An Introduction to the Textual History Tool

Diptesh Kanojia[†,♣,⋆], Malhar Kulkarni[†], Pushpak Bhattacharyya[†], Sayali Ghodekar[†],
Irawati Kulkarni[†], Nilesh Joshi[†], Eivind Kahrs[♠]
[†]IIT Bombay
[♣]IITB-Monash Research Academy
[⋆]Monash University
[♠]University of Cambridge
[†]{diptesh,pb,malhar}@iitb.ac.in, [†]sayalighodekar26@gmail.com,
[†]irawatikulkarni@gmail.com, [†]joshinilesh60@gmail.com, [♠]egk1000@cam.ac.uk

Abstract

This paper describes a digital tool called the Textual History Tool in detail. This tool captures the historical evolution of a text through various temporal stages, and inter-related data culled from various types of related texts. This tool also provides a historical view of the transmission of a text through the manuscript tradition. This tool provides an online interface which allows philologists to enter manuscript data for a text. It also provides an online interface which helps philologists compare the variants in a separate mode. It allows the user to generate phylogenetic trees, for the text, based on distance methods using the data entered in the tool. It also contains the facility to generate critical edition using a semi-supervised approach. This tool also divides the text into meaningful functional units and helps achieve a better comparison among the manuscripts. The text of the KV and its textual history is mentioned as a specific example to demostrate the features of this tool.

1 Introduction

In the twentieth century, before computers came to be used in the effort of preparing the critical edition of a text, philologists used paper-based methods for various purposes viz. collation, description of manuscripts, inter-relation of manuscripts, apparatus creation etc. With the advent of computers and development in technology, we can now have tools with us, that can facilitate the data entry, storage and display of the aforementioned functions, all on the same interface. The tool described in this paper is of the same kind.

A text is, generally a structured verbal expression of intellectual processes. This definition is derived from:

बुद्धिस्तद्व्यक्तिर्निबद्धा ग्रन्थ इत्यभिधीयते । स द्विधा मौखिक आद्यो लिखितोऽन्यश्च कीर्त्यते ।।

It can exist and get transmitted in both forms, oral as well as written.

वायुरूपो मुखे तिष्ठन् ग्रन्थो मौखिक उच्यते । पत्रे मश्यादिभिश्चिह्नै ग्रन्थो लिखित उच्यते ।।
मौखिकः कर्णनिर्ग्राह्यो लिखितश्चक्षुदर्शनः । मुखान्मुखं मौखिको हि प्रसरत्यथ कालतः ।।
लिखितश्च पुनर्लेखैः स वै सङ्क्रमः उच्यते । एवं ग्रन्थे पुनर्दृष्टमैतिह्यं सङ्क्रमस्य वै ।।
विकासस्य पुनर्बुद्धेः कालतो देशतोऽपि वा । ग्रन्थैतिह्यमहायन्त्रे योजितं स परिश्रमम् ।।[1]

Oral transmission led to the development of various vikṛtis i.e., methodologies used to memorize Vedic lore, based on cognitive features. Written transmission is carried out through copies of the text, also known as manuscripts. Historically, manuscripts were written or copied by one or more scribes. Transmission of the text from one source to another generates variants

[1] This definition comes from an Unpublished Sanskrit Work ग्रन्थेतिहासोद्योगः by Malhar Kulkarni.

which differ significantly when compared to each other. In terms of expression, the text undergoes various changes in terms of spellings, word replacements etc. Texts are used as the primary sources by scholars in reconstructing the History. The texts assume more significance as a source when it comes to reconstructing the history of an intellectual tradition. These texts represent important stages in the development of thought that contributes to the continuation of the intellectual tradition. What makes the process of reconstruction of intellectual history more complex and therefore, perhaps, more interesting as well as challenging, is the fact that these texts, themselves, are part of a historical process, also known as transmission, and have evolved in certain typical manners and ways in the course of time. It becomes necessary, therefore, in order to study the history of intellectual tradition, the history of the text used as a primary resource.

In the Indian context, we know that the transmission of texts happened in two major ways: oral and written. Texts like Vedas were transmitted from one generation to another, primarily, orally and were written down eventually. So is the case of Epic poems like Ramayana and Mahabharata[2]. In the case of Vedas, though, there is no scope of evolution of the text as such, as it was orally transmitted in a regulated manner with components of the texts noted down in great details up to the level of single letters and accent marks. In the case of Epics, however, the evolution of the text was observed by scholars and traditionally as well, it is believed that Mahabharata, for example, originally consisted of merely 10000 verses which grew in the course of time and has now become a text of one hundred thousand verses (Satasahari Samhita).

When we study the texts in the Indian grammatical tradition, that too, the paninian one, traditional commentators like Madhava and Bhattoji Dikshita etc. (Kulkarni, 2002b; Kulkarni and Kahrs, 2015), and modern scholars like Kielhorn (1887) and Kulkarni (2012a) observe that the text of the Aṣṭādhyāyī (AST) has evolved in the course of time. The text of the sutras that Patanjali had in front of him is not the same as we have it today. As shown by Kulkarni (2015b) and Kulkarni (2016), the traditional commentators quoted above, consider the text of the KV as an important stage of evolution of the text of the AST because the KV brought about numerous modifications in the text of the AST, by sometimes adding a word or two in the sutra, splitting one sutra into two, converting a later vārttika into a sutra etc. Joshi et al. (1995) also state that the KV also preserved a tradition of interpretation of the AST, independent of Patanjali. Bronkhorst (2009) showed that the KV also has an interface with other, non-paninian, Sanskrit grammatical traditions. Therefore it becomes important for scholars interested in the development of an intellectual tradition of linguistic thought in India to study the evolution of the text of the KV seriously through various sources like commentaries and manuscripts[3]. In order to study this stage of evolution further, when we turn to the printed text of the KV as available to us through more than 10 editions, as of now, we notice that the printed editions do not present to us a picture of a uniform text and rather suggest that this text of the KV that we have with us today, must have evolved in a particular manner historically. Kulkarni (2012c) studied the ‘ganapathas’ and after analyzing the data from manuscripts showed how the number of words in a ‘gana’ increased in the course of time and also formulated the stages of this historical development[4].

[2] When Malhar Kulkarni delivered his lecture on ’Text and Transmission with special reference to Classical Sanskrit Texts’ in Almaty, Kazakhstan on 25th August 2015, some members of the audience remarked that there exist texts even in Kazakhstan, which were committed to memory and were handed down from one generation to the next orally. For oral traditions of India, see (Falk, 1993) and for more recent discussions, see (Kulkarni, 2015a).

[3] There is no evidence that the KV was ever handed down orally. So oral transmission cannot be used as a resource in the reconstruction of the evolution of the KV. A modern counterexample will also make this point more clear: The text of the VaiyakaranaSiddhantaKaumudi was handed down orally, and even Malhar Kulkarni memorised it as part of his traditional education. In fact, it can be said that the primary focus of the structure of the text of the VSK is oral transmission.

[4] Also, Kulkarni presented another paper at the WSC 2018 studying in detail the printed editions of the KV on various Ganas (accepted for publication).

A Brief History of the Critical Edition of the KV in the Post-1990 era

It is this state of affairs with reference to the printed editions of the KV that led to Johannes Bronkhorst and Saroja Bhate to undertake the project of critically editing the text of the KV. Malhar Kulkarni joined this project in 1994 and collected manuscripts from various parts of India and successfully defended his dissertation submitted to the University of Pune in 2000 in which he prepared a critical edition of the KV on A 2.2. Following suit, Deo (2001) submitted her dissertation on the critical edition of the KV on A 3.1 and Dash (2004) on the KV on A 4.1. Malhar Kulkarni also published a sample edition of the KV on A 2.2.6 in a 2005 volume of a journal published by Bharatiya Vidya Bhavan, Mumbai. He also published his studies about the interrelation of groups of manuscripts of the KV (manuscripts written in Sharada script in 2003 (Kulkarni, 2003) and 2008 (Kulkarni, 2008) and manuscripts written in Malayalam script in 2012). In 2010, Eivind Kahrs and Malhar Kulkarni jointly got awarded by British Academy for their proposal to restart editing of the text of the KV critically. Kahrs and Kulkarni worked on preparing the critical edition of the KV on A 1.1 and also collected manuscripts for the same. This effort was further supported by the University of Cambridge through its funds and also by IIT Bombay. Through these funds, they paid their assistants[5] and assigned various tasks to prepare data for the purpose of critically editing the text of the KV. Through these funds, they could also get the entire manuscript collection earlier stationed at the University of Lausanne, Switzerland shipped to IIT Bombay. The outcome of this support was in the form of a book entitled "Material for the critical edition of the KV" published in April 2018. In 2018, Malhar Kulkarni was awarded another grant by Rashtriya Sanskrit Sansthan, India to critically edit the text of the KV on A 1.1. These grants are the base of our work for the purpose of critically editing the text of the KV. Textual history tool is part of our work to edit the text of the KV critically.

1.1 Functional Divisions of the text of KV

The text of KV, as mentioned above, can be, generally, divided into its functional parts. There are two basic divisions in the text of KV, one that of the sūtra and other of the KV. Within the KV, the text can further be divided according to its functional properties based on the type of sūtra it is commenting upon. We present below the functional divisions in the KV on the saṁjñā sūtra. Functional parts of the KV on vidhi sutra is described in (Kulkarni, 2012b).

- saṁjñā: this type of sūtra introduces a technical term, and hence the KV on this sūtra contains the following functional parts:
  1. Introduction of the words in the sūtra and meaning of the sūtra.
  2. Examples.
  3. Mention of other sūtras in which this technical term appears.

An example of the functional division of a sūtra is presented in Table 1.

1.1.1.	Sutra	वृद्धिरादैच्। (१॥१॥१)
1.1.1.1	Introduction & Meaning	वृद्धिशब्दः संज्ञात्वेन विधीयते प्रत्येकमादैचां वर्णानां सामान्येन तद्भावितानामतद्भावितानां च। तपरकरणमैजर्थं तादपि परस्तपर इति खट्वैरकादिषु त्रिमात्रचतुर्मात्रप्रसङ्गनिवृत्तये।
1.1.1.2	Examples	आश्वलायनः। ऐतिकायनः। औपगतः। औपमन्यवः। शालीयः। मालीयः।
1.1.1.3	Other Occurences of the term	वृद्धिप्रदेशाः। सिचि वृद्धिः परस्मैपदेषु इत्येवमादयः।।

Table 1: Example of Functional Unit based Division of the KV on AST 1.1.1

1.2 Motivation

The Textual History Tool is required because at one go it can present to a reader, the entire history of a text. A text in the Indian context can have a predecessor text as well as a successor

[5] Mukta Tilak, Prajakta Deodhar, Anuja Ajotikar, Trupti Kulkarni, Tanuja Ajotikar and Samhita Joshi.

text. It is an outcome of the intellectual activity based on one or more predecessor texts as well as textual traditions. It becomes a part of intellectual discourse and is commented upon by critical scholars within the same tradition. It gets quoted in the successor texts of the same tradition as well as other traditions and disciplines. It gets copied down in written form for various generations across different geographical regions and in different scripts. In this process, the text itself undergoes various stages of evolution, which can be marked as historical landmarks in the development of thought. Capturing the history of this intellectual world, at a glance, is the aim of this tool.

Currently, available tools do not present the historical information in a form which is coherent, and they do not provide an efficient data-entry interface which can help computational phylogenetics. There are multiple toolkits available which perform computational phylogenetics given the data is formatted in their required input format; none of them takes raw manuscript data to automate the complete pipeline which is the eventual aim of this tool. We allow users to enter raw manuscript data and create functional divisions to easy the task of phylogenetics which is a novel contribution of our work.

The key contribution of our work is:

'Building a comprehensive tool for visualizing the transmission and history of a text - a tool which can,
(i) Visualize the multiple versions of the same text which also allows data entry for manuscript versions and thus, helping one compare these versions with each other and aids one in adapting them to a graphical model viz. a phylogenetic tree.
(ii) Visualize the data from earlier texts.
(iii) Visualize the data from testimonia.
(iv) Visualize the data from commentaries.'

2 Related Work

Currently, a lot of texts written in Sanskrit are available in the electronic format available at SARIT[6], GRETIL[7], DCS[8] etc. Many of them are in searchable format. DCS presents texts with various other applied tools like Morphological Analyzer, POS tagger etc. However, no tool presents historical information the way it is needed i.e., with manuscript versions which can be compared/edited at the same time. KWIC is an acronym for Key Word In Context (KWIC) and is the most common format for concordance lines. DCS employs KWIC to be used in the concordance functionality it provides on its interface. Some tools for visualization of data are available online. Csernel and Patte (2007) discuss the LCS algorithm for preparing a critical edition of Sanskrit texts and provide a method for comparison of Sanskrit manuscripts using XML and HTML formats. BabelNet (Navigli and Ponzetto, 2010) is an important lexical resource as far as computational aspects are concerned. Navigli and Ponzetto (2012) design an explorer to visualize its database. It uses the tree layout for visualization which, in the convention, is similar to the phylogenetic visualization of texts. Visuwords[9] is an online graphical dictionary designed for accessing Princeton WordNet and uses a force-directed graph layout for visualizing the synset structure. Nodebox visualizer[10], on the other hand, provides a very static layout. WordTies (Pedersen et al., 2013) is a WordNet visualizer designed for Nordic and

[6] http://sarit.indology.info/

[7] http://gretil.sub.uni-goettingen.de/

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

[9] https://visuwords.com/

[10] https://www.nodebox.net/code/index.php/WordNetwo

Baltic wordnets. Chaplot et al. (2014) present such a visualizer for IndoWordNet- which is a lexical resource for Indian language WordNets.

Overlapping textual structures can be accurately modelled either as a minimally redundant directed graph, or, more practically, as an ordered list of pairs, each containing a set of versions and a fragment of text or data (Schmidt and Colomb, 2009). On a similar note, Hanneder (2010) writes about text genealogy and textual criticism. Maas (2009) discusses the textual versions of Carakasaṃhitā Vimānasthāna and uses computer stemmatics to aid them in the construction of a Phylogenetic tree later (Maas, 2010). Sathaye (2017) present an analyses of Vetālapañcaviṃśati, in the context of 'fluid' textual dynamics and discuss the differences in oral folklore when compared to written text. Phillips-Rodriguez et al. (2009) discuss the transmission of the Mahābhārata and the bifurcations within the diagrams about its written transmission. Kulkarni (2002a) discuss the transmission of KV and conclude that there seems to be no Vt (version) on 2.2.6 in the KV. Kulkarni (2015a) discuss the pespectives on how memory acts as an important device in the tradition of oral transmission of texts.

The TEI Critical Edition[11] Toolbox is a tool for preparing a digital TEI critical edition which allows you to check for the encoding of the text. It also facilitates the parallel look-up of the manuscript version by visualizing them on a web-based GUI. Although the software is not available for download and offline use, yet. In the current state, it accepts only TEI format XML files but does not allow one to generate versions. A technique for textual criticism is also provided by West (1973). Classical Text Editor[12] allows one to build a critical edition and critical apparatus manually. It also allows one to prepare the phylogenetic trees but does not provide a visualization interface. It allows one to collate the textual versions and edit them on an offline interface. Our work is significantly different from CTE as our online interface allows multiple users to collaborate and enter data for the same text. It allows the users to create functional divisions in the sutra text being entered and thus helps our novel phylogenetic methodology. In philosophy, our tool is focussed on the entire textual history of which manuscripts are an important part. Our tool preserves testimonia, printed editions, commentaries etc. which the CTE does not. PAUP is a tool for Phylogenetic Analysis based on Maximum Parsimony (Fitch, 1971) and other related methodologies, has been created by Swofford (1999) and is available online[13]. To the best of our knowledge, there is no tool which presents a comprehensive picture of the history of a text by presenting various resources useful for the reconstruction of the history of a text like testimonia, commentaries, earlier texts, printed editions etc.

3 Tool Architecture and Description

The Textual History Tool[14] allows users/philologists[15] to register and the registration to be approved by the tool administrator, which is authenticated based on a username/password based login interface. It also provides philologists with a data entry interface which allows the creation of a text with multiple manuscript versions in the tool database, which is a novel contribution of this work. It also encompasses a view mode, a compare mode, and a tree visualization mode (Kanojia et al., incorporated in Kulkarni and Kahrs, 2018). We describe the tool interface in the form of these modes, in the following subsections.

[11] `http://ciham-digital.huma-num.fr/teitoolbox/`

[12] `http://cte.oeaw.ac.at/`

[13] `http://paup.sc.fsu.edu/`

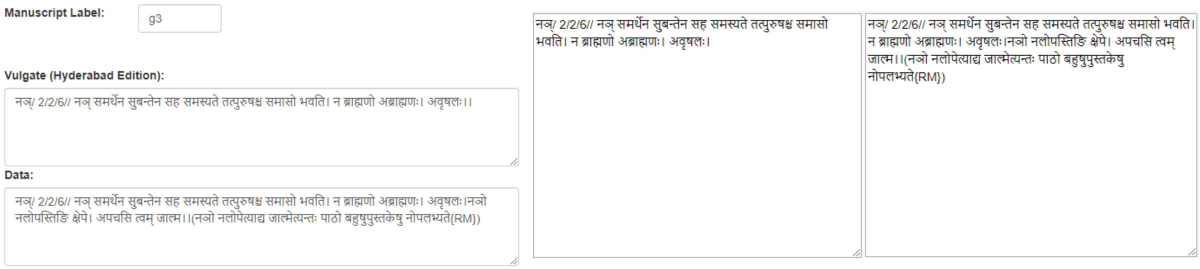
[14] The idea of developing such a tool was originally conceived by Malhar Kulkarni. He called it ग्रन्थेतिहास-यन्त्रम् in his Sanskrit work mentioned in Footnote 1. He thanks the other authors of this paper for the successful implementation. He wishes to dedicate this tool to the community of Indologists past, present and future. An earlier version of this tool was presented in the demo session at World Sanskrit Conference (2018), Vancouver, Canada.

[15] Further, we shall use users and philologists interchangeably depending on the usage of the tool.

3.1 Data Entry

The Data entry interface, based on the user login, allows the user to start with the creation of a new manuscript, or takes them back directly to the last entry they made in a previous manuscript they were working on. At any point, a user can choose to start a new manuscript creation. In such a case, the tool requests the entry of the manuscript label. Upon the entry of the manuscript label, the tool presents the user with an option to enter the manuscript data in a functional unit division or directly in a text box.

We provide this option because manuscripts are different in nature and may not contain that text or may contain the text in a different form. More importantly, the user can choose to enter text directly if they do not feel the need to divide the text into logical units. In such a case, the tool presents the users with text boxes with next and previous buttons, which allows the user to enter the text and move on the next text entry from the manuscript. In the case where the user chooses to enter the text in a functional unit division, they are presented with a text ID along with a text entry field for data. Such fields can be added or removed by the user as per the manuscript text. The user is allowed to create multiple logical divisions, and even leave a functional unit entry empty if the manuscript data requires them to do so. The tool requests the user to enter vulgate data which can be a basic building block for manuscript data for phylogenetic analysis, if the vulgate data is not present the user can ignore the request, and the phylogenetic analysis can then be carried out without it; although they can enter vulgate data at any point later in time. The data entry interface also allows a user to enter commentaries and quotations into the database. These optional entries can allow a philologist to evaluate the phylogenetic tree constructed, and can also aid the tree construction.



Manuscript Label: g3

Vulgate (Hyderabad Edition):

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।।

Data:

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।नञो नलोपस्तिङि क्षेपे। अपचसि त्वम् जाल्म।।(नञो नलोपेत्याद्य जाल्मेत्यन्तः पाठो बहुषुपुस्तकेषु नोपलभ्यते{RM})

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।नञो नलोपस्तिङि क्षेपे। अपचसि त्वम् जाल्म।।(नञो नलोपेत्याद्य जाल्मेत्यन्तः पाठो बहुषुपुस्तकेषु नोपलभ्यते{RM})

(a) View Mode Snapshot

(b) Comapre Mode Snapshot

Figure 1: Screenshot from the Textual History Tool

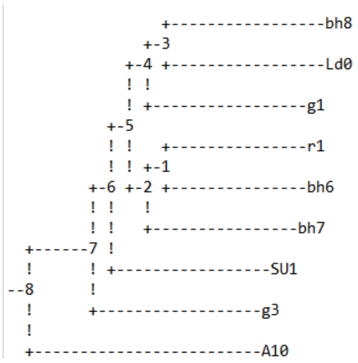
3.2 View Mode

In this mode, the user can view the manuscript version on the interface based on the label. They can select a label from the list labels in the database or search for a label and view the sūtra entries, one at a time; this mode also provides the option to correct an entry based on user privileges. We have added the functionality of viewing the sūtras in the form of functional unit division if they were created with one. This can also be used to instantaneously compare the current version with the Vulgate text, which appears on the top in view mode for each manuscript (if present in the database). A snapshot of the said mode is shown in Figure 1a.

3.3 Compare Mode

It allows a user to view different manuscript versions on the interface based on user selection. The data from Vulgate, if present in the database, is always shown on top for a base comparison. This mode does not facilitate editing of the manuscript versions but allows one to compare versions, the outcome of which can be utilized during a manual analysis later. It allows the user to select one to four versions for comparison. A snapshot of this mode is shown in Figure 1b.

3.4 Phylogenetic Tree Mode



```
                 +-----------------bh8
               +-3
             +-4 +-----------------Ld0
             ! !
             ! +-----------------g1
           +-5
           ! !   +---------------r1
           ! ! +-1
         +-6 +-2 +---------------bh6
         ! !   !
         ! !   +----------------bh7
  +------7 !
  !      ! +-----------------SU1
--8      !
  !      +------------------g3
  !
  +-------------------------A10
```

Figure 2: A sample tree produced in the Phylogenetic Tree Mode

This mode is a novel contribution of our work, where based on functional unit distances, a distance matrix can be created. These functional units are part of a text, and thus the user has a choice for selecting one or more texts wherein the functional unit division has been created in the Data Entry mode described in a subsection above. We use two different approaches to create a distance matrix. The baseline approach, which uses the notion of lexical similarity, uses Cosine Distance, Jaro-Winkler Distance, and Normalized Edit Distance to compute these distances. The second approach utilizes word-embeddings learned from Sanskrit corpora, which are stored in a model. These approaches are further detailed in Section 3.5.2.

Eventually, the distance matrix is used to cluster similar manuscripts in the same sub-group, and then the tree can be created using one of the distance based methods viz. Neighbor Joining or UPGMA. These methodologies are also explained in detail in Section 3.5.3. The tree visualization is shown on the interface in the form of manuscript labels being shown as leaf nodes, which can be seen in Figure 2. The user is allowed to view the tree on the interface as well as download it in PDF format for further analysis.

3.5 Technical Development Details

This section provides a detailed technical description of the tool interface frontend and backend. Along with the interface description, it also entails the methodologies used to create the distance matrix which is used for tree generation in the Phylogenetic Tree mode (Section 3.4). The tool architecture is shown as a diagram in Figure 3.

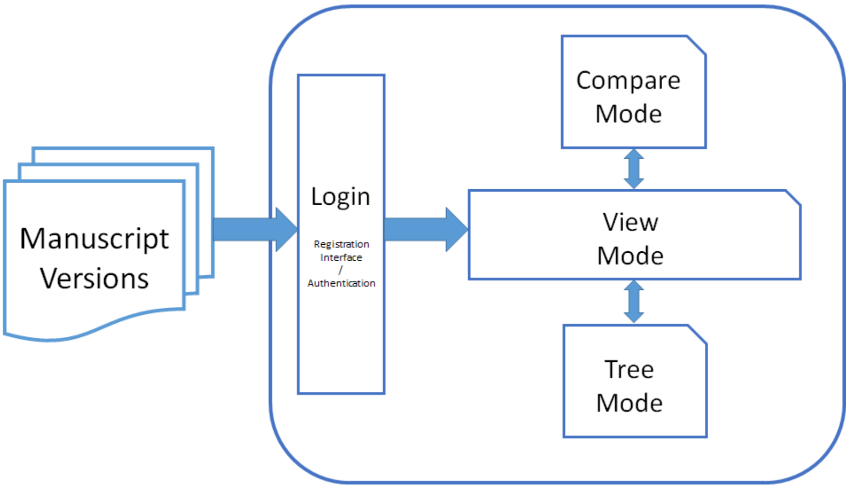


Figure 3: The basic architecture of our tool

3.5.1 Tool Interface

The tool is built as an online web-based interface[16] hosted locally on an Apache Server. It is built using PHP, Javascript and utilizes jQuery for querying the backend. The tool backend utilizes MySQL to efficiently store the manuscript data in a relational database format. MySQL queries from the tool frontend are sanitized before they are sent towards the backend to escape injection attacks. The tool comprises of an authentication interface which is based on username/password based login. The users have to be approved by an administrator after registration, which is available on the login page. The tool users can be granted different privileges based on their usage and expertise in the area. The tool source code can be downloaded and stored offline for local usage[17].

3.5.2 Methodologies for Distance Computation

The phylogenetic tree mode utilizes distance matrix creation based on code written in Python, which can be run for selected manuscripts. Our methodology requires as input the distance matrix between manuscript versions to infer the phylogenetic trees. This distance matrix is computed based on the distance among the functional units, which are divisions in the text as described in Section 1. In case of the unavailability of the division of functional units, the matrix can be computed based on the complete text acting as a single functional unit. The computation of this matrix can be done based on lexical similarity based measures as a baseline method. Our novel approach utilizes word embeddings from a large Sanskrit Corpus-based model, the details of which are below in this section.

Lexical Similarity-based Distance: Baseline Approach

The baseline approach utilizes three different metrics for the computation of lexical similarity. We use Cosine Distance, Normalized Edit Distance, and Jaro-Winkler Distance to compute three scores, which are later averaged into a single score. We also come up with a weighted average mechanism which provide 50% weight to NED, and 25% weight to each CoD and JWD to generate a more efficient tree.

- Normalized Edit Distance Method (NED): The Normalized Edit Distance approach computes the edit distance (Nerbonne and Heeringa, 1997) for all word pairs in a functional unit and then provides as output the average distance between all word pairs or 'Unit Distance'.

- Cosine Distance (CoD): The cosine similarity measure (Salton and Buckley, 1988) is another similarity metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this context, the two vectors are the arrays of character counts of two words. We calculate the cosine distance as (1 - Cosine Similarity).

- Jaro-Winkler Distance (JWD): Jaro-Winkler distance is a string metric measuring an edit distance between two sequences. It uses a prefix scale P which gives more favourable ratings to strings that match from the beginning, for a set prefix length L.

The above similarity metrics use different ways to compute similarity between each word pair and hence produces varying distance matrices. For computational purposes, we provide all the metrics equal weightages initially, and compute the distance matrix using the average score of all three methods. For manuscripts p and q, the average inter-manuscript distance is defined as:

$$LD_{pq} = \frac{(NED_{pq} + CoD_{pq} + JWD_{pq})}{3}$$

[16] Tool URL ANONYMIZED

[17] Tool Source Code ANONYMIZED

We experiment over weightages and later provide different weightages to each method. Empirically, we find best results by setting the weight as described above. For languages p and q, the weighted average inter-manuscript distance is defined as:

$$LD_{pq} = (NED_{pq} * 0.5) + (CoD_{pq} * 0.25) + (JWD_{pq} * 0.25)$$

Word embeddings based distance measures: Our Approach

We calculate the cosine distance between all word pairs belonging to the same functional unit from the embedding space. Thus, the average over the word pair distances gives us 'Unit Distance'. Similar to the baseline method, we average over all unit distances to find out the inter-manuscript distance for each manuscript pair and compute the distance matrix. Since angular cosine distance distinguishes nearly parallel vectors better (Cer et al., 2018), we also use angular cosine distance and calculate the inter-manuscript distance for each manuscript pair, in a similar fashion.

We train the models with the following hyperparameters. We create the SKIPGRAM model based on 100 dimensions due to a limited amount of the corpus collected[18]. We restrict the context window to 5 and use 0.1 as the learning rate. The maximum length of word n-gram we use is one word. We retain the sampling threshold at a default 0.0001. We use softmax as the loss function and train the models for five epochs[19].

3.5.3 Tree generation using distance-based clustering methods

We implement two distance-based methods for our work, namely, the Neighbor Joining method and the UPGMA method. We further describe these methods below, along with the reasons for choosing these methods.

Distance-based Methods

Distance analysis compares two aligned manuscripts at a time and builds a matrix of all possible sequence pairs. During each comparison, the number of changes (base substitutions and insertion/deletion events) is counted and presented as a proportion of the overall sequence length. These final estimates of the difference between all possible pairs of manuscripts are known as pairwise distances. A variety of distance algorithms are available to calculate the pairwise distance (between versions), for example, Proportional (p) distances. We use the baseline approach and our approach to compute these pairwise distances. Once the pairwise distances are calculated, they must be arranged into a tree. There are many ways to "arrange" the Taxa according to their distances. One way to cluster or optimize the distances is to join Taxa together according to their increasing differences, as embodied by their distances.

UPGMA Method

The Unweighted Pair Group Method with Arithmetic mean (UPGMA) method (Sokal and Rohlf, 1962) produces rooted trees and requires a constant-rate assumption, i.e., they assume an ultrametric tree in which the distances from the root to every branch tip are equal. At each step, the nearest two clusters are combined into a higher-level cluster. The distance between any two clusters $\mathcal{A}$ and $\mathcal{B}$, each of size (i.e., cardinality) $|\mathcal{A}|$ and $|\mathcal{B}|$, is taken to be the average of all distances $D(x, y)$ between pairs of objects x in $\mathcal{A}$ and y in $\mathcal{B}$, that is, the mean distance between elements of each cluster. In other words, at each clustering step, the updated distance between the joined clusters and a new cluster X is given by the proportional averaging of the distance between A given X and the distance between B given X.

[18] The standard number of dimensions for word embeddings, given a big corpus, is 300

[19] More epochs usually lead to a better learned/trained model; we retain the best epoch output with a minimum loss to be utilized for our work

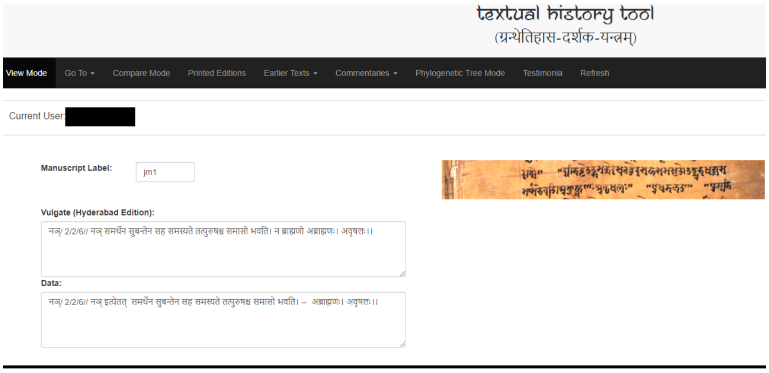
Neighbor Joining Method

Neighbour-Joining (Saitou and Nei, 1987) is a bottom-up (agglomerative) clustering method for the creation of phylogenetic trees. It applies general data clustering techniques to sequence analysis and uses genetic distance as a clustering metric. The simple version of the neighbour-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a constant timeline) across lineages.

4 Tool Features and Functionalities

The tool comprises of the following additional features and functionalities as described below:

4.1 Manuscript Pictures



Current User: [redacted]

Manuscript Label: jm1

Vulgate (Hyderabad Edition):

नञ्/ 2/2/6// नञ् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। न ब्राह्मणो अब्राह्मणः। अवृषलः।।

Data:

नञ्/ 2/2/6// नञ् इत्येतत् समर्थेन सुबन्तेन सह समस्यते तत्पुरुषश्च समासो भवति। -- अब्राह्मणः। अवृषलः।।

Figure 4: Screenshot of view mode displaying manuscript picture along with the text in the view mode

In addition to the tree generation and other salient features like a comprehensive data entry mode, the tool comprises of an additional feature where it enables the user to view the pictures of the manuscript document as a proof to substantiate the data. Philologists can attach pictures of the manuscript entry in the data entry mode as an option along with typing the manuscript data for the database entry. This picture (shown in Figure 4 as a screenshot), if uploaded by the philologist, is shown with the data entry in the view mode (Section 3.2).

4.2 Critically Edited Text

The tool also allows one to view the critically edited text in the view mode of the tool. The critically edited text allows a user to have a summarized view with additional opinions for the philologists. This helps a user decide which portion of the manuscript they want to consider for creating phylogenetics trees.

4.3 Critical Apparatus

The critically edited text is usually accompanied by a critical apparatus. The critical apparatus for a text consists of the set of variations made to the critically edited text. These changes are important to note down as they are an essential part of the preservation of historical texts. These changes allow one to notice the originally written text and how it changed over some time. The tool allows a user to view the critical apparatus in view mode as well.

4.4 Text Visualizer

Manuscripts can be envisioned as a tree in a hierarchical manner which helps philologists analyse them, conventionally. We propose a different method of viewing the manuscripts based on their distance. This text visualizer of the manuscripts allows one to view the manuscripts as leaf nodes connected using edges where one can manually change the leaf nodes in the visualizer setting. The visualizer uses the database and computes a distance matrix to visualize the graph. The graph is then creating using javascript based library which enlists all the manuscripts in an interactive way where one can manually change the leaf nodes and create their own version of a tree.

Additionally, we also implement the visualizer to depict the relation between the text and earlier texts. It can also display the inter-relations between the text and its commentaries along with the testimonia. It provides the user with an option to view these visualizations together and also as separate visualization. This feature allows the user to gather temporal information from the visualization as the database contains dated entries for the testimonia, commentaries, and some manuscripts. This will help the reader to study the evolution of the text as happened in the course of time.

4.5 Text Commentaries

There are some direct and indirect commentaries available which comment on the KV text. The two direct commentaries are Nyāsa (Ny) and Padamañjarī (Pm).

The tool allows a user to view these commentaries on each sūtra by providing a button, clicking on which, the commentary available for this sūtra is displayed to the user. This button acts dynamically on the page and is only visible as a clickable button if a commentary is available for the said sūtra which is under view on that page. This option provides additional insight into the text and allows a more holistic view of the work done on the KV text. Another button to view a sub-commentary is also provided. We also provide the option to view a consolidated version of the textual evidences available through the commentaries, as mentioned above.

Kulkarni (2002b) mentions the effort on the part of its author to collect information from the Ny and the Pm, which can act as an evidence to reconstruct the text of the KV. Kulkarni and Kahrs (2019b) enlist the variants of the text of the KV as found in the Pm through more than 300 quotations.

> "There are instances where both the Ny and Pm record the same pratīka. There we can say that both the commentaries received the text of the KV in a similar form. There are also cases when both these commentaries are silent about certain readings. And when they remain silent about certain important units of the text, say a vārttika, then it increases the probability that that vārttika might not have been there in the original text of the KV as received by these two commentaries. There are also cases when the pratīka recorded by the N and Pm vary. Such cases pose a problem for an editor. In these cases, the problem gets another dimension if the reading of both N and Pm is seen recorded in some number of mss."

Kulkarni and Kahrs (2019a) show that the textual evidence available in these two commentaries can be classified under two broad categories: Direct and Indirect. While Direct evidence is clearly visible in the text of the Ny and Pm, indirect evidence can be further classified under two categories: paroksha and atiparoksha. They, in turn, can further be classified into six and three categories, respectively. This categorization is shown below in Figure 5. The button in this tool does show all these categories of evidence, thereby displaying the text of the KV as known to these two commentaries.

Indirect commentaries are the commentaries on the direct commentaries. Tantrapradipa (Tp) is a commentary on Ny. Therefore, it becomes and indirect commentary on the KV. Some portions of Tp which are available are used in this work. Tp allows us to determine readings in the Ny, thereby indirectly helping reconstruct the text of the KV.

4.6 Earlier Texts

On the interface, we also provide an option to view the earlier texts. The purpose of this is to provide the reader with the historical view of the text. After clicking on the earlier texts button, the user is provided with an option to choose between "Paninian" and "Non-Paninian" texts. By choosing the option to view "Paninian" texts, the interface shows the earlier texts in the

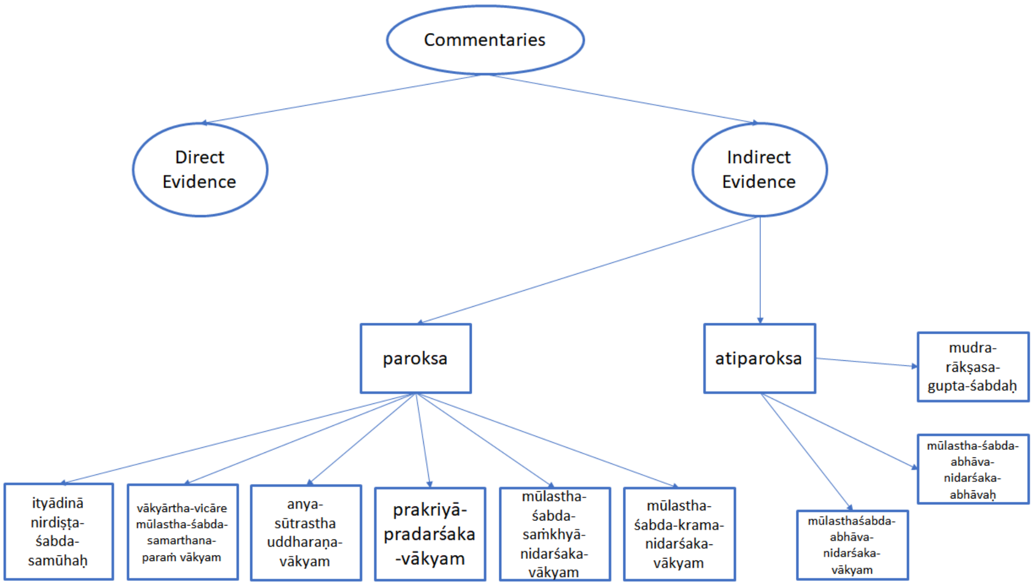

Commentaries
Direct Evidence
Indirect Evidence
paroksa
atiparoksa
mudra-rākṣasa-gupta-śabdaḥ
mūlastha-śabda-abhāva-nidarśaka-abhāvaḥ
ityādinā nirdiṣṭa-śabda-samūhaḥ
vākyārtha-vicāre mūlastha-śabda-samarthana-paraṁ vākyam
anya-sūtrastha uddharaṇa-vākyam
prakriyā-pradarśaka-vākyam
mūlastha-śabda-saṁkhyā-nidarśaka-vākyam
mūlastha-śabda-krama-nidarśaka-vākyam
mūlasthaśabda-abhāva-nidarśaka-vākyam

Figure 5: Classification of evidence from the commentaries on the KV(Kulkarni and Kahrs, 2019a).

Paninian tradition, in this context, the Vyakarana Mahabhashya (VMbh). This allows the user to see whether there is any historical connection between the KV and the VMbh. It is noted that VMbh is not available on at least more than 2300 sūtras. In those cases, obviously, the tool shows "Text Not Available".

When viewing "Non-Paninian" texts, the interface shows the earlier texts in the Non-Paninian traditions namely Katantra, Chaandra, etc. This allows the user to see whether there is any historical connection between the KV and these traditions. This historical connection is also presented in the text visualizer. The visualizer also provides and option to compare manuscript version in the database with the earlier texts. This allows the user to study the inter-relation of a particular version of the text of the KV and the earlier paninian and non-paninian texts.

4.7 Testimonia

The text of the KV is quoted in the later texts grammatical as well as non-grammatical. Kulkarni (2002b) collected and arranged chronologically more than 1000 such quotations as available from the later paninian grammatical tradition. Kulkarni (2002c) studied one quotation of the KV as found in the Shabdkaustubha and showed the inter-relation of KV manuscripts and Shabdkaustubha. The testimonia button displays all these quotations for the sūtra under study.

4.8 Printed Editions

The KV was printed for the first time in 1876. Kulkarni (2000) traced the manuscript sources of this edition. Ever since then, the text of the KV got printed more than ten times (See Footnote 4). When "Printed Editions" is clicked, the interface displays all the printed editions' text of the sūtra. This historical development in the printed editions is also presented in the text visualizer. It is hoped that the amount of variation available in the printed editions will serve as a basis to understand the manuscript variants.

4.9 Reverse Engineering and the Critical edition

This functionality allows a user to create the manuscript versions of the text based on the critical edition and the apparatus. We use the critical edition of the text and apply the variations mentioned in the apparatus to populate the manuscript versions. We believe that this function acts as a validator for the data present in the tool database.

5 Conclusion and Future Work

In this paper, we describe a tool which captures the historical evolution of a text and allows a user to view the transmission of a text through its history in a comprehensive manner. The tool allows a user to digitize a complete text and its versions through a data entry mode. The data entry mode allows one to partition the text data, based on functional units for a more accurate phylogenetic evaluation. The tool also comprises of view mode, and compare mode which can allow a user to view various parts in the text, along with the comparison of the parts in different manuscripts. Based on the data entry and/or division of functional units in the data, the tool also allows one to compute a distance matrix in the backend, which can be further used to compute a phylogenetic tree in the tree mode. The tool comprises of more features like showing manuscript pictures, visualization of manuscripts like a graph etc. In this paper, we show how this tool successfully digitizes one specific text, and we hope this can also be applied in a general domain. Utilizing all the features of the tool described above, it enables us to identify 19th Century as an important stage, in the evolution and development of this text, as the manuscripts belonging to this period add 2.2.6.3 to the main text. The justifications for this observation are noted by Kulkarni (2002a). The tool may have its technological advantages but still needs humans to interpret the text. We believe this tool can help the community digitize and view the manuscript data in a format which can be helpful to philologists for drawing further insights from the text and to understand the text for better.

In future, we would like more functionalities and different tree inferring methods to the tool. Currently, it only supports distance-based methods as described in the paper above. We would also like to provide options such as fuzzy matching between the text and the commentaries based on which a portion of the commentary can be aligned to a particular portion of the text. This automation can ease the philologists' work by automatically showing them alignments between the commentary portions and the main text. We would also like to implement generation of phylogenetic trees at the micro level (sūtras) as well as the macro level (padas, adhyayas and entire text).

References

[Bronkhorst2009] Johannes Bronkhorst. 2009. The importance of the kasika. Studies in the Kasikavrtti. The Section on Pratyaharas: Critical Edition, Translation and Other Contributions, pages 129–140.

[Cer et al.2018] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

[Chaplot et al.2014] Devendra Singh Chaplot, Sudha Bhingardive, and Pushpak Bhattacharyya. 2014. Indowordnet visualizer: A graphical user interface for browsing and exploring wordnets of indian languages. In Proceedings of the Seventh Global Wordnet Conference, pages 338–345.

[Csernel and Patte2007] Marc Csernel and François Patte. 2007. Critical edition of sanskrit texts. In Sanskrit Computational Linguistics, pages 358–379. Springer.

[Dash2004] Sasmita Dash. 2004. Critical edition of kashika 4.1.

[Deo2001] Pooja Deo. 2001. Critical edition of kashika 3.1.

[Falk1993] Harry Falk. 1993. Schrift im alten Indien: ein Forschungsbericht mit Anmerkungen, volume 56. Gunter Narr Verlag.

[Fitch1971] Walter M Fitch. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Biology, 20(4):406–416.

[Hanneder2010] Jürgen Hanneder. 2010. Text genealogy, textual criticism and editorial technique: guest ed. Jürgen Hanneder... Verlag der Österr. Akad. der Wiss.

[Joshi et al.1995] Shivram Dattatray Joshi, JAF Roodbergen, et al. 1995. The Aṣṭādhyāyī of Pāṇini, volume 4. Sahitya Akademi.

[Kielhorn1887] Franz Kielhorn. 1887. Notes on the mahabhashya, 6. the text of panini's sutras, as given in the kasika-vritti, compared with the text known to katyayana and patanjali. Indian Antiquary, 16:178–184.

[Kulkarni and Kahrs2015] Malhar Kulkarni and Eivind Kahrs. 2015. On unah um and evidence for one undivided sutra in the text of the Kāśikāvrtti. pages 1–14.

[Kulkarni and Kahrs2019a] Malhar Kulkarni and Eivind Kahrs. 2019a. Some more reflections on the role of the Nyāsa and the Padamañjarī in reconstructing the textual history of the transmission of the Kāśikāvrtti. pages 35–48.

[Kulkarni and Kahrs2019b] Malhar Kulkarni and Eivind Kahrs. 2019b. Variant readings in the text of the Kāśikāvrtti as noted by the Padamañjarī. Journal of the Oriental Institute (Vadodara), 67:143–159.

[Kulkarni2000] Malhar Kulkarni. 2000. On identifying the manuscript(s) at the base of the first printed edition of the Kāśikāvrtti. pages 203–212.

[Kulkarni2002a] Malhar Kulkarni. 2002a. On a vārttika on p. 2.2.6 in the Kāśikāvrtti. Annals of the Bhandarkar Oriental Research Institute, 83:201–205.

[Kulkarni2002b] Malhar Kulkarni. 2002b. A study of quotations of the Kāśikāvrtti in the late paninian grammatical tradition. pages 73–78.

[Kulkarni2002c] Malhar Kulkarni. 2002c. A study of quotations of the Kāśikāvrtti in the late paninian grammatical tradition. pages 73–78.

[Kulkarni2003] Malhar Kulkarni. 2003. The sharada manuscripts of the Kāśikāvrtti. pages 353–364.

[Kulkarni2008] Malhar Kulkarni. 2008. The sharada manuscripts of the Kāśikāvrtti: Part ii. pages 419–428.

[Kulkarni2012a] Malhar Kulkarni. 2012a. Franz kielhorn and the text of aṣṭādhyāyī as given in the Kāśikāvrtti: A study. 84:31–50.

[Kulkarni2012b] Malhar Kulkarni. 2012b. The malayalam manuscripts of the Kāśikāvrtti: A study. 6:103–112.

[Kulkarni2012c] Malhar Kulkarni. 2012c. Some issues in editing the ganapathas in the Kāśikāvrtti. 6:213–258.

[Kulkarni2015a] Malhar Kulkarni. 2015a. Memory: a device in traditional sanskrit learning. Memory and Human Wellbeing: Interdisciplinary Perspectives, pages 57–72. Edited by Yahei Kanayama, Malhar Kulkarni and Toshiya Unebe.

[Kulkarni2015b] Malhar Kulkarni. 2015b. Quotations in grammatical texts and the tradition of manuscript transmission of the Kāśikāvrtti. 43:182–190.

[Kulkarni2016] Malhar Kulkarni. 2016. Franz kielhorn and the text of the aṣṭādhyāyī as given in the Kāśikāvrtti: A study - ii. 32:205–212.

[Maas2009] Philipp A Maas. 2009. Computer aided stemmatics-the case of fifty-two text versions of carakasaṃhitā vimānasthāna 8.67-157. Wiener Zeitschrift für die Kunde Südasiens/Vienna Journal of South Asian Studies, 52:63–119.

[Maas2010] Philipp A Maas. 2010. On what became of the carakasaṃhitā after dṛḍhabala's revision. eJournal of Indian Medicine, 3(1):1–22.

[Navigli and Ponzetto2010] Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 216–225. Association for Computational Linguistics.

[Navigli and Ponzetto2012] Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250.

[Nerbonne and Heeringa1997] John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology.

[Pedersen et al.2013] Bolette Pedersen, Krister Linden, Kadri Vider, Markus Forsberg, Neeme Kahusk, Jyrki Niemi, Lars Nygaard, Mitchell Seaton, Heili Orav, Lars Borin, et al. 2013. Nordic and baltic wordnets aligned and compared through "wordties". Proceedings of NODALIDA 2013.

[Phillips-Rodriguez et al.2009] Wendy J Phillips-Rodriguez, Christopher J Howe, and Heather F Windram. 2009. Some considerations about bifurcation in diagrams representing the written transmission of the mahābhārata. Wiener Zeitschrift für die Kunde Südasiens/Vienna Journal of South Asian Studies, 52:29–43.

[Saitou and Nei1987] Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution, 4(4):406–425.

[Salton and Buckley1988] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5):513–523.

[Sathaye2017] Adheesh Sathaye. 2017. The scribal life of folktales in medieval india. South Asian History and Culture, 8(4):430–447.

[Schmidt and Colomb2009] Desmond Schmidt and Robert Colomb. 2009. A data structure for representing multi-version texts online. International Journal of Human-Computer Studies, 67(6):497 – 514.

[Sokal and Rohlf1962] Robert R Sokal and F James Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon, pages 33–40.

[Swofford1999] DL Swofford. 1999. Phylogenetic analysis using parsimony (and other methods) paup* 4.0. Sinauer, Sunderland.

[West1973] Martin L West. 1973. Textual criticism and editorial technique applicable to Greek and Latin texts. Walter de Gruyter.