The Vedic corpus as a graph. An updated version of Bloomfield's Vedic Concordance

Oliver Hellwig
Heinrich-Heine-Universität Düsseldorf
Oliver.Hellwig@uni-duesseldorf.de

Sven Sellmer
Heinrich-Heine-Universität Düsseldorf
Adam Mickiewicz University Poznań
sellmer@uni-duesseldorf.de

Kyoko Amano
Kyoto University
amano.skskrt@kcn.jp

Abstract

Bloomfield's *Vedic Concordance* is a valuable resource for understanding the history of Vedic literature, but its structure, which is designed for human readers, prevents its straightforward application in quantitative studies. This paper introduces a new version of the *Vedic Concordance* representing Bloomfield's text in an XML structure that facilitates the programmatic access to this data. We describe the steps necessary for creating the new representation and report about the ongoing extension of the Concordance with late Vedic material. In addition, we present four case studies that illustrate how the new version of the *Vedic Concordance* can be utilized in textual studies.

1 Introduction

The Vedic corpus is thematically largely restricted to a single topic, namely the Vedic sacrifice.[1] While the early metrical Saṃhitās can best be understood as hymns accompanying various sacrifices, the later Vedic prose explains these sacrifices and prescribes how they should be performed. An important trait of these exegetical texts is the frequent citation of mantras from the earlier metrical texts: For a sacrifice to be successful it is crucial to recite the appropriate mantras, which alone guarantee that the sacrifice achieves its purpose (see e.g. Staal (1996) and Patton (2006)). As a consequence, the Vedic literature abounds in mantras quoted in full, in parts and with different degrees of variation. In addition to these citations, the early metrical Saṃhitās share a large number of mantras and pādas that occur with different degrees of variation, a phenomenon that Witzel (1989) explains with a pre-Rigvedic reservoir of "floating mantras".

The importance of mantra citations for reconstructing the history of the Vedic literature has been understood early in Vedic studies, and in 1906 Maurice Bloomfield published the *Vedic Concordance* (VC), a systematic collection of all mantra citations found in the texts edited at that time (Bloomfield, 1906). The VC was re-published in digital form by Marco Franceschini (Franceschini and Bloomfield, 2007), who also extended the inventory of mantras on the basis of several texts published in the meantime. Since its first publication, Bloomfield's Concordance has been used extensively in scholarly research, but only, so to say, on a micro-level: While numerous research papers from the fields of Vedic Studies or Linguistics have based their argumentation on individual entries of the VC, no paper has so far studied what these data as a whole can tell us about the history and structure of the Vedic corpus. Apart from the fact that quantitative methods such as network analysis and graph theory that can provide a global view of Bloomfield's data are rarely applied in this field, there is one very practical reason for this obvious shortcoming: In its present form, the VC is human-, but not machine-understandable. Consider, as an example, line 1,676 of the electronic version of the VC:

```
agne prehi prathamo devayatām (AVŚ. devatānām; MS.KS. devāyatām) #
AVŚ.4.14.5a; AVP.3.38.3a; VS.17.69a; TS.4.6.5.2a; 5.4.7.1; MS.2.10.6a:
```

[1] We use the abbreviations given by Bloomfield when discussing examples from the Vedic Concordance, and those given by Griffiths (2009, 453-456) in all other cases. Citations from the VC are printed in typeface.

Proceedings of the Computational Sanskrit and Digital Humanities, 18^{th} *World Sanskrit Conference 2023*, pages 188–200

```
138.4; KS.18.4a; 21.9; ŚB.9.2.3.28. P: agne prehi Vait.8.17; 15.9;
Kauś.63.9; 137.27.
```

Bloomfield admittedly formatted this entry with maximum brevity in mind, but the reader must replace *devayatām* with *devatānām* for constructing the text of the AVŚ, and with *devāyatām* for constructing that of the MS and KS, as indicated by the content of the round brackets inserted into the mantra text. In addition, the second mantra variant occurs twice in the KS. This must be inferred from the semicolon-separated string `KS.18.4a; 21.9`, which must be resolved into KS 18.4a and KS 21.9. The occurrences of the pratīka `agne prehi` are encoded analogously without giving full citations.

While these steps are mostly not problematic for a human reader (with some important exceptions, on which see Sec. 2.1), this format is clearly not suited for automatic processing, which may explain why Bloomfield's data have not been used for a network analysis of Vedic texts so far. In this paper, we present a new machine-understandable version of the VC and give some initial pointers as to how the transformed data can be used for studying the development of the Vedic corpus.

In Sec. 2, we describe the new storage logic of the VC and the steps required for transforming the current digital version into this new format (2.1, 2.2). We report about ongoing work in expanding the coverage of the VC with texts from the late Vedic period (2.3) and give an overview of basic properties of the co-occurrence graph constructed from the data (2.4). Section 3 presents the results of four case studies that demonstrate which research questions can be addressed when using the new form of the VC and how the data can be visualized. Section 4 summarizes the paper. – The new version of the VC is available for download at `https://github.com/OliverHellwig/sanskrit/tree/master/papers/2023wsc`.

2 Transforming and extending the Vedic Concordance

As the example on page 1 has shown, parsing the VC in the form published by Franceschini and Bloomfield (2007) requires implicit knowledge that is not encoded in its structure. In this section, we describe how such implicit information is transformed into explicit one (Sec. 2.1). In addition, we introduce the new data structure (Sec. 2.2), report on the ongoing extension of the VC (Sec. 2.3) and describe basic properties of the co-occurrence graph (Sec. 2.4). We used standard software and short Python scripts for carrying out the steps described in this section. All transformations described in the following were carried out on the data published electronically by Franceschini and Bloomfield (2007).

2.1 Transforming the data

Each entry of the VC offers information about various facets of a mantra. This section describes these facets and their computational processing. – The strong variability of Vedic mantras is well known in Vedic Studies, and Bloomfield himself was often sceptical which form of a mantra should be accepted as its **basic form** (Bloomfield, 1906, xiv-xv). In the new version, the basic form is the mantra text of Bloomfield's entry, from which all variant information (in round brackets, see below) has been removed. If a mantra occurs in more than one text, its basic form is the one found in the oldest text according to Bloomfield's temporal arrangement of the Vedic literature (Bloomfield, 1906, xvi). This arrangement is certainly correct in most cases, esp. in those involving the RV and other early Saṃhitās, but needs a critical revision given the full graph emerging from the VC. We are currently working on a mathematical formulation of this problem.

Regarding the **citations** of a mantra, we face the challenge that the name of the text is only given once even if a mantra occurs more than one time in a text (see the example on p. 1). We solve this issue by repeatedly applying regular expressions to the original version of the VC. Citations of the form 'VS.8.4d; 33.68d' are, for example, resolved by a regular expression of the form

```
([^0-9 \.]+)\.([0-9]+\.[0-9]+[a-z]*); ([0-9]+\.[0-9]+[a-z]*)([\.;])\b
```

with replacements

```
\1\2; \1.\3
```

This regular expression leaves the first citation 'VS.8.4d' unchanged, but generates 'VS.33.68d' from the part after the semicolon. A few hundred cases that could not be handled using this approach (mostly due to inconsistencies in the encoded data) were resolved manually.

Minor variants of mantras that do not change the alphabetical order of an entry are certainly the most problematic part of the previous version of the VC. Having access to the full form of variants is, however, important for obtaining a clearer picture of the history of the Vedic corpus because they throw light on the processes of textual transmission. As shown in the example on p. 1, such variants are enclosed between round brackets in the main text of a mantra. Contrary to what Bloomfield claims (Bloomfield, 1906, xiv-xv), reconstructing the text of such variants can be challenging even for a human reader without knowledge of Vedic Sanskrit. Consider, for instance, line 70442: `vājajic ca bhava (VSK. caidhi) …`, where the two strings *ca bhava* must be replaced with one sandhied string *caidhi* (also note that the verb changed from *bhū* to *as*); or line 57802: `mā no agne'va (MG. vi)` where the sandhied preverb *ava* must be replaced with *vi*, resulting in two strings instead of one (*mā no agne vi …*). We distinguish the following conventions used in the VC (uppercase letters A, B stand for text names):

- …(A.B x): One or multiple strings to the left of the round bracket are replaced with string(s) x in texts A, B. The encoding does not indicate which strings must be replaced, the surface forms can vary significantly, and sandhi can change the number of strings involved. Manual intervention is required for resolving these cases.
- …(A.B @x) or …(A.B x@): A suffix or prefix of a word to the left of the bracket must be replaced with x in texts A, B. The same precautions as in the preceding case apply.
- …(A.B add x): One or more strings must be inserted in place of the bracket.
- …(A.B. omit(s) x): String(s) x must be deleted to the left of the round bracket. Note that this encoding is not consistent, even inside a single mantra: `… devo daivyo (omitted in ApŚ, BaudhŚ., ApŚ, HirŚ) … amuvad amuvat (ŚB. omits amuvad amuvat)` (line 999).

Fuzzy string matching algorithms such as Smith-Waterman could solve many, though not all of these cases, because Bloomfield's notation is not consistent and the scopes of some round brackets are unclear even for a human reader. In order to make the resource as accurate as possible, we therefore decided to manually process a csv file created in the following way:

- We extract the text names from the round brackets and add one line for each text name to the csv file.
- If the replacement x is a single string and if the string preceding the bracket does not contain a visarga, we replace the string to the left of the bracket with x for all texts mentioned in the bracket. The resulting string is written into the respective line of the csv file.
- In all other cases, the content of the round bracket without the text names is copied verbatim in the respective line of the csv file.

One author of this paper manually corrected the resulting csv file, paying special attention to the correct resolution of the variant information.

Apart from these minor variants, the VC features **cross-references** between "variants involving more than one alphabetic entry", which are introduced using the keywords 'See' and 'See under' (see Bloomfield (1906, xv) for details). Finding the targets of such cross-references poses similar problems as the resolution of minor variants because active knowledge of Vedic is again required in many cases. We generate another csv file that contains such cross-references and manually resolve cases in which a unique target line could not be determined automatically. While the original VC contains, for example, the entry `aṃhārir asi bambhāriḥ # ŚŚ.6.12.20. See aṅghārir`, making it necessary that the reader looks up a mantra starting with the word *aṅghārir*, the new XML version maps `See aṅghārir` to the correct line *aṅghārir asi bambhāriḥ* (line 2305).

The VC also features **pratīkas**, i.e. abbreviated citations of a mantra used preferably when an author cites from a Saṃhitā of his own school (Renou, 1947, 26-68). One of the nearly 8,000 pratīkas in the VC is found in line 79,937 where the mantra *saṃ te payāṃsi sam u yantu vājāḥ* (RV 1.91.18a) is abbreviated as *saṃ te payāṃsi* in ŚāṅkhŚS 1.15.4. Pratīkas are introduced with P in the VC and followed by a list of their occurrences which is structured in the same way as that of standard occurrences (see above). After resolving citation information in the same way as for the basic form, pratīkas are extracted automatically using regular expressions.

The (digital version of the) VC contains further pieces of information such as cross-references introduced by 'Cf.' (loose parallels), variae lectiones and occasional pointers into the secondary literature. This supplementary information, which is not central for understanding the structure of the Vedic corpus, has not been included in the present updated version of the VC, but could easily be integrated in future releases of the resource.

2.2 The new storage format

The transformations described in the preceding section modify the original plain text file of the VC (citations) or generate new files (variants). We now need to merge these data into a format that is platform independent and easily processed by a wide range of programming languages. Each entry in the VC can be thought of as a hierarchical structure combining global information (e.g. the original line, the text of the mantra) with a varying number of occurrences in the Vedic corpus, of variants and of cross-references to other lines of the VC. Due to this hierarchical structure, XML recommends itself as storage format, but other structured representations such as JSON would be equally feasible. Using standard libraries of popular programming languages, transforming the XML representation into other formats poses no difficulties.

The following list gives an overview of the new structure and its elements whose names are printed in bold. Indentation indicates hierarchy levels in the resulting XML file, data types are printed in italics, asterisks (*) mark optional elements, and the names of XML attributes start with @:

original-text [*string*] The original line as given in Franceschini and Bloomfield (2007)

number [*integer*] The line number in the text file published by Franceschini and Bloomfield (2007)

mantra [*string*] The basic form of a mantra (see p. 2 of this paper).

citations [*list*] The list of citations, each of which has the following data members:

- **@type** : 'default' (i.e. identical with the basic form), 'variant', 'pratika'
- **@source** This attribute distinguishes between entries found in Franceschini and Bloomfield (2007) and new occurrences added by authors of this paper: 'F(ranceschini)', 'HS' (Hellwig, Sellmer).
- **text** [*string*] The source text that contains the citation
- **chapter** [*string*] Where the mantra is found in the text. For 'AŚ.2.16.19a', for example, `text` is set to 'AŚ' and `chapter` to '2.16.19a'.
- ***string** [*string*] If the occurrence is a variant or a pratīka, this member gives its full text.

***see** [*list*] The list of mantra variants that change the alphabetic order.

- **targetNumber** [*integer*] The line number of the cross-referenced entry.
- **@type** [*string*] 'child' (introduced with 'See under' in Bloomfield (1906) or the second entry of a binary cross-reference), 'parent'

2.3 Extending the Vedic Concordance

We are extending the VC with material from late Vedic texts. We started with the mantras in the BhārŚS for which we performed the following steps:

- From a digital version of BhārŚS 1-12 that encloses mantra citations in pointed brackets[2] we extracted all strings marked as mantras.

[2]Many thanks to Francois Voegeli for providing this text.

Mantras	87,885
Occurrences	212,908
Mantras with more than one occurrence	38,606
Pratīkas	10,762
Minor variants	15,398
Cross-references	6,718

Table 1: Composition of the Vedic Concordance

- We searched the strings thus obtained in our version of the VC using the Levenshtein algorithm for string comparison. If the mantra in the BhārŚS is identical with the basic form of a mantra contained in the VC, its occurrence was added automatically to the VC. All other mantras and their context were written in a csv file.
- Two authors of this paper revised the entries in the resulting csv file, labelling each one either as a variant or a pratīka of an existing mantra, or adding a new mantra to the VC. As in the case of variant resolution (see above), this manual approach was chosen to create a resource as accurate as possible, and extensions using machine learning methods are easily conceivable.

In the same way we added the citations found in the complete BaudhŚS, which increased the number of occurrences by more than 4%.

In order to facilitate the manual revision, we designed a lean user interface based on PHP and XPath that can be used for querying and extending the XML version of the VC. Here the user can search for mantras either in their basic forms, in variants or in both. The retrieved mantras are presented as a list, each entry being formatted with a hyperlink for editing the information associated with it. By clicking on such a link a dialog window is opened in which the type and text of a mantra can be changed manually. This interface can be extended in multiple ways, e.g. by implementing fuzzy string search or searching for further types of information. The source of this UI is published along with the XML version of the VC so that such changes can be implemented by interested users.

2.4 Properties of the VC and its citation graph

Table 1 gives the frequencies of some important elements of the updated VC. The mere size of almost 88,000 entries is impressive, especially when we keep in mind that Bloomfield collected this material without the help of a computer, a fact also emphasized by Jamison (2010). However, only 37,877 from among these entries occur more than once in the extant Vedic canon. The number of entries would also be reduced considerably if full text lines would be considered instead of pādas, as was done by Bloomfield. Updating the VC with such text-line information should not be too challenging because occurrences of pādas are terminated with lowercase letters (e.g. `1.36.17c`).

The order in which Bloomfield recorded the occurrences of a mantra reflects how he understood the chronology of the Vedic corpus, the first occurrence given usually indicating the source of the mantra in Bloomfield's opinion. While it is possible to create a directed graph based on this arrangement, the actual transmission history of individual mantras is far less certain than suggested by these arrangements. For a substantial number of mantras, Bloomfield's ordering is nothing more than an educated guess.[3] To better understand the structure of the VC, we therefore construct an undirected co-occurrence graph. Instead of dividing the texts into citing and cited ones, all texts in which a given mantra appears are indiscriminately connected to each other by this co-occurrence. Let us now have a look at some basic properties of the resulting undirected graph, some of which are expected, others perhaps less so. The graph consists of 121

[3]See e.g. line 22,368 of the VC which records a mantra that is shared by three Gṛhysūtras with unclear chronological relationship: `ūrjaṃ prajām amṛtaṃ dīrgham āyuḥ (AG. amṛtaṃ pinvamānaḥ) # AG.2.4.14c; PG.3.3.6c; MG.2.8.6c.`.

Text 1	Text 2	Weight
MS	KS	6058
TS	MS	4986
VS	MS	4898
TS	KS	4775
RV	AVŚ	4678

(a) Highest edge weights (i.e. numbers of mantras shared by two texts) in the undirected co-occurrence graph built from the new version of the VC

Text	Betweenness
AVŚ	385
RV	302.3
TA	235.9
VS	213
GDh	153.9
GB	139.3
MahānU	134.5
ŚvetU	117.4
ApŚ	108.5
MŚ	108.3

(b) Highest betweenness centrality values

Table 2: Graph-theoretical properties of the VC network

vertices that represent the texts included in the VC and 3270 edges that can represent any of the possible connections discussed in Sec. 2.1 (exact citation, variant, or pratīka). The weights of the edges in this graph, i.e. the number of co-occurrences linking two given texts, vary strongly. The four highest values are assigned to edges between Saṃhitās of the Black and White Yajurveda, the fifth to the edge linking the RV and the AVŚ (see Table 2a). Notably the *Ṛgveda-Saṃhitā* is only part of the fifth pair, which shows that it does not have the same privileged position in an undirected graph as it would in a directed one. The edge weight distribution is heavily left-skewed, so that there is a very high number of low-weight edges (as many as 2856 with a weight < 100). This fact is largely due to the small size of many of the texts involved, e.g. most Upaniṣads. In order to nevertheless save the information contained in these weight attributions, which tends to be marginalized by the highest few percent, we employ the natural logarithm of the weight values for tasks such as community detection. In this way, the skew is sufficiently minimized, so as to allow finer structures to emerge.

The most basic property of the vertices is their degree, i.e., the number of edges they possess. Because the VC graph is undirected, among the ten vertices with the highest degrees (between 100 [RV and VS] and 94 [AVŚ]) there are both texts that are predominantly, or even exclusively—as is the case of the RV—*cited*, and such, typically long and younger ones, that contain many citations taken from other texts. To the first category belong RV, VS, MS, TS, KS, and AVŚ, to the latter one ApŚS, MŚS, ŚB, and BaudhŚS. A more sophisticated measure of the importance of a vertex is obtained by applying the PageRank algorithm (Brin and Page, 1998), which considers the degree of a given vertex as well as those of its neighbors, so that it more adequately shows the real influence of a vertex in a graph. Although a high degree does not necessarily correspond to a high PageRank count, in our case the top ten texts according to both measures are in fact almost identical. Notably the AVŚ takes second place just after the RV according to PageRank, whereas in the degree ranking it comes only at position no. 9, which can be attributed to the large number of mantras linking it to the first-placed RV.

Another interesting measure of centrality is the so-called betweenness (Freeman, 1977), for which the number of edges of a vertex is not decisive. Instead, betweenness favors vertices that lie on the shortest path between many other vertices, that is, in functional terms, vertices that act as a kind of bridge between different communities in a network. The texts with the ten highest betweenness centrality values in the unweighted graph are given in Table 2b. Here, the two oldest texts of the Vedic corpus, the Saṃhitās of the RV and AV, occupy the top two positions. Such a result could be expected because RV and AVŚ are cited by numerous other texts from different traditions and thereby constitute a link between them. More remarkable is the quite

high betweenness value of a very late (and rather short) text like the *Śvetāśvatara-Upaniṣad*.[4] It can be explained by the syncretic character of this Upaniṣad, which becomes apparent in the wide range of sources on which it draws.

3 Research problems and applications

3.1 Pratīkas

In order to convey an idea of which questions can be addressed with the new format of the VC, let us assess a hypothesis about the usage of mantras in Vedic texts that was brought forward by Gonda (1977, 505-506). Gonda claims that younger Sūtra texts preferably cite mantras as pratīkas, although exceptions to this rule are widely found in Vedic literature. Such a claim can be tested by making use of the new structure of the VC. We extract mantra citations found in all Śrautasūtra texts contained in the VC and record if each mantra is cited in full or as a pratīka.[5] We order the resulting list of texts by increasing pratīka ratios, i.e. by the proportion $\frac{\#\text{pratīka}}{\#\text{full}+\#\text{pratīka}}$ per text. Next, we perform pairwise G-tests (Agresti, 2007) that compare the counts of full and pratīka citations in texts i and $j : j > i$. If such a test gets significant at a level of 0.01 (1%), we consider the pratīka ratio of i as significantly lower than that of j and derive a pairwise ordering $i < j$ from the test result. These partial orderings are used to construct a directed acyclic graph (DAG) whose acyclicity results from the fact that the texts were initially ordered by increasing pratīka ratios and from $j > i$. A topological ordering of the nodes of this DAG produces the following arrangement of Śrautasūtras where a colored box around multiple texts indicates that the G-test did not become significant and that therefore no partial ordering could be induced for the respective group:

$$\text{BhārŚ} < \boxed{\text{BaudhŚ HirŚ ApŚ}} < \text{MŚ} < \text{AŚ} < \boxed{\text{Vait ŚŚ LŚ KŚ}}$$

This arrangement agrees in parts with the temporal order considered probable by Gonda (1977, 482-483), who claims that BaudhŚS < BhārŚS < ĀpŚS < HirŚS, that KātyŚS and ĀśvŚS are not among the oldest texts, and that the ŚāṅkhŚS is from about the same time as the ĀśvŚS. The graph-based evaluation of pratīka citations thus provides another piece of evidence that helps to understand the history of the late Vedic literature.

3.2 The Random Kavi model

In the second case study we throw a look on what the graph structure of the VC can tell us about the relationship between the early Saṃhitā texts. Witzel (1989, §4.3.2-3) brought up the idea that the Saṃhitās drew their texts from a repository of "floating mantras". In an admittedly over-simplistic setting, such a process can be approximated using a probabilistic urn model without replacement. We think of the whole VC as one huge urn containing $N = 87,885$ mantras (see Tab. 1) from which one Kavi randomly draws mantras for composing his text A;[6] just to repeat, this is an over-simplification. After this Kavi has drawn his $|A|$ mantras, all mantras are placed back in the urn, and the second Kavi draws $|B|$ mantras for his text B. The probability that two sets A, B drawn in this way have k elements in common follows a hypergeometric distribution that can be derived from the Venn diagram in Fig. 1a:

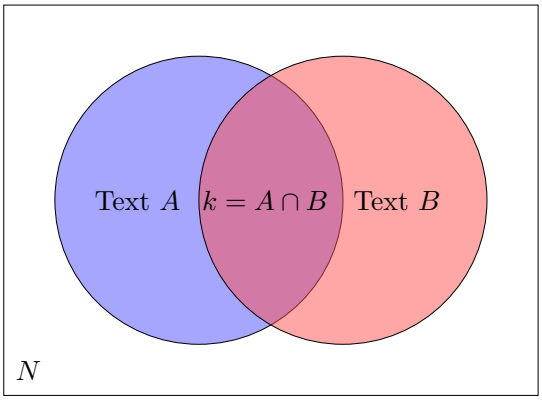
$$p(A, B, k, N) = \frac{\binom{|A|}{k} \cdot \binom{N-|A|}{|B|-k}}{\binom{N}{|B|}} \tag{1}$$

The probability of drawing text B in its transmitted form is the inverse of the binomial coefficient $\binom{N}{|B|}$. Taking the composition of A as given, there are $\binom{|A|}{k}$ ways to choose a subset of k mantras from its $|A|$ mantras. The second term in the numerator of eq. 1 gives the number of choices

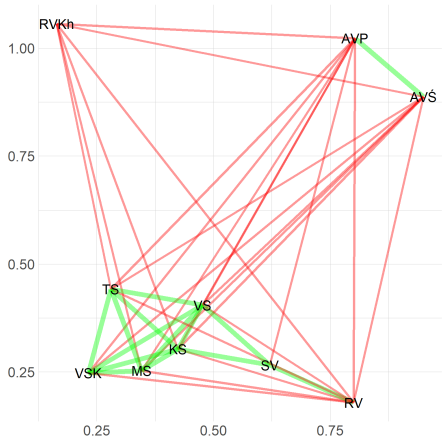
[4] The equally late *Mahānārāyaṇa-Upaniṣad* is a special case as it largely is identical with TĀ 10.

[5] Note that variants are counted as full citations and that we only consider citations from texts of the same school.

[6] The term mantra is used somehow sloppily here as most entries of the VC are actually pādas.



(a) Venn diagram for the Random Kavi model



(b) Results of the Random Kavi model for the Saṃhitās. Green: right tail; red: left tail. The distance between a pair of texts is approximately proportional to $1 - p(x \geq k)$.

Figure 1: The Random Kavi model: Venn diagram and results for the Saṃhitās

the second Kavi has after the first one has chosen his $|A|$ mantras: Kavi number two can choose from among $N - |A|$ mantras, and he chooses only $|B| - k$ mantras, because the intersection k is already contained in text A. The distribution in eq. 1 has one free parameter, which is the size of the intersection k. To obtain a proper probability distribution we form the sum of eq. 1 for all sizes of intersections $0 \leq x \leq \min(|A|, |B|)$ (the intersection k can maximally have the size of the shorter of the two texts) and divide eq. 1 by this sum. The probability of obtaining an intersection equal to or larger than k is $p(x \geq k) = \sum_{x=k}^{\min(|A|,|B|)} p(A, B, x, N)$. If this probability is less than 0.001,[7] we assume that the involved texts A, B share many more mantras than could be expected under the Random Kavi model and that they are therefore closely related to each other. Inversely, if $p(x < k) = \sum_{x=0}^{k-1} p(A, B, x, N)$ is less than 0.001, we assume that texts A, B share much less mantras than could be expected under the hypergeometric model.

Figure 1b visualizes the resulting probabilities for the early Saṃhitā texts, indicating the right tails by green lines (size k is larger than expected) and the left tails by red ones. The most obvious feature of this figure is the close connection between the texts of the Yajurvedic schools the texts of which share much more material than expected under the Random Kavi model. Such a result is not surprising given the close connections between these texts in terms of their internal structure, content and ritual application. The SV is loosely connected to this cluster and forms the only strong connection with the RV – again, this result is not surprising when one considers the nature and school association of these two texts. The two recensions of the AV are strongly associated with each other, but do not share strong connections with any of the other Saṃhitās. This special position reflects the status of the AV in comparison to the three other Vedic schools (on which see e.g. Gonda (1975, 267-270)). In our opinion, Fig. 1b thus highlights the fact that the Vedic schools separated early in the history of Vedic literature and that the Atharvaveda Saṃhitās occupy a special position in the early history of the Vedic corpus although they share a substantial amount of mantras with other early texts.

[7] We choose a very small threshold of 0.1% to obtain conservative results.

3.3 Citations from the RV

Citations of RVic mantras occur throughout the Vedic literature, and the importance of the RV as a source of mantras has been emphasized repeatedly in previous research; see e.g. Gonda (1975, 368-370) on the use of RVic mantras in the Brāhmaṇas. It is known that different schools cite from different parts of the RV as, for instance, the Sāmaveda preferably uses material from the Soma hymns in RV 9 (see the large rectangle in the first row of Fig. 2a). Such school-specific preferences interact with text-specific ones which may arise from the topic of a text or the preferences of its author(s). Given that a single text may cite hundreds or even thousands of RVic mantras, distinguishing between text- and school-specific preferences is tedious and time consuming when done in a traditional philological way. We therefore propose a latent variable admixture model that disentangles text- and school-specific contributions to the 60,675 citations of RVic passages found in the VC.[8]

The design of this model is inspired by Latent Dirichlet Allocation (Blei et al., 2003). It considers each citation from one of the ten books of the RV as an atomic feature whose presence is due either to citation habits of the school of a text or to special preferences of the text itself. More formally, there are N mantra citations m_i from the ten books of the RV, $m_i \in \{1, 2, \ldots, 10\}$. For each observed data point m_i we have the citing text $t_i \in T$ and its school $s_i \in \{\text{RV}, \text{YV}, \text{SV}, \text{AV}\}$. Each text t is associated with a text-specific Bernoulli distribution $\text{Bern}(\mu_t)$. The latent variable z_i is drawn from the text-specific Bernoulli distribution $\text{Bern}(\mu_{t_i})$. If $z_i = 1$, the observed RVic book is drawn from a text-specific multinomial distribution $\text{Cat}(\boldsymbol{\Theta}_{t_i})$, $\boldsymbol{\Theta}_t \in \mathbb{R}^{10}$; if $z_i = 0$, the RVic book is drawn from a school-specific multinomial with parameters $\boldsymbol{\Phi}_{s_i} \in \mathbb{R}^{10}$. Using a beta prior with uniform hyperparameters $\boldsymbol{\alpha}$ for the Bernoulli distributions and Dirichlet priors with flat hyperparameters $\boldsymbol{\beta}$ for the text- and school-specific multinomials, we obtain the following joint distribution of this model:

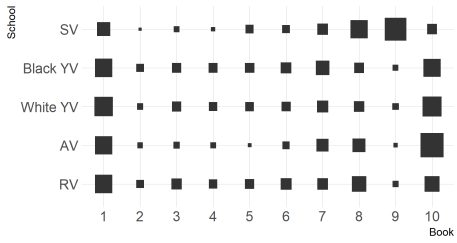
$$p(\mathbf{m}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \mathbf{t}, \mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{t \in T} \text{Beta}(\mu_t | \boldsymbol{\alpha}) \cdot \prod_{t \in T} \text{Dir}(\boldsymbol{\Theta}_t | \boldsymbol{\beta}) \cdot \prod_{s \in S} \text{Dir}(\boldsymbol{\Phi}_s | \boldsymbol{\beta}) \cdot \prod_{i}^{n} \Big[\text{Bern}(z_i | \mu_{t_i}) \cdot (z_i \cdot \text{Cat}(m_i | \boldsymbol{\Theta}_{t_i}) + (1 - z_i) \text{Cat}(m_i | \boldsymbol{\Phi}_{s_i})) \Big] \quad (2)$$

With $\mathbb{I}[.]$ denoting the Dirac delta function, let $a_{t1} = \sum_j^N \mathbb{I}[t_j = t, z_j = 1]$ (number of cases in which a citation in text t_i is labelled as text-specific), $a_{t0} = \sum_j^N \mathbb{I}[t_j = t] - a_{t1}$, $b_{tm} = \sum_j^N \mathbb{I}[t_i = t, z_i = 1, m_i = m]$ (number of cases in which a citation of the RVic book m_i is labelled as text-specific in text t_i), $c_{tm} = \sum_j^N \mathbb{I}[t_i = t, z_i = 0, m_i = m]$, and a_{t1}^{-i} denote the value of a_{t1} excluding the current data point i. Using this notation, we can derive the following collapsed Gibbs sampler:

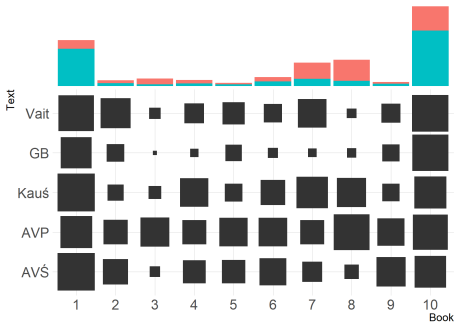
$$p(z_i = k | \mathbf{z}^{-i}, \mathbf{m}^{-i}, \mathbf{t}^{-i}, \mathbf{s}^{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \begin{cases} (a_{t1}^{-i} + \alpha) \cdot \frac{b_{t_i m_i}^{-i} + \beta}{\sum_u^{10} b_{t_i u}^{-i} + \beta} & \text{if } z_i = 1 \\ (a_{t0}^{-i} + \alpha) \cdot \frac{c_{t_i m_i}^{-i} + \beta}{\sum_u^{10} c_{t_i u}^{-i} + \beta} & \text{if } z_i = 0 \end{cases} \quad (3)$$

The sampler from Eq. 3 is run for 30,000 epochs. After a burn-in period of 10,000 epochs we record the state of the hidden variables $\mathbf{z}$ after each 1,000th epoch in order to obtain uncorrelated samples. We use the sampled values of $\mathbf{z}$ for calculating posterior estimates of the variational parameters $\boldsymbol{\Theta}, \boldsymbol{\Phi}$. The values of the hyperparameters often exert a strong influence on the result. We therefore search for good values of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ using posterior predictive checks (Mimno et al., 2015). For a given pair α, β, we run the sampler as described above, calculate the posterior estimates of $\boldsymbol{\Theta}, \boldsymbol{\Phi}$, and randomly draw $N = 60,675$ samples from the model described in Eq. 2. We obtain a probability distribution q over the ten books of the RV by normalizing the counts

[8] This number of citations does not contradict the number of 38,606 mantras with more than one occurrence reported in Tab. 1 because individual RVic mantras are often multiply cited in more than one text.



(a) Proportional frequencies of citations from the ten books of the RV (x-axis) in the five Vedic schools (y-axis).



(b) Proportions of RVic citations (x-axis) labelled as school-specific in major Atharvavedic texts (y-axis). The smallest square represents 2.2% and the largest one 91.7%. The stacked bar plots on top visualize the marginal proportions with which each RVic book is cited (blue: school-, red: text-specific).

Figure 2: Citations from the RV in the Vedic literature. Left: Observed values. Right: Results of the admixture model defined in eq. 2 for the Atharvanic texts in the VC.

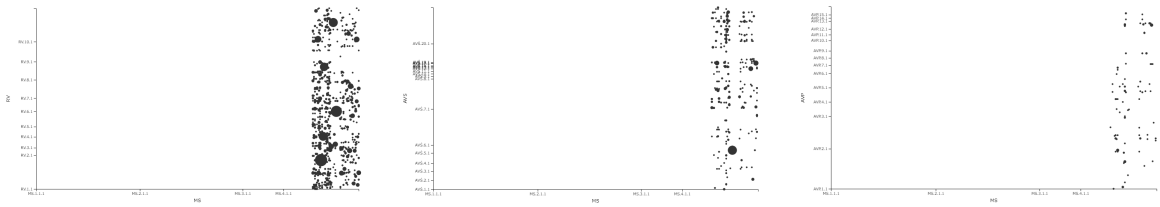
in the sample, and compare q with the observed global distribution p using the Kullback-Leibler (KL) divergence. When increasing the values of α, β in exponential steps in a grid search, the lowest values of $\mathrm{KL}(p, q)$ (i.e. the best generative results) are obtained for $\alpha = 1, \beta = 10$.

Figure 2b shows the model output for major Atharvanic texts when the optimal hyperparameter values are applied. The x-axis of the plot lists the ten RVic books, and its y-axis shows texts from the AV tradition that use mantras from the RV. The squares in the main plotting area visualize the proportions of data points m_i that are labelled as school-specific for each combination of citing text t and cited RVic book m, the edge lengths of the squares being proportional to $\frac{c_{sm}+\alpha}{b_{tm}+c_{sm}+2\alpha}$. These proportions are in general highest for the first and the tenth books of the RV, this means those RVic books that contain most of the material considered as Atharvanic (see e.g. Arnold (1897); Bloomfield (1899b, 45ff.); Gonda (1975, 28)). Contrary to that, the model tends to explain citations from RV 2-9 by text-specific citation preferences. Such preferences can, for instance, be observed for mantras from RV 8 cited in the AVŚ (see the respective tiny rectangle in the bottom row of Fig. 2b), but they are most obvious for the *Gopathabrāhmaṇa* (GB) which displays a very peculiar style of citing from the RV. This may be another indication of the late date and syncretistic character of this text (see e.g. Bloomfield (1899a, 101-104)) whose author(s) may have had access to the fully canonized version of the RV.

3.4 Rituals and internal structure in the MS

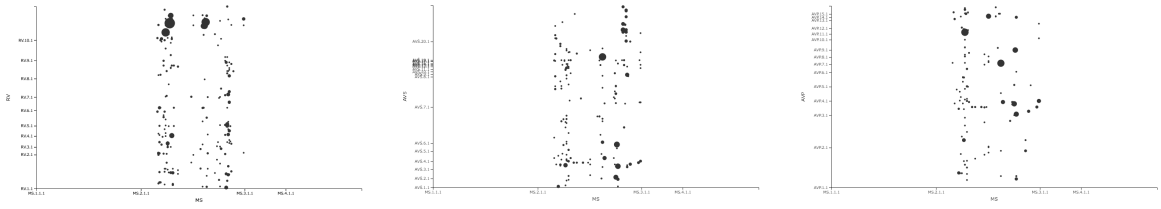
As the fourth case study we present a web-based interface that visualizes mantra co-occurrences in the Vedic corpus. This visualization system which is accessible at `http://34.146.175.179/` is implemented using standard web technologies (HTML, CSS, Javascript) and D3.js. Data are obtained from a relational SQLite database which currently contains mantra occurrences in nineteen Vedic texts, as recorded in the VC. As an example of what can be achieved with such a system, let us have a look at the *Maitrāyaṇī-Saṃhitā* (MS), a foundational text of the Black Yajurveda. The MS groups rituals in dedicated chapters, and each chapter shows its own peculiarities of language and style (Amano, 2020). This state of affairs probably is due to the fact that the development of the rituals and the composition of the respective chapters took place in chronologically and culturally different circumstances.

To see how mantra co-occurrences can help us to understand these processes and especially



(a) RV (b) AVŚ (c) AVP

Figure 3: The mantra citations from RV, AVŚ, and AVP (y-axes) in the MS (x-axis) *yājyānuvākyā* chapter



(a) RV (b) AVŚ (c) AVP

Figure 4: The mantra citations from RV, AVŚ, and AVP in the MS *agniciti* chapter; see Fig. 3

the relationship between the MS and the early metrical Saṃhitās (RV, AVŚ, AVP), we focus on two rituals:

1. The *agniciti* (MS 2.7–13) is the ritual of building a huge fire altar. This ritual is extended by numerous, partly esoteric ritual components. As a whole, it has probably been composed in the middle period of the MS editing process.
2. The *yājyānuvākyā* (MS 4.10–14) collects hymns recited by the *hotṛ* priest. This part of the MS has probably been composed in the late period of MS editing.

Figures 3 and 4 show visualizations generated with the web interface. One text is shown along the vertical axis and the other along the horizontal axis, and the corresponding frequencies of co-occurring mantras are displayed as a scatter plot, with sizes of points being proportional to the number of co-occurring mantras. Figure 3 shows where the mantras found in the *yājyānuvākyā* chapters of MS co-occur in RV, AVŚ, and AVP. The plots show that the MS incorporates many mantras from the RV, fewer from the AVŚ, and even fewer from the AVP. Figure 4 shows the co-occurrence frequencies of mantras from RV, AVŚ and AVP in the *agniciti* chapters of MS. From this plot it becomes apparent that the sources of such mantras are much more balanced than in the case of the – presumably late – *yājyānuvākyā* (see Fig. 3) because mantras from all three Saṃhitās are taken over to almost the same extent, thereby increasing the relative importance of the Atharvaveda material. A recent paper dealing with the influence of the Atharvaveda on rituals of the MS also emphasizes the significant impact that the Atharvaveda exerts on the *agniciti* ritual and that is visible in Fig. 4 (Amano, 2022). A detailed comparison of the RV-MS relationship in the *yājyānuvākyā* (Fig. 3) and the *agniciti* (Fig. 4) also shows that the *yājyānuvākyā* chapters contain many quotations from RV 1–8, whereas the influence of RV 10 is prominent in the *agniciti* chapters; note that a similar general preference for using material from RV 10 also becomes apparent in Fig. 2, where the late Gopatha-Brāhmaṇa (GB) draws freely from all parts of the RV. Considering the different sources of the mantras in each chapter of the MS thus yields insights into the history of the formation of the MS. At the same time such an approach is also helpful in understanding the processes involved in the formation and dissemination of the old metrical Saṃhitās.

4 Summary

In this paper, we have described an updated and extended version of Bloomfield's *Vedic Concordance* which was transformed from a weakly structured text-based format into strict XML. This new storage format makes it possible to explore Bloomfield's data from the perspective of graph and network theory and thereby gain insights into the structure and development of the Vedic corpus. We presented three case studies of this kind in Sec. 3 that are primarily meant to demonstrate how to utilize the resource and thus to stimulate network-based studies of the Vedic canon. In addition, this chapter also demonstrates how to visualize Bloomfield's data dynamically in an easily accessible user interface that can serve as a starting point for in-depth textual studies of post-Rigvedic Sanskrit. While we have mainly considered plain co-occurrence information in this paper, possible research applications multiply when additional aspects of the VC are taken into account. Future research could, for example, study what the types of the links (citations, variants, pratīkas) reveal about the relationship between Vedic texts and schools; or it could try to infer the direction of the edges in the graph which could provide a clearer picture of how the Vedic corpus has evolved over time.

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. Kyoko Amano was funded by JSPS KAKENHI Grant Number 21KK0004.

References

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

Kyoko Amano. 2020. What is 'knowledge' justifying a ritual action? Uses of yá eváṃ véda / yá eváṃ vidvān in the Maitrāyaṇī Saṃhitā. In Céline Redard, editor, *Aux sources des liturgies indo-iraniennes, Collection Religions, Comparatisme - Histoire - Anthropologie*, pages 39–68. Presses Universitaire de Liège, Liège.

Kyoko Amano. 2022. Influence of the Atharvaveda on rituals in the Maitrāyaṇī Saṃhitā. In *The Atharvaveda and its South Asian Contexts: 3rd Zurich International Conference on Indian Literature and Philosophy (ZICILP).*

Edward V. Arnold. 1897. Sketch of the historical grammar of the Rig and Atharva Vedas. *Journal of the American Oriental Society*, 18:203–353.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Maurice Bloomfield. 1899a. *The Atharva-Veda and the Gopatha-Brāhmaṇa.* Trübner, Strassburg.

Maurice Bloomfield. 1899b. *The Atharvaveda.* Verlag von Karl J. Trübner, Strassburg.

Maurice Bloomfield. 1906. *A Vedic concordance, being an alphabetic index to every line of every stanza of the published Vedic literature and to the liturgical formulas thereof, that is, an index to the Vedic mantras, together with an account of their variations in the different Vedic books.* Harvard University Press, Cambridge, Mass.

Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World-Wide Web Conference, Brisbane, Australia, April 1998.*

Marco Franceschini and Maurice Bloomfield. 2007. *An updated Vedic concordance: Maurice Bloomfield's A Vedic Concordance enhanced with new material taken from seven Vedic texts.* Harvard University Press, Cambridge, Mass.

Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Jan Gonda. 1975. *Vedic Literature (Saṃhitās and Brāhmaṇas).* Otto Harrassowitz, Wiesbaden.

Jan Gonda. 1977. *The Ritual Sūtras*. Otto Harrassowitz, Wiesbaden.

Arlo Griffiths. 2009. *The Paippalādasaṃhitā of the Atharvaveda. Kāṇḍas 6 and 7. A New Edition with Translation and Commentary*. Egbert Forsten, Groningen.

Stephanie W. Jamison. 2010. Review of: Marco Franceschini: An Updated Vedic Concordance: Maurice Bloomfield's A Vedic Concordance Enhanced with New Material Taken from Seven Vedic Texts. *Indo-Iranian Journal*, 53(1):35–36.

David Mimno, David M. Blei, and Barbara E. Engelhardt. 2015. Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences*, 112(26):E3441–E3450.

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

Louis Renou. 1947. *Les Écoles Védiques et la Formation du Veda*. Imprimerie Nationale, Paris.

Frits Staal. 1996. *Ritual and Mantras. Rules without meaning*. Motilal Banarsidass Publishers, Delhi.

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