**In literature studies interest in the Rgveda (RV) is typically based on the observation of the regular occurrence in hymns of certain textual features, treated as evidence of homogeneity of their genre. It is argued that verbal formulas are "the vehicles of themes and... in the totality of these we find the doctrine, ideology, and culture of the Indo-Europeans" (Watkins 1995), including the Indo-Aryans. Results of a recent computer-aided study[1], however, suggest that the hymns show a less conformity in terms of their lexical structure than previously believed.**

# Ancient corpus under digital scrutiny:

## *deciphering lexical structure of the Veda*

Alexandre Sotov

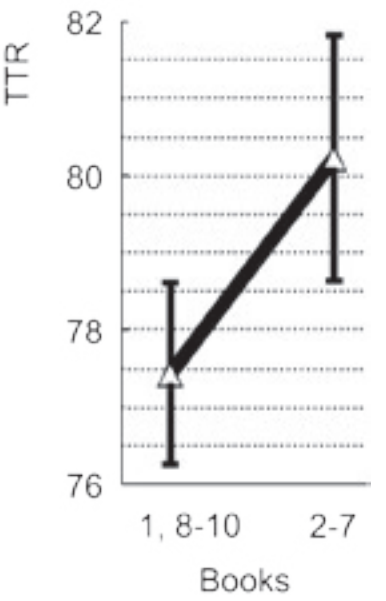
The qualitative/quantitative divide is not at all foreign to Vedic studies. In the traditional approach style emerges as the mastery of figures of speech: the purpose of the hymns was "to give delight, to both the deity addressed and the listener in general" (Mainkar 1966). This intention, along with the belief in the power of the spoken word as a means of attaining human goals, formed a poetical practice in which different kinds of repetition were central. According to Gonda's study (1959), a stylistic analysis of the *Veda* should go beyond such devices, since their use is constructive rather than ornamental. This laid a foundation for the functional method, which tries to link stylistic features and the production of texts. Elizarenkova analyses the hymns in relation to the structure of the poetic message (Elizarenkova 1995). The formal devices, studied by Jakobson (1960) within his theory of self-orientation of the poetical language, thus acquire a communicative purpose: the hymns dealt with the situation of a gift exchange between the poets and the divinity (Elizarenkova 1995). Similarly, Watkins (1995) treats the technique and the purpose of literary creativity "in the Indo-European times" as a part of the social function of the Indo-European poet as "the custodian and the transmitter" of the tradition.

On the other hand, linguistic features can be analysed irrespective of their function. Bloomfield et al. (1934) describe grammar in recurring *mantras* in terms of formal and, notably, stylistic variants. Quantitative data presented by Wüst (1928) is also descriptive: the distribution of countable features in the collection is its important empirical characteristic, although it is meaningless without a valid category of comparison. Furthermore, corpus approach to lexical analysis was established in Vedology long before the arrival of digital humanities (cp. (Grassman 1964) and more recently (Lubotsky 1997)). Together with the Indological tradition of quantitative research (Fosse 1997) this suggests the necessity of a data-driven analysis of the Vedic lexis, semantics, genre and discourse.

### Lexical analysis of the ancient corpus

Uncovering whether or not there is empirical evidence of a relationship between word usage and an interpretable typology of hymns may help to better understand the phenomena, which influenced the choice of vocabulary by the Indian rhapsodes. The hymns are part of a unique corpus, which features traditional lists of subject matter and authors, as well as certain schemes of internal arrangement (Witzel 1997). The texts' subject matter and indications of poetic family attribution, refrains and the location of a text in the 'family core', are obvious candidate categories of comparison of the lexis of the hymns. The latter can be studied with the help of various statistics of lexical diversity or richness, such as the type-token (TTR) and *hapax* (HR) ratios, which show how inclined the authors were to repeat the same words and to use rare, highly specialised vocabulary (see Biber 1995). Another such measure is the ratio of frequent content word tokens (FCWR), since common lexical items correspond to the vocabulary of the typical formulas and mythological representations. In a sample of size-adjusted *pāda* texts, around 25 per cent of the collection, lexical diversity differs significantly between the hymns to 'popular' deities, Indra, Agni, and Soma, where repeated text fragments (clusters) were found, and those dealing with other topics and void of repetitions. The former texts exhibit a higher rate of frequent content words and contain fewer hapaxes. The differences are minute, yet statistically significant: Figures 1-3 show results of the ANOVAs comparing means in the respective groups of hymns. Books appear to differ as well: family core scores higher on TTR (see Figure 4)[2].

Fig. 4




Another point of interest are the so-called collocations (Stubbs 1995) of high-frequent content words: *índra-*, 'the head of the pantheon', *agní-*, 'fire and its personification'; *soma-*, 'the Soma plant, ritual beverage, and a deity', *áp-*, 'deified water', and *dyú-*, 'the sky'. These last two are important cosmological concepts, while *índra-*, *agní-*, and *soma-*, are the actors of the creation myth (Kuiper 1960). Statistics has it that association of words with these lexemes is accounted for by a single factor. A highly positive loading on it is shown for the verb roots 'to purify' and 'to kill', nouns 'a hero', 'an enemy' and 'a cow'. They are seen more often with the 'deity' headwords. In contrast, a highly negative loading on this factor is exhibited by 'the earth', 'the sun', 'a plant' and 'a descendant', which scored more on the association with 'nature'. This could be due to the attribution of the lexemes to various components of mythology rather than to characters. Jamison notices that in the Vedas 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). If this roughly Proppian model (Propp 1968) is adopted, the division between the agon and etiology would seem to be pivotal.

### Aryan verbal contest and strategies of discourse

Altogether, such facts suggest a systematic variety within the genre of the *RV*. The poets may have practiced different creative strategies that shaped the complexity of the genre. They tended to adjust vocabulary to major topics, although there was probably a striving for a freer choice of subject matter and lexis, represented by magical charms, occasional or 'abnormal' hymns, i.e. *RV* 10.106. 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. Pictured by Kuiper (1960) and Thompson (1997), Aryan verbal contests were a grand spectacle of the force of words. In such a situation lexical choice must have been strategic, while conservative handling of discourse by the poets might account for the stability of their favourite themes.

Many important questions remain unanswered, as getting more interpretable statistical results is problematic in a corpus of circa 165,000 tokens. The diachronic nature of the collection also has to be reconsidered: hymns in the family books are viewed by scholars as the oldest. Is the wider repertoire of vocabulary (and grammar forms) in such books connected with factors of time or geography? And yet it remains to be demonstrated empirically that an approach, which embraces cultural categories and natural data, presents an alternative to a literary theory based on deductive constructs, such as the poetic function.

**Alexandre Sotov**
St Petersburg State University
*a.sotov@yahoo.co.uk*

Fig.1

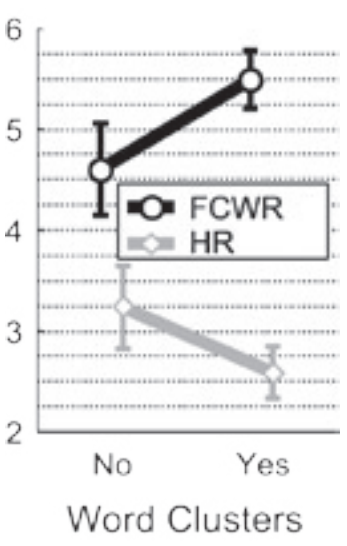



Fig.2

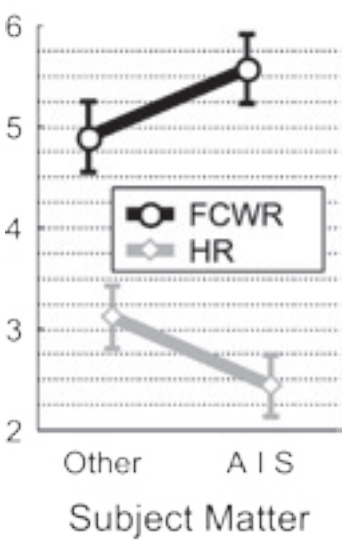
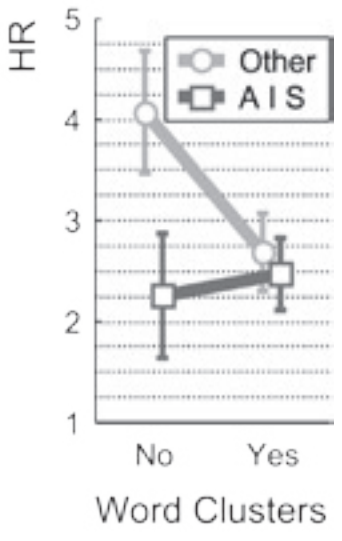



Fig.3




**Notes**

1 The research 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 Figures 1-4 points represent the means for each group; vertical lines indicate the 95% confidence limits; n=255.

**References**

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*, Cambridge: Cambridge 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*, Philadelphia: University of Pennsylvania.

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

Fosse, L. M. 1997. *The crux of chronology in Sanskrit literature : statistics and Indology, a study of method*, Oslo: Scandinavian University Press.

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

Grassman, H. G. 1964. *Worterbuch zum Rig-veda*, Wiesbaden: Harrassowitz.

Jakobson, R. 1960. Linguistics and Poetics. In Sebeok, T.A. (ed). *Style in Language*, Cambridge: Technology Press, pp. 350-377

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.

Kuiper, F. B. 1960. The ancient Aryan verbal contest. *Indo-Iranian Journal*, 4(4), pp.217-281.

Lubotsky, A. 1997. *A Rgvedic word concordance*, New Haven, Conn.: American Oriental Society.

Mainkar, T. G. 1966. *Some poetical aspects of the Rgvedic repetitions*, Poona: University of Poona.
Propp, V. 1968. *Morphology of the folktale*, Austin: University of Texas Press.

Stubbs, M. 1995. Collocations and semantic profiles: On the cause of trouble with quantitative studies. *Functions of Language*, 2(1), pp. 23-55.

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

Watkins, C. 1995. *How to kill a dragon: aspects of Indo-European poetics*, New York: Oxford University Press.

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.