Lexical Diversity in a Literary Genre: A Corpus Study of the *Ṛgveda*

Alexandre Sotov
St. Petersburg State University

Abstract

This research[1] evaluates the extent to which lexical diversity, measured by frequent content words, *hapax legomena*, and type-token ratios (TTRs), is dependent on three features of the genre of the oral Indo-Aryan cultic poetry represented by the literary corpus of the *Ṛgveda* (ca. 165,000 tokens): characteristic choice of subject matter, usage of refrains, and the attribution of hymns to distinct poetic collectives. Analysis of 255 texts of 200 tokens showed that hymns on popular topics and where refrains were attested have a significantly higher rate of high-frequency content words and a lower ratio of once-occurring types. A higher TTR is observed in the hymns of specific family origin. Complexity of genre can be interpreted as a result of different discourse strategies of the poets. Overall, conservative mythological texts are characterized by regularity in word usage. Occurrence of content words, in the entire corpus, with lexemes denoting 'deities' on the one side and 'nature' on the other is accounted for by the factor of semantics, which deals with the structure of narrative.

Correspondence:
Alexandre Sotov
St. Petersburg State
University, St. Petersburg,
Russia
E-mail:
a.sotov@yahoo.co.uk

1 Introduction

Addressed mostly to various deities of the Vedic pantheon, the hymns of the *Ṛgveda* (*ṚV*) were created by ancient Indian poets and priests (*kaví*-, *ṛ́ṣi*-, *brahmán*-) in the condition of orality (Ong, 1982) during a lengthy period. They were put together as a *saṃhitā*, 'an arranged collection of texts', in ca. 1000–600 BC (Gonda 1975, p. 15; Deshpande, 1993, p. 134; Witzel, 1997, p. 263). The central, yet not infrequently implied rather than articulated, topic of the hymns is cosmogony, a set of concepts on the origin of the Universe and on the place of human beings in it. Indo-Aryans believed in the magical force of a skilfully pronounced word and revered their *saṃhitās* as the divine revelation (*śrúti*-). They entrusted the privilege of accessing their sacred knowledge (*veda*-) to seers, whose prestige relied on victories in ritualized verbal contests (Kuiper, 1960; Thompson, 1997, p. 20). The collection of 'praises' (*ṛc*-) reflects the dominant, institutionalized discourse of the Vedic epoch (Smith, 1994). The *ṚV* can be viewed as a diachronic genre-specific corpus representing the ideology of its creators. Following Biber (1995), 'genre' in this article corresponds to a named and culturally distinguished category of texts (p. 9), while 'discourse' is understood as a linguistically mediated 'technology of knowledge' (Duranti, 1997, p. 12) in its particular historic conditionality, i.e. 'conventional ways of talking, . . . which embody . . . social values and views of the world' (Stubbs, 1996, p. 158). The term 'narrative' is used broadly to identify the ways discourse is habitually structured into separate texts.

doi:10.1093/llc/fqn044 Advance Access published on 12 January 2009

In literature studies interest in the genre of the *ṚV* is reflected in various approaches to the structure and style of hymns, which typically observe the regular occurrence of certain textual features, i.e. 'formal devices . . . all of which would help shape the attentive listening of an audience of poetic adepts' (Jamison, 2004, p. 247). It is useful to distinguish aesthetic, functional and descriptive methods of analysis.

In the aesthetic approach to the genre of the hymns style emerges as the creative mastery of figures of speech. Codified in normative poetics, they presuppose a specific mode of perception: the recognition and evaluation of these figures. In Indian context, the tradition of aesthetic stylistics is rooted in a number of treatises referred to as *alaṃkāraśāstra*, theory of poetic devices (*alaṃkāras*). Indian poetology is traced back by some scholars to the epoch of the *ṛ́ṣis* (Kane, 1951), while others put emphasis on an inherently receptive aspect of the Vedic poetry. It is often argued that its 'main purpose' was 'to give delight, to both the deity addressed and the listener in general' (Mainkar, 1966, p. 3). This intention, along with the belief in the power of the spoken word as a means of attaining human goals, formed a poetical practice to which different kinds of repetition were central.

According to Gonda (1959), a stylistic study of the Veda should not be limited to poetic devices. It is necessary to establish the interrelationship between stylistic features and the production of texts in an archaic oral tradition. Since 'Vedic literature, and especially the mantras, is highly formulaic, and increasingly conventional, in character' (p. 155), the use of stereotypic expressions and repetitions in it is constructive, not simply ornamental. Gonda's interest in stylistic repetition laid a foundation for the functional method. Elizarenkova analyses the *ṚV* explicitly in relation to the purpose of the poetic message (Elizarenkova, 1995, p. 9). Vocabulary, metrical form, 'sound-picturing' and syntax are kinds of 'supra-text' (p. 121) employed in a laudation as channels of transmitting information to the addressee, a Vedic deity. The communicative factor is said to structure the texts to such a degree of conformity that a standard hymn model is introduced (p. 9). The formal devices, studied by Jakobson (1960) within his theory of self-orientation of the poetical language, acquired in Elizarenkova's functional analysis a communicative purpose: the hymns dealt with the situation of a gift exchange between the *ṛ́ṣi* and the divinity (Elizarenkova, 1995, p. 3–4; 1997, p. 49).

Similarly, Watkins (1995) treats the interrelation of the technique and the purpose of literary creativity 'in the Indo-European times' in respect of the function of the Indo-European poet, 'the custodian and the transmitter' of the tradition (p. 68).

> . . . [T]he art of the Indo-European poet is to say something wholly traditional in a new and interesting, but therefore more effective way. It is verbal activity . . . directed toward a . . . concrete goal . . . [F]ormulas are the vehicles of themes and . . . in the totality of these we find the doctrine, ideology, and culture of the Indo-Europeans.

Watkins maintains an indiscriminate view of stylistics and poetics, referring by it to 'all the linguistic devices which in Jakobson's phrase are ''what makes a verbal message a work of art'' '(p. 21). Elizarenkova and Watkins understand poetics as a system of textual features and presuppose the Saussurian dichotomy of *langue* and *parole* (cf. Elizarenkova, 1995, p. 2 ff) between the reconstructed poetics and its specific, individual realization in hymns, which are viewed as the functional whole.

Unlike the functional method, the descriptive approach is aimed at the analysis of linguistic features irrespective of their purpose. Bloomfield *et al.* (1934) describe flexion of nouns, pronouns, and adjectives in recurring *mantras* in terms of formal, syntactic, and, notably, stylistic variants. The latter is an *a posteriori* category; it includes variants, defined as follows (p. 20).

> . . . [They] have no real relation to the syntactic uses of the variant morphemes, but . . . [their] interest consists in the light they throw on the process of Vedic tradition. They illuminate the ways in which the whole stock of mantra material was reworked in the course of centuries, but do not illustrate points of Vedic grammar.

'Based entirely upon linguistic and stylistic evidence' (Fosse, 1997, p. 42), a study by Wüst (1928) follows the descriptive method, although it deals primarily with the internal chronology of the hymns rather than their genre and structure. Quantitative data collected by Wüst is descriptive: the distribution of countable formal features in the books of the *ṚV* is an important empirical characteristic of the corpus. Yet it is meaningless without a sound category of comparison, whereas his factor of 'stylistic history' leads astray (Gonda, 1959, p. 12). The quest for the internal chronology of the corpus might altogether be a folly when it comes to establishing a cause of linguistic variation, for, indeed, 'it has long been recognised that a time difference that correlates with linguistic differences does not in itself explain the linguistic difference' (Jucker, 1992, p. 24).

Aesthetic, functional, and descriptive methods of analysis differ as much as the researchers' objectives are different. On the one side is the investigation of the figures of speech and the reconstruction of the 'mytho-poetic model of the universe', as Toporov would put it (1992, p. 161), of which the *ṚV* 'presents one variant' (Elizarenkova, 1995, p. 10); on the other is the description of textual regularities which 'do not illustrate points of Vedic grammar' (Bloomfield *et al.*, 1934, p. 20) or allegedly deal with language change (Wüst, 1928). Criteria for selecting features for analysis are also different. An *a priori* knowledge of poetic devices, *alaṃkāras*, may or may not be required for a stylistic investigation of genre from the aesthetic perspective. In the latter instance an empirical criterion of poeticity is the repetition of linguistic features. Functional analysis on the ground of Jakobsonian linguistic poetics empirically relies on repetitions (phonetic, syntactical, or lexical), although an *a priori* non-linguistic factor is presupposed. The genre of cultic poetry, its communicative nature or the 'grammar of poetics' are categories that cover the entire *saṃhitā*, not really its parts. In contrast, extensive registers in Bloomfield *et al.* (1934) give an idea of stylistic variation in similar contexts after 'purely grammatical' instances are exhausted, but the question of factors that influenced the diversity is not explicitly raised.

However, the method of corpus-based lexical description is well established in Vedology, a discipline relying almost exclusively on collections of fixed texts. It is consistently employed in the monumental dictionary of Grassman ([1873] 1964), and in the concordances of Bloomfield (1906) and Lubotsky (1997). Corpus approach to lexical analysis suggests the possibility of an empirical study of the Vedic lexis (and semantics) in relation to the genre of the corpus and the regularity of discourse it presumably carried. This agenda is encouraged by the tradition of quantitative analysis in Indology (Fosse, 1997). Various attempts in the fields of corpus semantics and discourse analysis emphasize the importance of genre studies, especially with regard to ideology, rhetoric, style, and poetics (Stubbs, 1996, 2001; Biber and Conrad, 2001), while lexical representation of social discourse attracts the persistent attention of linguistic anthropologists (Duranti, 1997). This makes the task of describing lexis in a speech event specific, socially constructed collection of texts particularly relevant.

2 Purpose of the Study

The purpose of this article is to provide empirical evidence of patterns of word usage in the *ṚV* in support of the thesis of the complexity and heterogeneity of its genre. The research questions are as follows: (1) is there a dependency between lexical diversity and a typology of hymns? (2) Do key concepts of mythology relate to each other in a regular and consistent manner in transmitting the subject matter of the hymns? Both questions are viewed in the present research as a part of the problem of how lexical choice by the Vedic rhapsodes was influenced by their conventions and ideology.

3 Method and Data

3.1 Data description

Vedic tradition developed several techniques of invariable oral transmission of its literary heritage, two of which provided data for the present study: the linguistic analysis of *padapāṭha*, the modification of the continuous text (*saṃhitā*) which split it up into constituent word forms, and the *anukramaṇī* indexes. Since the type-token distinction is critical for a corpus study, only *padapāṭha* texts, as they appear in the computer-readable database of the *Śākala Śākhā Ṛ̣V* (Gippert, 2000), were taken into account.[2] An occurrence of a specific word-string in the word frequency list of the corpus is hereafter understood as the word 'type', while 'token' refers to its instance in a particular *padapāṭha* context. According to the word frequency list of the *Ṛ̣V*, compiled automatically with the help of concordance software, the collection of hymns ($N = 1028$) counts 164,757 tokens and 29,199 types. The mean length of text is 160.3 tokens. Rare and frequent type lists were taken from the word frequency list in order to measure lexical diversity as will be clarified in the next section. Word frequency list, collocations, and word clusters were extracted from the entire corpus, while the analyses of lexical diversity were based on a sample of 255 texts.

The *anukramaṇī* indexes identify deities, poetic meters, number of verses, and the names of the poets for hymns and their parts (Gonda, 1975, p. 34 ff). The *Sarvānukramaṇī* is based on anterior indexes, 'embodying the substance of all of them' (MacDonell, 1886, p. VIII), and contains this information for every hymn of the available recension, the *Śākala Śākhā Ṛ̣V*. The emergence of the indexes is attributed to the codification of hymns in a canon; although authorship statements in *Sarvānukramaṇī* are considered 'quasi-historical' (Bloomfield, 1916, p. 634), its definition of the hymns' subject-matter is used in this study as early evidence of their perception within the tradition. It has been noticed that the *anukramaṇī* characteristics partly correspond to the arrangement of hymns in the recension (Witzel, 1997, p. 261; Bryant, 2001, p. 66). This is especially true for the so-called family books (*maṇḍalas* 2–7), the exclusive collections of certain priestly clans, which are viewed by scholars, following Oldenberg (1888), as the core (Gonda, 1975, p. 9).

3.2 Data analysis

Lexical diversity was described as three variables [type-token ratio (TTR), frequent content words ratio (FCWR), *hapax* ratio (HR)] which were compared with three categorical factors: attestation of repeated word clusters, position of a hymn in the collection, and its topic (C1, C2, C3). Since the measures are sensitive to the lengths of texts, they were calculated for each hymn on the basis of the first 200 tokens only. Shorter hymns were excluded.

The TTR is the 'average number of tokens per types' (Baker, 2006, p. 52):

$$\mathrm{TTR} = \frac{\text{Number of types} \times 100\%}{\text{Total number of tokens}} \qquad (1)$$

It shows how inclined the authors of the hymns were to repeat the same words in their poetry. Generally, the higher the TTR, the more varied is the verbal repertoire. Lower TTR may indicate 'a high degree or repetition and reduced complexity of a text' (Scharl, 2004, p. 28), as well as the use of 'standardised terminology' (p. 29).

The FCWR is the percentage of frequent content word tokens:

$$\mathrm{FCWR} = \frac{\text{Number of frequent content tokens} \times 100\%}{\text{Total number of tokens}} \qquad (2)$$

Frequent types, i.e. words attested one hundred and more times in the corpus, occurring twice as often than any others in each hymn, correspond to 0.5% of the word frequency list of the *Ṛ̣V*. Lexemes which contributed to the largest number of frequent types are function words, verbs of motion and existence, names of the deities (*índra*-, *sóma*-, *agní*-; high-frequent are forms of V., N., A., G., I.), as well as *áp*- and *dyú*-, the deified Waters and Heaven (V., N., G. and V., N., A., respectively), and, not surprisingly, the word *deva*-, 'a deity, god' (N., A., G, L., V.). Consisting of seventy-eight word types (nouns, verbs, adjectives, adverbs, and numerals), the frequent content word list corresponds primarily to the vocabulary of the typical formulaic expressions of the *ṛ́ṣis* and other common means used in order to talk about conventional topics: the gods and nature.

The HR is the ratio in percent between once-occurring types (*hapax legomena*) and the vocabulary size:

$$\mathrm{HR} = \frac{\text{Number of once-occurring types} \times 100\%}{\text{Total number of types}} \tag{3}$$

Hapaxes are words, often of obscure meaning (Gonda, 1971, p. 172–73), which were used in the corpus only once. They comprise 56% of the word frequency list of the *ṚV*. Maximum relative number of such types, 40%, was found in hymn 10.163, which is practically repeated in the *Atharvaveda*, the book of magical chants, as its purpose was to charm away illness. The hymn contains some 'medical' terminology, such as six out of nine Ṛgvedic attestations of yákṣmam, from *yákṣma-*, 'a name of a disease', vṛhāmi, from $\sqrt{}$*vṛh*, $\sqrt{}$*bṛh*, 'to pull out', etc. Hymn 1.162, which praises a sacrificial horse and where up to 25% of types are once-occurring, gives an illustration of yet another source of unique lexis: craftsmanship terminology (i.e. páḍbīśam and saṃdá ä nam, from *páḍbīśa-*, 'a fetterlock', and *Saṃdā́na-*, 'a bond'), which is rare in the collection of priestly hymns. Although *hapax* frequency was introduced by Wüst as a sign of lateness (Wüst, 1928, p. 14), a high value of HR essentially indicates the originality of the hymn's topic or a situation it presumably dealt with (Edgerton, 1929, p. 278). It may generally imply the use in a text of 'highly characteristic and specialised vocabulary suited to . . . subject matter' (Gamberini, 1983, p. 450) or 'higher level of exactness of message content' (Fox and Fox, 2004, p. 117).

Family seal (C1) is a dichotomous category, which presents an occurrence (or absence) in a hymn of a repeated text fragment, defined as a cluster of five words attested in the corpus at least twice. Word clusters were viewed in the present study as empirical traces of the family seals of the *ṛ́ṣis*, refrains and catch-phrases of varying length which laid down the claims of certain clans or 'corporations', *sakha-* (lit. 'a friend'), of poets (Elizarenkova, 1999, p. 472) to certain books of the collection. 'These clans were not willing to part with their ancestral and secret knowledge. They indicate their "copyright" by a "clan seal": refrains, poets' own names, openly or disguised' (Witzel, 1997, p. 261). The most popular of such refrains is the phrase 'yūyám pāta svastíbhiḥ sádā naḥ' ('protect us always with your blessings'), which is attested in eighty-two verses, mostly in book 7 (hymns 7.1, 7.3–7.4, 7.7–7.9, 7.11–7.14, 7.19–7.30, 7.34–7.37, 7.39–7.43, 7.45–7.48, 7.51, 7.53, 7.54, 7.56–7.58, 7.60–7.65, 7.67–7.73, 7.75–7.80, 7.84–7.88, 7.90–7.93, 7.95, 7.97–7.101, 9.90, 9.97, 10.65–10.66, 10.122); an inventory of repetitions is given by Bloomfield (1916).

Maṇḍala (C2) characterizes the position of a hymn in the collection and describes it as belonging either to the family books (2–7) or the others (1, 8–10). The recension is traditionally divided into ten books, or 'circles' (*maṇḍala-*). The hymns of the family books were created by 'poets of the same family which handed them down as its heritage' (Gonda, 1975, p. 9). Book 9 contains only hymns to Soma and is 'ascribed to more than sixty poets' (p. 11), while books 1, 8, and 10 'were not composed each by a distinct family of *ṛ́ṣis*, but consist of groups of hymns based on identity of authorship' (p. 10). Hymn placement category is non-linguistic and is derived from the traditional arrangement of the corpus; it can arguably be described as the time factor, since the family books contain the oldest texts. But 'all we can say with confidence is that book 10, as such, is late but judgment must be exercised for each individual hymn. Some in book 8, sometimes even in book 1 and 10, can be as early as the "family books" '(Witzel, 1995, p. 310). In some cases analysis results for individual books will be given.

Topic (C3) is a dichotomous category which deals with the subject matter of the collection according to the traditional definition. About one-half of the hymns are dedicated either to Agni, Indra, or Soma, the most frequently mentioned gods. Such texts were united in this study in the group of hymns to 'popular' deities. It should be noted that hymns to Indra and Agni typically precede any others in individual books, while Soma hymns form a separate part of the collection (*maṇḍala* 9).

The scores for TTR, FCWR, and HR for each text of the sample were calculated and tested for significant differences between means within categories C1, C2, and C3 as determined by analysis of variance (ANOVA). Categorical variables were further analysed using the chi-squared statistic.

Collocations of five important content words were retrieved from the corpus in order to reveal, with the help of factor analysis, the interrelation of ideas they represented. A collocation is ‘a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text’ (Stubbs, 2001, p. 24). Collocations were extracted from the corpus with ANTCONC 3.1.2 concordancer using the Mutual Information (MI) coefficient which is effective for identifying units of meaning, as it ‘picks out lexical collocations’ (Stubbs, 1995, p. 39). Only collocations with high value of the coefficient ($\mathrm{MI} \geq 5$), following Stubbs (p. 40), attested not less than five times (Church and Hanks, 1990, p. 24) in the window of five tokens around the node, were taken into consideration. Context window size complies with the recommendations of Sinclair *et al.* (1997, p. 241 ff) and is usually equal to two *padas*, i.e. half a verse of the *ṚV*.

4 Results

Table 1 shows results of the one-way ANOVAs comparing means in the groups of hymns classified according with the occurrence of repeated clusters (C1), placement in the collection (C2), and subject matter (C3).

Books appear to differ according to the TTR, irrespective of the attestation of either rare or frequent words. Family *maṇḍalas* generally score a higher TTR, 80.22% versus 77.43%, which is a significant difference ($F=7.72$, $P=0.005$), see Fig. 1A and B for individual books.[3] Book groups are indiscriminate in respect of frequent content word and HRs, although hymns in *maṇḍala* 9 ($n=13$), if compared with all the others combined, scored a higher FCWR (7.61–5.11%, $F=20.71$, $P=0.00001$) and a lower HR (1.33–2.85%, $F=9.14$, $P=0.003$); both results are significant. However, no significant result emerged for TTR in between book 9 and the others ($F=0.14$).

On Fig. 2 one can see that lexical diversity, scored as FCWR and HR, differs significantly between hymns with or without a family seal. More frequent content word tokens and fewer *hapaxes* were used in texts with repeated word clusters ($F=10.53$, $P=0.001$ and $F=6.85$, $P=0.009$, respectively). Yet the difference was not significant for this factor on an ANOVA for TTR.

Concerning their subject matter, hymns do not differ much in TTR score, but are distinguishable in HR and FCWR (Fig. 3). *Hapaxes* occur at a significantly lower rate in hymns dedicated to Agni, Indra,

Table 1 Means (M) and standard deviation (SD) for the scores of lexical diversity tested with one-way ANOVAs

	M	SD	M	SD	F
C1	No clusters attested (*n*=72)		Clusters attested (*n* = 183)		
TTR	79.47	4.67	78	8.66	1.86
FCWR	4.60	1.75	5.49	2.04	10.53*
HR	3.23	2.19	2.59	1.57	6.85**
C2	Books 1, 8–10 (*n* = 165)		Books 2–7 (*n* = 90)		
TTR	77.43	8.34	80.22	6.22	7.72**
FCWR	5.20	2.18	5.30	1.63	0.13
HR	2.84	1.87	2.64	1.62	0.70
C3	Hymns to other deities (*n* = 125)		Hymns to Agni, Indra, Soma (*n* = 130)		
TTR	78.05	8.72	78.77	6.72	0.54
FCWR	4.90	2.15	5.57	1.79	7.31**
HR	3.11	2.05	2.44	1.43	9.38*

*$P \leq 0.002$, **$P \leq 0.009$; $N = 255$.

or Soma ($F = 9.38$, $P = 0.002$), while the laudation of other gods seemingly required poets to focus on rare lexis. Frequent content types are seen more in the hymns to popular deities ($F = 7.31$, $P = 0.007$) rather than in any others.

A combined relation of the topic (C3) and repeated cluster (C1) categories to HR was revealed by a two-way ANOVA ($F = 9.99$, $P = 0.001$). Regardless of their subject matter, hymns have about the same percent of unique types to the size of vocabulary if a repeated word cluster is attested, but there are significantly fewer *hapaxes* if a hymn is

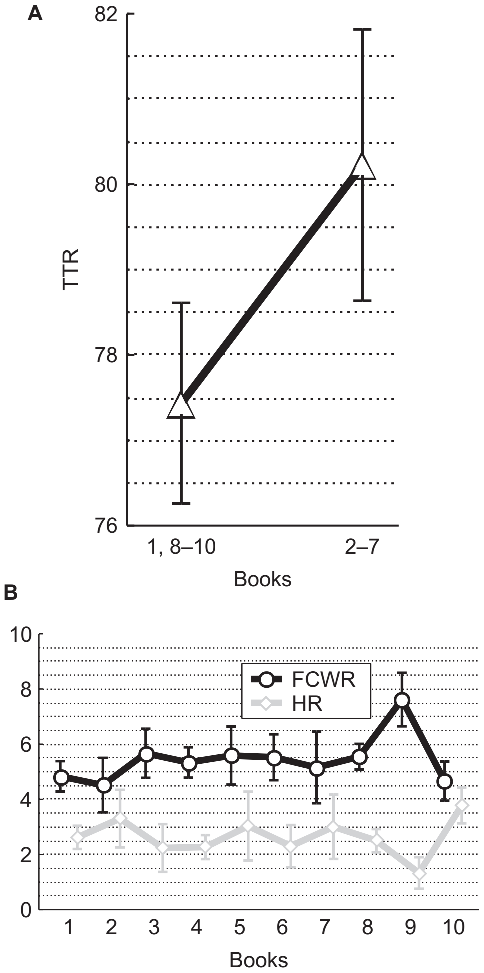

Fig. 1 (**A**) TTR in book groups. (**B**) Frequent and rare words in individual boooks: the outstanding position of the Soma book (9) can be seen

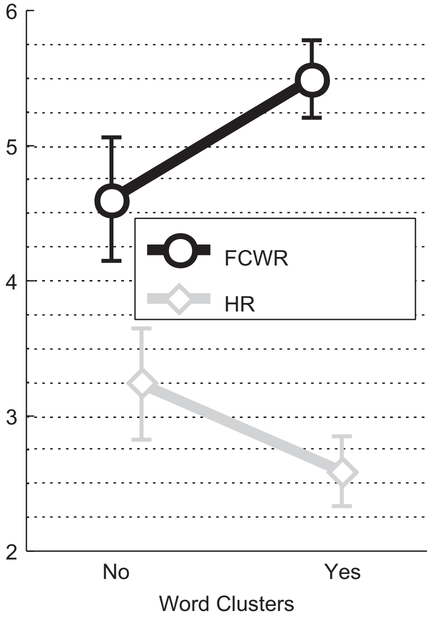

Fig. 2 Frequent and rare word in hymns with or without repeated word clusters

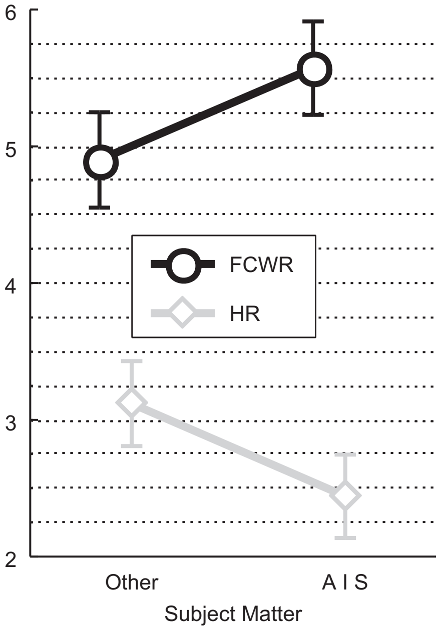

Fig. 3 Frequent and rare words in hymns to popular deities, Agni, Indra or Soma (AIS), and others

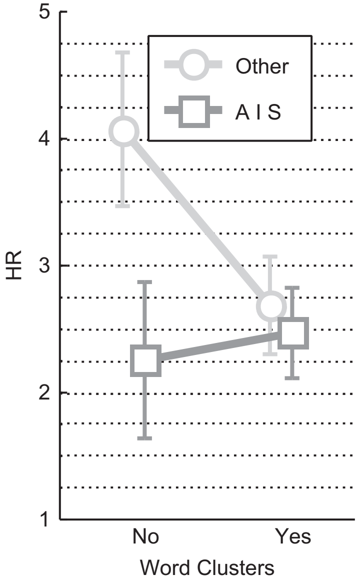

Fig. 4 Rare words in four groups of hymns: interaction of C1 and C3

dedicated to a popular deity and does not contain repetitions (Fig. 4). No further ANOVA interaction effects were found.

In terms of interaction of categorical variables, a significant mutual dependence of C2 and C1 was discovered. Repeated word clusters tend to appear more in family books: 176 hymns without repetitions to 253 with repetitions in *maṇḍalas* 2–7, which is a significant difference ($\chi^2 = 10.16$, $P = 0.001$). *Maṇḍalas* 2–7 were exclusively associated with the collectives of poets who viewed such 'family *saṃhitās*' as their exclusive intellectual heritage (Witzel, 1997, p. 261).

One of the significant results so far is that the occurrence of *hapaxes* and frequent content words in a hymn is dependent on its addressee. Based on the traditional data, this factor is both semantic and receptive. It deals with the meaning put into the texts by the members of the tradition. Since meaning of a word is dependent on its use, we examined contextual usage and interrelation of several high-frequent content words, hereafter referred to as 'headwords': *índra-*, 'the head of the Vedic pantheon', *agní-*, 'fire and its personification' *soma-*, 'the Soma plant, ritual beverage made of it, and a Vedic god', *áp-*, 'deified Water', and *dyú-*, 'the Sky'. Each headword produced an above average number of word forms (not less than three) in the frequent types list. According to the method described in the previous section, 614 node-collocate pairs featuring the headwords were retrieved. Collocation sets contained 267 types for *índra-*, 212 for *agní-*, 121 for *sóma-*, 111 for *áp-* and *dyú-* combined. Table 2 shows their relation.

Table 2 The relation of headword collocations

Headwords	Number of collocates	
	Union	Intersection (%)
índra-, *soma-*	346	42 (12.13)
índra-, *agní-*	457	22 (4.81)
agní-, *soma-*	320	13 (4.06)
áp- and *dyú-*, *índra-*	367	10 (2.99)
áp- and *dyú-*, *soma-*	222	10 (4.50)
áp- and *dyú-*, *agní-*	315	8 (2.53)
índra-, *sóma-*, *agní-*	528	5 (0.94)
índra-, *sóma-*, *áp-*, and *dyú-*	440	4 (0.90)

It can be seen in Table 2 that there are more collocates common to *índra-* and *sóma-* than to any other headwords. Indra, the most praised deity, was repeatedly mentioned by seers as the drinker of the sacred beverage, of which the god Soma is the personification (i.e. in 2.11.11: 'píbā-pibéd indra śūra sómam mándantu tvā mandínaḥ sutā́saḥ', 'Drink, drink, o hero Indra, soma! Let the joy-giving pressed [juices] make you drunk'). Agni and Indra (twenty-two common collocates) were also occasionally paired by the poets, sometimes forming a noun compound (*indrāgní-*). For example, although the initial verses in hymn 6.60 refer to the cosmogonic feat (v.1) which is typically claimed as Indra's, the myth is extended to Agni, without, however, describing his specific function (v.2 ff.):

śnáthad v ṛtrám utá sanoti vā́jam índrā yó
agnī́ sáhurī saparyā́t ‖1a ‖. . .
tā́ yodhiṣṭam abhí gā́ indra nūnám apáḥ svàr
uṣáso agna ūḷhāḥ |
díśaḥ svàr uṣása indra citrā́ apó gā́ agne
yuvase niyútvān ‖2‖

He kills Vṛtra and gets the reward, who worships Indra and Agni, the victors . . .

And now you both, Indra and Agni, fight for cows, Waters, the Sun, the Uṣas', taken away.

The Sky, the Sun, the bright Uṣas', o Indra, Waters, cows, o Agni, you harness.

In an attempt to find out what vocabulary use can reveal about the *ṛ́ṣis*' presuppositions of cosmogony, we analysed collocates of *áp*- or *dyú*- ('nature' headwords) intersecting with those of *índra*-, *agní*- or *sóma*- ('deity' headwords), as the former represent two important cosmological concepts, and the latter are the actors of the creation myth (Kuiper, 1960). Omitting function words and forms of the headwords, the search resulted in the following collocation pairs.

- ahan 55 (√*han*, 'to smite, slay, hit, kill') <apáḥ 10.9%,[4] índraḥ 20%>
- áhim 40 (*áhi*-, 'a snake') <apáḥ 22.5%, índraḥ 12.5%>
- upásthe 50 (*upástha*-, 'lap, middle or inner part of anything') <apā́m 32%, agníḥ 24%>
- óṣadhīḥ 51 (*óṣadhi*-, 'a herb, plant') <āpaḥ 25%, apáḥ 9.8%, agníḥ 9.8%>
- janitā́ 30 (*janitṛ*-, 'a begetter, parent') <índrasya 16.6%, divaḥ 26.6%>
- gā́ḥ 112 (*go*-, 'a cow, pl. cattle, kine') <apáḥ 10.7%, índraḥ 10.7%>
- gā́vaḥ 103 (*go*-) <āpaḥ 4.8%, sómam 4.8%>
- napāt 31 (*nápat*-, 'descendant, offspring, son') <apā́m 51.6%, agne 16.1%>
- nápātam 20 (*nápat*-) <apā́m 40%, agním 25%>
- nr̥̄n 44 (*nṛ*-, 'a man, hero, person; mankind, people') <divaḥ 15.9%, indra 20.4%>
- pátiḥ 97 (*páti*-, 'a master, owner, possessor, lord') <sómaḥ 6.1%, divaḥ 10.3%>
- pavate 66 (√*pū*, 'to make clean or pure or bright, cleanse') < sómaḥ 40.9%, índrāya 12.1%, divaḥ 9%>
- pṛthivyā́ḥ 98 (*pṛthivī́*-, 'the earth or wide world') <divaḥ 60.2%, agníḥ 7.1%>
- pṛthivyā́m 31 (*pṛthivī́*-) <diví 35.4%, agníḥ 16.1%>
- bṛhatáḥ 50 (*bṛhát*-, 'lofty, high, tall, great, large, wide, vast') <divaḥ 28%, agníḥ 10%>
- vṛtrám 83 (*vṛtrá*-, 'an enemy, foe; N. of demon, lit. resistance') <apáḥ 14.4%, índraḥ 21.6%>
- śūra 96 (*śū́ra*-, 'a strong or mighty or valiant man, warrior, hero') < sómam 6.2%, apáḥ 5.2%, indra 27%>
- sū́ryam 91 (*sū́rya*-, 'the sun or its deity') <diví 19.7%, apáḥ 6.5%, agním 7.6%>

Those words are habitually used with both groups of headwords. Cross-comparison of the referent verses highlighted hymns containing what seems to be an instructive 'summary' of the discourse, i.e. 1.32 (an extensive narration of the feats of Indra, esp. vv.1–5), 1.103 (discovery of creatures and plants by Indra, vv.2, 5), 5.29 (Indra and Soma, vv.2, 3, 8, also 4.28.1, 5), 2.35 (Agni as the son of waters, esp. vv.1–3, 7), 9.97 (Soma creating wide space, v.10 and gathering poetic thoughts, v.34–35), 9.72 (Soma as the master of cows, in vv.4–5, who is poured in rivers, v.7), etc.

The aggregate of all contexts of headwords can lead to the detection of latent structure of interrelation between the respective concepts. For that purpose collocation data listed above was summarized as three variables: the absolute frequency of each node and the frequency of its co-occurrence with the two groups of headwords. Factor analysis reduced the three variables to two factors. Factor 1 (F1) explains 47% of the variance observed in the data and attributes to the frequency of a word. It shows a highly positive loading of lexemes *pṛthivī́*- and *go*-, which are, indeed, very frequent in absolute terms. The association of a word with 'deities' on the one hand and 'nature' on the other is accounted for by Factor 2 (F2), which explains 37% of variance. The factor loading on F2 was 0.83 for the 'deity' and –0.66 for the 'nature' headwords. A highly positive loading on F2 is shown for the root √*pū*, lexemes *śū́ra*-, *vṛtrá*-, and *go*-, seen more often with the 'deity' headwords. In contrast, a highly negative loading on F2 is exhibited by *pṛthivī́*-, *sū́rya*-, *óṣadhi*-, and *nápat*-, which scored more on the association with 'nature' (Fig. 5). The words occurring in the 'deity' contexts would be less relevant in the 'nature' contexts, and vice-versa.

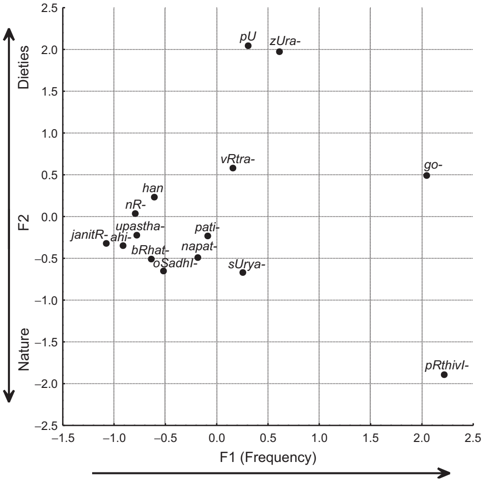

Fig. 5 Factor 1 (frequency) against Factor 2 (semantics); the transliteration follows the Harvard–Kyoto convention

5 Discussion and Further Research

One of the main issues in Vedology is what phenomena account for hymn structure. Two such factors are the situation of text production and features of the poet. The former includes the addressee of a particular text and its topic, while the latter includes the poet's world view and stylistic conventions, as well as the attribution of a hymn to a particular poetic collective. This article examined empirical evidence of a relationship between the vocabulary use and the texts' subject matter (1) and both linguistic and non-linguistic indications of poetic family attribution: attestation of repetitions (2) and the location of a text in the family core of the collection (3). Following a descriptive stance, the present research has shown that lexical diversity differs significantly between the hymns to 'popular' deities and where repeated text fragments were found, and those dealing with other topics and free from repetitions. The texts of the former kind exhibit a higher rate of frequent content words and contain fewer rare or obscure words. Although these differences are small, they show a high statistical significance. It is fair to say that the present study employed only a part of the available linguistic information, while the extraction of more data is problematic. Nevertheless, the facts suggest a considerable, systematic variety within the genre of the *Ṛ̣V*.

A preliminary interpretation suggested here is that the poets practiced different creative strategies that shaped the complexity of the genre. The *ṛ́ṣis* tended to adjust vocabulary to major topics and to say things 'pretty much the same way every time they're said' (Jamison, 1997, p. 127). Yet there was also a striving for a freer choice of subject matter and lexis, seemingly represented by magical charms, occasional, or 'abnormal' hymns, i.e. 10.106. A considerable part of it is 'mantras without meaning', containing over seventy once-occurring tokens, which qualify it as untranslatable according to Elizarenkova (1999, p. 508). Gonda writes about book 10 as exhibiting 'marked deviations from the usual contents of the corpus' (1975, p. 12). Perhaps the trend is comprehensive: the ability of the genre to contain heterogeneous texts, conservative on the one side and challenging on the other, may be due to the competitive nature of this form of poetry. The characteristics of the speech situation, essentially a verbal contest, should be taken into account, especially the setting and the norms of interaction (Duranti, 1997). Pictured by Kuiper (1960) and Thompson (1997), Aryan verbal contests were a grand spectacle of the force of words. As Thompson puts it, it was predominantly 'a means of self-display, . . . of one's mastery of the exoteric lexicon and, on the other hand, of one's personal authority and power' (1997, p. 20). In such a situation lexical choice must have been strategic.

This study has shown that books are indiscriminate in terms of frequent content word and HRs, although book 9, a liturgical *maṇḍala* dedicated entirely to the god Soma in his ritual aspect, occupies a somewhat different position. The observation that it has more frequent words and less *hapaxes* than the others agrees with the results of Wüst (1928, p. 34 ff) and confirms Bloomfield's notion that book 9 'for the most part repeats itself' (Bloomfield, 1916, p. 644). But contrary to Wüst (1928, p. 14), the present research points out that the trend is related to topic, as suggested by Edgerton: 'Ṛgvedic poets when dealing with identical situations, tend strongly to use identical language' (1929, p. 278). Indeed, this explanation is 'not related to the age of texts' (Fosse, 1997, p. 45). However, according to the present research, lexical diversity judged by TTR differs significantly between hymns in the family books and in the others. That observation is not easy to interpret. The diachronic nature of the corpus has to be taken into account: hymns in family *maṇḍalas* are commonly considered by scholars as the oldest. Is wider repertoire of vocabulary (and grammar forms) in books 2–8 connected with the factor of time or, perhaps, geography? Does it suggest that at some stage a degree of uniformity was reached (i.e. in books 1 and 10)? More data is required to answer these questions.

Conservative handling of discourse by the poets might account for the stability of their favourite themes, of which it can be fairly said that they 'maintained their identity remarkably unchanged' (Jamison, 1997, p. 138) through the entire period of the floating oral tradition (Deshpande, 1993, p. 134). Factor analysis has shown that regardless of any particular difference in the poets' conventions, association of words with the lexemes denoting popular deities and the elements is accounted for by a single factor. This could be due to the attribution of the lexemes to various components of mythology rather than simply to individual characters. Jamison notices that in Vedic narrative there are 'thematic building blocks that function as episodes in a number of different myths', and that 'in these the action or situation remains constant, but the participants vary' (Jamison, 1997, p. 133). If such a roughly Proppian model (Propp, 1968) is adopted, the division between the agon (represented by *śū́ra*-, √*han*, *vṛtrá*-, *go*-), and etiology (*pṛthivī́*-, *sū́rya*-, *óṣadhi*-, *nápat*-, *janitṛ*-) would seem to be pivotal (Fig. 5). The picture becomes clearer when referent verses are compared. For example, in hymn 4.28 Indra is praised as 'an ally' and 'a friend' of Soma (v.1: 'tvā́ yujā́ táva tát soma sakhyá ...'). Indra is the hero who 'made waters flow, slew the serpent, released seven rivers' (v.1: 'índro apó ... sasrútas kaḥ ... áhann áhim áriṇāt saptá síndhūn'), and 'pressed down the wheel of Surya', i.e. the Sun (v.2: 'ní khidat sū́ryasyéndraś cakrā́ṃ...'). In the same hymn Agni is said to have destroyed the enemies together with Indra, 'Indra killed, Agni burned ... the *dasyus* in the collision' (v.3: 'áhann índro ádahad agnír ... dásyūn ... abhī́ke ...'), and in the closing verse Soma is also given a tribute as Indra's partner who helped to release 'horses and cows from a hidden stall' (v.5: 'índraś ca somorvám ā́śvyaṃ góḥ ... riricáthuḥ'). In contrast, in hymn 8.36 Indra is not only praised as the victor of all battles ('víśvāḥ sehānáḥ pŕ̥tanā ...'), a refrain going through the entire hymn, but also is admitted to be 'the begetter of the sky and earth, horses and cows' (v.4–5: 'janitā́ divó janitā́ pṛthivyā́ḥ ... janitā́śvānāṃ janitā́ gávām asi ...'). More or less the same things are said of Soma in hymn 9.96: 'Soma purifies, the begetter of thoughts, ... of the sky, earth, Agni, Surya, Indra, and also Viṣṇu' (v.5: 'sómaḥ pavate janitā́ matīnā́ṃ janitā́ divó janitā́ pṛthivyā́ḥ ... janitā́gnér janitā́ sū́ryasya janiténdrasya janitótá víṣṇoḥ ...'), even though the motif of his military chiefdom is explicitly pronounced in the opening verse: 'Forth goes the hero, the chief, leading the chariots ...' (v.1: 'prá senānī́ḥ śū́ro ágre ráthānāṃ ... eti ...'). The militant aspect of Agni, referred to as 'the hero' (śū́ra), is clearly implied in hymn 4.3: 'Protect us, Agni ... kill the evil demon, even when he strengthens' (v.14: 'rákṣā ṇo agne ... jahí rákṣo máhi cid vāvṛdhānám ...'), while a much less articulated idea of his creative potency is linked with the Waters, i.e. in hymn 3.1: 'The parent, who begot cows, the Child of Waters, ... Agni' (v. 12: 'úd usríyā jánitā yó jajā́nāpā́ṃ gárbho ... agníḥ ...'); cf. in hymn 2.35: 'The Son of Waters, ... noble, begot all creatures' (v.2: 'apā́ṃ nápād ... víśvāny aryó bhúvanā jajāna ...'). In some sense Indra, Soma, and Agni were equally assumed to be the protagonists of the cosmogonic feat and the creators of the differentiated world. A relative freedom of the poet's conceptual bricolage and of his lexical choice must have been constrained by the need to put ideas together in terms of birth or struggle. However, one should keep in mind that this generalization concerns not vocabulary use in general, but just in one aspect—a strategy of the Vedic poet in unfolding cosmologic discourse.

The idea behind this study was that regularity on the lexical level might be more informative about genre when analysed in the scope of cultural (and receptive) categories. They were given preference over deductive constructs and this brings about a theoretical issue. In Jakobsonian, stylistics poetic language functions ‘to point to the message’ (Watkins, 1995, p. 29) and is considered ‘a sort of grammar’ (p. 28). Elizarenkova stresses the importance of a hymn’s ‘formal side’ and ‘the formal construction of a piece of poetry’ (Elizarenkova, 1995, p. 9) in a way that ‘the poetic function, the self-orientation of language’ (p. 9) becomes essential for the study of the *ṚV*. There remains a danger with such constructs as the poetic (aesthetic, indexical, etc.) function, that ‘instead of referring to the historical and structural concept of the literary system (as an institutionalized set of discursive norms and practices governing the production on new texts), they can tend to hypostatize a quality which resides in the text, to treat an analytic fiction as an essential property’ (Frow, 1986, p. 95). It remains to be demonstrated that an empirical narratological approach (cf. Klapproth, 2004), in the tradition of Belyj (1934) and Propp (1968), presents a plausible alternative to deductive stylistics. Further research is required to analyse lexical choice in a culturally specific genre in connection with narrative components and reception. In my view such research should be driven by corpus data.

References

Aufrecht, T. (1877). *Die Hymnen des Rigveda*. Bonn: A. Marcus.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Belyj, A. (1934). *Masterstvo Gogolja: issledovanie*, Moskva: Gosizdat Chud. Lit.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. and Conrad, S. (2001). Register Variation: A Corpus Approach. In Schiffrin, D., Tannen, D., and Hamilton, H. (eds), *The Handbook of Discourse Analysis*. Oxford: Blackwell, pp. 175–96.

Bloomfield, M. (1906). *A Vedic concordance*. Cambridge, MA: Harvard University.

Bloomfield, M. (1916). *Rig-veda repetitions*. Cambridge, MA: Harvard University Press.

Bloomfield, M., Edgerton, F. and Emeneau, M. B. (1934). *Vedic Variants: a Study of the Variant Readings in the Repeated Mantras of the Veda: Noun and pronoun inflection*. Philadelphia: University of Pennsylvania.

Bryant, E. (2001). The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate. Oxford: Oxford University Press.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1): 22–29.

Deshpande, M. (1993). *Sanskrit and Prakrit*. Delhi: Motilal Banarsidass.

Duranti, A. (1997). *Linguistic Anthropology*. Cambridge: Cambridge University Press.

Edgerton, F. (1929). Stilgeschichte und Chronologie des Rgveda. Von Walther Wuest. *Journal of the American Oriental Society*, **49**: 276–82.

Elizarenkova, T. (1995). *Language and Style of the Vedic Rsis*. Albany: State University of New York Press.

Elizarenkova, T. (1997). Problems of a Synchronic Description of Language and Style in the Rgveda. In Witzel, M. (ed.), *Inside the Texts, Beyond the Texts: New Approaches to the Study of the Vedas*. Harvard: Harvard University, pp. 49–59.

Elizarenkova, T. (1999). *Rigveda: Mandaly IX-X*. Moskva: Nauka.

Fosse, L. M. (1997). *The Crux of Chronology in Sanskrit Literature: Statistics and Indology, a Study of Method*. Oslo: Scandinavian University Press.

Fox, R. and Fox, J. (2004). *Organizational Discourse: A Language-Ideology-Power Perspective*. Westport, CT: Praeger.

Frow, J. (1986). *Marxism and Literary History*. Harvard: Harvard University Press.

Gamberini, F. (1983). *Stylistic Theory and Practice in the Younger Pliny*. Hildesheim: Georg Olms Verlag.

Gippert, J. (2000). TITUS Rg-Veda. Based on the edition by Th. Aufrecht, Bonn 1877 (2.Aufl.); arranged with the metrically restored version by B. van Nooten and G.Holland and the ‘Padapātha’ version by A.Lubotsky. *Thesaurus Indogermanischer Text- und Sprachmaterialien*. http://titus.uni-frankfurt.de/texte/etcs/ind/aind/ved/rv/mt/rv.htm (accessed 29 February 2008).

Gonda, J. (1959). *Stylistic Repetition in the Veda*. Amsterdam: Noord-Hollandsche Uitg. Mij.

Gonda, J. (1971). *Old Indian*. Leiden: E. J. Brill.

Gonda, J. (1975). *Vedic Literature: Saṃhitās and Brāhmaṇas*. Wiesbaden: Harrassowitz.

Grassman, H. G. (1964). *Wörterbuch Zum Rig-veda*. Wiesbaden: Harrassowitz.

Jakobson, R. (1960). Linguistics and Poetics. In Sebeok, T. A. (ed.), *Style in Language*. Cambridge: Technology Press, pp. 350–77.

Jamison, S. (1997). Formulaic Elements is Vedic Myth. In Witzel, M. (ed.), *Inside the Texts, Beyond the Texts: New Approaches to the Study of the Vedas*. Harvard: Harvard University Press, pp. 127–138.

Jamison, S. (2004). Poetry and Purpose in the Rgveda: Structuring Enigmas. In Griffiths, A. and Houben, J. (eds), *The Vedas: Texts, Language & Ritual: Proceedings of the Third International Vedic Workshop, Leiden 2002*. Groningen: Egbert Forsten, pp. 237–50.

Jucker, A. (1992). *Social Stylistics: Syntactic Variation in British Newspapers*. Berlin: Mouton de Gruyter.

Kane, P. V. (1951). *History of Sanskrit Poetics*. Bombay: Kane.

Klapproth, D. (2004). *Narrative as Social Practice: Anglo-Western and Australian Aboriginal Oral Traditions*. Berlin, New York: Mouton de Gruyter.

Kuiper, F. B. (1960). The Ancient Aryan Verbal Contest. *Indo-Iranian Journal*, **4**(4), pp.217–81.

Lubotsky, A. (1997). *A Rgvedic Word Concordance*. New Haven, CT: American Oriental Society.

Mainkar, T. G. (1966). *Some Poetical Aspects of the Rgvedic Repetitions*. Poona: University of Poona.

MacDonell, A. (1886). Kātyāyana's Sarvānurkamaṇī of the Rigveda. Oxford: Clarendon Press.

Oldenberg, H. (1888). *Die Hymnen des Rigveda: Metrische und Textgeschichtliche Prolegomena*. Berlin: Wilhelm Hertz.

Ong, W. (1982). *Orality and Literacy: The Technologizing of the Word*. London: Methuen.

Propp, V. (1968). *Morphology of the Folktale*. Austin: University of Texas Press.

Scharl, A. (2004). *Environmental Online Communication*. New York: Springer.

Sinclair, J., Mason, O., Ball, J. and Barnbrook G. (1997). Language Independent Statistical Software for Corpus Exploration. *Language Resources and Evaluation*, 31(3): 229–55.

Smith, B. (1994). *Classifying the Universe: The Ancient Indian Varna System and the Origins of Caste*. New York: Oxford University Press.

Stubbs, M. (1995). Collocations and Semantic Profiles: On the Cause of Trouble with Quantitative Studies. *Functions of Language*, **2**(1): 23–55.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Thompson, G. (1997). The Brahmodya and Vedic Discourse. *Journal of the American Oriental Society*, **117**(1): 13–38.

Toporov, V. (1992). Model' mira. In Tokarev, S. (ed.), *Mify Narodov Mira*. Moskva: Sovetskaja Enciklopedija, pp. 161–63.

Watkins, C. (1995). *How to Kill a Dragon: Aspects of Indo-European Poetics*. New York: Oxford University Press.

Witzel, M. (1995). Rgvedic History: Poets, Chieftains and Polities. In Erdosy, G. (ed.), *The Indo-Aryans of Ancient South Asia: Language, Material Culture and Ethnicity*. Berlin: Walter de Gruyter.

Witzel, M. (1997). The Development of the Vedic Canon and its Schools: The Social and Political Milieu. In Witzel, M. (ed.), *Inside the Texts, Beyond the Texts: New Approaches to the Study of the Vedas*. Harvard: Harvard University, pp. 257–346.

Wüst, W. (1928). *Stilgeschichte und Chronologie des Rgveda*. Leipzig: Deutsche morgenländische gesellschaft.

Notes

[1] This article was made possible by the Jan Gonda Foundation, which granted me a fellowship in January–June 2006, and the International Institute for Asian Studies (IIAS), which provided facilities and assistance throughout the research.

[2] In quotations of the Vedic verse I have used the edition of Aufrecht (1877) and presented the text in the *saṃhitā* form (otherwise *padapāṭha* instances are given).

[3] In Figs 1–4 points represent the means for each group; vertical lines indicate the 95% confidence limits; $N = 255$.

[4] Occurrence (%) of a collocate, in the span of 5:5 around the node, to the absolute frequency of the node.