Etymological trends in the Sanskrit vocabulary

Oliver Hellwig
Südasien-Institut, Universität Heidelberg, Germany

Correspondence:
Oliver Hellwig,
Südasien-Institut,
Universität Heidelberg,
Germany.
E-mail: hellwig7@gmx.de

Abstract

The article examines how the etymological composition of the Sanskrit lexicon is influenced by time and whether this composition can be used to date Sanskrit texts automatically. For this purpose, statistical tests are applied to a corpus of lexically analyzed texts. Results reported in the article may contribute to the diachronic lexicography of Sanskrit and help to develop computational methods for analyzing anonymous and undated Sanskrit texts.

During the last three millennia, the vocabulary of Sanskrit has been influenced by numerous languages. While the early influences from non-Indo-Aryan (non-IA) languages have been studied intensively (see, e.g. Witzel (1995) for a recent survey), a different picture emerges for later texts. It has long been noted that post-Vedic Sanskrit[1] has drawn heavily on languages from 'lower' strata since the times of the epics (see, e.g. Bloch, 1965, p. 14). Some scholars claim that the convergence of IA and non-IA vocabulary supports the theory of a South Asian 'Sprachbund' (e.g. Emeneau, 1980, p. 168). However, when it comes to the details of this convergence, the statements often become vague. Emeneau, for instance, who refers to Burrow's survey of the Sanskrit vocabulary (Burrow, 1959, pp. 376–88), claims that most borrowings from Dravidian 'appear in the early stages of Sanskrit literature, including the grammarians and the epics' (Emeneau, 1980, p. 184) and that these borrowings virtually stop 'by the end of the pre-Christian era' (Emeneau, 1956, p. 6). Although such statements are certainly grounded in an intensive study of the source texts, they are purely qualitative. It remains unclear how big these changes were, whether they show a systematic pattern, and whether they were equally important in all parts of the Sanskrit literature and for all etymological classes. In addition, such theories cement the image of classical Sanskrit as a static language untouched by the regular mechanisms of linguistic development in post-Pāṇinean times, which has been criticized by researchers such as Salomon (1989). In this article, I will examine the external influences on the Sanskrit vocabulary in terms of time by applying some basic statistical procedures to an electronic corpus of Sanskrit texts. The study addresses the question of whether the proportions of IA, Middle IA (MIA), Dravidian, and Austro-Asiatic (AA) words have changed significantly during the development of Sanskrit literature. The results presented in this study may be used to clarify some long-known Indological theories, and may also shed light on a notoriously difficult area of Sanskrit philology. If it is possible to create a mathematical model that describes the influences of other languages on Sanskrit in terms of time, such a model can also be used to date texts on the basis of their etymological composition.

From the perspective of computational philology, this article addresses two questions. First, it discusses a feature of genuinely philological nature. Juola (2008, p. 73) notes that mainstream scholars fail 'to value and participate in the digital humanities community'. This complaint is definitely true. However, while searching for the 'killer app' that may increase the acceptance of computational methods in traditional scholarship,

doi:10.1093/llc/fqp034 Advance Access published on 8 October 2009

Juola (2008, pp. 78–81) mainly proposes tools for searching, indexing, and text annotation. In my opinion, the central barrier between traditional and computational humanities is located elsewhere: many computational approaches use simple features such as sentence length or the frequency of function words. These features can easily be extracted from digital texts, but they are almost meaningless for traditional philologists. Using features of a higher scientific order may, therefore, narrow the gap between the two philological paradigms. The second question is closely connected with the first one. The results of computational studies are frequently not correlated with the linguistic or cultural history, which makes it difficult for traditional scholars to appreciate the findings of these studies. Therefore, this article allots much space to the question of whether the etymological trends found with statistical methods reflect the development of Indian religions (cf. p. 8ff.).

The article is divided into three sections. The first section focuses on the question of whether we can detect statistically significant changes in the vocabulary of Sanskrit. For this purpose, texts contained in an electronic corpus of Sanskrit are assigned to one of five important periods of Sanskrit literature (Section 1.1). Next, the etymological composition of each text is calculated based on Mayrhofer's *Concise Etymological Sanskrit Dictionary* (Mayrhofer, 1956–1976). The frequency distributions of IA, MIA, Dravidian, and AA[2] words are then examined using analysis of variance (ANOVA) and some related methods (Section 1.2). In Section 2, we will report how the linguistic trends found in the first section can be used to date Sanskrit texts. Section 3 evaluates the findings critically and discusses perspectives for future research.

1 Developments in the Sanskrit Vocabulary

1.1 The Sanskrit corpus

This study is based on a corpus of tagged Sanskrit texts that has been compiled during the last 10 years (ct. Hellwig, 2009b). The corpus currently contains more than 2,500,000 lexical items in about 200 texts, some of which are analyzed only in parts. Since the study investigates the influence of time on the vocabulary of texts, the texts contained in the corpus must first be dated. The study views the development of the Sanskrit vocabulary from a bird's eye perspective. To detect trends in the etymological composition of texts, we do not need an exact date for each text, but can content ourselves with approximate datings.[3] Therefore, I operate with so-called time slots (TSs). First, I tried to obtain an acceptable dating for each text from the standard Indological literature (e.g. Winternitz (1968), Frauwallner (1953), Meulenbeld (2000), and the volumes of the *History of Indian Literature*). Based on these frequently preliminary datings, each text was inserted into one of five TSs (TSs 1–5), which correspond to important periods of Indian literature. In addition, the texts are labeled with one of three content types (CTs). The CT serves as a control variable because the etymological composition of a text may also be influenced by its content. Especially, obvious examples are texts dealing with Indian flora or agriculture. As many of the plants described in such texts grow only in India, the IAs may have adopted their local names, which came either from Dravidian or Munda in most cases (cf. Emeneau (1954, p. 286), Zide and Zide (1976), and Masica (1979, pp. 129–37) for proportions in modern Indian languages). If, for instance, texts about Indian flora occurred only in late parts of the Sanskrit literature, a high rate of non-IA words in the late literature could simply be due to its content. To keep the number of variable levels low, each text was assigned to one of the three classes '*r*(eligious)', '*n*(arrative)', or '*s*(cientific)'. Of course, this grossly oversimplifies the content of these texts, and many examples of unclear classifications can be adduced (see, for instance, the labeling of the Mahābhārata and the philosophical texts). As the study is exploratory and operates with approximate datings, these inaccuracies seem excusable at the current phase of research. The place of origin of each text could have been another important control variable. However, each additional variable increases the minimal number of texts on which the statistical part of the study must be based. Since the corpus is still rather limited in extent and, in addition, unbalanced in its temporal distribution, examining the relationship between place and etymological composition must be relegated to a later study. The following overview of the corpus is structured by the five TSs. After a short definition of each slot, it lists the texts contained in the slot and the Indological source from which this information was obtained. An asterisk (*) indicates that I estimated the date of the text (c. = century).

Time slot 1: Late Vedic literature (≤500 BC): This slot contains the old ritual literature, the Dharmasūtras, the Brāhmaṇas and the Upaniṣads. Although the dates of these texts are hard to fix (cf. Mylius, 1983, pp. 84–91), they are probably earlier than the epics (Gonda, 1977b, p. 478). All texts contained in this slot are labeled as *r*(eligious)—AŚVALĀYANAGṚHYASŪTRA, ĀPASTAMBHAGṚHYASŪTRA, GOPATHABRĀHMAṆA, GOBHILAGṚHYASŪTRA, CHĀNDOGYOPANIṢAD, JAIMINIGṚHYASŪTRA, PĀRASKARAGṚHYASŪTRA, BAUDHĀYANAGṚHYASŪTRA, BHĀRADVĀJAGṚHYASŪTRA, MĀNAVAGṚHYASŪTRA, VĀRĀHAGṚHYASŪTRA, VAIKHĀNASAGṚHYASŪTRA, -DHARMASŪTRA, ŚVETĀŚVATAROPANIṢAD, and ŚĀṄKHĀYANAGṚHYASŪTRA

Time slot 2: Early Sanskrit literature (500 BC–300 AD): This period begins with the origin of non-Brāhmaṇic religions such as Buddhism and Jainism and spans the empire of the Mauryas and the following political decline. The large Sanskrit epics MAHĀBHĀRATA and RĀMĀYAṆA and early medical and philosophical works were composed in this period, and the BUDDHACARITA illustrates the early phase of Sanskrit poetry–ARTHAŚĀSTRA (300 BC–300 AD, s, Mylius (1983, pp. 285–6)), ṚTUSAṂHĀRA (100–300, n, Lienhard (1984, p. 108)), CARAKASAṂHITĀ (100 BC–300 AD?, s, Meulenbeld (2000, IA, p. 105ff.)), TANTRĀKHYĀYIKĀ (3./4. c. AD, n, Mylius (1983, p. 167)), NYĀYASŪTRA (similarities with the CARAKASAṂHITĀ, r, Ruben (1966, p. xiv–xvi)), PĀŚUPATASŪTRA (100 AD?, r, Gonda (1977a, pp. 216–19)), BUDDHACARITA (50 BC–100 AD, r, Johnston (1936, p. 1, xvii)), MANUSMṚTI (200 BC–200 AD, r, Mylius (1983, p. 294)), *MAHĀBHĀRATA, MŪLAMADHYAMAKĀRIKĀḤ (2. c., r, Winternitz (1968, p. II, 253)), YĀJÑAVĀLKYASMṚTI (2./3. c. AD, r, Mylius (1983, p. 295)), *YOGASŪTRA, RĀMĀYAṆA (5. c. BC–3. c. AD, n, Brockington (1984, p. 329)), *VIṢṆUSMṚTI (r, Mylius (1983, p. 292)): 'sehr alt', VAIŚEṢIKASŪTRA (200 BC–0, r, Matilal (1977, p. 54)).

Time slot 3: Classical literature (400–800): During this short period, which partly coincides with the rule of the Guptas, the classical works of Sanskrit poetry and philosophy were composed. The first documents written in Apabhraṃśa appear near 500 AD (von Hinüber, 1986, p. 24)).—ABHIDHARMAKOŚA, ABHIDHARMAKOŚABHĀṢYA (≤500, r, Winternitz (1968, p. II, 256–7)) AMARUŚATAKA (7. c., n, Lienhard (1984, p. 92)), AṢṬĀṄGAHṚDAYASAṂHITĀ, AṢṬĀṄGASAṂGRAHA (<850, s, Meulenbeld (2000, IA, p. 631)), KĀMASŪTRA (3.-5. c., s, Mylius (1983, p. 314)), KĀVYĀDARŚA (7. c., s, Mylius (1983, p. 177)), KĀVYĀLAṂKĀRA (650 AD, s, Mylius (1983, p. 176)), KUMĀRASAṂBHAVA (400 AD, n, Lienhard (1984, p. 171)), KŪRMAPURĀṆA (8. c.?, n, Rocher (1986, p. 186)), GAṆAKĀRIKĀ, GAṆAKĀRIKĀṬĪKĀ (younger than Kauṇḍinya, r, Gonda (1977a, pp. 220–1)), PAÑCĀRTHABHĀṢYA (400–600 AD, r, Gonda (1977a, p. 219)), BODHICARYĀVATĀRA (7. c., r, Winternitz (1968, pp. II, 259–60)), BHĀGAVATAPURĀṆA (600–900?, n, Rocher (1986, pp. 147–8)), MATSYAPURĀṆA 500?, n, Rocher (1986, p. 199)), MEGHADŪTA (see KUMĀRASAṂBHAVA), YOGASŪTRABHĀṢYA (500?, r, Frauwallner (1953, p. 1, 288)), LAṄKĀVATĀRASŪTRA (≤443, r, Winternitz (1968, pp. II, 243–44)), LIṄGAPURĀṆA (6.–9. c., n, Rocher (1986, pp. 187–8)), VAIŚEṢIKASŪTRAVṚTTI (8./9. c., r, Matilal (1977, p. 74)), SĀṂKHYAKĀRIKĀ (4./5. c., r, Hulin (1978, p. 127)), SĀṂKHYAKĀRIKĀBHĀṢYA (≥550, r, Frauwallner (1953, p. 1, 287)), SUŚRUTASAṂHITĀ (first half of the first millenium AD, s, Meulenbeld (2000, IA, pp. 342–4)), SŪRYAŚATAKA (7. c., n, Lienhard (1984, p. 135)).

Time slot 4: Medieval literature (900–1400): This period witnesses the rise of sectarian Hinduism and its literature. The MIA dialects were probably replaced by new IA during this period ((Bloch, 1965, p. 24)). The dates of the alchemical texts (beginning with RASA- and RASENDRA-) are based on my own recent research (cf. Hellwig, 2009a)—AGNIPURĀṆA (800–1100?, n, Rocher (1986, pp. 136–7)), ĀNANDAKANDA (12./13. c., s, Meulenbeld (2000, IIA, p. 592)), ĀYURVEDADĪPIKĀ (11. c., s, Meulenbeld (2000, IIA, pp. 92–3)), ĀYURVEDARASĀYANA (1250, s, Meulenbeld (2000, IA,

p. 668)), Kālikāpurāṇa (10. c.?, n, Rocher (1986, p. 162)), Gītagovinda (1200, n, Lienhard (1984, p. 204)), Tantrasāra (11. c., r, Gonda (1977a, p. 212)), Nāṭyaśāstravivṛti (Abhinavagupta), Nibandhasaṃgraha (12. c., s, Meulenbeld (2000, IA, p. 378)), Mṛgendratantra, Mṛgendraṭīkā (900–1400, r, Gonda (1977a, pp. 183–5)), Rasaprakāśasudhākara (13./14. c., s, Hellwig (2009a)), Rasamañjarī (14./15. c., s, Hellwig (2009a)), Rasaratnasamuccaya (13./14. c., s, Hellwig (2009a)), Rasaratnākara (11./12. c., s, Hellwig (2009a)), Rasahṛdayatantra (10./11. c., s, Hellwig (2009a)), Rasādhyāya, Rasādhyāyaṭīkā (11./12. c., s, Hellwig (2009a)), Rasārṇava (10./11. c, s, Hellwig (2009a)), Rasendracintāmaṇi (13./14. c., s, Hellwig (2009a)), Rasendracūḍāmaṇi (12.–14. c., s, Hellwig (2009a)), Rasendrasārasaṃgraha (14./15. c., s, Hellwig (2009a)), Varāhapurāṇa (10.–12. c.?, n, Rocher (1986, p. 242)), Vetālapañcaviṃśatikā (≥10. c., n, Mylius (1983, p. 211)), Śārṅgadharasaṃhitā (14. c., s, Meulenbeld (2000, IIA, p. 207)), Śivapurāṇa (8.–14. c., n, Rocher (1986, p. 222ff.)), Śivasūtra (9. c., r, Gonda (1977a, p. 209)), Śivasūtravārtika (11. c., r, Gonda (1977a, pp. 209–10)), Sāṃkhyatattvakaumudī 850 AD, r, Frauwallner (1953, p. 1, 287)), Spandakārikā(nirṇaya) (9. c.?, r, Gonda (1977a, p. 210)), Hitopadeśa (9.–14. c., n, Mylius (1983, p. 206)).

Time slot 5: Late literature (≥1500): This slot coincides with the Mughal and the British rule in India—Agastīyaratnaparīkṣā (>13. c., s, Finot (1896, p. xi)), Gūḍhārthadīpikā (16./17. c., s, Meulenbeld (2000, IIA, p. 209)), Gokarṇap urāṇasāra (20. c., n, communication by K. Schier, Berlin)), Gorakṣaśataka (after 13. c., r, Gonda (1977a, p. 222)), Nāḍīparīkṣā (≥16. c., s, Meulenbeld (2000, IIA, p. 622)), Bhāvaprakāśa (1550, s, Meulenbeld (2000, IIA, p. 264)), Mugdhāvabodhinī (≥16. c., s, Meulenbeld (2000, IIA, p. 622)), Rasakāmadhenu (17. c., s, Meulenbeld (2000, IIA, p. 634)), Rasataraṅgiṇī (1923, s, Meulenbeld (2000, IIA, p. 698)), Rasaratnasamuccayaṭīkā, -dīpikā, -bodhinī (19./20. c., s, Meulenbeld (2000, IIA, pp. 671–2)), Skandapurāṇa (Revākhaṇḍa, ≈1450, n, communication by J. Neuss, Berlin)), Haṭhayogapradīpikā (after 13. c., r, Gonda (1977a, p. 222)), Haṃsadūta (15. c., n, Lienhard (1984, p. 125)).

1.2 Etymological trends in Sanskrit

The study is based on the digital corpus contained in the database of the program `SanskritTagger`. The dictionary of this program has been extended to store etymological information. If, for example, the Sanskrit lexeme *nīra* has been marked as Dravidian, the program is able to locate all occurrences of this lexeme in the text database (corpus) and calculate a distribution of this etymological unit. I used Mayrhofer's dictionary (1956–1976) as the sole source of etymological information. The work may have conceptual drawbacks, the gravest of which is Mayrhofer's reluctance to state his own opinion clearly in questionable cases (cf. Tedesco, 1960). Nevertheless, it is one of the best and most comprehensive etymological references for Sanskrit. Each text in the database is divided into contiguous '*text windows*' consisting of 3,000 lexical units. There are two reasons for splitting texts into text windows. First, large texts such as the epics or the Purāṇas frequently consist of several text layers that show different linguistic features. If we calculated a single proportion for each etymological class per text, the features contained in these layers would be leveled into one global variable. This approach obviously obscures the true linguistic composition of the text. Second, dividing large texts into smaller portions increases the amount of data available for statistical examination. Although the maxim 'the more, the better' is not valid without restrictions in statistics (see p. 5), a larger number of records strengthens the reliability of statistical results. Next, the absolute numbers of IA, MIA, Dravidian, and AA words in each window are transformed into percentage values by dividing them by the size of the window. In the following, these four rates (IA, MIA, Dravidian, and AA), combined with the basic information about the text from which the text window is extracted, are called a '*record*'. Table 1 shows the number of these records sorted by TSs and content types. Note that there are no records for the combination TS = 1 and CT = *s* or *n* and that the frequencies in the remaining cells vary strongly.

ANOVA answers the question of whether an independent variable or factor influences the

Table 1 Number of text records per slot and content type

	1	2	3	4	5
r	27	32	40	39	18
n		298	135	19	51
s		39	106	149	38

group means of a dependent variable. In our case, the TSs are the five levels of the independent variable 'time', while the rates of words from the four etymological classes constitute the dependent variables. We choose a comparatively high level of significance, $\alpha = 0.1$ or 10%, because the study is exploratory and it should not be too difficult to reject the null hypothesis. Of course, the result of an ANOVA is not a proof, especially not of causality. There may be other causal mechanisms that influence both the independent and the dependent variables, and there may be interactions between the variables that are not covered by the ANOVA. In addition, while the ANOVA evaluates the model time slot → etymology, interactions working in the opposite direction (etymology → time slot) are introduced by the 'impressionistic' mode of dating usual in Indology (see p. 16ff).

Since the ANOVA belongs to the class of parametric statistical techniques, important statistical measures of the data must conform to predefined mathematical distributions. If these preconditions are violated, the result of a parametric test may be invalid. In the case of ANOVA, we must check three conditions (Bortz 2005, pp. 284–7):

(1) The dependent variable(s) should be normally distributed in the basic population (i.e. the Sanskrit texts in the corpus). This condition can be assessed either by a graphical or a computational test. Visual inspection of the data using the `R` function `qqnorm` shows that the non-IA distributions are skewed due to many very small values. The visual impression is supported by a computational goodness-of-fit χ^2-test (see, e.g. Bortz, 2005, pp. 164–5).

(2) The variances of the dependent variable(s) must be equal for every factor (homoscedasticity). A Bartlett test demonstrates that this condition is violated for all etymological classes.

(3) Each sample should be assigned to a single (combination of) factor(s). This means that the classification of texts concerning their dates and content types must be unequivocal, which is fulfilled for our data.

Most researchers agree that the violation of the first two conditions has only an insignificant influence on the test result if the number of samples per factor level is equal. Therefore, we draw random samples from each level of the independent variable(s). The maximal number of samples per level is given by the respective cell with the lowest frequency in Table 1. In addition, we perform a nonparametric Kruskal–Wallis test to corroborate our results cf., e.g. Bortz and Lienert (2003, pp. 154–5) on the theoretical background of this test. Using only small, randomly selected subsamples of our data also increases the reliability of the test results. Statistical tests tend to yield more significant results when the size of the sample is increased. Therefore, optimal sample sizes should be chosen before starting a statistical test. The exact sample size depends on the size of the effect that is expected to occur in the data. Since we expect medium to large effects, sample sizes between 20 and 50 records should be appropriate to detect the expected effects in the group means (see, e.g. Bortz, 2005, pp. 258–9).

A final point before presenting the results: the tests only assess whether there are differences between the group means of the etymological composition. They do not indicate whether these differences follow a trend. Since we are interested in the temporal development of the etymological composition, which is indeed a trend, we perform pairwise comparisons of the group means if the ANOVA has yielded a significant result. We use a nonparametric test (Kruskal–Wallis test) because the etymological values per level suffer from similar distortions as the data for all texts.

Table 2 reports the results of the tests performed on different subsets of the data. Next to the abbreviation of the etymological class examined in each test, *F* gives the value of the test statistics of the ANOVA, *P* is the corresponding *P*-value, and

Table 2 Results of ANOVAs (F) and Kruskal–Wallis tests (χ^2) assessing the influences of time and content type on the etymological composition of Sanskrit texts

	ANOVAs				**Kruskal–Wallis test**	
	***F*-value**	***P*-value**	η^2	**Rate of words**	χ^2	***P*-value**
Differences between TSs 2–5 (content ignored)						
IA	27.493***	<0.001	0.302	$0.5863 \gg 0.5405 \gg 0.5178 \ll 0.5560$	69.523***	<0.001
MIA	5.729***	<0.001	0.085	$0.0045 \ll 0.0061 \leq 0.0070 > 0.0059$	14.585**	0.002
Drav.	10.073***	<0.001	0.138	$0.0073 < 0.0095 \ll 0.0145 \gg 0.0105$	25.562***	<0.001
AA	7.998***	<0.001	0.104	$0.0027 \ll 0.0042 < 0.0060 \geq 0.0053$	20.284***	<0.001
Differences between content types (TSs ignored)						
IA	14.476***	<0.001	0.148	$0.5721 \leq 0.5841 \gg 0.5272$	36.788***	<0.001
MIA	30.095***	<0.001	0.286	$0.0029 \ll 0.0049 \ll 0.0069$	49.114***	<0.001
Drav.	12.482***	<0.001	0.134	$0.0060 < 0.0074 \ll 0.0116$	22.761***	<0.001
AA	17.696***	<0.001	0.174	$0.0024 < 0.0031 \ll 0.0061$	32.924***	<0.001
Differences between TSs with fixed content types						
Religious						
IA	25.874***	<0.001	0.516	$0.6718 \gg 0.5736 > 0.5509 > 0.5135 \leq 0.5145$	50.614***	<0.001
MIA	4.863**	0.001	0.194	$0.0030 < 0.0042 \leq 0.0053 \gg 0.0019 \ll 0.0035$	20.189***	<0.001
Drav.	2.609*	0.041	0.092	$0.0050 \leq 0.0065 \gg 0.0037 < 0.0084 \geq 0.0080$	11.391*	0.023
AA	3.181*	0.017	0.141	$0.0006 < 0.0017 \leq 0.0018 \leq 0.0024 \leq 0.0035$	20.043***	<0.001
Narrative						
IA	3.296*	0.025	0.110	$0.5901 \gg 0.5548 < 0.5818 \geq 0.5675$	13.628**	0.003
MIA	3.018*	0.035	0.116	$0.0049 < 0.0060 \geq 0.0059 < 0.0073$	8.594*	0.035
Drav.	7.232***	<0.001	0.237	$0.0052 < 0.0082 \leq 0.0092 < 0.0123$	21.178***	<0.001
AA	1.576	0.203	0.063	$0.0026 \leq 0.0032 \leq 0.0033 \leq 0.0043$	9.069*	0.028
Scientific						
IA	8.345***	<0.001	0.138	$0.5556 \gg 0.5191 \geq 0.5117 < 0.5282$	24.486***	<0.001
MIA	4.590**	0.004	0.079	$0.0055 \leq 0.0058 \ll 0.0077 > 0.0064$	13.679**	0.003
Drav.	7.794***	<0.001	0.125	$0.0098 \ll 0.0145 < 0.0165 \gg 0.0115$	27.129***	<0.001
AA	7.124***	<0.001	0.114	$0.0037 \ll 0.0068 \leq 0.0068 \leq 0.0075$	23.995***	<0.001

the number of asterisks indicates the level of significance of the result (***: very significant, no asterisk: not significant for the given level of significance, $\alpha = 0.1$). Since P cannot be used to compare the results of several ANOVAs (Rietveld and von Hout, 2005, pp. 119–20), the value η^2 gives the effect size of each ANOVA (Bortz, 2005, pp. 259–60). Any η^2 above 0.1 can be considered to describe a strong effect. χ^2 and the following values record the respective information for the nonparametric Kruskal–Wallis test. The second line reports the results of pairwise sequential Kruskal–Wallis tests between the group means. The symbols between the group means encode the significance of one pairwise comparison ($\gg$, $\ll$: significant on the level $\alpha = 0.01$; $>$, $<$: significant on the level $\alpha = 0.1$; $\geq$, $\leq$: not significant on the level $\alpha = 0.1$). The results of some trend tests are, in addition, visualized in Fig. 2, where the z-standardized means of each etymological class are plotted against the TSs to make the development of the group means comparable. Strong line segments indicate that the t-test has yielded a significant result. To give an example of how to read Table 2: the third test describes the distribution of Dravidian words in TSs 2–5. Since both the parametric ($F = 10.073^{***}$) and the nonparametric tests ($\chi^2 = 25.562^{***}$) are highly significant, we can conclude that there are differences between the group means that cannot be explained by random factors at the given level of significance. $\eta^2 = 0.138$ indicates that the test has detected a rather strong effect. The second line gives a detailed image of the temporal development (refer Fig. 1 for a graphical representation). While the rate of Dravidian words in all texts increases significantly from the early to the medieval literature ($0.0073 < 0.0095 \ll 0.0145$), a clear decrease of this rate can be observed in late texts ($0.0145 \gg 0.0105$; see also Fig. 2, left side, line 'Drav.'). In general, the proportion of IA words decreases significantly from the first (late Vedic) to the fourth (medieval) TS and increases again in the late Sanskrit literature. The other three etymological groups show an opposite development (see also Fig. 2, left half). Their rates increase from the first to the fourth TS and decrease in the fifth one (or remain equal, as in the case of AA).

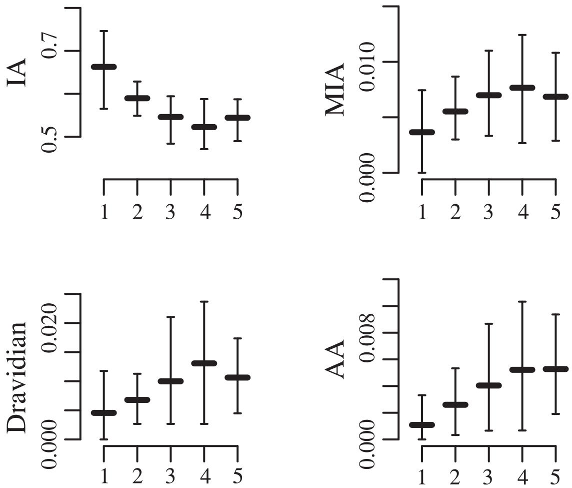

Fig. 1 Averages (strong horizontal lines) and 10–90% quantiles (vertical lines with whiskers) of etymological groups. The *x*-axis records the TS. Note that the y-axes have different scales

A second group of ANOVAs tests the influence of the content type on the etymological distributions (Table 2, second group of tests). The highest rates of non-IA words occur in scientific texts, with medium rates in narrative and low rates in religious texts. On the contrary, IA words dominate in narrative texts and have their lowest rates in the scientific texts. Since the influence of content on the etymological distribution is highly significant, the interaction between time and content is examined in detail. The lower part of Table 2 displays the results of three groups of one-way ANOVAs that were performed on records from texts with fixed content type.[4] By far, the largest effect can be observed for the IA words contained in religious texts (first group, first line, $\eta^2 = 0.516$), whose rates form a (not always significant) descending sequence. As can also be seen on the right half of Figure 2, the distribution of Dravidian and AA words in religious texts is less intelligible. After a clear increase from late Vedic to early Sanskrit literature, the rates of these words decrease in the classical literature, only to increase again in late Sanskrit.

Since the effects of time are especially obvious for religious texts, we should have a closer look at the changes that occur in the IA, Dravidian, and AA religious terminology (MIA terms are too rare to be

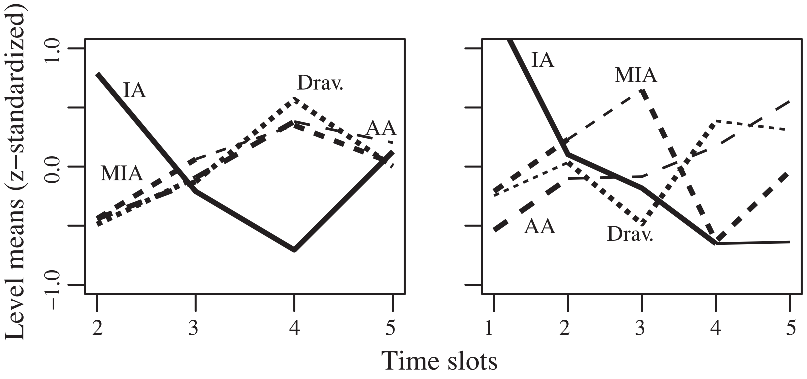

Fig. 2 z-standardized group means for all content types (left) and scientific texts (right); cf. p. 8

examined in detail). While we were working with text records up to now, we have to switch to word records for this detail study. First, frequency profiles for all words contained in religious texts are created, which look like

slot	1	2	3	4	5
frequency	0	1	15	180	30

This means that word x occurs once in the second slot, 15 times in the third slot, etc. To obtain words occurring in a certain period of time, we calculate a weighted mean μ_W of the frequency profile of each word and filter out the candidates by indicating the desired time range.[5] Note that the fact that a word is assigned to slot s_i does not mean that this word cannot be found in earlier or later texts. It only means that the word has its maximum relative frequency in s_i.

As can be expected from the general religious development in India, the IA vocabulary of old Brāhmaṇism is mainly confined to the first slot. This old stratum contains words such as *abhimṛś*, which is used in the domestic rituals of the Gṛhyasūtras, the name of the god Pūṣan, and the term *stoma*. References of these words are also found in later literature and would certainly have been multiplied if texts about ritualism (e.g. Mīmāṃsā) had been included in the corpus. However, their importance in genuinely religious texts diminishes rapidly after the late Vedic literature. The IA vocabulary of the remaining four slots is less clearly confined to single slots, but 'smears' over several periods in most cases. Nevertheless, it is still possible to trace the general religious development in these word records. Judging from its most typical words, religious texts from early Sanskrit literature are mainly concerned with questions of caste (*rājan* ('king' or 'Kṣatriya'), *vaiśya* and the duty of a person towards his ancestors *pitṛ* ('father'), *putra* ('son'), *ṛṇa* ('debt [towards one's ancestors]'), *pitrya* ('worship of deceased ancestors')). Especially, interesting is the distribution of the word *brāhmaṇa* as a caste name. The following table shows the absolute frequencies of *brāhmaṇa* and its relative frequencies per text record:

slot	1	2	3	4	5	1	2	3	4	5
r	125	**322**	57	39	7	4.81	**10.06**	1.43	1.00	0.39
n		1050	288	15	246		3.52	2.15	0.79	4.82
s		39	23	21	10		1.03	0.22	0.14	0.26

Although the high values of *brāhmaṇa* in the early religious literature (slot 2) may be biased because the *dharma* literature is *in toto* labeled as '*r*(eligious)', the fact that its relative frequency decreases rapidly after TS 2 may indicate a deep ('anti-Brahmanical'?) change in later religious literature. The classical religious literature discusses the structure of the human mind (*manas* ('mind'), *indriya* ('sense organ'), *buddhi* ('intellect')) and soteriological questions (*karman*, *sukha* ('joy'), *duḥkha* ('suffering')). This discussion, which is influenced by the classical systems of philosophy, has already begun in TS 2 and declines in importance soon after the classical period. While the IA vocabulary of the medieval religious literature testifies the rise of sectarian Hinduism (e.g. *mantra*, *bheda* ('difference [between ontological levels]'), *māyā*, *īśvara* ('supreme deity')), the picture changes again in the late religious texts, where words denoting body parts (e.g. *rasanā* ('tongue') and *pārṣṇi* ('heel')) and words for 'cleaning' (*kṣālay*, *dhautī*) indicate the growing importance of Yogic discipline (at least in my corpus!). Among the few Dravidian words occurring in religious texts from the TSS 1 and 2, only *ulūkhala* ('mortar') and *cīra* ('[an ascetic's dress made of] bark') have a clear religious connotation. *kuṭumba* ('family') pertains to socio-religious questions represented by words such as *pitṛ* in the IA vocabulary. In the classical period words such as *mālā* ('wreath'), *candana* ('sandalwood'), and even the popular *phala* ('fruit') are used for describing the new mode of religious worship. *pūjā*, the most important term for this kind of worship, is a good example for the close connection between etymology and theses about cultural history. As Mayrhofer points out, scholars tried to find 'an un-Aryan name . . . for the probably un-Aryan rite of devotion'

(Mayrhofer 1956–1976, p. II, 320), which may result in an unavoidable vicious circle. AA words become important only in the last two phases of religious literature. The etymologically unclear word *kañcuka* ('cover [separating man from his divine nature]') belongs to the medieval religious debate while words such as *cibuka* ('chin') and *kūrpara* ('elbow') come from the same religious stratum as the IA Yogic terminology.

The following table illustrates the development of the religious vocabulary in absolute numbers. It reports the absolute frequencies of distinct words from the etymological classes IA, Dravidian, and AA that have their maximal weighted means μ_W (cf. note 5) in the respective TS:

	1	2	3	4	5
IA	122	283	269	176	52
Drav.	1	9	4	10	3
AA	0	2	1	8	4

To assess whether the three etymological classes are distributed similarly over the TSs, we perform pairwise comparisons of the row data using the nonparametric Fisher–Yates test (a χ^2 test cannot be applied due to the low frequencies in the Dravidian and AA data; cf. Bortz and Lienert (2003, pp. 82–3)):

IA–Dravidian	$P = 0.046$*
IA–AA	$P < 0.001$***
Dravidian–AA	$P = 0.345$

The test results show that religious terms from Dravidian and AA have similar distributions, which differ significantly from the distribution of IA words. The test results, however, do not answer the question of whether Dravidian and AA words have been included in the Sanskrit vocabulary by the same socio-linguistic mechanisms, although they may support such a supposition.

After the general etymological development and its details in religious texts have been studied, we will finish the explorative part of the article by examining rates of lexical productivity.

The temporal development of etymological proportions does not cover the question of lexical productivity as, for instance, even in TSs with a low rate of IA words, many new IA words may have been introduced into Sanskrit. Therefore, we will determine the number of words that occur in slot i for the first time (n_i^*) and the number $n_i^\dagger$ of words that die out in this slot. Since n^* and $n^\dagger$ are influenced both by the number of words contained in the etymological class and by the number of records per slot, both values are transformed into relative frequencies per text record and then normalized to sum up to 1[6] (see Table 3 and Fig. 3 for the plotted relative values). The four plots in Fig. 3 show a similar general tendency: after comparatively high values in slot 2, the lexical productivity, i.e. the rate of new words introduced into the Sanskrit vocabulary, decreases sharply. It should, however, be noted that Dravidian and AA words show considerably higher values of n'^* in medieval and late Sanskrit than the IA and MIA words. This visual impression can be supported by performing pairwise Fisher–Yates tests that compare the rates of new words in slots 2–5. These tests yield the following P-values:

	MIA	Drav.	AA
IA	0.007**	$\ll 0.001$***	$\ll 0.001$***
MIA		0.234	0.112
Drav.			0.863

The trend found in the IA words differs clearly from the development of the other etymological classes (significant values in the first line), while the non-IA classes show highly similar patterns of temporal distribution (esp. the pair Dravidian–AA). These results may again be a support for the thesis that the different groups of non-IA words were incorporated into the Sanskrit vocabulary by similar linguistic and cultural mechanisms. At the same time, the relative rates and the statistical tests clarify Burrow's sketch of foreign influences on the Sanskrit vocabulary, which claimed that the inclusion of non-IA words almost came to an end in TS 2 (Burrow 1959, pp. 386).

Table 3 Source data for lexical productivity; see Figure 3

	Abs. frequencies						Rel. frequencies				
Etymology	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**
IA											
n^*		2934	667	186	63	n'^*		0.673	0.201	0.076	0.05
$n^\dagger$	145	792	1077	750		$n'^\dagger$	0.359	0.143	0.256	0.242	
MIA											
n^*		93	27	13	6	n'^*		0.54	0.206	0.134	0.12
$n^\dagger$	1	14	27	25		$n'^\dagger$	0.127	0.13	0.329	0.414	
Drav.											
n^*		185	84	40	11	n'^*		0.457	0.273	0.176	0.094
$n^\dagger$	0	19	70	85		$n'^\dagger$	0	0.072	0.35	0.577	
AA											
n^*		112	53	29	5	n'^*		0.447	0.278	0.206	0.069
$n^\dagger$	1	11	50	39		$n'^\dagger$	0.085	0.069	0.411	0.435	

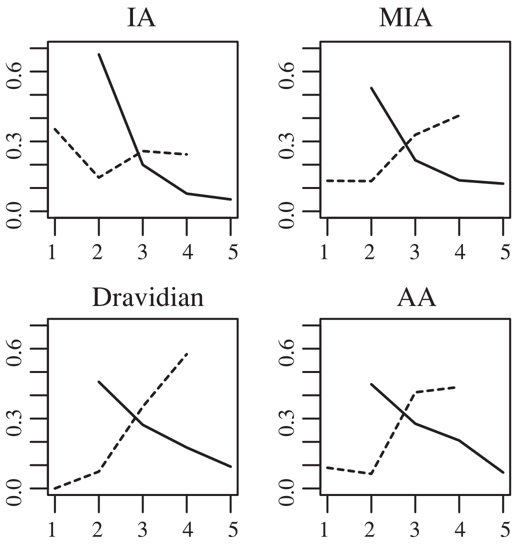

Fig. 3 Rates of lexical productivity in the Sanskrit vocabulary (data in Table 3)—Solid line = n'^*, dotted line = $n'^\dagger$

2 Dating Sanskrit Texts Automatically?

Now that the explorative part of the article has demonstrated that the etymological composition of Sanskrit texts is influenced significantly by its time of origin (and its content); we will examine whether it is possible to assign a text to a certain TS when only its etymological composition is known. Developing such a method is not merely a vain exercise in statistics. On the contrary, its practical importance for Sanskrit philology becomes obvious as soon as we consider the numerous unclear points in the history of Sanskrit literature and the frequently inadequate philological methods available to tackle these problems. We will compare the performance of two multivariate methods to date Sanskrit texts on the basis of their vocabulary. The first technique is linear discriminant analysis (LDA), which is, for instance, used for authorship identification in Juola and Baayen (2005).[7] LDA is capable of separating classes by linear functions. In some cases, it may, however, be more promising to separate the classes by using a nonlinear function. Since some neural networks (NNs) are able to approximate such nonlinear functions (see, e.g. Bishop, 2000, p. 130ff.), we use a simple two-layer NN with three hidden units and a sigmoidal activation function as the second classifier.

Table 4 reports the accumulated results of twenty runs of the two classification methods. In each run, both classifiers were trained on the same random subset of 50% of the text records and tested using the remaining half. In general, the NN performs better than the LDA. This impression can be corroborated by a statistical test. When an LDA model and an NN are applied repeatedly to the

Table 4 Accumulated classification results of twenty repetitions of LDA (left) and NN (right)

	1	2	3	4	5	μ_{ri}	1	2	3	4	5	μ_{ri}
1	**175**	58	10	7	1	0.70	**126**	96	23	6	0	0.50
2	345	**2516**	398	164	259	0.68	24	**2882**	613	140	23	0.78
3	48	684	**818**	696	517	0.30	0	742	**1331**	664	26	0.48
4	16	303	391	**1156**	229	0.55	0	325	687	**1073**	10	0.51
5	15	249	247	237	**341**	0.31	10	308	529	215	**27**	0.02
μ_{ri}	0.29	0.66	0.44	0.51	0.25		0.79	0.66	0.42	0.51	0.31	

The rows give the true TS of the records and the columns the result of the classification. Correct classifications are printed in bold characters.

same random subsets of the text records, we get an estimation of their average classification accuracies μ_r. Twenty repetitions resulted in $\mu_r(\text{LDA}) = 50.7\%$ and $\mu_r(\text{NN}) = 55.1\%$. A *t*-test that compares the two sets of twenty recognition rates with significance assigned at the 10% level yields $P \ll 0.001$ and, therefore, a highly significant result. Two points need further clarification. First, recognition results between 50 and 55% may not seem noteworthy for researchers working in the area of pattern recognition.[8] A look at Table 4, however, reveals that most of the misclassified cases are assigned to a slot directly neighboring the true slot of the text record. In these cases, the classification results are estimations of the true TS, since the TSs can be interpreted as a pseudo-cardinal scale. Second, classification quality differs strongly among both TSs (rows) and classifiers (columns). While the best results are achieved for the first four TSs, the recognition rate decreases in the fifth slot and for the fifth classifier. The main reason for this phenomenon can be observed in Figure 1. In the first four slots, the etymological classes develop with high rates, which results in clearly distinguishable group means and, therefore, good classifications. On the contrary, the etymological composition of the fifth slot can hardly be distinguished from that of the preceding slots, which causes the majority of the misclassifications. In addition, the NN performs especially badly in the fifth slot. Since NNs need many examples to approximate their nonlinear discriminative function, the bad performance in the fifth slot may also be due to the low number of training records in this slot.

3 Discussion and Perspectives

The article has shown that considerable development has occurred in the vocabulary of post-Vedic Sanskrit. Three linguistic results should be noted:

(1) While the rate of IA words decreases significantly until the end of the medieval period, it increases again in the last TS, which covers the period from 1500 AD to the present day. The rates of MIA, Dravidian, and AA words show an opposite development. The etymological rates in the last slot may be interpreted as a kind of purification of the Sanskrit vocabulary as described, for instance, in Deshpande (1979, p. 73).[9]

(2) The etymological development of Sanskrit was not complete in the centuries BC Instead, the language continued to assimilate material from Dravidian and AA until the end of the medieval period.

(3) Central distributional patterns of non-IA word classes and especially of Dravidian and AA closely resemble each other. This fact may be a remarkably clear indication either of similar socio-linguistic mechanisms by which these words were included into Sanskrit or of some modern scientific axioms that assume such mechanisms.

The etymological change has, of course, been mentioned before by scholars such as Burrow and Emeneau. However, its existence has not yet been demonstrated on the basis of a large Sanskrit corpus and with statistical methods. The second point deserves special attention since the traditional philological approach tends to mention the occurrence of linguistic phenomena without specifying their exact temporal distribution and significance. Supporting philological argumentation with simple statistical tests may help to avoid this purely qualitative approach, which is certainly responsible for the 'fuzziness' of some philological theories.

Concerning the details of the etymological development, I do not claim that the article has drawn a final picture of the change of etymological proportions and the rates of lexical productivity. First, the corpus is neither comprehensive nor well balanced. While there are plenty of scientific and early narrative texts, it contains only a few late religious and narrative texts. Second, the dates of many texts used for the study cannot be fixed at the moment, which directly influences the quality of the (statistical) results. Errors-in-variables models as described, for example, in Fuller (1987) may help to circumvent this problem. Third, many etymological attributions in Mayrhofer's dictionary are unsafe, which again influences the etymological distributions. A further problem is the great number of words that Mayrhofer labels as 'unclear'. Future etymological research will certainly be able to assign many of these cases to one of the non-IA classes, thereby changing the etymological rates on which the study is built. A less obvious source of errors is hidden in the general presuppositions of etymological research, which claim that unclear words should be assigned to a non-IA class if they occur only in supposedly late texts. However, since texts are not infrequently dated based on an impression of their vocabulary (obscure words → late text), this presupposition may result in a vicious circle that is hardly resolvable with the current state of Indology.

The second result of the article concerns the automatic dating of texts. First, I must emphasize that all precautions brought forward against the explorative part of the article apply to its second part to an even higher degree. In addition, the time scale used in the study is much too coarse. The result that an undated text may come from the time interval 500 BC–300 AD is certainly not very detailed and may also be found with traditional philological methods. Nevertheless, the dating algorithms open up interesting perspectives for future research since they show that it is possible in principle to date Sanskrit texts on the basis of their linguistic features. The points at which the dating algorithm fail indicate the direction of further work. First, Indology should try to detect other linguistic features that can be used for dating. Such features may be helpful to distinguish between texts from TSs with similar etymological distributions (slots 3–5 in this article). Second, a larger and more diversified corpus is needed to increase the number of TSs and the accuracy of the dating procedure. Third, more sophisticated mathematical models for evaluating large numbers of (disparate) features have to be developed. Such models that integrate philological information from different sources into one global dating should, however, be the last stone in building a statistical philology of Sanskrit.

References

Bishop, C. M. (2000). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bloch, J. (1965). *Indo-Aryan from the Vedas to Modern Times*. Paris: Librarie d'Amérique et d'Orient.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*, 6th edn. Heidelberg: Springer Medizin.

Bortz, J. and Lienert, G. (2003). *Kurzgefasste Statistik für die klinische Forschung*. Heidelberg: Springer Medizin.

Brockington, J. (1984). *Righteous Rāma. The Evolution of an Epic*. Delhi: Oxford University Press.

Burrow, T. (1959). *The Sanskrit Language*. London: Faber and Faber.

Deshpande, M. M. (1979). *Sociolinguistic Attitudes in India*. Ann Arbor: Karoma Publishers.

Emeneau, M. (1980). Dravidian and Indo-Aryan: the Indian linguistic area. In Language and Linguistic Area. Essays by Murray B. Emeneau. Stanford: Stanford University Press, pp. 167–196.

Emeneau, M. B. (1954). Linguistic prehistory of India. *Proceedings of the American Philosophical Society*, **98**(4): 282–92.

Emeneau, M. B. (1956). India as a lingustic area. *Language*, **32**(1): 3–16.

Finot, L. (1896). *Les lapidaires indiens*. Paris: Émile Bouillon.

Fosse, L. M. (1997). *The Crux of Chronology in Sanskrit Literature*. Oslo: Scandinavian University Press.

Frauwallner, E. (1953). *Geschichte der indischen Philosophie*. Salzburg: Otto Müller.

Fuller, W. F. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Gonda, J. (1977a). *Medieval Religious Literature in Sanskrit, A History of Indian Literature*, **Vol. 2**, Fasc. 1, Wiesbaden: Otto Harrassowitz.

Gonda, J. (1977b). *The Ritual Sūtras. A History of Indian Literature*, **Vol. 1**, Fasc. 2, Otto Harrassowitz.

Hellwig, O. (2009a). A chronometric approach to Indian alchemical literature. *Literary and Linguistic Computing* Advance Access published on January 8, 2009; doi:10.1093/llc/fqn043.

Hellwig, O. (2009b). SanskritTagger, A Stochastic Lexical and POS Tagger for Sanskrit. In Huet, G. and Kulkarni, A. (eds), *Sanskrit Computational Linguistics. First and Second International Symposia'. Lecture Notes in Artificial Intelligence*. 5402. Berlin: Springer, pp. 266–277.

Hulin, M. (1978). *Sāṃkhya Literature. A History of Indian Literature*, **Vol. VI**, Fasc. 3, Wiesbaden: Otto Harrassowitz.

Johnston, E. (ed.) (1936). *The Buddhacarita or, Acts of the Buddha*. Lahore: University of the Punjab.

Juola, P. (2008). Killer applications in digital humanities. *Literary and Linguistic Computing*, **23**(1): 73–83.

Juola, P. and Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, **20**: 59–67.

Lienhard, S. (1984). *A History of Classical Poetry: Sanskrit, Pali, Prakrit. A History of Indian Literature*, **Vol. 3**, Fasc. 1, Wiesbaden: Otto Harrassowitz.

Masica, C. P. (1979). Aryan and non-Aryan elements in North Indian agriculture. In Deshpande, M. M. and Hook, P. E. (eds), *Aryan and non-Aryan in India*. Ann Arbor: The University of Michigan, pp. 55–151.

Matilal, B. K. (1977). *Nyāya-Vaiśeṣika. A History of Indian Literature*, **Vol. VI**, Fasc. 2, Wiesbaden: Otto Harrassowitz.

Mayrhofer, M. (1956-1976). *Kurzgefaßtes etymologisches Wörterbuch des Altindischen*. Heidelberg: Carl Winter Universitätsverlag.

Meulenbeld, G. J. (2000). *A History of Indian Medical Literature*. Groningen: Groningen Oriental studies, Egbert Forsten.

Mylius, K. (1983). *Geschichte der Literatur im alten Indien*. Leipzig: Reclam.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Wien: R Foundation for Statistical Computing.

Rietveld, T. and von Hout, R. (2005). *Statistics in Language Research: Analysis of Variance*. Mouton de Gruyter.

Rocher, L. (1986). *The Purāṇas. A History of Indian Literature*, **Vol. II**, Fasc. 3, Wiesbaden: Otto Harrassowitz.

Ruben, W. (1966). *Die Nyāyasūtra's. Text, Übersetzung, Erläuterung und Glossar*. Nendeln: Kraus Reprint Ltd.

Salomon, R. (1989). Linguistic variability in post-Vedic Sanskrit. In Caillat, C. (ed.), *Dialectes dans les littératures Indo-aryennes*. Paris: Collège de France, pp. 275–94.

Tedesco, P. (1960). Notes to Mayrhofer's etymological Sanskrit dictionary. *Journal of the American Oriental Society*, **80**(4): 360–6.

von Hinüber, O. (1986). *Das ältere Mittelindisch im Überblick*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Winternitz, M. (1968). *Geschichte der indischen Literatur*, K.F. Koehler.

Witzel, M. (1995). Early Indian history: linguistic and textual parametres. In Erdosy, G. (ed.), *The Indo-Aryans of Ancient South Asia. Language, Material Culture and Ethnicity*, **Vol. 1**, Berlin, New York: Walter de Gruyter, pp. 85–125.

Zide, A. R. K. and Zide, N. H. (1976). Proto-Munda cultural vocabulary: Evidence for early agriculture. *Oceanic Linguistics, Special Publications,* Austroasiatic Studies Part II (13), pp. 1295–334.

Notes

[1] In the following, the term 'Sanskrit' will be used in the restricted meaning of 'post-Vedic Sanskrit'. The statistical calculations were performed with R (R Development Core Team, 2007).

[2] The terms IA, MIA, and AA do not fully correspond to Mayrhofer's uses. IA words comprise Mayrhofer's words with traceable Indo-European (e.g. *dyaus*) or

Aryan etymologies (e.g. *duh*) and words that are clearly of Indo-European origin, but are not accompanied by comparative material in Mayrhofer (1956–1976) (e.g. *nagnatā*). MIA words comprise words directly derived from a MIA language and Hypersanskritisms. The class AA contains generic AA words and those from a Munda language.

[3] Exact dates would, of course, be preferable; cf. Fosse (1997, pp. 154–6) for an introduction into this problematic area.

[4] Similar results can be found when a single two-way ANOVA is performed on the data. Since the results of such an ANOVA, and especially the interactions between the factors, are frequently hard to interpret, we use several one-way ANOVAs with fixed factors instead.

[5] If m_i is the number of text records in slot i, μ_{Wi} is calculated by dividing the sum of relative frequencies multiplied with slot numbers by the sum of relative frequencies:

$$\mu_W = \frac{1}{\sum_{i=1}^{5} \frac{f_i}{m_i}} \cdot \sum_{i=1}^{5} i \cdot \frac{f_i}{m_i}.$$

This results in values between 1 (occurs only in the first slot) and 5 (occurs only in the last slot).

[6] If m_i is the number of text records in slot i, the plotted values n' are given as

$$n'^{*}_{i} = \frac{1}{\sum_{j=1}^{5} \frac{n^*_j}{m_j}} \frac{n^*_i}{m_i}.$$

[7] To be exact, classification is performed using the result of a LDA. This approach resembles the 'quadratic classification function' (cf. Bortz (2005, pp. 618–24) and the open source code of the `R` package `MASS` (`lda.R`)). The prior probabilities of each TS are set to 1/5 to counteract the bias caused by the unbalanced temporal composition of the corpus.

[8] The test described in Bortz (2005, p. 625) shows that the influence of random effects on the hit rates can be neglected in our experiments. We expect $e = 1976$ cases of random hits for the LDA, and observe $o = 5008$ cases of correct classification out of a total of $N = 9880$ cases. Using formula 18.45 in Bortz (2005, p. 625), we calculate $z_{LDA} = (o - e) \cdot \sqrt{N} / \sqrt{e \cdot (N - e)} = 76.209$ and $z_{NN} = 87.099$. Both values are highly significant.

[9] Thanks to J. Houben for calling my attention to this publication.