DRAFT

ANNOTATING AND ANALYZING THE AṢṬĀDHYĀYĪ

Wiebke Petersen

petersen@phil.uni-duesseldorf.de

Oliver Hellwig

hellwig7@gmx.de

1. Introduction

The paper introduces the new research project '*Aṣṭādhyāyī 2.0*' that aims at developing a digital edition of the *Aṣṭādhyāyī* - Pāṇini's nearly 2,500 years old grammar of Sanskrit, the ancient Indian language. For modern linguists this grammar is interesting for two reasons. First, its Western (re-)discovery in the 19th century had an enormous influence on contemporary linguistics. For example, the Indologist Maurice Bloomfield wrote about the grammar: "The descriptive grammar of Sanskrit, which Pāṇini brought to its perfection, is one of the greatest monuments of human intelligence and an indispensable model for the description of languages." (Bloomfield 1929). This admiration is mainly due to the fact that Pāṇini made use of many modern linguistic concepts such as thematic roles, abstract derivation levels, and rewrite rules.

Second, the *Aṣṭādhyāyī* uses sophisticated formal techniques to encode the rule system of the grammar in text form. In this way, the *Aṣṭādhyāyī*, though written in Sanskrit, can be interpreted as a compiled code of the grammatical system of Sanskrit. It is generally assumed that Pāṇini aimed at a compilation that maximizes compactness and conciseness by making use of inheritance structures, a sophisticated meta-language, and a marker system (Staal, 2006). Modern linguists could benefit from a deep study of Pāṇini's precise description and concise encoding methods, but regularly lack the ability of understanding the original Sanskrit text. Therefore, modern Western investigations have been based mainly on translations

[1] University of Düsseldorf. Email: petersen@phil.uni-duesseldorf.de

[2] University of Düsseldorf. Email: hellwig7@gmx.de

and classical commentaries. A rigid text-based analysis from a modern linguistic point of view is, however, still a desideratum.

The project '*Aṣṭādhyāyī 2.0*' develops a research tool that allows such a text-based analysis (Petersen and Soubusta, 2013). Providing a simple translation of the grammar is not sufficient for a deep understanding of the *Aṣṭādhyāyī*. While translating the sentences (*sūtras*) of the *Aṣṭādhyāyī*, one frequently has to add information that is not provided by the "optimized" Sanskrit text itself, but is inherited implicitly from other *sūtras*. As a consequence, one loses control over the function of the single text components by reading a translation. We tackle this dilemma by building a digital edition with annotations on all text components (*sūtras*, words, morphemes, and even phonemes). The annotated text is stored in a database and accessed by a web-interface that facilitates to browse the text and to search for particular patterns within the grammar. Our main aim when designing the digital edition was that it requires little prior knowledge and stays as close as possible to the text of the *Aṣṭādhyāyī*. Thus, by entering and analyzing *sūtras* we build a powerful research tool for understanding the deep formal and intellectual structure of this groundbreaking linguistic treatise.

The paper focuses on the annotation task which is complex for three reasons. First, the rich morphology of Sanskrit and the omnipresent Sandhi phenomena make automatic text processing of Sanskrit texts difficult in general. Second, the condensed style of Pāṇini's grammar adds additional obstacles to the task, because its language is not identical with standard Sanskrit, for which computational tools are available. Third, most text components contained in the *Aṣṭādhyāyī* are highly ambiguous. Therefore, we start in Section 2 by sketching the morpho-lexicographic complexity of Sanskrit before we introduce the peculiarities of the text of Pāṇini's grammar. The central annotation task will be described in Section 3. Section 4 gives a short overview of the newly developed *Aṣṭādhyāyī* web interface, and Section 5 indicates what kind of research questions can be tackled with the new tool.

2. Sanskrit and Pāṇini's grammar

In this section, we concentrate on those aspects of standard Sanskrit and of the text of Pāṇini's grammar that confronted us with special problems in the annotation process. More information on the language Sanskrit can be found in standard grammars like Macdonell (1926) or Whitney (1950). For details about the structure of Pāṇini's grammar see Cardona (1976) or

Kiparsky (2009), or refer to one of the translations of the *Aṣṭādhyāyī* (Katre, 1987; Sharma, 2003).

2.1 The Sanskrit language

Sanskrit is a morphologically rich and highly inflected Indo-Germanic language. It has three numbers, eight cases, three genders, and a complex verbal system. Altogether, each noun accounts for 24 inflected word forms and each verb for more than 150 forms.[3] As any highly inflected language, Sanskrit shows many instances of syncretism, where a single form serves two or more morpho-syntactic functions. Furthermore, Sanskrit shows a strong tendency for the formation of nominal compounds many of which consist of more than ten concatenated words. In addition, Sanskrit has a complex system of euphonic rules (*saṃdhi*, Sandhi). These rules induce sound changes at the boundary of words (outer Sandhi) and morphemes (inner Sandhi) triggered by sounds of the immediate context. Finally, Sanskrit has developed a rich vocabulary during the last 2,500 years, including numerous homophonous lexemes derived from different substrate languages, and countless polysemous words.

All these four linguistic properties of Sanskrit — syncretism, compounding, Sandhi, and the rich vocabulary — are responsible for the fact that Sanskrit expressions can be highly ambiguous, making its automatic processing a challenging task (see Section 3). Furthermore, it is one of the reasons why a translation can never replace the study of an original text, as the disambiguation of ambiguous terms is not deterministic and often not invertible.

For a simple example illustrating these phenomena, consider the string *sarvaivātmasampad* ("[this is,] indeed, the full perfection of the soul"). A few possible Sandhi analyses of this string are:

(1) sarvā+ eva + ātma + sampad (ā + e = ai, a + ā = ā)
(2) sarvā + evā + tma + sampad
(3) * sarva + evā + atma + sampad (a + e = ai)
(4) ...

[3] See http://sanskrit.inria.fr/DICO/grammar.html for a Sanskrit form generator.

The first two analyses consist of sequences of valid Sanskrit lexemes, although the second one contains the highly unusual lexeme *tman-* (“soul”) and will be assigned a low score by a language model trained on standard Sanskrit texts. Analysis 1, which is the correct one, also demonstrates the formation of compounds, as *ātma* and *sampad* build up a single syntactic unit in which only the last word *sampad* is declined, while *ātma* is the compositional and, therefore, undeclined form of the noun *ātman* (“soul”). Finally, the word *sampad* is morphologically ambiguous in the given context, as it can be analyzed as a nominative (preferred solution) or vocative singular in any of the three genders.

2.2. The language and structure of Pāṇini’s Sanskrit grammar

The *Aṣṭādhyāyī* (‘eight books’), Pāṇini’s grammar of Sanskrit, consists of approximately 4,000 *sūtras* divided into eight books with four chapters each. The core grammar is accompanied by three special lists:[4]

- *Śivasūtra*: list of the 42 sounds of Sanskrit with intermittent marker elements.
- *Dhātupāṭha*: list of verbal roots organised by the ten verb classes.
- *Gaṇapāṭha*: list of primitive nominal stems.

Together with these lists, the *Aṣṭādhyāyī* is more than a pure grammar. It is a full generative description of Sanskrit including the lexicon, rules for its pronunciation, and its use in different sociolinguistic contexts. It is worth noting that the *Aṣṭādhyāyī* is rooted in the oral tradition of Indian grammar. Over centuries it has been circulated mainly by memorization. Thus, the division into eight books and the numbering of the single *sūtras* by a triple consisting of the book, the chapter, and the position therein has been added later. Created for memorization and recitation, the *Aṣṭādhyāyī* is remarkably short for a grammar of its coverage. This shortness is reached by several techniques of which only some will be demonstrated by analyzing an example *sūtra*.

(5) sūtra 6.1.77: iko yaṇaci

[4] The authorship of the *Dhātupāṭha* and the *Gaṇapāṭha* is unclear. Similar lists predate Pāṇini, but the lists referred to nowadays have probably been extended by later grammarians.

In a traditional *Aṣṭādhyāyī* edition *sūtra* 6.1.77 is translated as: "The semivowel *y, v, r, l* are the substitutes of the corresponding vowels $i, u, ṛ, ḷ$ (long and short) when followed by a vowel" (Vasu, 1891). However, such a 'translation' is much more than a standard translation. It is rather an evaluation or decompilation of a *sūtra* in its context in the *Aṣṭādhyāyī*. In the following, we will see how this evaluation procedure works for sūtra 6.1.77. None of the 'words' in *sūtra* 6.1.77 is a word in standard Sanskrit.

So, we first need to clarify what is expressed by the elements of this *sūtra*. When we resolve the Sandhi, we get

(6) *sūtra* 6.1.77: *ik-aḥ | yaṇ | ac-i |* (de-sandhified and split into morphemes)

Here, *-aḥ* and *–i* are standard Sanskrit case morphemes, *k*, *ṇ*, and *c* are markers from the *Śivasūtras*, and *a* is the unstressed neutral vowel inserted for easier pronunciation. Hence, the *sūtra* can be analyzed as

(7) sūtra 6.1.77: $[\text{iK}]_{\text{GEN}}[\text{yṆ }]_{\text{NOM}}[\text{aC}]_{\text{LOC}}$ (metalinguistically analyzed)

We will first evaluate the elements *iK*, *yṆ* and *aC*. They are pairs, so-called *pratyāhāras*, consisting of a sound and a marker both of which are contained in the *Śivasūtras*. For the evaluation we need *sūtra* 1.1.71:

(8) *sūtra* 1.1.71: A sound-marker pair denotes all sound elements in the interval from the sound to the marker. (interval rule)[5]

From this *sūtra* and the list of sound elements in the *Śivasūtras* we get that *iK* denotes the simple vowels apart from *a* (*iK* = $\{i, u, ṛ, ḷ\}$), *yṆ* denotes the semivowels ($yṆ = \{y, v, r, l\}$) and *aC* denotes all vowels ($aC = \{a, i, u, ṛ$

[5] Note, all *sūtras* but *sūtra* 6.1.77 are given in evaluated or decompiled form to simplify matters. Furthermore, we have reduced the translations for the purpose of our evaluation procedure in focus.

, ḷ, e, o, ai, au}).[6] *Sūtra* 1.3.10 ensures that the right correspondence between vowels and their semivowels is established.

(9) *sūtra* 1.3.10: List elements correspond by their position.

Next, we focus on the evaluation of the case markers in *sūtra* 6.1.77. Two *sūtras* tell us how to interpret them.

(10) *sūtra* 1.1.49: The genitive case marks the substituent.

(11) *sūtra* 1.1.66: The locative case marks the right context of a substitution.

Taking together the four meta-rules in *sūtra* 1.1.49, *sūtra* 1.1.66, *sūtra* 1.1.71, and *sūtra* 1.3.10, we can interpret *sūtra* 6.1.77 as: "Simple vowels apart from *a* are replaced by their corresponding semivowels before a vowel". Finally, *sūtra* 6.1.72 states an important precondition for the application of *sūtra* 6.1.77.

(12) *sūtra* 6.1.72: In close contact.

Sūtra 6.1.72 is not a meta-rule, but a header rule. Its words are inherited by *sūtra* 6.1.72, so that we can confine its interpretation to: "Simple vowels apart from *a* are replaced by their corresponding semivowels immediately before a vowel". Actually, even more *sūtras* are needed to adequately restrict the application of *sūtra* 6.1.77. For example, the two words *asti* and *iha* become *astīha* in close contact and not *astyiha* as *sūtra* 6.1.77 would predict. Here, a complex procedure of rule blocking takes place, for details refer to Kiparsky (2009). The example evaluation demonstrates how much of the structure of the *Aṣṭādhyāyī* one loses if one only reads the translation of a *sūtra*.

---

[6] In Petersen (2009) it has been proven that the list of the *Śivasūtras* is a minimal solution to the problem of ordering the sounds of Sanskrit and interrupting it with marker elements such that all necessary sound classes can be denoted by sound-marker pairs as intervals. Minimality here means that the number of duplicated sounds is minimal and the number of marker elements cannot be reduced. Due to their minimal length, the *Śivasūtras* follow the economy principle (*lāghava*) that is assumed to underlie the whole construction of the *Aṣṭādhyāyī* (cf. Kiparsky, 1991).

In our example analysis of *sūtra* 6.1.77 we saw only one instance of inheritance, namely from *sūtra* 6.1.72 to *sūtra* 6.1.77. Here, *sūtra* 6.1.72 has served as a header for *sūtra* 6.1.77, which inherited all expressions of the former. However, such header *sūtras* are less frequent in Pāṇini's inheritance system. More often we find inheritance instances where only some parts of a former *sūtra* are inherited by its followers. These inherited parts can be small, but convey crucial information like a negation particle.

One common — yet still unproven — hypothesis is that Pāṇini arranged his *sūtras* in a way that maximizes conciseness by exploiting inheritance of *sūtra* parts. There is strong evidence that the *Aṣṭādhyāyī* follows a principle of economy (*lāghava*), but so far it could be only formally proven for the arrangement of the *Śivasūtras* (see footnote 6). For example, in his *sūtras* Pāṇini does not make use of verbs, the word order is often chosen in such a way that Sandhi processes minimize the number of syllables, and the meta-language consists of particular short expressions (see the generated monosyllabic names for phonological classes *ik, yaṇ* and *ac*).

If Western linguists try to get a grasp on the fascinating structure of the *Aṣṭādhyāyī*, they usually face several problems. First, they need to get acquainted with Sanskrit, and they have to gain some mastery of Pāṇini's meta-language, which is rooted in the Indian grammar tradition. Even speakers of Sanskrit (or trained machines, as we will see in the following section) without any special training are unable to read the *Aṣṭādhyāyī*, because it is not written in standard Sanskrit. Second, the *Aṣṭādhyāyī* is structured in such a way that for the analysis of a single *sūtra* one has to apply *sūtras* which are spread over several books and chapters. The linear order of the *sūtras* does neither follow a thematic classification nor a functional classification, e.g., phonological as well as meta-rules are spread over all books. In particular, questions concerning the economic organization of the *Aṣṭādhyāyī* cannot be tackled by sticking to a pure translation of it (see our example evaluation of *sūtra* 6.1.77). That is the reason why we have decided to set up a new digital research tool that aims at facilitating the study of the *Aṣṭādhyāyī* for linguists with no or little previous knowledge of Sanskrit and Indian grammar theory.

3. Annotation of the *Aṣṭādhyāyī*

We are building a multi-layered, richly annotated electronic version of the *Aṣṭādhyāyī* that contains linguistic (morphological, lexical, semantic) and

structural information about the text. This section describes the tools and resources used for the annotation. It focuses especially on domain specific adaptations of existing software tools and knowledge bases, and on the collaborative annotation in an intercultural setting.

3.1. Linguistic annotation

As indicated in the previous section, Sanskrit poses several problems to an automatic linguistic analysis and annotation. Most importantly, single words are merged into larger strings using a fixed set of euphonic rules (*saṃdhi*). While the application of these rules is deterministic, reverting these rules leads to ambiguous analyses in most cases, which need to be resolved using a language model. In addition, Sanskrit has a huge vocabulary and tends to construct large compound nouns (*samāsa*), which can be transformed further into adjectives if the sentence context requires (*bahuvrīhi* formation). The language has a highly unregularized orthography and uses few, if any, punctuation marks.

An edition of the *Aṣṭādhyāyī* from the GRETIL web directory[7] was used as the starting point for the annotation. This e-text was proofread following the printed edition of the text found in Katre (1987). In the next step, the software *Sanskrit-Tagger* (Hellwig 2009) was used to perform joint tokenization, lemmatization, and morphological analysis of the *Aṣṭādhyāyī*. This software achieves an accuracy of over 97% in unsupervised tokenization of narrative texts (Hellwig 2010), which drops to significantly lower rates for texts from rare domains such as the grammatical Sanskrit literature, for which only few training data are available. Therefore, a team of Indian and European experts manually checked the tokenization and the lexical and grammatical analysis of each *sūtra* that was produced by the tagger. The first correction stage was performed by the European team members directly on the raw output of the tagger. The corrected results were sent to the Indian colleagues and revised by an expert in Pāṇinian grammar. Any necessary changes were discussed intensively and integrated into the corrected result.

This annotation procedure differs considerably from the usual state of the art approach for standard Western texts. However, as illustrated in the

[7] http://gretil.sub.uni-goettingen.de/; input of the digital version by Mari Minamino

previous section, the Sanskrit of the *Aṣṭādhyāyī* shows several fundamental linguistic differences when compared with "natural" standard Sanskrit. First, many *sūtras* use linguistic features of the standard language such as noun cases for marking formal relations between the words by which they are constituted. Statistical estimations of the distributions of such linguistic features that were learned on texts of domains other than grammar cannot capture their distribution in the *Aṣṭādhyāyī*, which leads to an increased error rate in morphological analysis. Second, most *sūtras* are not complete sentences, which complicates an n-gram based lexical analysis that was again trained on language from non-grammar domains. Third, the *Aṣṭādhyāyī* introduces numerous lexemes that do not occur in non-grammatical literature. Several of these domain-specific lexemes consist of less than three letters as, for instance, single letters used as markers in *pratyāhāras* or grammatical terms such as *it*. Including these lexemes in the standard dictionary led to a massive overgeneration of possible analyses, as can be observed in the case of the sample word *sukhī*. When only the regular dictionary without the extended grammatical vocabulary is applied, the (correct) analysis *sukhin*[nom. sg. masc.] is proposed as the first of three different analyses. When, however, the additional grammatical vocabulary is activated, the tagger produces six additional analyses such as *su*[compound]+*kh*[nom. du. n.], which are linguistically valid, though meaningless in most contexts. As a consequence, the computational models for the lexicographical and morphological analysis had to be adapted carefully to the requirements of the grammatical literature. How to perform this adaptation in detail was one of the main issues in the discussion of the team of annotators. The main improvements in the tagger were changes in the structure of the dictionary (grammatical terms can be excluded completely from the analysis of a text), relaxed stopping criteria for the generation of possible analyses, and a computationally more efficient implementation of the core algorithms.

Because annotating the *Aṣṭādhyāyī* required an intensive adaptation of the lexical resources and of the computational methods, we were not able to measure the inter-annotator agreement of the annotation. We are, however, confident to have built a resource that conforms as closely as possible to the traditional Indian understanding of the *Aṣṭādhyāyī*.

3.2 Semantic annotation

In the next layer, word semantic annotations were added to each lexical unit in the *Aṣṭādhyāyī*. Word semantic meanings were taken from a hierarchic semantic inventory integrated in the tagging software. The semantic inventory was heavily expanded during the word semantic annotation, because the *Aṣṭādhyāyī* contains numerous semantic classes and items that are not found in normal Sanskrit. For example, *sūtra* 1.1.68[8] defines one of these new classes: In the *Aṣṭādhyāyī*, numerous language expressions do not point to their normal referent, but denote the textual string by which they are constituted. The *sūtra* 1.2.70 is an example for this phenomenon: In *pitā mātā*[9] the semantic information of the nouns *pitṛ* and *mātṛ* refers to the form of the strings, but not to a father or a mother. We have defined a new semantic class "Sanskrit noun denoting itself" in the semantic inventory, and assigned all instances of the phenomenon described in 1.1.68 as subclasses to this new class. 1296 subclasses had to be created, which were assigned to 2021 different words in the *Aṣṭādhyāyī*. Other, though less prominent, cases in which the semantic inventory had to be expanded included grammatical marker words and verbal roots. Apart from semantically unclear words for which no suitable semantic category could be established, the complete *Aṣṭādhyāyī* has been annotated with semantic meanings.

4. The *Aṣṭādhyāyī* web-interface

For building a research tool that allows the study of the *Aṣṭādhyāyī* for linguists without specific training on Sanskrit or Pāṇini's techniques, we decided to develop a digital edition of the *Aṣṭādhyāyī* that is based on an SQL-database.[10] This installation allows us to account for the non-linear structure of the *Aṣṭādhyāyī* as a rule system with a high amount of rule interactions. The order in which the rules are presented can be adapted to

---

[8] Sanskrit text: *svaṃ rūpaṃ śabdasyāśabdasaṃjñā*; translation in Katre (1987): "An expression denotes itself (...) unless it is the name of a linguistic technical term"; see also Cardona (1976)[203] for a discussion of this rule.

[9] "The nominal stem *pitṛ-* 'father' [alone subsists 64] when conjoined with *mātṛ-* 'mother' [optionally 69]." Katre (1987)[49]

[10] The web interface is publicly available at http://panini.phil.hhu.de.

individual research questions. The user interface is designed to optimize our four main aims: knowledge-independency, literality, flexibility, and extendibility.

The basic units in our database scheme are the sound tokens occurring in the *Aṣṭādhyāyī*. Inspired by Pāṇini's use of sound intervals in the *Śivasūtras*, we use intervals of these sound tokens for identifying components. This gives us a maximum of flexibility, as we can decompose expressions in the *Aṣṭādhyāyī* to any desired degree. Remember our example *sūtra* 6.1.77, which we first decomposed into the morpheme level '*i-aḥ yaṇ ac-i*'. Expressions like '*ac*' act like free morphemes on the language level, but as a generated technical term in Pāṇini's system they can be further decomposed into the sound sign '*a*' and the marker sign '*c*'. Our system allows us to annotate the elements on each level of decomposition individually. As an earlier version of our database scheme and the web-interface has been described in greater detail in Petersen and Soubusta (2013), we only give a short introduction into its main browsing features and concentrate on the new search function.

4.1. Browsing

The simplest way to access the data in our digital edition is by browsing through the *sūtras*. Figure 1 shows the current version of our *Aṣṭādhyāyī* browser. Its functionality is similar to the one described in Petersen and Soubusta (2013). The main improvement is that the design is more sober and that readability is improved. The browser is realized as a PHP application that dynamically changes the display according to the user's input. The figure shows the default view on the data (several different tabular views are provided as well).

The interface is flexible and easily extendable. Users can determine which *sūtras* they want to inspect by setting the appropriate interval of *sūtras*. Furthermore, they can decide what kind of information they are interested in by selecting the appropriate checkboxes. At the moment, we provide a translation, the expressions inherited from other *sūtras*, and keywords ('Topics') on the *sūtra* level. On the level of the components within a *sūtra* we provide grammatical information such as part-of-speech, inflectional information, and the corresponding lexeme and meaning for each expression. Compounds and technical terms like *pratyāhāras* are decomposed into their subparts. Via hyperlinks one can navigate through the structure of the *Aṣṭādhyāyī* by jumping to those *sūtras* from which expressions are inherited or in which a lexeme co-occurs.

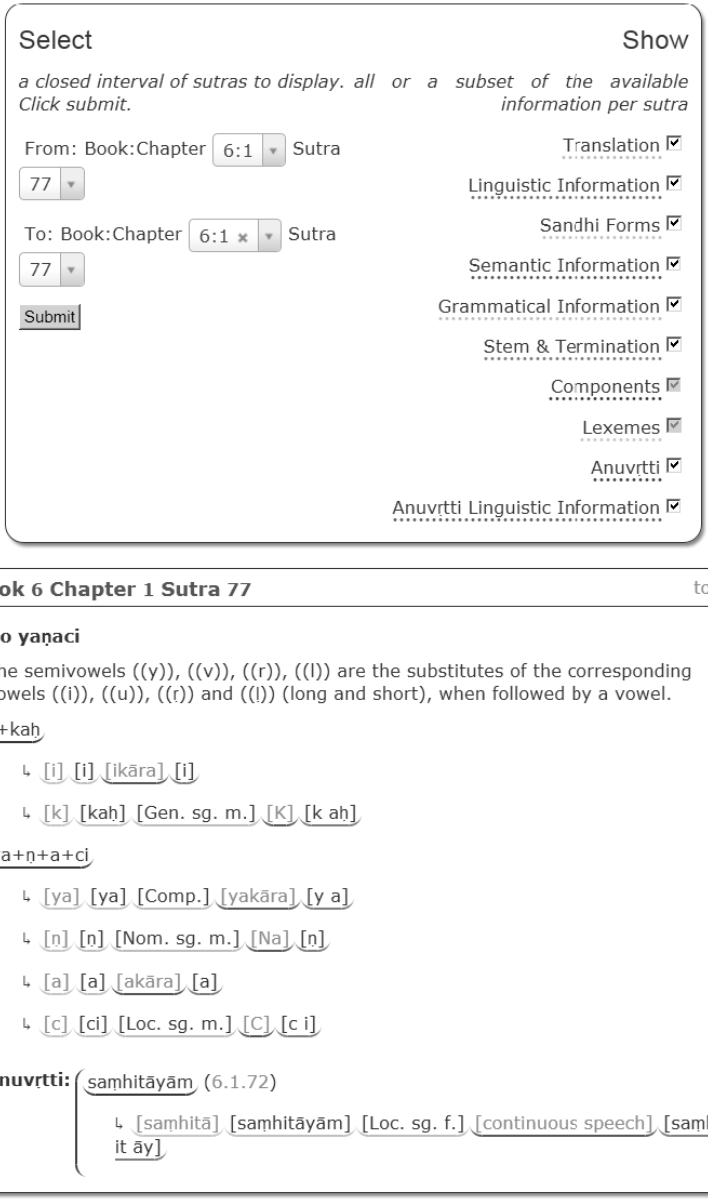

Figure 1. Pāṇini Browser.

4.2. Searching

When structured appropriately, digital editions offer more efficient search facilities than printed ones. When developing our digital edition of the *Aṣṭādhyāyī*, the aim of providing powerful search functionality was our main motivation. It should be possible to search for specific topics, words in particular grammatical contexts or of specific meanings, and for expressions occurring in the text of the *Aṣṭādhyāyī* or its translation. The search function should be easy to use, but flexible enough to allow the formulation of complex queries. Therefore, we decided to use an SQL-database and not a digital text as the backbone of our system. When it comes to searching, the database solution offers concept-based retrieval options as its main advantage. Within an unstructured digital text one is limited to full-text search. In our conceptually organized database we can, for example, search for 'c' used as a marker or for all *sūtras* which are classified as belonging to the domain of phonology.

Having our data stored in an SQL-database, it would in principle be sufficient to offer users of our web-interface the possibility to formulate their search queries immediately in SQL. However, we have decided against this option for two reasons. First, the system would be vulnerable by SQL injections. Second, users would have to learn SQL syntax and to familiarize themselves with our database design in order to utilize SQL queries. In order to avoid these problems, we first experimented with a search interface that allowed users to search the database by filling in a search form with predefined fields (see Petersen and Soubusta, 2013, for details). Judging from the feedback of a few test users, it turned out that this approach was not flexible enough for our needs. As a consequence we developed our own query language that has a simple syntax and allows the use of boolean operators (AND, OR, NOT) and bracketing. Furthermore, it is possible to restrict (parts of) the search to specific fields or clusters of fields in our database. For example, with the expression 'grammar(gen) AND meaning(word)' one searches for *sūtras* in which a genitive word form and a word meaning 'word' occurs. The same query in SQL is much more complex due to the inner structure of the database:

```
SELECT c.componentid , s.sutra , s.chapter , s.book
FROM component c
JOIN atom a ON a . a tomid = c.componentid
JOIN  sutra s ON s.book  =  a.book AND s.chapter =  a.chapter
AND s.sutra =  a.sutra
JOIN cmeaning cm ON cm.meaningid = c.meaningid
```

JOIN grammar g **ON** g.componentid = c.componentid
WHERE cm.meaningname **LIKE** ' word ' **AND** g.grammartype **LIKE** 'decl'
GROUP BY c.componentid , s.sutra, s.chapter , s.book
ORDER BY s.sutra **ASC** ;

In our web-interface users can formulate queries in our query language which are automatically translated into SQL-queries and transferred to the database. The results are displayed in the form that the user is familiar with from browsing the *Aṣṭādhyāyī*.

5. Conclusion

The design principles that Pāṇini followed when he encoded his grammatical description of Sanskrit in the actual text of the *Aṣṭādhyāyī* are still not fully understood. The *sūtras* are neither ordered by linguistic topics (phonology, morphology, etc.) nor by pedagogical considerations (from simple frequent forms to more complex exceptional forms). It is generally assumed that the principle of *lāghava* (economy) guided the design of the *Aṣṭādhyāyī*: The number of meta-linguistic elements is kept small, and the rules are ordered such that by using inheritance and rule blocking mechanisms the text length is minimized. So far, this could only be proven for the *Śivasūtras*, a very small part of Pāṇini's system (see footnote 6). With the digital edition *Aṣṭādhyāyī* 2.0 we have developed a powerful research tool that facilitates the study of the design structure of the *Aṣṭādhyāyī*. At the same time, our web interface is a helpful tool for learners of Pāṇini's system.

As it is based on a database, the digital edition can be easily extended. One extension we are planning to add is a more fine-grained analysis of the components belonging to Pāṇini's meta-language. By adding personal notes to *sūtras* and individual *sūtra* components users have the possibility to personalize their edition. These notes are stored in the database as well and, if desired, can be shared with other users, who can access them while searching and browsing the data. Thus, *Aṣṭādhyāyī* 2.0 provides an interactive research tool for exchanging knowledge and ideas about the analysis of Pāṇini's grammar.

Acknowledgments

Research on this paper was conducted in the research project *Aṣṭādhyāyī* 2.0, which has been funded by the Ministry for Innovation, Science and Research of the State of North Rhine Westphalia, Germany. We are grateful to our project colleagues, in particular Anil Kumar, Norbert Endres, Valentin Heinz, and Patrick Simon for their collaboration.

References

Bloomfield, L. 1929. Review of Liebich. *Language* 5: 267–276.

Cardona, G. 1976. *Pāṇini. A Survey of Research*. The Hague – Paris: Mouton.

Hellwig, O. 2009. “SanskritTagger, a stochastic lexical and POS tagger for Sanskrit.” In *Sanskrit Computational Linguistics. First and Second International Symposia, Lecture Notes in Artificial Intelligence 5402*, ed. by G. Huet, A. Kulkarni and P. Scharf, 266–277, Berlin: Springer Verlag.

Hellwig, O. 2010. “Performance of a lexical and POS tagger for Sanskrit.” In *Proceedings of the Fourth International Sanskrit Computational Linguistics Symposium*, ed. by G. Jha, 162–172. Berlin: Springer Verlag.

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

Kiparsky, P. 2009. “On the architecture of Pāṇini’s grammar.” In *Sanskrit Computational Linguistics, volume 540 of Lecture Notes in Computer Science*, ed. by G. Huet, A. Kulkarni and P. Scharf, 33–94. Berlin, Heidelberg: Springer.

Kiparsky, P. 1991. “Economy and the construction of the Sivasutras.” In *Paninian Studies: S.D. Joshi felicitation volume, no. 37 of Michigan papers on South and Southeast Asia*, ed. by M. M. Deshpande and S. Bhate. Ann Arbor, Michigan: Ann Arbor, Michigan: Center for South and Southeast Asian Studies, University of Michigan.

Macdonell, A. A. 1926. *A Sanskrit Grammar for Students*. Oxford: Oxford University Press.

Petersen, W. 2009. “On the construction of Śivasūtra-alphabets.” In *Sanskrit Computational Linguistics, volume 5406 of Lecture Notes in Computer Science*, ed. by A. P. Kulkarni and G. P. Huet, 78–97. Springer.

Petersen, W. and S. Soubusta. 2013. “Structure and implementation of a

digital edition of the Aṣṭādhyāyī." In *Recent Researches in Sanskrit Computational Linguistics - Fifth International Symposium IIT Mumbai, India, January 2013 Proceedings*, ed. by M. Kulkarni, 84–102. D.K: Printworld.

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

Staal, F. J. 2006. "Artificial languages across sciences and civilizations." *Journal of Indian Philosophy* 34(1-2): 87–139.

Vasu, S. C. 1891. *The Aṣṭādhyāyī of Pāṇini*. Allahabad. 2 volumes.

Whitney, W. D. [1889] 1950. *Sanskrit Grammar*. Harvard University Press.