; $N = \sum_i^5 n_i$; $\mathbf{m}_i \in \mathbb{Z}_{\geq 0}^5$: counts for constituent i):

$$\mathbf{d}_i^t = \mathbf{o}_i^t - \mathbf{e}_i^t = \frac{\mathbf{m}_i}{\sum_{j=1}^5 m_{ij}} - \frac{\mathbf{n} - \mathbf{m}_i}{N - \sum_{j=1}^5 m_{ij}} \quad (1)$$

While pre-processing the data, we use a G-test (Agresti, 2007) that assesses if the distribution of $\mathbf{o}_i^t$ differs significantly from $\mathbf{e}_i^t$ at an error level of 0.01. If it does not, the respective constituent is discarded from the data set because its chronological distribution cannot be said to differ from the corpus distribution at the given error level. This step reduces the number of constituents from 5,153 to 3,605 and thus helps to concentrate on chronologically relevant phenomena.

4.2. Constructing and partitioning the graph

We are interested in grouping constituents that describe related syntactic surface phenomena and that have similar diachronic distributions (see Sec. 1). To achieve the first aim, we construct an undirected graph G each vertex of which is one constituent. G has an edge e_{ij} between vertices i, j if the constituents i, j occur at least once as analyses of the same surface form (see

N number of distinct constituents
K current number of partitions inferred by the model
$\mathbf{z} \in \mathbb{Z}_+^N$ partition assignments of the N constituents
n_k number of constituents assigned to partition k
$\mathbf{g} \in \mathbb{Z}_2^{N(N-1)/2}$ binary edges in G
$\mathbf{\Theta} \in \mathbb{R}_{[0,1]}^{K(K-1)/2}$ parameters of Bernoulli distributions that model the presence of edges $\mathbf{e}$ in G
$\mathbf{A}_{kl}$ number of edges in G that connect constituents assigned to groups k and l
$\mathbf{B}_{kl}$ number of cases in which two constituents assigned to groups k and l are not connected by an edge in G
$\mathbf{a}_l^i$ number of cases in which G has a connection between constituent i and another constituent which is assigned to partition l
$\mathbf{b}_l^i$ number of cases in which G does not have a connection between constituent i and another constituent assigned to partition l
$\boldsymbol{\mu} \in \mathbb{R}^{K\times 5}, \boldsymbol{\sigma} \in \mathbb{R}_+^{K\times 5}$ parameters of the partition specific chronological Normal distributions
$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\mu}_0, \boldsymbol{\sigma}_0$ parameters of the prior distributions of the Dirichlet process, the edge Betas and the Normals on the partitions

Figure 2: Notation for the Gibbs sampler (eqs. 3 and 4)

Sec. 4.1). Edges are unweighted, because the selection of texts in the VTB as well as the Vedic corpus as a whole are biased samples from the Vedic language, and absolute counts may rather represent scholarly and ritualistic preferences. – Using the notation from Fig. 2, the generative process of our model can be described as follows:

$$\begin{aligned} z_i &\sim \mathrm{DP}(\boldsymbol{\alpha}) \\ \Theta_{ab} &\sim \mathrm{Beta}(\boldsymbol{\beta}),\ g_{ij} \sim \mathrm{Bern}(\Theta_{z_i z_j}) \\ d_{ik} &\sim \mathcal{N}(\mu_{z_i k}, \sigma^2_{z_i k}) \end{aligned} \tag{2}$$

The model draws the latent assignment z_i of constituent i from a Dirichlet process with concentration parameter α. The value of the edge g_{ij} between constituents i and j is drawn from a Bernoulli distribution whose parameter depends on the partitions assigned to i and j. Finally, the univariate Normal distributions determine how well the chronological profile of constituent i fits that of the partition to which i is assigned. It should be noted that we use five univariate Normals instead of one five-dimensional Normal because we have no a priori intuition about how the covariance matrix between the five diachronic layers defined in Sec. 3 should be structured. While we could infer the posterior of the covariance from the data using an inverse Wishart distribution, we choose the comparatively easier univariate approach for this exploratory model. For the same reason, we set the univariate precision values to constant small values (0.1) in order to obtain clear chronological profiles.

To obtain a collapsed Gibbs sampler, we remove all information about constituent i from the data (counts $\mathbf{A}_*^{-i}, \mathbf{B}_*^{-i}$) and calculate the product of the posterior (for $k \leq K$) and prior predictives (for $k = K+1$) of the Dirichlet process, the actual IRM (on which see e.g. Ishiguro et al. (2014)) and the univariate Normal distributions. A histogram of the difference values calculated using eq. 1 shows that the values of $\mathbf{d}$ are normally distributed around zero. Setting $\mu_0 = 0$, the posterior predictive of a Normal distribution for layer v given all constituents assigned to group k therefore has the following parameters (see e.g. Bishop (2006, 98); $\mathbb{I}[\ldots]$ is the indicator function):

$$m_{kv} = \frac{\sigma_0^2 \sum_{j=1}^N (\mathbb{I}[z_j = k] d_{jv}^t)}{n_k \sigma_0^2 + \sigma^2},\ s_{kv}^2 = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + n_k \sigma_0^2} + \sigma_0^2 \tag{3}$$

For the respective prior predictive we obtain $m'_{kv} = 0,\ s'^2_{kv} = \sigma^2 + \sigma_0^2$. – In the following update equation the upper row gives the posterior, the lower one the prior predictive, and $\mathrm{B}(x, y)$ is the Beta function:

$$\begin{aligned} &p(z_i = k | \mathbf{z}^{-i}, \mathbf{e}^-, \mathbf{\Theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \alpha, \beta, \mu_0, \sigma_0) \\ &\quad \propto \left.\begin{matrix} n_k \\ \alpha \end{matrix}\right\} \cdot \prod_l \frac{\mathrm{B}(A_{kl}^{-i} + a_l^i + \beta, B_{kl}^{-i} + b_l^i + \beta)}{\mathrm{B}(A_{kl}^i + \beta, B_{kl}^i + \beta)} \\ &\qquad \cdot \left\{ \begin{matrix} \prod_{v=1}^5 \mathcal{N}(d_{iv} | m_{kv}, s_{kv}^2) \\ \prod_{v=1}^5 \mathcal{N}(d_{iv} | m'_{kv}, s'^2_{kv}) \end{matrix} \right. \end{aligned} \tag{4}$$

4.3. Weighting the members of the inferred partitions

In the last step, we order the members of each partition. Our aim is to find constituents whose chronological distributions deviate clearly from their expected ones as estimated from the corpus, while their distributions over schools and registers conform to the respective expected values as closely as possible. A natural way for formalizing this notion of closeness is provided by the Hellinger distance between the expected and observed distributions for each of the three factors. The Hellinger distance is confined to $[0, 1]$, with zero meaning no difference between $\mathbf{e}$ and $\mathbf{o}$ and one meaning maximum dissimilarity. We calculate the Hellinger distances for time (h_i^t), school (h_i^s) and register (h_i^r) and use the following expression for weighting constituent i:

$$w_i = (h_i^s + h_i^r) - h_i^t \tag{5}$$

w_i becomes less than zero if the observed chronological distribution differs strongly from the expected one, but the observed distributions over schools and registers conform to their expected values.

5. Evaluation

This section provides a qualitative evaluation of some salient trends detected by the proposed model after it was trained for 100 epochs with hyper-parameters $\alpha = 1$, $\beta = 0.5$, $\sigma = \sigma_0 = 0.1$.[2] Due to lack of

[2]The results are not really sensitive to the choice of α and β which is probably due to the fact that the constituents generate strongly connected components in G.

space, we concentrate on selected partitions detected by the model. Those not discussed here either have high weights (eq. 5) and are therefore correlated with the bias terms, or capture well known diachronic trends such as an increasing preference for elliptic constructions (UD label orphan) in the Sūtra literature.

5.1. Compounds

Some of the strongest chronological signals h^t come from partition #11 which contains constituents found in predominantly long nominal compounds. Nominal composition is one of the few areas in which the annotation scheme of the VTB extends the UD standard (Hellwig and Sellmer, 2021) because nominal compounds in late Vedic and especially in classical Sanskrit can encode a wide range of syntactic functions that would be expressed with verbal sentences in other languages (Lowe, 2015). Vedic linguistics have early noticed the chronologically increasing preference for complex compounds (Wackernagel, 1905, 24-26), and our model thus discovered a known diachronic trend.

Partition 11 represents three major aspects of compounding. First, coordinative compounds (UD label compound with sublabel coord) correspond to the class of dvandva ('pair') compounds in traditional Sanskrit grammar which enumerate concepts by juxtaposing their stems. The usage of such compounds is especially widespread in the Gṛhya- and Dharmasūtras and becomes the preferred mode of coordination in classical Sanskrit.

Second, compounds involving nominal (nmod) modifiers of nouns also get more prominent towards the end of the Vedic period. Many of these compounds express a possessive relation as in *sūkta-anta*, lit. "hymn-end", i.e. "end of the hymn", and thus belong to the class of tatpuruṣa compounds in indigenous terminology. Figure 3 plots the ratios of individual levels of the two influencing variables time and register for the constituent NOUN/Cpd/nmod/∅. Each point in the left part of the plot gives the ratio e^t_j/o^t_j for layer j (see eq. 1), and the whiskers indicate the 95% confidence interval of this ratio. The dashed horizontal line indicates equal proportions, i.e. $e^t_j = o^t_j$; if the whiskers intersect with this horizontal line, the ratio cannot be said to differ from 1 at an error level of 5%. The plot shows that the proportion of this constituent increases monotonically over the five layers of the VTB and additionally makes a large jump in the last one. However, other influence variables must be considered as well. In this case, the distribution over the registers (Fig. 3, right) replicates the chronological trend because such compounds are underrepresented in the early metrical texts and over-represented in prose. As the ratio of compounded nominal modifiers already increases slightly from the first to the second metrical layer (RV → MA) and clearly between the older and younger prose layers (PO → PL), it seems plausible that the register is not the central influencing variable and a real chronological trend is discernible here.

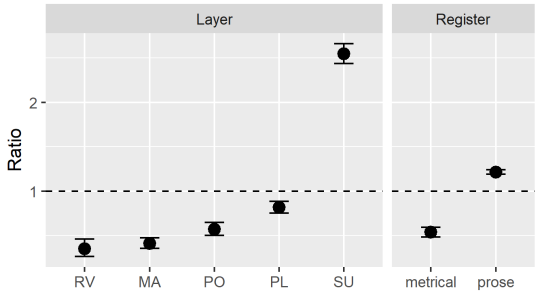


Figure 3: Ratio plots (see p. 5) for the constituent NOUN/Cpd/nmod/∅; see the listing in Sec. 3 for the chronological labels (RV etc.).

Third, closely connected with the two preceding categories are long compounds in general. Their presence can be deduced from members of partition 11 that represent long syntactic arcs in compounds such as ∅/Cpd/∅/4.

Partition 11 also demonstrates how endeavors to control for bias variables are hampered in corpora with a limited coverage. Apart from a strong chronological signal h^t, the top entries in this partition have a strong signal h^s indicating an uneven distribution over the Vedic schools. Mainly responsible for the strength of h^s are texts belonging to the school of the Rigveda. Closer inspection of the data reveals that the predominance of Rigvedic material is due to long compounds found in the Gautama-Dharmasūtra, a late Rigvedic. Other Vedic schools have composed such Dharmasūtras as well, and we are currently working on including samples from them in the VTB. An apparent school-related skew therefore can be explained with selection bias in this case.

Another interesting type of compounds is part of partition 90 which contains clauses modifying nouns (acl) and verbs (advcl). Constituents of the type VERB/Cpd/acl/∅ typically involve verbal nouns in their stem forms compounded with a governing noun as in the phrase *jaritāraḥ suta-somāḥ* 'singers who have pressed Soma' (Rigveda 1.2.2bc) where the head *soma-* is modified by the past participle *suta-* (from *sav* 'press'). As Fig. 4 shows, such constructions are well attested in the oldest metrical levels of Vedic (RV, MA), but lose popularity in the two subsequent layers of Vedic prose (PO, PL), a trend already mentioned by Wackernagel (1905, 315-321). The preference for these constructions increases strongly in the last layer of the Vedic corpus where we find complex, sometimes irregular compound formations as at Āśvalāyana-Gṛhyasūtra 1.17.2: *... vrīhi-yava-māṣa-tilānām ... pūrṇa-śarāvāni nidadhāti* 'he puts down vessels filled with rice, wheat, beans and sesame'. Here the noun *śarāva* 'vessel' is modified by the verbal noun *pūrṇa* (from *par*^{i}*/prā* 'to fill') which is in turn modified by a dvandva compound enumerating different types of

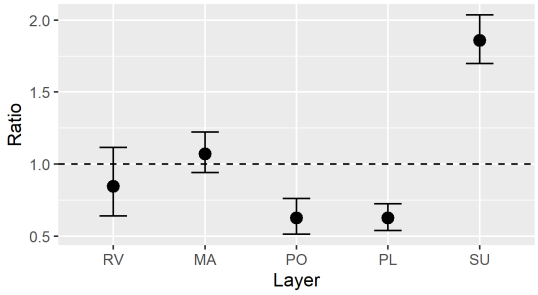


Figure 4: Ratio plot of verbal nouns functioning as clausal modifiers in compounds

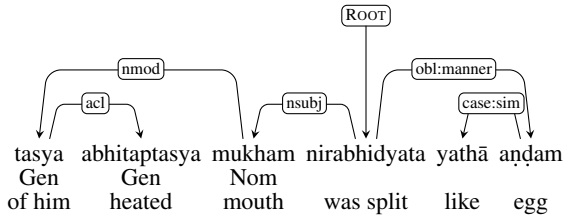


Figure 5: Proto-*genitivus absolutus* at Aitareya-Upaniṣad 1.1.4: "When he was heated, his mouth was split like an egg."

food.

5.2. Precursors of the *genitivus absolutus*

Partition 108 consists almost exclusively of clauses (acl) that modify nouns in nominative, accusative and genitive case in any relative placement. The highest coefficients w (see eq. 5) in this partition are reported for verbal nouns in genitive case that are preferably found to the right of their congruent heads. An example of this construction is displayed in Fig. 5. Here, the past participle *abhitaptasya* (from *abhi tap* 'heat') modifies the pronoun *tasya* which in turn expresses the possessor of the subject *mukham* 'mouth'. While the past participle is preferred in the Vedic prose, analogous constructions with a present participle are found in early metrical texts as well (see e.g. Rigveda 10.38.2c).

Both types of constructions have in common that the modified noun stands in a possessive relation to another word in the sentence, typically its subject (*mukham* in Fig. 5). Oertel (1926, 101ff.) interpreted such cases as precursors of the *genitivus absolutus*. This idea is supported by the fact that these acl constructions are especially frequent in the two oldest layers (Fig. 6) whereas possible replacements such as the *locativus absolutus* and regular adverbial clauses become more popular in prose (see the left half of Fig. 7). It should, however, be noted that the corresponding ratio plots of the register (Fig. 7, right) point to pronounced differences between metrical and prose texts which is why the model does not mention them among the top rated constituents according to eq. 5. We may therefore just face a register split in the proper Vedic corpus (layers RV–PL) and early echoes of classical Sanskrit in the last layer.

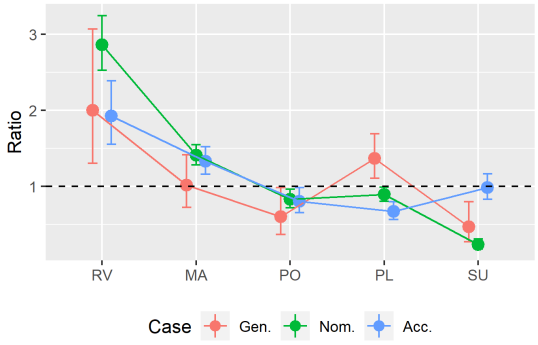


Figure 6: Ratio plots for the proto-*genitivus absolutus* (VERB/Gen/acl/∅) and related adnominal constructions of participles

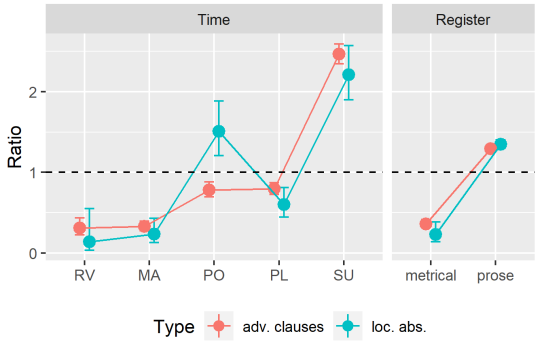


Figure 7: Ratio plots for the *locativus absolutus* (VERB/Loc/advcl/∅) and adverbial clauses in general (VERB/IV/advcl/∅)

As was mentioned above, partition 108 also contains clausal modifiers in nominative and accusative case. The structural link that connects these constituents with those in genitive case are more abstract representations such as VERB/∅/acl/|1| also found in this partition. The data in Fig. 6 shows that the frequency ratios of these two types decrease over the first three layers of the VTB in a similar way as those of the genitive. While, however, the construction in the nominative case becomes dispreferred in the last layer, constructions in the accusative become more popular again towards the end of the Vedic period. This trend is mainly due to accusative participles placed directly in front of their heads as at *Śāṅkhāyana-Gṛhyasūtra* 3.3.10: *abhyaktam aśmānam . . . nikhanet* 'he may bury an anointed stone' where the accusative noun *aśmānam* 'stone' is modified by the past participle of *abhi añj* 'anoint'. Note that Delbrück (1878, 41) interprets such prenominal placement of verbal nouns as indicating that they had assumed an adjectival function.

5.3. Clausal subjects

Partition 22 combines two types of constituents that are structurally and functionally unrelated at first view. The highest values of w are reported for verbal nouns

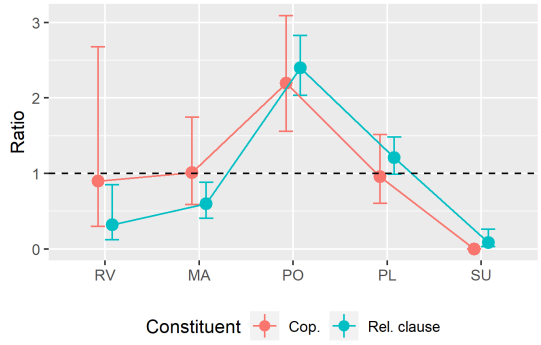


Figure 8: Ratio plots for verbal nouns derived from copulae (VERB/Nom/cop/∅) and relative clauses functioning as subjects (VERB/∅/csubj/∅; see Sec. 5.3)

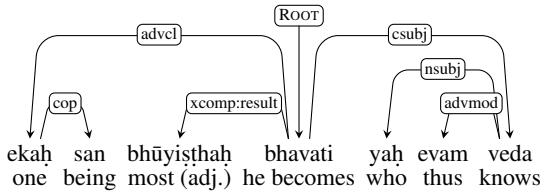


Figure 9: Combination of a clausal subject and the participle of a copula at *Maitrāyaṇī-Saṃhitā* 1.9.5: "He who knows thus obtains most although he is only one."

of copulae in nominative case that function as clausal components (acl) of a noun, directly followed by clausal subjects (csubj) placed (far) to the right of their – mostly verbal – heads. Most of the clausal subjects occur in stereotyped expressions of the form *yaḥ evam veda* 'who knows thus' and variants thereof which describe what a sacrificer must know in order to make a ritual successful (see e.g. Freedman (2012)). The copulae functioning as acl have a similar distribution over the chronological layers (see Fig. 8), and they are occasionally part of the same sentence as the clausal subjects. The example in Fig. 9 shows that both constructions functionally resemble adverbial clauses: While the copula construction has a concessive sense, the clausal subject gives the condition, i.e. the right knowledge of the ritual, for achieving the intended aim. Amano (2009, 121-125) explains the zero subject in such statements by the fact that these exegetical texts assume the sacrificer as the agent if not stated otherwise. The clear chronological distribution in Fig. 8 may therefore be caused by changes in style and content rather than by chronological changes although clausal subjects that have an adverbial sense can already be found in the *Rigveda* (Hettrich, 1988, 575,615) and are therefore not ad hoc formations in the Vedic prose.

Because the csubj constructions preferably occur in stereotyped phrases that can be detected by string search, they also allow to estimate how well the distribution of this constituent in the VTB approximates its distribution in the digitized parts of the Vedic corpus. We form the ratios of csubj constructions and syntactically annotated words per text (r_1) and compare them with the ratios of the string *ya evaṃ veda* and the number of characters per text (r_2).[3] Kendall's rank correlation of r_1 and r_2 yields $\tau = 0.245$ ($Z = 1.4062$, $p = 0.16$) and thus a weak correlation that is not statistically significant at an error level of 10%. The distribution found in the VTB obviously does not fully reflect the corpus distribution, a caveat one should keep in mind when using treebanks of limited coverage for diachronic studies.

5.4. The rise and fall of oblique pronominal arguments

In the last case study, we consider partition 114 which assembles pronouns in any relative placement that function as oblique arguments of verbs. At first view, the ratio plots in Fig. 10 suggest that the members of this partition show clear diachronic developments while register seems to play no role. As pronouns are a closed class of words and have received much scholarly attention in the past (Gotō, 2013), a more detailed inspection of this partition appears worthwhile. After discarding quantifiers with pronominal inflection, we are left with four classes of pronouns for which Fig. 11 gives chronological ratio plots.

Apparently, the four classes have very different diachronic distributions, and the chronological profile in Fig. 10 is a superimposition of them. The profile of the personal pronouns (Skt. *mad-* 'I', *tvad-* 'you') in Fig. 11 is due to register differences: The earliest metrical texts directly address deities and thus make use of pronouns of the second and first person. More interesting is the distribution of the relative pronoun (*ya-*) the ratios of which decrease slowly, but constantly. It may well be the case that we observe a genre-specific phenomenon here that Hock (1992) explains with greater restrictions that didactic Vedic prose imposes on the use of apposite relative clauses: Vedic prose texts focus on imparting knowledge about the ritual in the most effective way and therefore refrain from giving elaborate side information which could be encoded in relative clauses. Although the amount of data is limited (there are only 125 relative pronouns used as oblique arguments in the VTB), such a conclusion receives further support from the fact that almost half of the occurrences in the latest layer come from the *Śvetāśvatara-Upaniṣad*, a speculative text in verses that represents a completely different genre than the Sūtra texts typical for this layer. Genre-related mechanisms may also explain the unexpected distribution of the oblique forms of interrogative pronouns. The peak in Fig. 11 is caused by stereotyped pairs of questions and answers that deal with aspects of the sacrifice and are at least

[3] Note that Sandhi, i.e. phonetic merging of words, prevents a straightforward word count in Vedic texts stored in plain text format.

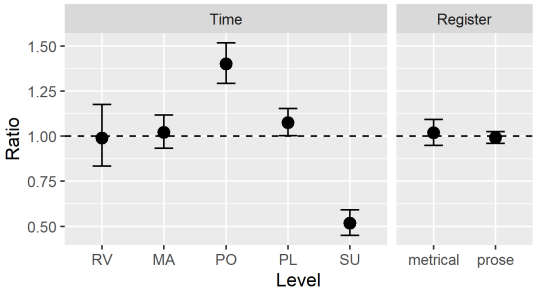


Figure 10: Ratio plots for pronouns functioning as oblique arguments (PRON/∅/obl/∅)

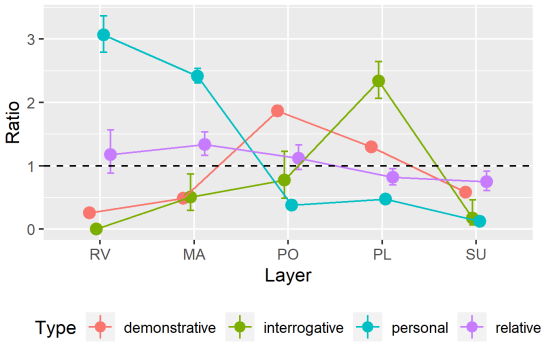


Figure 11: Chronological plot from Fig. 10 (left), split by four classes of pronouns

structurally related to ritualized question-answer contests called *brahmodya* (Thompson, 1997).

The largest class of oblique pronominal arguments consists of impersonal pronouns (labelled 'demonstrative' in Fig. 11). Only four of them occur more than ten times as oblique arguments in the VTB: *sa-/ta-* (anaphoric; see Amano (2009, 55ff.)),[4] *eṣa-/eta-* (discourse deixis, points to something known to hearer and speaker; see Kümmel (2014)) and the enclitic anaphora *a-/ena-* (Amano, 2009, 64ff.). In addition there are instances of the proximal deictic pronoun *ayam-*. Most of the oblique forms of this pronoun are identical with the respective forms of *a-/ena-* in unaccented texts and for this reason not differentiated from this paradigm in the VTB. These cases are subsumed under the class 'unassigned' in Fig. 12. The four types of pronouns show a similar diachronic distribution in Fig. 12: While instances in the two oldest, metrical layers are rare, they occur most frequently in the oldest prose (PO) just to slowly disappear again. The few occurrences of *eṣa-/eta* in the Sūtra layer, for example, are mostly confined to stereotyped uses of the instrumental feminine which refers to mantras accompanying ritual acts.

[4]The frequent use of the accusative singular neuter of this pronoun as a local or temporal adverb is not recorded in Fig. 12 because these instances are syntactically labelled with advmod.

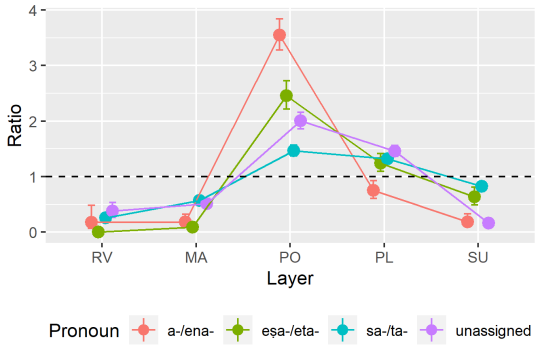


Figure 12: Details for the demonstrative pronouns from Fig. 11

6. Summary

Studying diachronic syntactic changes typically takes its way from qualitative to quantitative research because richly annotated treebanks make it possible to follow the chronological trajectories of syntactic structures marked as noteworthy in qualitative studies. In this paper we have taken the opposite direction because the qualitative work available for Vedic syntax is limited. We have developed a framework that proposes, in a purely data-driven fashion, abstract syntactic structures whose frequencies probably change with time while controlling for influence variables such as register and intellectual affiliation. One obvious way to extend this framework is to allow for combinations of multiple constituents; and another one to account for the content of the source passages as derived, for instance, from contextualized word embeddings.

The method developed here also serves as an intermediate step towards a better understanding of the chronology of Vedic: We aim at finding constituents with a clear chronological profile that can help in clarifying the dates of chronologically disputed Middle and Late Vedic texts. The four case studies in Sec. 5 have shown that some diachronic variation can be explained with differences in style and content after a closer inspection. Such an outcome would not be *per se* problematic for studying the chronology of Vedic – a stylistic change is as good for determining the age of a text as a syntactic one. The central challenge is, however, that the preliminary chronology of the Vedic corpus (see Sec. 3) is ultimately grounded on style and content so that such an approach runs the risk of circularity. Controlling for the content of text passages and above all a careful qualitative scrutiny of possible chronological signals are therefore indispensable for obtaining a clearer picture of this important historical corpus.

Acknowledgments

Oliver Hellwig and Sven Sellmer were funded by the German Federal Ministry of Education and Research, FKZ 01UG2121, when doing research for this paper.

7. Bibliographical References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley and Sons, Hoboken, New Jersey.

Amano, K. (2009). *Maitrāyaṇī Saṃhitā I–II. Übersetzung der Prosapartien mit Kommentar zur Lexik und Syntax der älteren vedischen Prosa*. Hempen, Bremen.

Biagetti, E., Hellwig, O., Ackermann, E., Widmer, P., and Scarlata, S. (2021). Evaluating syntactic annotation of ancient languages. Lessons from the Vedic Treebank. *Old World*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cohen, S. (2008). *Text and Authority in the Older Upaniṣads*. Brill, Leiden.

Delbrück, B. (1878). *Die altindische Wortfolge aus dem Śatapathabrāhmaṇa*. Verlag der Buchhandlung des Waisenhauses, Halle.

Freedman, Y. (2012). Altar of words: Text and ritual in Taittirīya Upaniṣad 2. *Numen*, 59(4):322–343.

Frermann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Gonda, J. (1971). *Old Indian*. Handbuch der Orientalistik, Zweite Abteilung, Erster Band, Erster Abschnitt. E.J. Brill, Leiden.

Gotō, T. (2013). Pronouns. In Toshifumi Gotō, et al., editors, *Old Indo-Aryan Morphology and its Indo-Iranian Background*, pages 66–78. Verlag der Österreichischen Akademie der Wissenschaften.

Haase, C., Anwar, S., Yimam, S. M., Friedrich, A., and Biemann, C. (2021). SCoT: Sense clustering over time: A tool for the analysis of lexical change. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 198–204.

Hellwig, O. and Sellmer, S. (2021). The Vedic treebank. In Erica Biagetti, et al., editors, *Building New Resources for Historical Linguistics*, pages 31–40. Pavia University Press.

Hellwig, O., Scarlata, S., Ackermann, E., and Widmer, P. (2020). The treebank of Vedic Sanskrit. In Nicoletta Calzolari, et al., editors, *Proceedings of the LREC*, pages 5139–5148.

Hellwig, O. (2020). Dating and stratifying a historical corpus with a Bayesian mixture model. In *Proceedings of LT4HALA*, pages 1–9.

Hettrich, H. (1988). *Untersuchungen zur Hypotaxe im Vedischen*. de Gruyter, Berlin.

Hilpert, M. and Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In Merja Kytö et al., editors, *The Cambridge Handbook of English Historical Linguistics*, pages 36–53. Cambridge University Press.

Hock, H. H. (1984). (Pre-)Rig-Vedic convergence of Indo-Aryan with Dravidian? Another look at the evidence. *Studies in the Linguistic Sciences*, 14(1):89–108.

Hock, H. H. (1992). Some peculiarities of Vedic-prose relative clauses. *Wiener Zeitschrift für die Kunde Südasiens*, 36:19–29.

Hock, H. H. (1997). Chronology or Genre? Problems in Vedic Syntax. In Michael Witzel, editor, *Inside the Texts – Beyond the Texts: New Approaches to the Study of the Vedas*, pages 103–126. Harvard University, Cambridge, MA.

Hock, H. H. (2015a). A bibliography of Sanskrit syntax. In Peter M. Scharf, editor, *Sanskrit syntax. Selected papers presented at the seminar on Sanskrit syntax and discourse structures, 13-15 June, 2013, Université Paris Diderot*, pages 399–470. The Sanskrit Library.

Hock, H. H. (2015b). Some issues in Sanskrit syntax. In Peter M. Scharf, editor, *Sanskrit syntax. Selected papers presented at the seminar on Sanskrit syntax and discourse structures, 13-15 June, 2013, Université Paris Diderot*, pages 1–52.

Ishiguro, K., Sato, I., and Ueda, N. (2014). Collapsed variational Bayes inference of infinite relational model. *arXiv preprint arXiv:1409.4757*.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the AAAI*, volume 3, pages 381–388.

Kümmel, M. J. (2014). Pāīiṇi 5.3.5 and the function of sanskrit etád. In Hans Henrich Hock, editor, *Vedic Studies: Language, Texts, Culture, and Philosophy*, pages 39–56. Rashtriya Sanskrit Sansthan and D.K. Printworld, New Delhi.

Kümmel, M. J. (2000). *Das Perfekt im Indoiranischen. Eine Untersuchung der Form und Funktion einer ererbten Kategorie des Verbums und ihrer Entwicklung in den altindoiranischen Sprachen*. Reichert, Wiesbaden.

Lehmann, W. P. (1974). *Proto-Indo-European syntax*. University of Texas Press, Austin.

Lowe, J. J. (2015). The syntax of Sanskrit compounds. *Language*, 91(3):71–115.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Oertel, H. (1926). *The Syntax of Cases in the Narrative and Descriptive Prose of the Brāhmaṇas. I. The Disjunct Use of Cases*. Carl Winter's Universitätsbuchhandlung, Heidelberg.

Patton, L. L. (2006). *Bringing the Gods to Mind: Mantra and Ritual in Early Indian Sacrifice*. University of California Press, Berkeley.

Renou, L. (1937). *La décadence et la disparition du*

subjonctif. Number 1 in Monographies sanskrites. Adrien-Maisonneuve, Paris.

Renou, L. (1947). *Les écoles védiques et la formation du Véda*. Imprimerie Nationale, Paris.

Renou, L. (1956). *Histoire de la Langue Sanskrite*. Edition IAC, Lyon.

Schätzle, C., Dennig, F. L., Blumenschein, M., Keim, D. A., and Butt, M. (2019). Visualizing linguistic change as dimension interactions. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 272–278.

Thompson, G. (1997). The brahmodya and Vedic discourse. *Journal of the American Oriental Society*, 117(1):13–37.

Wackernagel, J. (1905). *Altindische Grammatik. Band II, 1: Einleitung zur Wortlehre. Nominalkomposition*. Vandenhoeck & Ruprecht, Göttingen.

Wezler, A. (2001). Zu der Frage des 'Strebens nach äußerster Kürze' in den Śrautasūtras. *Zeitschrift der Deutschen Morgenländischen Gesellschaft*, 151:351–366.

Witzel, M. (1989). Tracing the Vedic dialects. In Colette Caillat, editor, *Dialectes dans les littératures indoaryennes*, pages 97–265. Collège de France, Paris.

Wüst, W. (1928). *Stilgeschichte und Chronologie des Ṛgveda*, volume XVII of *Abhandlungen für die Kunde des Morgenlandes*. Deutsche Morgenländische Gesellschaft, Leipzig.