VedaWeb – the Role of Annotations in Analyzing Ancient Indo-Aryan Texts

Daniel, Kölligan¹, Claes, Neußeind², Börge, Kiss¹,
Francisco, Mondaca³, Uta, Reinöhl¹, Patrick, Sahle³

University of Cologne, ¹Dept. of Linguistics, ²Institute for Digital Humanities, ³Cologne Center for eHumanities

The VedaWeb platform provides access to ancient Indian texts written in Vedic Sanskrit, which are enriched with morphological and metrical annotations, making them searchable according to lexicographical and corpus-linguistic criteria. The cornerstone of this DFG-funded project is the Rigveda, one of the oldest and most important texts of the Indo-European language family. Its origin can be traced back to the late second millennium BC. With an extent of ca. 160,000 words (comparable to Homer's Iliad and Odyssey combined totalling ca. 190,000 words), it represents an extremely rich corpus. The project is meant to advance research in all areas of Vedic Sanskrit, for example in syntax (e.g. referential null objects, [6], non-configurationality, [9]), morphology (e.g. the Vedic vr̥kī-type, [12], ya-presents, [7]) or word formation (e.g. compounds, [11]).

The starting point of the project is a complete morphological annotation of the Rigveda, which was mainly carried out at the University of Zurich. These annotations are the basis for a combined search function for linguistic parameters (including lemma, word form, morphological and metrical information) that allows for linguistically motivated search queries across multiple levels of annotation by means of complex, combined criteria. The annotations are of key importance for opening up novel perspectives on the data, e.g. by means of quantifying feature combinations or context-dependent phenomena.

A central feature of VedaWeb is the enrichment of the Rigveda with links from each word to entries in the standard dictionary for the Rigveda by Hermann Grassmann [2], an online version of which is part of the Cologne Digital Sanskrit Dictionaries (http://www.sanskrit-lexicon.uni-koeln.de), a central resource for international Sanskrit research hosted at the University of Cologne. Based on a TEI data model, the word forms in the text are linked to the respective lemmata in Grassmann's dictionary and vice versa. The data model also allows linking the lemmata to other Sanskrit dictionaries hosted in Cologne or elsewhere.

Relevant for the technical implementation and crucial for data persistence is the TEI model adopted for VedaWeb, which is used for the representation of both the annotated text and the lexical resources. During the project, various further research and analysis tools will be integrated into the platform. These include the display of translations (including [3], [4], [1], [10]) as well as commentaries to the Rigveda ([8], [10]), and the possibility of exporting sections of annotated text according to criteria selected by the user. In addition, metrical information (Kevin Ryan, Harvard University, http://www.meluhha.com/rv/) and syntactic information from various completed and ongoing research projects ([5]) will be available in VedaWeb.

We will give a compact overview of the possible uses of the VedaWeb platform. Our focus is on the possibilities and challenges of embedding multiple searchable annotation layers and ways in which
these can be linked to external sources of information including dictionaries, with the aim of creating a rich web-based research platform for historical corpus linguistics.

References


