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

Daniel, Kölligan[1], Claes, Neuefeind[2], Börge, Kiss[1],

Francisco, Mondaca[3], Uta, Reinöhl[1], Patrick, Sahle[3]

University of Cologne, [1]Dept. of Linguistics, [2]Institute for Digital Humanities, [3]Cologne 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 *vṛkī*-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

[1]   Geldner, Karl Friedrich (2003)[1951-57]: *Der Rig-Veda. Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen von Karl Friedrich Geldner*. Cambridge (Mass.): Harvard University Press.

[2]   Grassmann, Hermann (1873): *Wörterbuch zum Rig-veda.* Wiesbaden, O. Harrassowitz.

[3]   Grassmann, Hermann (1876): *Rig-veda. Übersetzt und mit kritischen und erläuternden Anmerkungen versehen von Hermann Grassmann*. Leipzig: F.A. Brockhaus.

[4]   Griffith, Ralph T. H. (1896): *The Hymns of the Rigveda*. Benares: Lazarus.

[5]   Hellwig, Oliver and Heinrich Hettrich and Ashutosh Modi and Manfred Pinkal (2018): *Multi-layer Annotation of the Ṛgveda*. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7.-12.5. 2018. Paris: European Language Resources Association (ELRA).

[6]   Keydana, Götz and Silvia Luraghi (2012): *Definite referential null objects in Vedic Sanskrit and Ancient Greek*. Acta Linguistica Hafniensia 44 (2):116–28. https://doi.org/10.1080/03740463.2013.776245.

[7]   Kulikov, Leonid (2012): *The Vedic -ya-Presents: Passives and Intransitivity in Old Indo-Aryan*. Amsterdam, Netherlands: Rodopi.

[8]   Oldenberg, Hermann (1909/1912): *Ṛgveda. Textkritische und exegetische Noten*. Berlin: Weidmann.

[9]   Reinöhl, Uta (2016): *Grammaticalization and the Rise of Configurationality in Indo-Aryan*. Oxford: Oxford University Press.

[10]   Renou, Louis (1956-1969): *Études védiques et paninéennes*. Paris: Boccard.

[11]   Scarlata, Salvatore and Paul Widmer (2015): "Vedische exozentrische Komposita mit drei Relationen", *Indo-Iranian Journal* 58(1):26-47.

[12]   Widmer, Paul (2007): "Der altindische *vrkī́*-Typus und hethitisch *nakki-*: Der indogermanische Instrumental zwischen Syntax und Morphologie", *Zeitschrift für Sprachwissenschaft* (1-2):190-208.