What's Pāṇini got to do with it?
The use of *gaṇa*-headers from the Aṣṭādhyāyī in Sanskrit literature from the perspective of corpus linguistics

Oliver Hellwig
University of Düsseldorf
Wiebke Petersen
University of Düsseldorf

Abstract

The paper presents strategies for evaluating the influence of Pāṇini's Aṣṭādhyāyī on the vocabulary of Sanskrit. Using a corpus linguistic approach, it examines how the Pāṇinian sample words are distributed over post-Pāṇinian Sanskrit, and if we can determine any lexicographic influence of the Aṣṭādhyāyī on later Sanskrit. The primary focus of the paper lies on data exploration, because the underlying corpus shows imbalances in the data distribution.

1 Introduction

1.1 Motivation and Previous Work

The paper investigates if and how we can quantitatively trace the influence of Pāṇini's Aṣṭādhyāyī on the literary production in Sanskrit. Indologists as early as Whitney (1869 70) have examined which state of Sanskrit is described in the Aṣṭādhyāyī and if this grammar reflects a spoken idiom of Sanskrit.[1] This discussion is far from being satisfactorily settled, as, for instance, Kulikov has shown with a detail study of -*ya* present forms in Sanskrit (Kulikov, 2013). Far less discussed is the question how strongly the literary production of later, post-Pāṇinian Sanskrit texts has been influenced by the rules and the content of the Aṣṭādhyāyī, because most researchers who work in the field of Ancient Indian grammar concentrate on the grammatical tradition of Ancient India itself and on its internal discussions. This is, to a certain degree, in accordance with the impression of self-focusedness of the later grammatical tradition in India. As Houben (2008, 568) has formulated it, the grammatical tradition "largely stands on its own; that is, it addresses intellectual problems in its own tradition and tries to answer criticisms brought by competing traditional disciplines. The problems are mainly internal to the tradition (. . .)". While a lot of work has been invested in reconstructing these intellectual discussions, it seems to be more or less common sense in Indological research that later authors used the Aṣṭādhyāyī or a similar grammar to learn and produce correct Sanskrit.[2]

To simplify matters – actually, to oversimplify them –, we may distinguish between two different levels of influence the Aṣṭādhyāyī had on Sanskrit literature:

1. "Analytic": The analytical influence is found in numerous areas of Sanskrit literature when given word forms are analyzed using the formal system proposed in the Aṣṭādhyāyī.

2. "Synthetic": This term denotes the influence of the Aṣṭādhyāyī on the actual production of Sanskrit texts. The synthetic influence is more difficult to detect than the analytic one, because most authors do not declare that they have constructed a word form according to the rules of the Aṣṭādhyāyī. Therefore, the synthetic influence needs to be estimated from linguistic peculiarities of Sanskrit texts. For this sake, we further split this level into:

[1] Refer to Cardona (1976, 239) for a survey of research in this field.

[2] This does not mean that modern research has completely ignored this problem. For example, while examining parallels between Pāṇini's Sanskrit and Pālī, Oberlies (1997, 19-21) notes that the meaning of the two Pali nouns *koleyyaka* and *gīveyya*, whose Sanskrit equivalents *kauleyaka* and *graiveyaka* are only attested in the classical literature, can be derived using *sūtra* 4.2.96.

Time	Text
250-150 BCE?	Kātyāyana’s Vārttika (135ff.)
150 BCE	Patãnjalis Mahābhāṣya (153)
450-510 CE	Bhartṛhari, the “first author after Patañjali whose work we still have” (170)
7th cent. CE	Kāśikavṛtti (174); commentaries on this work from the 8.-13. c. CE
10. c. CE	Rūpāvatāra “teaches Sanskrit in the form of a catechism arranged by grammatical topics” (174)
14.-17. c. CE	“Other such rearrangements” such as the Siddhāntakaumudī (174ff.)

Table 1: Some extant grammatical treatises; numbers in brackets refer to pages in Scharfe (1977)

(a) Formational synthetic influence, i.e. the influence of the rule-system of the Aṣṭādhyāyī on later literature.

(b) Lexicographic synthetic influence, i.e. the impressions that the vocabulary of the Aṣṭādhyāyī left in later Sanskrit texts.

This paper deals with level 2.b only: It examines if the sample vocabulary contained in the Aṣṭādhyāyī has left any measurable traces in the later Sanskrit literature.

1.2 Interactions between the Aṣṭādhyāyī and later Sanskrit literature

As Houben (2008) has pointed out, the identity of the public for which grammatical texts were composed needs to be inferred from indirect indications. According to Houben, only a small circle of readers actively used original grammatical texts like the Aṣṭādhyāyī, and these professional readers concentrated on discussions about minute details of the Pāṇinian grammar. At least for the later period of Sanskrit literature, Houben assumes that most authors outside these circles only commanded a rather low level of grammatical knowledge (Houben, 2008, 572) which they gained mainly from grammatical “schoolbooks” and not from the original resource. Apart from this unclear inner-Indian attitudes towards (scientific) grammar, the early history of the grammatical tradition is known to us only in fragments. Table 1 sketches the historical distribution of extant works on Sanskrit grammar according to Scharfe (1977). As with any literary tradition of ancient India, the gap in the centuries around CE should not be emphasized too much, because numerous works from this period may have got lost. Bronkhorst comes to the conclusion that “the early centuries following Patañjali saw a rather great activity in the Pāṇinian school of grammar” (Bronkhorst, 1983, 398). The distribution of the extant grammatical texts shows an increasing number starting from the midst of the first millenium CE. Around the same time we see a new bias in the texts towards a didactic adaptation of grammatical theories. In addition, the prestige of Sanskrit increased in the course of time as new forms of Middle and New Indo-Aryan languages were introduced into the “polyglossia in Ancient India” (Kulikov, 2013, 66). For later stages of Sanskrit grammatical literature rearranged versions of the Aṣṭādhyāyī are characteristic ranging from the Rūpāvatāra (Laddu, 1987) to the probably most famous representative of this genre, the Siddhāntakaumudī.

Even if we only take into account the three factors of the inner-Indian attitude, the rank in the polyglossia, and the reformulation in “schoolbooks”, we obtain a complicated, “non-linear” network of interactions in which “elitist” tendencies (grammar restricted to a small circle of savants, its status in the polyglossia) may have been counterbalanced or overridden by a decreasing knowledge of the language which lead to the need and success of “schoolbooks” such as the Siddhāntakaumudī. So, we may sketch some possible interactions as follows. Many researchers agree that Pāṇini described some kind of Sanskrit that was current at his time, though combined with elements from the Vedic language. Therefore, there should be a substantial intersection between linguistic data from his grammar and an early stage of Sanskrit. Starting from the second half of the first millennium CE, we observe a growing number of Sanskrit grammars and of works that were intended for teaching Sanskrit. If the number of extant

Sanskrit grammars is taken as an indication for the interest in the Pāṇinian system, we may suppose that the knowledge of the Pāṇinian system spread further along with these "schoolbooks". If the Aṣṭādhyāyī has actively influenced the language production and especially the vocabulary in later texts, we should be able to find a growing amount of the Pāṇinian vocabulary in the later Sanskrit literature, after a strong lexicographic intersection in the earliest post-Pāṇinian Sanskrit literature. To summarize these considerations, looking at the number of words occurring in Pāṇini's grammar which are used in Sanskrit texts of the different periods, one would expect that this number starts from a high level and constantly increases after a first short drop immediately after Pāṇini.

While this rudimentary model coincides well with the growing prestige of Sanskrit in the Indian polyglossia, as postulated by Kulikov, Houben's statement about the numbers of active users of grammars seems to openly contradict it. Therefore, we use a purely data-driven approach for estimating the lexicographic influence of the Aṣṭādhyāyī on post-Pāṇinian Sanskrit. Results found using this approach (Section 3) will be compared with the results of qualitative research in the conclusion in Section 4.

2 Data and Methods

In this section, we describe which data were selected for quantifying the interaction between the Aṣṭādhyāyī and post-Pāṇinian Sanskrit, and how the data were prepared.

2.1 Definition and Source of the Sample Words

As mentioned in section 1.1, we examine the use of the Pāṇinian "sample" vocabulary from a historical perspective. In this paper, the term "sample word" means a noun (*prātipadika*) that Pāṇini uses to exemplify grammatical rules or linguistic phenomena. In 1.1.68, it is stated that a word (*śabda*) "serves to denote itself (...) unless it is a technical term" (Cardona, 1976, 203). In this way, the term includes what may be called "*gaṇa* headers" such as the word *paila* in *pailādibhyaśca* (2.4.59), stand-alone sample words (*deva* or *brahman* in *devabrahmaṇoranudāttaḥ* (1.2.38)), and terms from the *nipāta* rules. The scope of the paper is restricted to those nouns that are found directly in the Aṣṭādhyāyī. Nouns that are only contained in the *gaṇapāṭha* are excluded from the study, because the date, the authorship and the exact composition of the *gaṇapāṭha* are strongly disputed issues.[3]

The sample words used in this paper are extracted from the semantic layer of the Pāṇinian database that is described in (Petersen and Soubusta, 2013). An edition of the Aṣṭādhyāyī from the GRETIL web directory[4] 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). The software SanskritTagger (Hellwig, 2009) was used to perform joint tokenization, lemmatization and morphological analysis of the Aṣṭādhyāyī. The results were checked repeatedly by a team from India and Germany, and questionable cases were mostly resolved in the spirit of traditional Indian grammatical analysis. Next, all Sanskrit words were annotated with semantic meanings from the semantic inventory of SanskritTagger. Each noun type that denotes itself according to the translation in Katre (1987) was added as a separate subclass to a semantic superclass $S = \{\text{Sanskrit noun}\}$. Finally, all nouns whose semantic meanings are subclasses of S were extracted from the database of SanskritTagger and labeled with one of the following three classes:

- s: single, stand-alone sample words

- g: "*gaṇa* headers", marked by the compound terminators *ādi* or *prabhṛti*[5]. If one of these terminators is found after a word w, we checked if w is interpreted as *gaṇa* header in Katre (1987). This step was necessary to distinguish these uses from expressions such as w-*ādi* "(a compound) beginning with w"[6].

- i: unclear, ignore. This class includes three relevant subtypes:

[3] Refer, for instance, to the introduction of Birwé (1961) and the summary of research in Scharfe (1977, 103/04) and Cardona (1976, 164).

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

[5] See, for instance, *avyayībhāve śaratprabhṛtibhyaḥ* (5.4.107).

[6] See, for example, 4.4.131: *veśoyaśāder bhagād yal* where *ādi* is used in the meaning of "prior member (of a compound)".

Type	*adhyāya*							
	1	2	3	4	5	6	7	8
g	0	10	1	93	43	17	4	7
s	16	52	220	529	416	435	85	93

Table 2: Distribution of the sample nouns over the *adhyāyas* of the Aṣṭādhyāyī

- Number words, which had to be annotated as noun in the data structure of SanskritTagger.
- Words that typically do not occur in non-grammatical texts and may be ad-hoc formations for deriving other forms.[7]
- Cases in which the tokenization or lexical analysis of a compound differ in Katre (1987) and our Pāṇini annotation. One of these comparatively rare cases is found in the *sūtra kuśāgrācchaḥ* (5.3.105), where the compound *kuśāgrāt* is analyzed as *kuśa*[comp.]-*agrāt*[abl. sg. n.] in our annotation, while Katre (1987, 609) interprets the word as a single lexical item "tip of the kuśa grass". From a lexicographic perspective, this compound should not be accepted as an independent lexicographic item, because its meaning is compositional.

It should be noted that nouns were also labeled as sample nouns – and thus included in this study – when the Aṣṭādhyāyī denies that a given rule applies to them.[8] Table 2 provides an overview of how the tokens tagged with one of the accepted types g and s are distributed over the Aṣṭādhyāyī.

2.2 Structure of the Corpus

The paper examines how the sample nouns from the Aṣṭādhyāyī are distributed over time slots and topics of the Digital Corpus of Sanskrit (DCS, Hellwig (2010 2014)). This section describes the structure of this corpus regarding topics and times of origin.

2.2.1 Temporal Structure of the DCS

Dating the older strata of Sanskrit literature up to the first half of the 1st century CE is complicated, and the datings proposed for one text in the research literature frequently differ by centuries. Applying the coarse grained temporal structure brought forward in Hellwig (2010)[9], the Sanskrit literature is split into the following five time slots:

- Time slot 1: Late Vedic literature ($\leq$500 BCE)
- Time slot 2: Early Sanskrit literature (500 BCE – 300 CE)
- Time slot 3: Classical literature (400 – 800)
- Time slot 4: Medieval literature (900 – 1400)
- Time slot 5: Late literature (≥ 1500)

Table 3 presents an overview of how the corpus is composed along the time axis, split up into tokens ("How many nouns are found?") and types ("How many distinct nouns are found?"). The distribution is strongly biased towards the slots 2 and 3, which contain most of the epic and Purāṇic texts in the DCS, while time slot 1 is strongly underrepresented. Moreover, assigning a voluminous and anonymous text such as the Mahābhārata to a single time slot certainly introduces a large amount of noise in the data. It should be noted that table 3 only contains the counts for nouns from the DCS.

[7] The word *eta* in *etetau rathoḥ* (5.3.4) would have been excluded from a list of pronominal sample words.

[8] Example: 5.1.121: *na nañp ūrv āt tatpuruṣād acaturasaṃgatalavaṇavaṭayudhakatarasalasebhyaḥ*

[9] Apart from the secondary sources given in Hellwig (2010), we use the following reference works for dating the texts: (1979) for Tantric texts; general: Winternitz (1908 1920); *dharma* literature: Kane (1962 75), Olivelle (2010); Purāṇas: Hazra (1975); Sāṃkhya: Hulin (1978)

	Number of		
Slot	Texts	Noun tokens	Noun types
1	23	28757	3842
2	15	545760	16259
3	56	436448	24249
4	66	370658	29070
5	39	153329	13155

Table 3: Temporal structure of the DCS

2.2.2 Topic Structure of the DCS

Most of the texts contained in the DCS are labeled with one subject identifier. These identifiers are derived from a system originally proposed by Scharf (forthcoming). For this paper, several of the subject identifiers have been merged into more general super-topics in order to mitigate the problem of data sparseness. The final system comprises the following categories:

- *bud* (Buddhist): Buddhist
- *dar* (darśana): Darśana, Nyāya, Vaiśeṣika, Sāṃkhya, Yoga, Karmamīmāṃsa, Vedānta
- *dha* (dharma): Śrautasūtra, Gṛhyasūtra, Dharmaśāstra, Dharmasūtra
- *gra* (grammar): Pratiśākhya, Śikṣā, Vyākaraṇa, Pāninīya, Apāninīya, Nirukta, Chandas
- *iti* (itihāsa): Itihāsa, Mahābhārata, Rāmāyaṇa, Purāṇa
- *kos* (kośa): Kośa, Saṃgraha
- *poe* (poetry): Kāvya, Kathā
- *rel* (religion): Bhakti, Tantra, Āgama, Mantra
- *sci* (science): Jyotiṣa, Upaveda, Āyurveda, Gāndharvaveda, Dhanurveda, Vāstuśāstra, Alaṃkāraśāstra, Nāṭyaśāstra, Śilpaśāstra, Arthaśāstra, Ratnaśāstra, Kamaśāstra, Rasaśāstra
- *shr* (śruti): Śruti, Saṃhitā, Ṛgveda, Sāmaveda, Yajurveda, Atharvaveda, Brāhmana, Āraṇyaka, Upaniṣad

Table 4 lists the frequencies of nouns split by the two variables time slot and topic. As in the case of temporal labeling, assigning only one topic to a text may not be adequate for text classes such as the Purāṇas, which are compilations from different intellectual domains. Equally critical is the distinction made between the classes *dha*, *shr* and *rel*. This distinction partly incorporates temporal information, because *shr* occurs only in the first time slot, and it may be made from a Western perspective distinguishing between "morals" (*dha*) and religion.

2.3 Temporal Distribution of Sanskrit Nouns

We needed to examine the general distribution of nouns over the time slots, because varying amounts of noun types influence the possibility that a word from Pāṇini occurs in a text just by chance. The first important factor is the number of noun types used in every time slot, which is plotted in Figure 1. For producing this plot, we split the full noun data into the five temporal layers and drew ten samples of 20.000 words from each of these layers. The plot demonstrates that the nominal vocabulary of Sanskrit becomes, in general, more diverse in the course of time. The drop in the last time slot may be due to the comparatively uniform vocabulary of alchemical texts, many of which are inserted in this last layer. The observation that the diversity of the nominal vocabulary increases over time is important for this study,

		Time slot				
		1	2	3	4	5
Topic	*bud*		8903	13546		
	dar		4310	17019	878	7309
	dha	11884	15502	32746	3357	641
	gra	239		1238		
	iti		463893	156639	82119	57334
	kos			4880	42207	1177
	poe			47294	20094	1749
	rel		195	10529	32972	31246
	sci		52099	147201	183585	51978
	shr	14765	794			

Table 4: Frequencies of noun tokens in the DCS, split by time slots and simplified topics

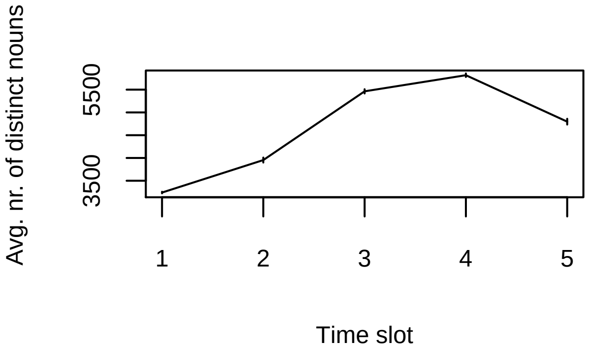

Figure 1: Averaged number of noun types per time slot; sample size: 20.000 words, sampling for each slot repeated ten times

because chances to find Pāṇinian sample words in the later, more diversified layers of Sanskrit literature should be higher than of finding them in the earlier literature, if these words are merely selected by chance.

A note on the sampling method: Drawing random samples of fixed sizes (instead of using the full subpopulations) was motivated by the fact that noun diversity grows approximately logarithmically with the sizes of the samples, which is due to Zipf's law of word frequencies.[10] If the number of noun types in all layers were compared using the respective full subsets, the distribution would generate too high values for the first time slot, for which only about 30.000 words are available (refer to Table 3). As a consequence, we drew samples of fixed size for all statistical evaluations in the rest of this paper.

The increasing diversity of the Sanskrit vocabulary shown in Figure 1 goes along with another trend that is plotted in Figure 2. This figure shows that the proportion of nouns increases with the course of time, while the relative frequencies of finite verbal forms and of adjectives drop after the first time slot. This distribution points to the increasing nominalization of Sanskrit, which can be observed, for instance, in the later scientific literature written in Sanskrit. In the context of the present study, the increasing nominalization means that we have to expect a higher ratio of nouns in later texts and, as a consequence, a higher chance to meet any of the sample words from the Aṣṭādhyāyī.

The results of Section 2 can be summarized as follows:

- According to Figure 1, the diversity of the Sanskrit vocabulary increases at least up to the fourth

[10] Piantadosi (2014) provides an introduction in this area, with a strong focus on cross-linguistic, data-driven evaluation.

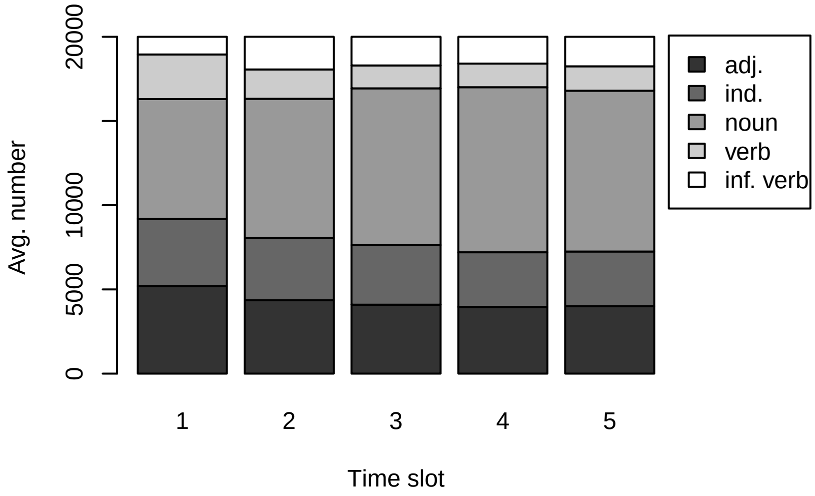

Figure 2: Proportions of grammatical classes per time slot. Data for each slot are averaged from 10 independently drawn samples of 20.000 words.

time slot ("medieval" texts).

- The number of noun types in a sample depends on the selected sample size (Zipf).
- The proportion of nouns increases in each time slot (increasing nominalization of Sanskrit; Figure 2).
- As a consequence, we used stratified samples of equal sizes from each time slot, and recorded the absolute numbers of sample word types in each sample. These numbers form the basis for the statistical evaluation in Section 3.

3 Evaluation

3.1 Did Pāṇini use Typical Words?

For a better understanding of how the Aṣṭādhyāyī and Sanskrit literature interacted on a lexicographical level, it should be asked if the frequencies with which sample words are mentioned in the Aṣṭādhyāyī are correlated to the frequencies with which these words are mentioned in non-grammatical Sanskrit texts. Does Pāṇini use typical vocabulary for his samples, and does he use popular Sanskrit words more often than less popular ones?

The frequencies of all sample words from the Aṣṭādhyāyī were split into five mutually exclusive bins, which contain the number of sample words that are dealt with once (bin 1), twice (bin 2), three to five times (bin 3), six to ten times (bin 4), and more than ten times (bin 5) in the Aṣṭādhyāyī. Each sample word was assigned to one of these bins based on its frequency in the Aṣṭādhyāyī, and the corresponding absolute frequencies of the words were retrieved from the DCS. Figure 3 shows boxplots for each frequency bin (x-axis). For a convenient display, the absolute frequencies in the DCS (y-axis) were transformed into the log space, with zero frequencies replaced by a small value of 0.001. As can be observed in Figure 3, the group means are increasing with the frequency in the Aṣṭādhyāyī. A non-parametric rank correlation test between the frequencies in Pāṇini and the frequencies in the DCS (Kendall's τ) supports this impression. The test produces $\tau = 0.3331$, which is highly significant at the 10% level. Thus we may note that there is a strong correlation between the frequencies of sample words in non-grammatical texts and in the Aṣṭādhyāyī, without making any further statement about the direction or nature of this correlation.

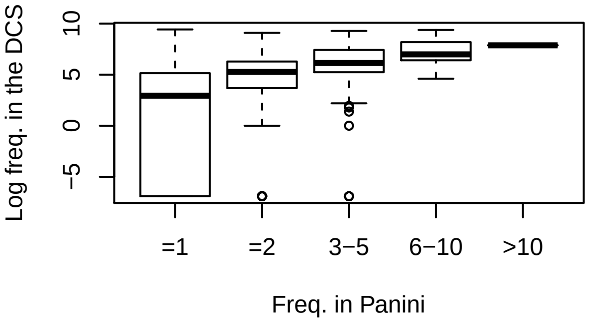

Figure 3: Binned frequency of sample words in Pāṇini (x-axis), compared with their frequencies in the DCS (y-axis)

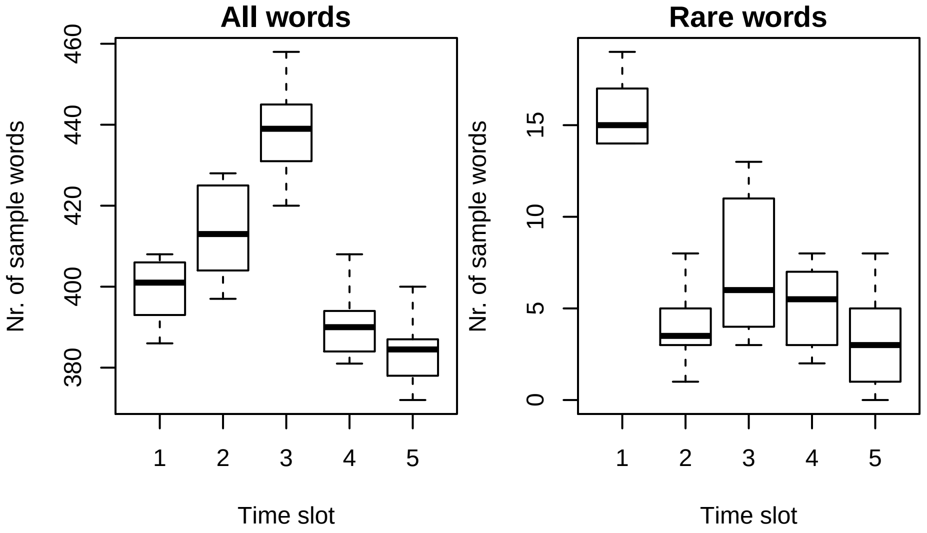

Figure 4: Number of sample word types per time slot; left: all sample words, right: sample words that occur 20 times at most in the DCS. Data for each slot are averaged from 10 independently drawn samples of 20.000 words.

3.2 The Influence of Time

Following the sampling method described above in Section 2.3, we drew 10 random samples of size 5.000 from each time slot, and calculated the average number of sample word types for each slot. Figure 4 (left) shows the result when all sample words from the Aṣṭādhyāyī are used. Most notably, the total number of sample words grows up to the third "classical" time slot. We performed a detail analysis to detect those words that occur most typically in time slot 3. While only few of the top scoring words were semantically rather generic (e.g., *sākṣin*, "witness"), a substantial part of them was found in Āyurvedic texts from the *bṛhattrayī*, and they denote parts of the body (*sakthi*, "thigh"), diseases (*atisāra*, "diarrhoea"), and items from flora and fauna (*ikṣu*, "sugar cane", or *varṣābhū*, "frog"). This finding may point to a growing interest in the Aṣṭādhyāyī up to the end of the first millennium CE, as described in Section 1.1.

A different picture emerges when only rare sample nouns (frequency threshold in the DCS: $\theta \leq 20$) are examined. As can be observed in the right subplot of Figure 4, the distribution for the slots 2-5 follows a similar pattern as the distribution in the left subplot for all words. By far the largest value is, however, found in the earliest time slot 1. A detailed analysis shows that words from the domain of sacrifice (*yajña*) such as *chandoga* and *uṣṇih* and proper names (*jābāla*, *gotama*) are dominant in the

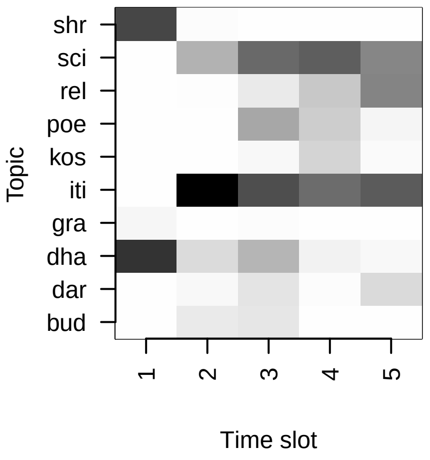

Figure 5: Data from Figure 4 (left subplot), split up by simplified topics. The darkest colour marks the highest value.

early time slot.

3.3 The Influence of Topics

As the detail analyses in Section 3.2 have suggested, the subplots in Figure 4 may present spurious correlations, because they don't take into account the topics of texts. Therefore, a more realistic evaluation should include the topic of each text as an additional independent variable in the statistical assessment.

Figure 5 displays a heatmap in which the darkest colour corresponds to the highest number of sample word tokens found in a combination of time slot and topic. The highest values are assembled in the first column, i.e. the first time slot, and in the rows corresponding to scientific and *itihāsa* literature. Each of the time slots in the columns of Figure 5 represents the averaged counts found in 20.000 words, but split over the simplified topics. This approach does not guarantee that each factor level, i.e. each combination of time slot and topic, contains the same number of words. Because unequal cell counts result in biased estimations of token frequencies (refer to Section 2.3 and to Table 4), we repeated the evaluation with equalized cell counts for each combination of time slot and topic. For this sake, we removed the following rows (topics) and columns (time slots) from the data plotted in Figure 5:

- Time slot 1: Because the main focus of the paper is on the later use of the Pāṇinian vocabulary, this low frequency slot is removed from the data.

- Topics *bud*, *gra*, and *shru*, for which not enough textual material is contained in the DCS. In addition, *gra* reflects the internal discussion of the Indian grammatical tradition and may, therefore, refer intensively to the examples given in the Aṣṭādhyāyī. Because this paper aims at evaluating non-grammatical Sanskrit, the scientific study of Sanskrit grammar is left out from the evaluation.

Although there are no data for time slot 2 and only few data for time slot 3 (Amarakośa), the topic "kos" was retained, because the lexicographic tradition may be a good candidate for incorporating rare, but "prestigious" words from the Aṣṭādhyāyī. Based on the frequencies recorded in Table 4, we chose a sample size of 500 words. The results are displayed in Figure 6. The left subplot (all words without frequency threshold) shows that the number of sample word tokens in general decreases in the course of time, and that the effects observed in the left subplot of Figure 4 are, most probably, due to another topic structure of the Sanskrit literature in slot 3. The right subplot of Figure 6 confirms the assumptions made about the role of the *kośa* tradition. Finally, Figure 7 shows the temporal distribution of Pāṇinian sample words as the column-wise average of the numbers used to create Figure 6. Once again, it supports the statements made about the decreasing use of the sample vocabulary of the Aṣṭādhyāyī.

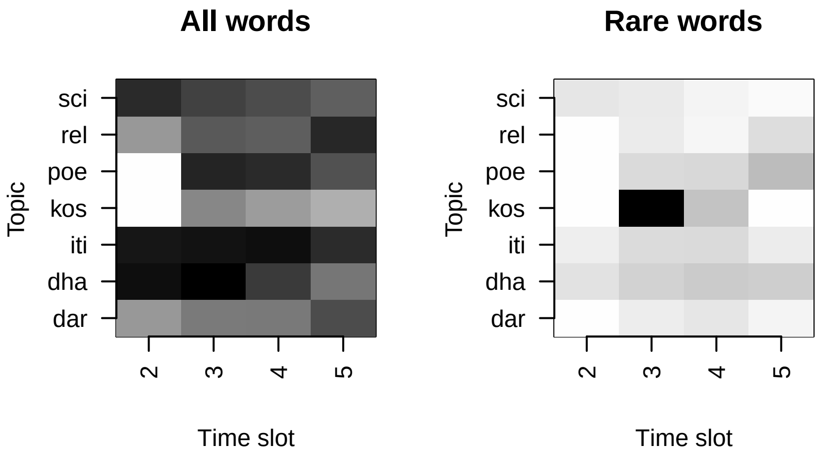

Figure 6: Heatmaps for all (left subplot) and rare ($\theta \leq 20$) sample words. Data for each slot are averaged from 100 independently drawn samples of 500 words.

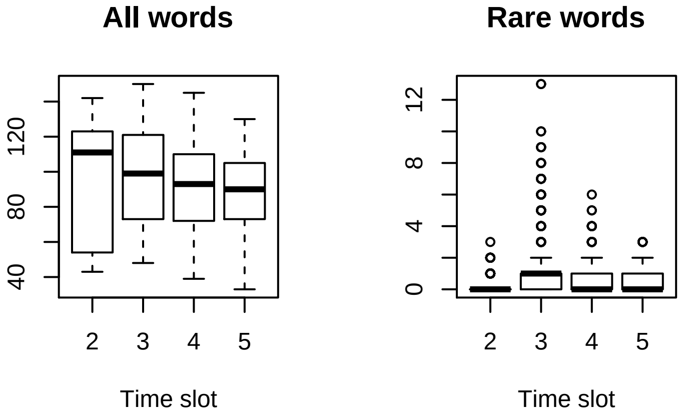

Figure 7: Frequencies for all (left subplot) and rare ($\theta \leq 20$) sample words, averaged per time slot; data source as for Figure 6.

4 Summary and conclusion

Our study has produced results that are contradictory at first view. Figure 4, for which the topic structure of the DCS is not taken into account, displays an increasing use of Pāṇinian sample words until and including the “classical” period of Sanskrit literature. This effect is also, though less clearly, discernible when only the rare words are examined. However, when text topics are added as a factor level, we obtain another picture for the post-Pāṇinian Sanskrit literature (cmp. Figure 6 and esp. Figure 7): There are only few domains such as medicine and especially indigenous lexicography in which relevant amounts of the Pāṇinian sample words are found. In the remaining domains, the lexicographic trace of the Aṣṭādhyāyī becomes increasingly unnoticeable. This result seems to support Houben’s ideas concerning the limited readership of the Aṣṭādhyāyī. At this point, we want to emphasize once more that our conclusions only concern the lexicographic influence.

It has become clear that this study opens several directions for future research and expansion. From the perspective of Corpus Linguistics, the database of the DCS needs to be expanded. As corpora used for studying modern languages are frequently larger by a factor of 100 at least, the only viable method is the unsupervised analysis of digital Sanskrit texts. Section 2.3 has shown the crucial importance of sampling methods and sample size. As a consequence, the corpus should not only become larger, but also better balanced with regard to additional information such as time, genre or the regional distribution, which has not been analyzed for this paper. Obviously, Indology needs to build up reliable, peer-reviewed resources for such types of metainformation. Tracing the influence that the Aṣṭādhyāyī has left on the formational level remains the second open problem. The polishing theory proposed by van Daalen (1980) has shown first, but mainly qualitative steps in this direction. In general, studies of the formational influence should first concentrate on a single grammatical phenomenon described in the Aṣṭādhyāyī, and examine its distribution in the Sanskrit literature using methods from Corpus Linguistics.

References

Robert Birwé. 1961. *Der Gaṇapāṭha zu den Adhyāyas IV und V der Grammatik Pāṇinis.* Otto Harrassowitz, Wiesbaden.

Johannes Bronkhorst. 1983. On the history of Pāṇinian grammar in the early centuries following Patañjali. *Journal of Indian Philosophy*, 11:357–412.

George Cardona. 1976. *Pāṇini. A Survey of Research.* Mouton, The Hague - Paris.

Sanjukta Gupta, Dirk Jan Hoens, and Teun Goudriaan. 1979. *Hindu Tantrism.* Handbuch der Orientalistik, Zweite Abteilung, Vierter Band, Zweiter Abschnitt. E.J. Brill, Leiden/Köln.

R.C. Hazra. 1975. *Studies in the* Purāṇic *records on Hindu rites and customs.* Motilal Banasidass, Delhi.

Oliver Hellwig. 2009. `SanskritTagger`, a stochastic lexical and POS tagger for Sanskrit. In Gérard Huet, Amba Kulkarni, and Peter Scharf, editors, *Sanskrit Computational Linguistics. First and Second International Symposia*, Lecture Notes in Artificial Intelligence, 5402, pages 266–277, Berlin. Springer Verlag.

Oliver Hellwig. 2010. Etymological trends in the Sanskrit vocabulary. *Literary and Linguistic Computing*, 25(1):105–118.

Oliver Hellwig, 2010–2014. *DCS - The Digital Corpus of Sanskrit.* Heidelberg.

Jan E.M. Houben. 2008. Bhaṭṭoji Dīkṣita’s “Small Step” for a Grammarian and “Giant Leap” for Sanskrit Grammar. *Journal of Indian Philosophy*, 36:563–574.

Michel Hulin. 1978. *Sāṃkhya Literature.* A History of Indian Literature, Volume VI, Fasc. 3. Otto Harrassowitz, Wiesbaden.

P.V. Kane. 1962–75. *History of Dharmaśāstra.* Bhandarkar Oriental Research Institute, Poona.

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

Leonid Kulikov. 2013. Language vs. grammatical tradition in Ancient India: How real was Pāṇinian Sanskrit? *Folia Linguistica Historica*, 34:59–91.

S. D. Laddu. 1987. How early were the handbooks on derivation (prakriyā) in Sanskrit grammar? *Annals of the Bhandarkar Oriental Research Institute*, 68:593–601.

Thomas Oberlies. 1997. Pali, Pāṇini and "Popular" Sanskrit. *Journal of the Pali Text Society*, 23.

Patrick Olivelle. 2010. Dharmaśāstra. In *Brill's Encyclopedia of Hinduism*, pages 56–71. Brill, Leiden.

W. Petersen and S. Soubusta. 2013. Structure and implementation of a digital edition of the Aṣṭādhyāyī. In M. Kulkarni, editor, *Recent Researches in Sanskrit Computational Linguistics*, pages 84–103. D.K. Printworld.

S.T. Piantadosi. 2014. Zipf's law in natural language: a critical review and future directions. *Psychonomic Bulletin and Review*.

Hartmut Scharfe. 1977. *Grammatical Literature*. A History of Indian Literature, Volume 5, Fasc. 2. Otto Harrassowitz, Wiesbaden.

Peter Scharf. forthcoming. Providing access to manuscripts in the digital age. In *Writing the East: History and New Technologies in the Study of Asian Manuscript Traditions*. Schoenberg Center for Electronic Text and Imaging, Philadelphia.

L.A. van Daalen. 1980. *Vālmīki's Sanskrit* Orientalia Rheno-Traiectina, 25. E.J. Brill, Leiden.

William Dwight Whitney. 1869–70. The study of Hindu grammar and the study of Sanskrit. *Transactions of the American Philosophical Society*, pages 20–45.

Moriz Winternitz. 1908–1920. *Geschichte der indischen Litteratur*. Amelang, Leipzig.