Structure and Implementation of a Digital Edition of the Aṣṭādhyāyī

Wiebke Petersen and Simone Soubusta

Heinrich-Heine-Universität Düsseldorf
wiebke.petersen@phil.uni-duesseldorf.de
simone.soubusta@uni-duesseldorf.de

Abstract. The Aṣṭādhyāyī, Pāṇini's grammar of Sanskrit, exhibits an unparalleled structure which to this day has not been fully understood. It encodes the grammar rules in a very concise manner, making use of inheritance and a sophisticated metalanguage. Modern linguistics could benefit from a deep study of its precise description methods. Unfortunately, due to the fact that they have little to no knowledge of Sanskrit and Indian grammatical theory and are unable to read Devanāgarī, many (western) linguists lack the necessary skills to understand the Aṣṭādhyāyī. In this paper, we present an approach towards a digital, web-based edition of the Aṣṭādhyāyī developed in order to unlock it for broader scientific research. We focus on the database scheme which provides the core of the digital edition.

Keywords: Panini, Aṣṭādhyayī, database, grammar, representation

1 Introduction

Sanskrit is the oldest Indo-European language that is still spoken. It distinguishes itself through a rich morphology and phonology. The Aṣṭādhyāyī, Pāṇini's approximately 2.500 year old grammar, developed into the standard grammar of Sanskrit. In the process, it shifted from a descriptive into a prescriptive grammar. The main reasons for its still outstanding position are its completeness and its compactness. With respect to its completeness, Bloomfield (1933, p. 11) writes: "It describes, with the minutest detail, every inflection, derivation, and composition, and every syntactic usage of its author's speech. No other language, to this day, has been so perfectly described."

Consisting in its core of nearly 4.000 very short *sūtras* (aphorisms), the Aṣṭādhyāyī is extremely compact for a grammar of its range. Pāṇini reaches this level of conciseness by developing sophisticated description methods which are based on inheritance and a metalanguage on which Staal (1995, p. 107) comments: "I believe that Pāṇini reached here a level of artificiality that neither the logicians nor the mathematicians attained. [...] He created artificial constituents and a mechanism through which these constituents could be integrated into a new language, the metalanguage of his grammar." Modern linguistics could benefit from a deep study of its precise description methods and of the way it is structured.

However, most linguists lack the necessary education to study the Aṣṭādhyāyī in the way it is usually presented. They do not understand Sanskrit, are not familiar with the traditional Indian grammar terminology and cannot read Devanāgarī.

In this paper we present an approach towards a new digital edition of the Aṣṭādhyāyī with the aim to unlock it for non-professionals. The key idea is to develop an edition that preserves the structure of the Aṣṭādhyāyī and does not offer a mere collection of interpreted and translated *sūtras*. The remainder of this paper is structured as follows: Section 2 elaborates in greater detail the question of why such a new structure-focused edition is useful. In Section 3 an outline of the technical and structural properties of our edition is given. The edition consists of a user interface (Subsection 3.1) and an underlying database (Subsection 3.2). The paper is concluded by an outlook in Section 4.

2 On the need of a structure-based edition of the Aṣṭādhyāyī

2.1 The non-linear structure of the Aṣṭādhyāyī

It is clear that neither the language described in the Aṣṭādhyāyī (i.e., Sanskrit), nor Pāṇini's description of it have an inherently linear structure. Pāṇini was confronted with the usual problem of any grammarian of how to present his language description in a linear text. Being situated in an oral tradition, his problem was even more severe, since he could not make use of explicit cross-references, indices or non-linguistic elements like tables and symbols and also because conciseness is especially desirable in a grammar which is meant to be learned by heart and recited regularly. Thus, the question arises of what determines the specific linear order of the *sūtras* in the Aṣṭādhyāyī.

The Aṣṭādhyāyī is not ordered according to the ordering principles commonly used in other grammars. The order reflects neither the conventional discrimination between linguistic subdisciplines (e.g., phonology, syntax, semantics,...), nor does it follow any didactic considerations like school grammars (e.g., starting with easier and more frequently used structures). Moreover, no strict order based on treated phenomena like, for example, vowel sandhi is recognizable. Although Pāṇini's grammar covers not only classical Sanskrit but also aspects of Vedic and of some dialectal and socio-linguistic variants, it is not ordered according to those subjects. Furthermore, the order does not follow any classification based on the descriptive or the operational structure of the grammar. A descriptive classification would distinguish between the tasks which are fulfilled by a *sūtra* (e.g., stating a rule of Sanskrit, stating a meta rule, stating a definition, ...). An operational structure would present the *sūtras* in the order in which they are applied in language processing like, for example, grammatical derivations.

As the Aṣṭādhyāyī follows none of the mentioned ordering principles, the question arises of what determines its structure. It is commonly assumed that the structure of the Aṣṭādhyāyī is governed by economical principles, referred to as lāghava. However, it is not clear what *lāghava* means with respect to the Aṣṭādhyāyī. Gérard Huet (p.c.) proposes an interesting analogy. He compares the way the Aṣṭādhyāyī is presented to compiled program code that does not explicitly express the logical structure of the program anymore, but follows efficiency considerations. In our opinion this is a very appropriate analogy. Thus, one should distinguish between Pāṇini's grammar itself as a description of Sanskrit (i.e., Pāṇini's system of grammatical rules which corresponds to the logically structured program source code) and the compiled code in the form of the actual text of the Aṣṭādhyāyī, which encodes the grammar in a compact manner by making use of inheritance, a metalanguage, markers, and so forth.

All the possible orders discussed above are determined by aspects of the grammar and not by its encoding. Previous studies support the thesis that *lāghava* governs not only Pāṇini's description of Sanskrit but even more its textual encoding. One example for this is Pāṇini's treatment of the phonological and morphophonemic aspects of Sanskrit. With respect to the grammatical description, the question is which phonological rules and which sound classes are assumed. With respect to the encoding of this description, the question is in what form the rules are presented and how the sound classes are defined and named. For the former, Pāṇini uses natural language case suffixes as metalinguistic markers in order to indicate the role of a sound class in a rule. For the latter, he developed the *pratyāhāra*-technique: it is a generative method to construct maximally short, namely monosyllabic, metalinguistic names for phonological classes. The class names refer to intervals in a marker-interrupted list of Sanskrit sounds listed in the Śivasūtras. In Petersen (2009) it was mathematically proven that there is no shorter solution than this list to the problem of ordering the sounds and placing the markers such that all the desired sound classes form intervals which end immediately before a marker. Thus one could say that Pāṇini compiled the set of sound classes into an interval based code of minimal length.

It is generally assumed that the compilation of the rest of the grammar is guided by similar principles, i.e., that Pāṇini uses inheritance and rule blocking mechanisms such that the text of the Aṣṭādhyāyī represents his Sanskrit grammar in the shortest possible form. However, a formal proof of this hypothesis is still missing. This thesis is more difficult to prove than the minimality of the Śivasūtras, as for the latter only 14 and not all of the nearly 4.000 *sūtras* have to be investigated. One of the main motivations for our digital Aṣṭādhyāyī edition is to support the verification of such general hypotheses about the structure of the Aṣṭādhyāyī.

2.2 Why traditional editions do not suffice

The structure of the Aṣṭādhyāyī as a compiled grammar makes it very difficult to use. Traditional editions try to support users of the Aṣṭādhyāyī in this task. If humans want to apply the grammar in the production of correct Sanskrit sentences, they first have to recompile it. A long line of commentaries on the Aṣṭādhyāyī tackle this problem. They explain how the Aṣṭādhyāyī encodes individual derivation procedures. Traditionally, the encoded text of the Aṣṭādhyāyī is learned by heart and the trained grammarian will jump through the *sūtras* while generating Sanskrit expressions. Due to its non-linear structure, it is necessary to have access to all of the nearly 4.000 *sūtras* at once, because *sūtras* from all 32 *pādas* (roughly: 'chapters') can be involved in the derivation of a single word. In order to support the reader, a written edition of the Aṣṭādhyāyī usually contains several indices (e.g., of the technical terms, the *anubandhas*,...) and enriches each *sūtra* with references to other *sūtras* in order to aid in its comprehension (e.g., Vasu, 1891; Katre, 1987; Sharma, 1987-2003; Joshi and Roodbergen, 1991-). Nevertheless, everybody working with the Aṣṭādhyāyī without remembering it by heart knows the problem of having an insufficient amount of fingers or bookmarks to mark all the other *sūtras* influencing the analysis of the *sūtra* that is currently being studied.

By establishing an appropriate digital edition of the Aṣṭādhyāyī these difficulties can be minimized. A digital edition allows the modeling of the dependencies between the individual *sūtras* through active hyperlinks. Additionally, the same *sūtra* data can be reordered according to different ordering principles. Hence, users can easily switch between, e.g., the didactically motivated *sūtra*-order given in the Siddhānta Kaumudī and the original order of the Aṣṭādhyāyī. The electronic edition by Shivamurthy Swamiji implements some of these possibilities (`http://www.taralabalu.org/panini/`). It allows users to search for expressions in the Aṣṭādhyāyī, uses hyperlinks to navigate through the *anuvṛtti*-inheritance and enables a reordering of the *sūtras*. However, Swamiji's digital edition follows the traditional line in that it focusses on the question of how the Aṣṭādhyāyī are to be applied in language production and not on its structure and how it is encoded in the plain text. Although providing much less functionality than Swamiji's, Peter Scharf's digital edition (`http://sanskritlibrary.org/`) is much closer to our aim. It provides the user with detailed grammatical and metalinguistic information on the *sūtra* components and a link to the Monier Williams Sanskrit-English Dictionary.

Traditional editions focus on the translations of and comments on the individual *sūtras*. However, a translation of a Pāṇinian *sūtra* is nearly never a mere literal, word-for-word translation, but also invariably includes an interpretation of its function in the context of the Aṣṭādhyāyī. The reason for this is that the *sūtras* do not form meaningful sentences in Sanskrit. They lack information that has to be inherited from other *sūtras* and they make use of metalinguistic elements not only as terminological expressions but also as simple markers. If someone who has not mastered Sanskrit tries to investigate the structure of the Aṣṭādhyāyī with a traditional edition, he or she will be confronted with interpretations that disambiguate the *sūtras*. The reader thus loses control over possible alternative interpretations and over the function of the single components of a *sūtra*.

2.3 Requirements of a structure-based digital edition

Our aim is to develop a digital edition that enables linguists with no special training to study the Aṣṭādhyāyī. Such a digital edition would not only support the application of the grammar and thereby contribute to the traditional question of the commentators of *how the Aṣṭādhyāyī works* but also to the modern question of *how it is structured* (cf. Kiparsky, 2009; Bhate and Kak, 1993). It could be a useful tool to decode Pāṇini's encoding, to recompile the source grammar and to investigate the description techniques used. The aim is not to explain the way the grammar functions, but to provide the means to investigate the principles governing the organization of the Aṣṭādhyāyī.

Our target users include those who have no or only insufficient command of Sanskrit. They are not familiar with the traditional Indian grammar terminology and they do not know the traditional commentaries. Instead, they have had basic training in modern linguistics. We therefore transcribe the Aṣṭādhyāyī in Latin script and use standard English linguistic terminology.

In order to enable users to follow their own research questions concerning the structure of the Aṣṭādhyāyī independent of former interpretations, we follow a text-immanent approach as much as possible and stick to the wording of Pāṇini. Thus, we provide a grammatical analysis and a translation of each word and each element of a compound. *Sūtra* translations and commentaries are provided as an additional help for the user, but not as the main source of information. However, although the text of the Aṣṭādhyāyī is very well preserved considering its age, some information has been lost over the millennia. By writing down the Aṣṭādhyāyī, accents and nazalization have been lost. As a result, the scope of *anuvṛtti*-inheritance and the identification of metalinguistic markers cannot be read off the Aṣṭādhyāyī text anymore, but must be intellectually reconstructed by referring to knowledge of the object language being described (Sanskrit) and the way it is described (cf. Joshi and Bhate, 1984). As some of our target users will not be able to do this reconstruction themselves, we provide them with this information taken from the classical commentaries.

Another important requirement for a new digital edition is to be flexible. Different users want to research different aspects of the grammar and are thus interested in different parts of the Aṣṭādhyāyī. We would thus like to provide different *views* on the same information, allowing users to *zoom in* on different aspects, e.g. see a detailed grammatical analysis of a *sūtra* or an uncluttered overview over all *sūtras* pertaining to 'phonology'. Being able to combine the existing data in numerous ways to provide these *views* is thus another important consideration.

Lastly, a very important advantage of a digital edition over a nondigital one is that it is generally easier to extend through additional data. Thus, already in the planning process of a new digital edition, one should try to design it such that it can be flexibly extended. For example, since the Aṣṭādhyāyī is still being investigated, new, more appropriate translations could be provided and added to the edition. Furthermore, since language is ambiguous, the grammatical analyses of the *sūtras* are not uncontroversial and it is desirable to represent the conflicting views in a digital edition, so that interested parties can access them. Due to ongoing research, the information about the Aṣṭādhyāyī is not static, but changes over time, an aspect that should be taken into consideration. Even beyond this evolving corpus of information, we might want to add further facets to our implementation in the future and should thus provide the technical means to do so.

To sum up, four requirements govern the development of our structure-based, digital edition of the Aṣṭādhyāyī which will be presented in the next section: (1) Require as little prior knowledge as possible, (2) stay as close as possible to the text of the Aṣṭādhyāyī and be (3) flexible and (4) extendable.

3 Implementation

3.1 User Interface

In principle, the data of our edition can be approached in two different ways: Either by browsing through the *sūtras* or by searching for occurrences of special patterns within *sūtras* or among groups of *sūtras*.

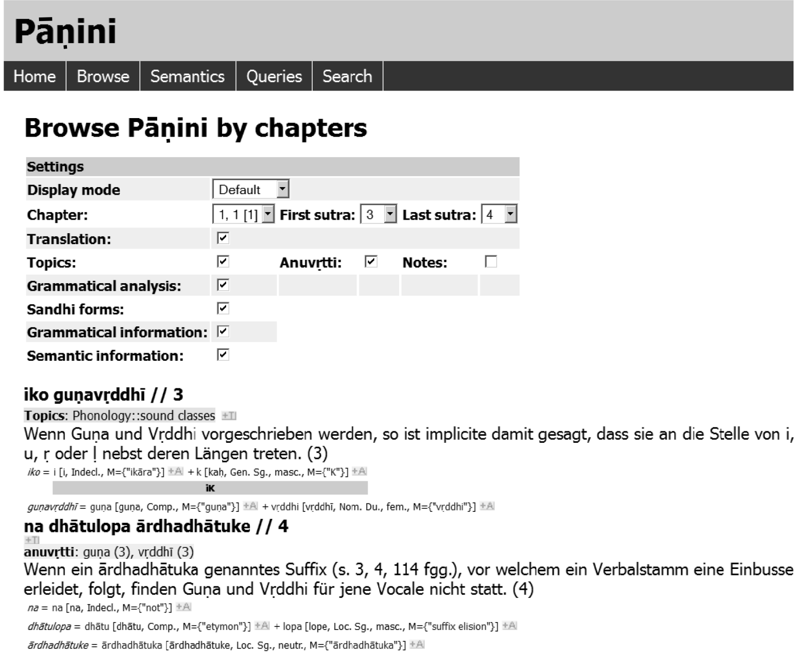
3.1.1 Browsing

One way of accessing the data in our digital edition is by browsing through the *sūtras*. Figure 1 shows the prototype of our Pāṇini browser developed in the research project "*Pratyāhāras* or features?" at the university of Düsseldorf[1] and implemented by Oliver Hellwig. The browser is realized as a PHP application that accepts user input and dynamically changes the display accordingly. The figure shows the default view on the data. The prototype already demonstrates how the four requirements identified in the last section, i.e., knowledge-independency, literality, flexibility, and extendability are implemented in the user interface.

The interface is flexible and easily extendable. Users can determine which *sūtras* they want to see by setting the appropriate interval of *sūtras* within one *pāda* of the Aṣṭādhyāyī. Furthermore, they can decide what kind of information they are interested in by selecting the appropriate checkboxes. The set of checkboxes offered depends on previous choices such that, for example, someone who is not interested in 'Grammatical analysis' does not see the boxes for 'Sandhi form', 'Grammatical information' and 'Semantic information'. Having a decision-dependent menu enables us to extend the database with additional information (e.g. translations, commentaries, notes) without being forced to overload the user with a confusing menu.

Currently, in the prototype, all the information is presented in Latin script and we use standard modern linguistic terminology such as 'singular' ('Sg.') instead of '*ekavacana*'. Furthermore, we provide a translation as well as grammatical information such as part-of-speech, inflectional information and the corresponding lexeme for each expression. Compounds and technical terms like *pratyāhāras* are decomposed into their subparts. On the *sūtra*-level, a translation, the expressions inherited by *anuvṛtti* and keywords ('Topics') are provided. Furthermore, the user has the possibility of adding personal notes to the *sūtras*. As our primary focus lies on a detailed analysis of the *sūtra* elements, we currently only provide the German translation from Böhtlingk (1887) which is not

[1] Project webpage: `http://panini.phil.uni-duesseldorf.de/`



Pāṇini

Browse Pāṇini by chapters

Settings					
Display mode	Default				
Chapter:	1, 1 [1]	**First sutra:**	3	**Last sutra:**	4
Translation:	☑				
Topics:	☑	**Anuvṛtti:**	☑	**Notes:**	☐
Grammatical analysis:	☑				
Sandhi forms:	☑				
Grammatical information:	☑				
Semantic information:	☑				

iko guṇavṛddhī // 3

Topics: Phonology::sound classes +T

Wenn Guṇa und Vṛddhi vorgeschrieben werden, so ist implicite damit gesagt, dass sie an die Stelle von i, u, ṛ oder ḷ nebst deren Längen treten. (3)

iko = i [i, Indecl., M={"ikāra"}] +A + k [kaḥ, Gen. Sg., masc., M={"K"}] +A

iK

guṇavṛddhī = guṇa [guṇa, Comp., M={"guṇa"}] +A + vṛddhi [vṛddhī, Nom. Du., fem., M={"vṛddhi"}] +A

na dhātulopa ārdhadhātuke // 4

+T

anuvṛtti: guṇa (3), vṛddhī (3)

Wenn ein ārdhadhātuka genanntes Suffix (s. 3, 4, 114 fgg.), vor welchem ein Verbalstamm eine Einbusse erleidet, folgt, finden Guṇa und Vṛddhi für jene Vocale nicht statt. (4)

na = na [na, Indecl., M={"not"}] +A

dhātulopa = dhātu [dhātu, Comp., M={"etymon"}] +A + lopa [lope, Loc. Sg., masc., M={"suffix elision"}] +A

ārdhadhātuke = ārdhadhātuka [ārdhadhātuke, Loc. Sg., neutr., M={"ārdhadhātuka"}] +A

Fig. 1. Pāṇini Browser

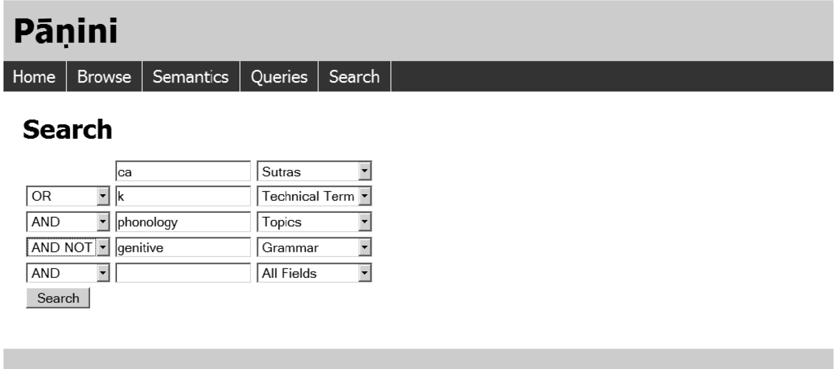
protected by copyright. However, our database design is flexible enough to add additional translations to other languages as well (cf. Section 3.2).

Using hyperlinks, it is possible to navigate through the non-linear structure of the Aṣṭādhyāyī by jumping to those *sūtras* that provide additional information through *anuvṛtti* or to *sūtras* in which a lexeme co-occurs or which are classified by the same keywords. As our data is stored in a relational database, additional hyperlinks can be added with no effort (cf. Section 3.2).

The figure shows the current default view on the data. Additionally, we provide the possibility of implementing personalized views to aid the users' research. Such a view could, for example, present the *sūtras* in a more compact tabular form containing only the data needed to work on a specific research question. At the moment, we have predefined two tabular views for members of our research group. We plan to develop more predefined views according to the needs of the users of our browser as soon as our GUI goes online and we hopefully receive valuable feedback.

3.1.2 Searching

One of the primary aims of this digital edition of the Aṣṭādhyāyī is searchability – allowing both those already familiar with the Aṣṭādhyāyī and those who are new to them to search for specific passages, grammatical features or certain topics. Because this search functionality should not be limited to content-based search



	ca	Sutras
OR	k	Technical Term
AND	phonology	Topics
AND NOT	genitive	Grammar
AND		All Fields

Fig. 2. Pāṇini Browser – search function

(i.e., fulltext search), we also provide concept-based retrieval options (e.g. search for the marker ‘L’ or for all *sūtras* dealing with ‘phonology’). Furthermore, we enable the formulation of complex queries using boolean operators (AND, OR, AND NOT).

For reasons explained below, our implementation uses a relational database for data storage. This offers professional users the advantage of being able to formulate queries directly in SQL, provided such inputs are sanitized to avoid SQL injection vulnerabilities. There are virtually no boundaries to the complexity of such queries, allowing experts to make full use of the available data. We are, however, conscious of the fact that not all researchers interested in the Aṣṭādhyāyī are familiar with SQL. Furthermore, users would have to familiarize themselves with our database design in order to utilize SQL queries. For this reason, we also offer an alternative, slightly reduced, search interface, a prototype version of which is shown in Figure 2 . The interface requires no expert knowledge to use (although familiarity with boolean logic is recommended for more complex queries). In comparison to the possibilities offered by SQL, the interface is slightly restricted. However, it should still be more than enough for most information needs.

Figure 2 shows an (artificial) example for a complex query. It would return all *sūtras* that contain either ‘ca’ or the technical term ‘k’ or both, but only if these deal with the topic of ‘phonology’ and do not contain any words in the genitive case. Searching for the technical term ‘k’, the topic of ‘phonology’ and words in the genitive case in this query are examples of concept-based retrieval. Even though this prototype version of the search interface already contains a lot of the desired functionality, we plan to extend the functions of the simplified interface further to include options like searching within a certain *pāda*, or phenomena that occur within a certain range of *sūtras*, among other things.

3.1.3 Personalizing

Another important feature of our digital edition of the Aṣṭādhyāyī is that users can personalize the application for their needs by adding notes to *sūtras*. Users will also have the ability to share not only these notes, but also saved searches and personalized *views* with other users (e.g. members of the same research group). A possible future expansion of this personalization would be to allow users to *tag sūtras*, to make it easier to re-find passages of interest to them. These tags could also be used as a further index by other users, as is done in many other web applications. The *views* could be used to establish links with other Aṣṭādhyāyī-related web services on the level of single *sūtras* like, e.g., the Aṣṭādhyāyī-specific Wiki currently being developed under the guidance of Navjyoti Singh at IIIT Hyderabad and similar projects.

3.2 Database

The requirements of a digital edition of the Aṣṭādhyāyī that we mentioned above – the ability to access different *views* of the information, searchability of the data, extensibility and the flexibility to adapt to changing data – have led us to the use of a relational database as the backbone of our implementation. A relational database stores data in so-called *relations* or *tables*, which are sets of tuples that share the same attributes (Codd, 1990, p. 2ff.). Databases are an obvious solution for our needs, since they were designed to provide these different *views* that we stipulated above. They are furthermore designed to be adaptable, allow for data control and allow concurrent access. Relational databases also feature flexible authority, i.e., users can be granted access to some parts of the database, while being denied access to others (Codd, 1990). Not to mention that relational databases come with their own powerful query language in the form of SQL, facilitating the above-mentioned searchability. All of these considerations make a relational database most suited for our application in comparison to alternative storage technologies like XML, which features no inbuilt technique to provide different *views*, does not allow concurrent access and features no access and data integrity control.

Moreover, our database adheres to the third normal form as specified by Codd (1971), which states that all attributes must be atomic, there must be no repeating groups, no partial functional dependencies and no transitive functional dependencies. The objectives of this normalization are to free the database from “undesirable insertion, update and deletion dependencies” by removing redundancies and to “reduce the need for restructuring the [database] as new types of data are introduced and thus increase the life span of application programs” (Codd, 1971, p. 34).

Despite the fact that databases are most suited for our digital edition of the Aṣṭādhyāyī, one implementation detail that requires special attention is the fact that instead of considering each *sūtra* as a whole, we want to attach information to different parts of a *sūtra* (e.g. a word, affix) on a flexible basis. Databases, however, do not provide this sort of relation, necessitating that we develop a way to enable this feature in our implementation, which shall be explained in detail in the next section. Section 3.2.2 then presents the database design as a whole.

3.2.1 Defining the atomic units

One important design consideration is that we want to provide analyses and commentaries for all (sub)parts of the *sūtras*: i.e., for word forms, words, compound segments and affixes as well as metalinguistic units and compounds. This is nontrivial as the object-language and the metalanguage are intermixed in the Aṣṭādhyāyī: Metalinguistic compounds, i.e, expressions composed of two or more metalinguistic expressions, form new technical terms which are inflected according to the object language rules; case markers of the object language serve as metalinguistic markers on object language expressions and so forth. For example, if we have a *sūtra* like *īdūdeddvivacanaṃ pragṛhyam*, we have inflectional information on the word form level, i.e., *īdūdeddvivacanaṃ* and *pragṛhyam*. On the word level, i.e., *īdūdeddvivacana* and *pragṛhya*, we have, for example, the semantic information that *īdūdeddvivacana* is a determinative compound. One level deeper we find the two segments *īdūdet* and *dvivacana*, the former of which is a copulative compound. On the next level we get the expressions *īt*, *ūt*, and *et*. Here, the deepest level of the object language is reached. However, *īt*, *ūt*, and *et* are not atomic units in the metalanguage. They are metalinguistic compounds that can be decomposed into *ī+T*, *ū+T*, and *e+T*. Each of those compounds consists of a vowel sound in the object language and the metalinguistic marker *T*.

The example demonstrates that there is no general fixed number for the levels of analysis and that at each level segments from the object language and from the metalanguage get intermixed. Therefore, we discarded the idea to build up a general lexicon of all language units such that each *sūtra* can be analyzed as a composition of those units. In order to keep our approach as flexible as possible, we decided to index the *sūtras by character* in the padapāṭha form and use this index to form intervals on the basis of which an analysis will be allocated and retrieved. Thus, in our previous example the *sūtra* in *padapāṭha* form is *īdūdeddvivacanaṃ pragṛhyam* and the component *pragṛhyam* would be addressed by the interval [18,26] (counting the blank as a character) and would be linked to the indices 18 and 26 in the database. Similarly, *ūt* forms an interval from 3 to 4. The rule is that segments that are changed by internal sandhi get the indices of the sandhified form. Counting the blanks offers the possibility of identifying even empty declension morphemes by intervals. In some rare cases of elision, dummy indices have to be added in order to enable the addressing of each desirable segment by its interval boundaries. Since intervals can overlap, we can identify arbitrary components of a *sūtra* as long as they correspond to a continuous interval in the *padapāṭha* form. Indexing by character allows us to attach information on various levels and to easily add new information (and levels) in the future.

Please note that encoding the Aṣṭādhyāyī in Latin script with diacritics simplifies our interval method. The alternative script, Devanāgarī, is an alphasyllabic script which expresses complex segments like consonant-vowel pairs and consonant clusters using single graphemes. The transcription in Latin script with diacritics shows a stronger correspondence between sounds and graphemes, although it uses several digraphs (e.g. for aspirated sounds, diphthongs, and 'sh'). The relation between sounds and graphemes in Devanāgarī is many-to-one while in Latin script it is one-to-many. Thus, using Latin script we tend to get more indices than necessary, which is harmless, while using Devanāgarī script, sounds belonging to different segments would often be clustered in a single grapheme and thus in a single index which would be very problematic for our indexing-by-character method. This does not, however, prevent us from being able to display the *sūtras* in Devanāgarī in the browser by converting the Latin script to Devanāgarī on the fly, should this be desired in a future version.

3.2.2 Database design

When planning an application with a database backbone, the database design is especially important. While a good design ensures that the database is adaptable and extendable, a badly designed database will likely soon outlive its usefulness. Having to redesign a database, however, also means that the surrounding program code has to be rewritten and the existing data has to be painstakingly migrated from the old database to the new one. Furthermore, if the old database was not normalized properly (see Section 2), the existing data may already have been corrupted. For this reason, we have designed our database to conform to the third normal form and with a view to future extensions.

Figure 3 shows the UML diagram of our database. The Sutra table is the central point of the database. It stores all *sūtras* in normal form and *padapāṭha* form with their numbers for easy reference. The connected table Translation contains a variable number of translations for each *sūtra*, referenceable by language and author (through an author ID). Note allows users to comment on *sūtras*, again referenceable by author ID. Topic assigns different categories (e.g. phonology) to *sūtras*. The Atom table contains the character-based index which we described above. Every row in this table contains a *sūtra* ID from the Sutra table, a position (e.g. 1 for the first character) and the character itself. The Component table then takes two atoms from the same *sūtra* (upper and lower bound of the interval) to form a component to which information can be attached.

Each component can be assigned grammatical features from a number of grammar categories (e.g. number, gender) in the Grammar table, as well as up to one lexeme and a variable number of meanings. Linking comments directly to components via its own ComponentNote table could be a useful extension. The Anuvrtti table represents a relation between the Sutra table and the Component table, listing the *sūtra* that inherits via *anuvṛtti* on one side and specifying which part of another *sūtra* it inherits on the other. TechnicalTerm and Marker are subclasses of Component, i.e., they contain those components that are technical terms and markers, respectively, attaching additional relevant information like term type and marker type. Lexeme also has an n-to-m relation to its own meaning table LMeaning, allowing a meaning to belong to several lexemes and a

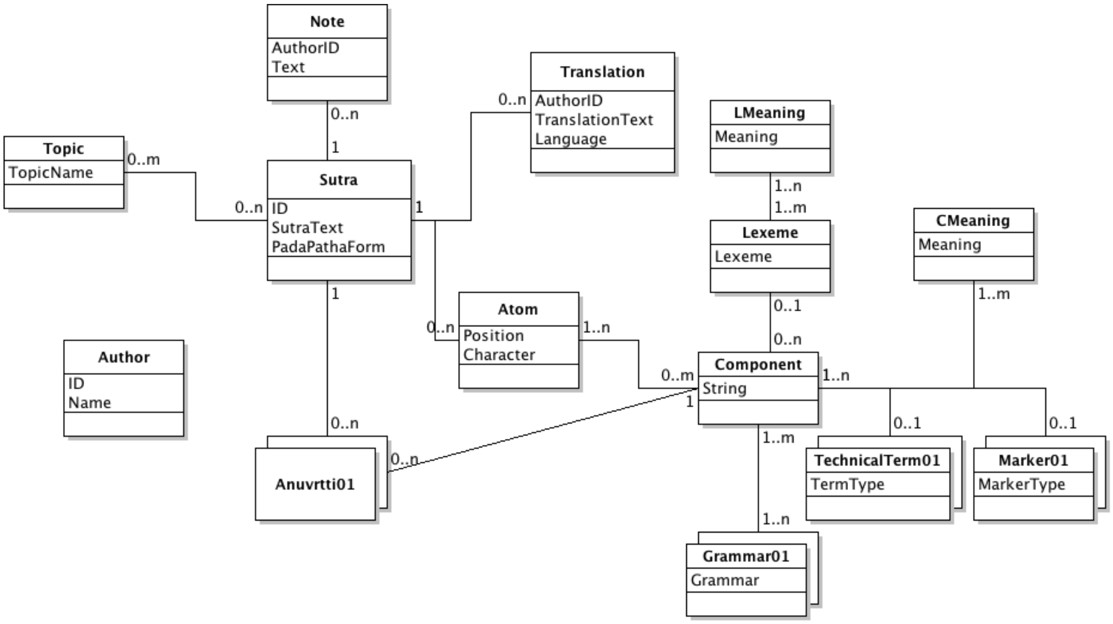


Fig. 3. The database scheme

lexeme to have several meanings. Our existing data has already been stored in the tables Sutra, Atom, Component, Grammar, Translation and Note. The Anuvrtti table is in the process of being filled, with Topic, TechnicalTerm and Marker to follow shortly. All the data will of course be proofread for errors.

Due to the ambiguity of the Aṣṭādhyāyī, the database will not contain a finished product so much as an ongoing discussion with conflicting views. Thus, the tables Anuvrtti and Grammar as well as Marker and TechnicalTerm will exists not just once, but in different 'versions', so to speak, representing the views and opinions of contributors to the database. These tables Anuvrtti01, Anuvrtti02, etc. will exist in parallel, fulfilling the same functions, and users will be able to choose between those different Pāṇini 'interpretations'. Besides choosing a specific view, users will also be able to view all interpretations or only the commonly shared ones. The 01 in Anuvrtti01 represents the author ID, which identifies the author whose views are stored in the table. The table Author lists all authors and their IDs, which are also in the tables Note and Translation, as mentioned above. Please note that we regard classical commentators and editors of existing editions as potential contributors or data 'authors'. After finishing the prototype version of the database and an intensive test period for debugging and receiving user feedback, we intend to ask for permission to integrate existing Aṣṭādhyāyī data and to make it available through our web-based GUI. Note that the aim of this paper is not to present the actual data in our database, which will hopefully be extended through other contributors anyway, but to discuss the database design,

$$\begin{array}{ll}
\text{1.1.1 vṛddhir ādaic} & \Rightarrow \text{vṛddhi} \mid (\text{āT} \circ \text{aiC}) \\
\qquad N^1_{sg,fem,nom} | (N^2 \circ N^3)_{sg,msc,nom} & \Rightarrow N^1 := N^2 \sqcup N^3 \\
\text{1.1.2 adeṅ guṇaḥ} & \Rightarrow (\text{aT} \circ \text{eN}) \mid \text{guṇa} \\
\qquad (N^1 \circ N^2)_{sg,msc,nom} | N^3_{sg,msc,nom} & \Rightarrow N^3 := N^1 \sqcup N^2 \\
\text{1.1.3 iko guṇavṛddhī} & \Rightarrow \text{iK} \mid (\text{guṇa} \circ \text{vṛddhi}) \\
\qquad N^1_{sg,fem,gen} | (N^2 \circ N^3)_{du,fem,nom} & \Rightarrow N^1 \rightarrow (N^2 \sqcup N^3)
\end{array}$$

Fig. 4. Formal analysis of the first three *sūtras*

which needs to be flexible enough to smoothly integrate data from different sources. For the future, it would be great to agree upon one standard database scheme which could provide the basis for a database which integrates all the Aṣṭādhyāyī-related information. With SQL as a powerful query language, researchers could compare different analyses by asking questions like "where do author A and author B differ in their grammatical or *anuvṛtti* analysis?" or they could stick to a single-author analysis and use SQL queries to investigate, for example, Pāṇini's use of negation particles.

4 Outlook

The digital edition presented here is being developed by the research group "*Pratyāhāras* or features?" at the university of Düsseldorf as a tool to support the task of uncovering the underlying structure of the Aṣṭādhyāyī. One of our central research questions is that if we allow all techniques which Pāṇini uses for the description of phonological classes, is it still true that he uses them in the most economical way possible? And what does 'in the most economical way' mean? If we could succeed in fully understanding the structural principles underlying the Aṣṭādhyāyī, i.e., the methods by which it is compiled from a source grammar, we should be able to (semi)-automatically compute a grammatical description in the style of the Aṣṭādhyāyī. In analogy to our previous work on the Śivasūtras, the hypothesis that the Aṣṭādhyāyī is maximally economic could be tested by comparing the computed description to the original one. In order to reduce the difficulty of the task, we start by restricting ourselves to phonological descriptions.

As a first step towards this goal we are currently developing a formal language to decode the *sūtras*, which shall be integrated into our database as well. A first impression of this language can be found in Figure 4. The first *sūtra* consists of a binary compound $(N^1 \circ N^2)$ in singular case and a simple noun (N^1) in singular case. The *sūtra* is interpreted as a definition of N^1 by the union of N^2 and N^3. *Sūtra* 1.1.2 follows the same pattern. *Sūtra* 1.1.3 states an operational rule in which the case markers are metalinguistic markers for the rule components. The other *sūtras* will be analyzed in an analogous manner. Please note that the formal *sūtra* language is still under development and that it is presented here only for illustrative purposes. Our assumption is that a full-fledged formal representation will reveal *sūtra* patterns which correspond systematically to interpretations. We hope that such a representation will help us to decode Pāṇini's compilation principles and to reveal the structure underlying the Aṣṭādhyāyī.

Acknowledgments

Research of this paper was conducted in the research project "*Pratyāhāras* or features?" which is funded by a strategic research fund from the university of Düsseldorf. We are grateful to our project colleagues, in particular Silke Hamann, Oliver Hellwig, Norbert Endres, and Anil Kumar for their collaboration. Moreover, we would like to thank, among others, Jan Houben, Gérard Huet, Amba Kulkarni, Anand Mishra, Peter Scharf, and Navjyoti Singh, for inspiring discussions on the research project. Finally, thanks to the anonymous reviewers for their valuable comments.

Bibliography

S. Bhate and S. Kak. Pāṇini's grammar and computer science. *Annals of the Bhandarkar Oriental Research Institute, Poona*, 72:79–94, 1993.

L. Bloomfield. *Language*. Holt, Rinehart and Winston, New York, 1933.

O. Böhtlingk. *Pāṇinis Grammatik*. Leipzig, 1887. Reprinted: Hildesheim 1964.

E. F. Codd. Further normalization of the data base relational model. *IBM Research Report, San Jose, California*, RJ909, 1971.

E. F. Codd. *The relational model for database management: version 2*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990. ISBN 0-201-14192-2.

S. D. Joshi and J. A. F. Roodbergen. *The Aṣṭādhyāyī of Pāṇini*. Sahitya Akademi, 1991-. So far up to volume 14/2011.

S.D. Joshi and S. Bhate. *The Fundamentals of Anuvṛtti*. Publication of the Centre of Advanced Study of Sanskrit, University of Poona, 1984.

S. M. Katre. *Aṣṭādhyāyī of Pāṇini*. University of Texas Press, Austin TX, 1987.

P. Kiparsky. On the architecture of Pāṇini's grammar. In Grard Huet, Amba Kulkarni, and Peter Scharf, editors, *Sanskrit Computational Linguistics*, volume 5402 of *Lecture Notes in Computer Science*, pages 33–94. Springer, Berlin, Heidelberg, 2009.

W. Petersen. On the construction of Sivasutras-alphabets. In Amba P. Kulkarni and Gérard P. Huet, editors, *Sanskrit Computational Linguistics*, volume 5406 of *Lecture Notes in Computer Science*, pages 78–97. Springer, 2009. ISBN 978-3-540-93884-2.

R. N. Sharma. *The Aṣṭādhyāyī of Pāṇini*. Munshiram Manoharlal Publishers, New Delhi, 1987-2003. 6 volumes.

F. J. Staal. The Sanskrit of science. *Journal of Indian Philosophy*, 23(1):73–127, 1995.

Ś. Ch. Vasu. *The Aṣṭādhyāyī of Pāṇini*. Allahabad, 1891. 2 volumes.