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 (1955), a stylistic analysis of the Vedas 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 metres in terms of formal and, notably, stylistic variants. Quantitative data presented by West (1934) 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 Veldigology long before the arrival of digital humanities (cp. Grassman 1964) and more recently (Lobotsky 1997). Together with the Indological tradition of quantitative research (Fosse 1997) this suggests the necessity of a data-driven analysis of the Vedic texts.

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 rhapsoids. 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 information. Thus, the texts can be studied with the help of various statistics of lexical diversity or richness, such as the type-token (TTR) and hyper (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 pablo texts, around 25 per cent of the collection, lexical diversity differs significantly between the hymns to ‘popular’ deities, Indra, Agni, and Soma, where repeated test 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, 2 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).

Many important questions remain unanswered, as getting more interpretable statistical results is problematic in a corpus of nearly 100000 lines. 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 (grammatical 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

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 95% confidence limits; n=255.

References


