It Takes a Village: Co-developing *VedaWeb*, a Digital Research Platform for Old Indo-Aryan Texts

Börge Kiss (Institute for Digital Humanities)[1], ✉ Daniel Kölligan (Dept. of Linguistics – Historical-Comparative Linguistics)[1[0000-0002-3134-8398]], Francisco Mondaca (Cologne Center for eHumanities)[1], Claes Neuefeind (Institute for Digital Humanities, Data Center for the Humanities)[1], Uta Reinöhl (Dept. of Linguistics – General Linguistics)[1, 2[0000-0003-2829-682X]] and Patrick Sahle (Cologne Center for eHumanities)[1[0000-0002-8648-2033]]

[1] University of Cologne, D-50923 Köln, Germany
[2] Johannes Gutenberg University Mainz, D-55099 Mainz, Germany

b.kiss@uni-koeln.de
✉ d.koelligan@uni-koeln.de
f.mondaca@uni-koeln.de
c.neuefeind@uni-koeln.de
uta.reinoehl@uni-koeln.de
sahle@uni-koeln.de

Presenters: Daniel Kölligan (humanities scholar), Claes Neuefeind (digital expert)

1 Research Goal: the "Humanities Problem"

The traditional linguistic approach in dealing with large text corpora includes the writing of concordances (or word indexes) to make co-occurrences of forms visible and to enable researchers to detect the frequency of usage patterns. The determination of meanings, functions, syntactic patterns etc. are mostly based on the individual assessment of a specialist drawing their conclusions from their intuition based on personal reading experience and more or less implicit knowledge of the texts (cf. e.g. [12, 20]). However, the more the quantity and quality of data increase, the more intractable these resources become for such traditional "tools" and approaches. At the same time, the recent wave of digitizing textual data makes the application of computational methods unavoidable in order to ensure the highest possible scientific standard in quantitative terms. The more information resources are made accessible, systematically prepared and annotated, the more urgent the need for their formal analysis becomes. The project described below is the interdisciplinary endeavour to prepare such a resource for the linguistic analysis of a large corpus of richly annotated textual data.

The objective of the *VedaWeb* project is to develop a web-based research platform for linguistic and philological work on Old Indo-Aryan texts. To this aim, multiple textual, lexical and other types of linguistic data are integrated in a single digital research environment. Various layers of texts and annotations are made searchable

and are linked with lexicographic information. At a later stage, the platform will offer personalized work spaces for researchers to enable collaborative work on the texts. Long-term sustainability of data and software will be ensured through cooperation with the Data Center for the Humanities at the University of Cologne (http://dch.phil-fak.uni-koeln.de). In order to build *VedaWeb*, a group of historical linguists, general linguists and computational linguists have teamed up with specialists in digital humanities at the University of Cologne. In addition, the project team collaborates with several international co-operation partners (e.g., from the University of Zurich and from Harvard University).

For illustration, concrete examples of humanities research questions that will be possible to address working with the *VedaWeb* platform include the following: In contrast to the rather intuitive description of meanings as described above, linking lexical units with and measuring their distance to all other units will allow to form networks of co-occurrences, enabling a more systematic and better-controlled semantic analysis. Linear distances on the one hand and co-occurrence frequencies on the other serve as variables for the configuration of these networks. This possibility will also further the study of lexical fields that so far had to be carried out "by hand" (cf. [9, 10, 17]). In the realm of morphology, given the data stored in *VedaWeb,* it becomes possible to study allomorphic variation within Vedic Sanskrit, e.g. the competing forms of the nominative plural masculine of stems in ending in *-a-* (e.g. *áśv-a-* 'horse'), viz. *-ās* (*áśvās* 'horses') and *-āsas* (*áśvāsas,* also 'horses'). With the help of *VedaWeb* these are currently studied in a project at the University of Cologne (http://ifl.phil-fak.uni-koeln.de/36486.html?&L=1), which tries to establish their synchronic distribution in the Rigveda according to various criteria such as agency and topic-hood and to propose hypotheses regarding their diachronic development.

2 Background and Starting Points

VedaWeb provides access to ancient Indo-Aryan texts written in Vedic Sanskrit, which are enriched with morphological and metric annotations as well as integrated with lexicographic information, making them searchable according to corpus-linguistic criteria. The pilot text of this project is the Rigveda, one of the oldest and most important texts of the Indo-European language family and the oldest extant one composed in Indo-Aryan, whose origin can be traced back to the late second millennium BC. With an extent of c. 160.000 words in nearly 40.000 lines of verse (comparable to the Homeric epics *Iliad* and *Odyssey* combined totalling c. 190.000 words), the Rigveda presents an extremely rich text corpus given its considerable ancestry. At a later stage of the project, further texts such as the *Atharvaveda* (c. 170.000 words), *Yajurveda* and Vedic prose texts such as the *Aitareya Brahmana* (c. 100.000 words) and the *Maitrayani Samhita* (c. 120.000 words) are also to be integrated into the *VedaWeb* platform to allow complex searches across several texts. Further possible features include user-specific annotation tiers, semantic searches, and the embedding of audio and video data of traditional recitals of the texts. The project aims to advance research in all areas of Vedic linguistics and philology, for example

in syntax (e.g. referential null objects [11], non-configurationality [15]), morphology (e.g. the Vedic *vr̥ kī*-type [21], *ya*-presents [13]), metrics [18] and word formation (e.g. compounds [19]).

The starting point of the project is a complete morphological annotation of the Rigveda, which was carried out at the University of Zurich over several years, where each word form has been annotated according to morphological categories (nouns: case, number, gender, verbs: tense, mood, number, person, voice, etc.). In addition, metrical information (provided by Kevin Ryan, Harvard University, http://www.meluhha.com/rv/) and syntactic information from different completed and ongoing research projects will be made available in *VedaWeb*. Metrical data for each word form note the frequency of attestation, the number of syllables, the metrical weight template (heavy and light syllables), any metrical variants (e.g. *indra* vs. *indara*) and the number of occurrences in different positions in the verse. Syntactic information comes from two sources: H. Hettrich (University of Würzburg) and O. Hellwig (Heinrich Heine University Düsseldorf) [8] have annotated all verbs and their arguments (e.g., direct objects) of the *Rigveda* as part of Hettrich's studies on the syntax of case in Vedic. This will allow searches for semantic categories and their morphological surface expressions (e.g. "Which morphological cases are used to denote instruments?"). Based on the Haig/Schnell annotation scheme GRAID (*Grammatical Relations and Animacy in Discourse* [7]), a project team lead by PI Reinöhl (http://sfb1252.uni-koeln.de/b03.html) is currently annotating selected Vedic texts for syntactic and pragmatic information which will allow searches according to types of referents (+/- human, +/- animate), clause types (e.g. main vs. relative clause), etc.

In the course of the project, multiple research and analysis tools are integrated into the platform. These include a combined search function for linguistic parameters (including lemma, word form, morphological and metric information), links for each word to entries in the standard dictionary of the Rigveda by Hermann Grassmann [3], the display of translations (including [2, 4, 6, 16]) as well as commentaries [14, 16], and the possibility of exporting sections of annotated text according to criteria selected by the user.

A key feature is to link the Rigveda with the C-SALT (*Cologne South Asian Languages and Texts*, http://c-salt.uni-koeln.de/) application programming interfaces (APIs) for Sanskrit Dictionaries (https://api.c-salt.uni-koeln.de/). Based on a TEI (Text Encoding Initiative) data model, the word forms in the text are linked to the respective lemmas in Grassmann's dictionary. In this way, it will also be possible to create links to lemmas in further Sanskrit dictionaries, for example to enable comparative, cross-dictionary searches (such as the DFG-funded *Nachtragswörterbuch des Sanskrit* at the University of Halle with whom a collaboration to this aim is planned, http://nws.uzi.uni-halle.de/). The direct linking of text and dictionary is a unique feature of the *VedaWeb* platform compared to existing resources of Old Indo-Aryan texts, e.g. the *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS, http://titus.uni-frankfurt.de).

3 Realization: the “Technical Solution”

The *VedaWeb* platform is implemented as a web application based on the Spring framework (https://spring.io). Spring serves as middleware providing controllers for data handling, service orchestration and different views on the data. Since the *VedaWeb* search should allow for linguistically motivated queries on different levels of annotations by means of complex, combined search criteria, a tailored search logic was implemented on the basis of the established search server Elasticsearch (https://www.elastic.co/). This solution provides fast search functionality and enrichment of search results (e.g. highlighting of lemmata vs. word forms). The frontend is implemented as a single-page application (SPA) based on React (https://reactjs.org) for building interactive front-end components. It makes use of Ant Design (https://ant.design/) for a clear and effective interface.

The TEI model adopted for *VedaWeb* plays an important role both with regard to the technical implementation as well as for data sustainability. As a cooperative project, the various sets of information to be integrated into *VedaWeb* have different sources and formats. For this reason, the various data resources had to be standardized in order to be employed in the *VedaWeb* application and made available for external projects. TEI provides a well-documented model and the necessary flexibility to achieve this goal, especially with regard to long-term sustainability.

The Rigveda has been modelled mainly according to an inline-markup paradigm: all of its text versions, annotations and translations are available in one TEI document. Versions are aligned on the level of the stanza while linguistic annotations are added to the single word forms of the chosen ‘Leittext’ via feature structures [22]. The exception to this markup-style are the references to dictionaries, which are encoded as stand-off markup (https://wiki.tei-c.org/index.php/Stand-off_markup). In the TEI document, we only annotate the reference to a lemma in Grassmann’s dictionary, which is also available in TEI format through an API. The motivation for modeling all data in one document is, on the one hand, to show all the data per stanza in one single source and, on the other hand, to simplify the process of exporting the data in TEI format. To enable data export for users is a key feature that is planned to be implemented in *VedaWeb* in the next project phase. In the web application, TEI data is imported and stored in a document-oriented database (mongoDB, see https://www.mongodb.com/), largely reflecting the original structure of the TEI model. This allows for fast data delivery and flexible data manipulation. Additionally, the use of a database is a prerequisite for implementing personalized work spaces for researchers to add, edit and share data, which requires at least a combination of versioning functionality (revision history) and user management.

To ensure the sustainability of the data integrated and software created within the project, *VedaWeb* collaborates with the Data Center for the Humanities (DCH), which is a “CLARIN Knowledge Centre for linguistic diversity and language documentation” (see http://ckld.uni-koeln.de/). Furthermore, the DCH is a CLARIN C-Centre (https://centres.clarin.eu/centre/47) and has applied for B-Centre status

(currently in the certification process). As such, the DCH hosts the data of the *VedaWeb* project, ensures their long-term archiving and provision, and will also take care of the data and software produced within the project beyond the end of the funding period.

4 Background and Starting Points

Before we address more specifically the challenges and opportunities arising in this collaboration, we sketch the general structures and dynamics of collaboration within the *VedaWeb* team, which play a key role for the success of the project.

The project team has been on a steep learning curve with regard to reaching a shared language and understanding of the tasks and challenges arising in the project for the different team members. This is not only true of the chasm between the "linguistic" and "technical" camp respectively, but applies on a more fine-grained level. The following diagram illustrates the fields of competence (narrow bars) as well as the de facto work areas (thick bars) of the various team members and primary collaboration partners. The diagram makes clear that the team members do not actually fall into two neatly separated domains, but that the range of competence areas are carved up in different ways with overlapping responsibilities.

Although the project has a rather small staff basis financed by external funding (DFG, German Research Foundation) with only one full researcher position (shared among two DH specialists, BK and FM), two student assistant positions (NK and JH) and a limited amount of money for the transition into a permanent hosting and curation by the data center (JB), it is supported by a large team of researchers involved either as principal investigators (UR, DK, PS, CN) or as collaborators in Cologne (FR) or abroad (PW). Further colleagues such as K. Ryan (Harvard) and D. Gunkel (Richmond) and consultants for special issues such as M. Gödel (Cologne) for the TEI data model have also contributed to the project, but are not included in the diagram for space constraints.

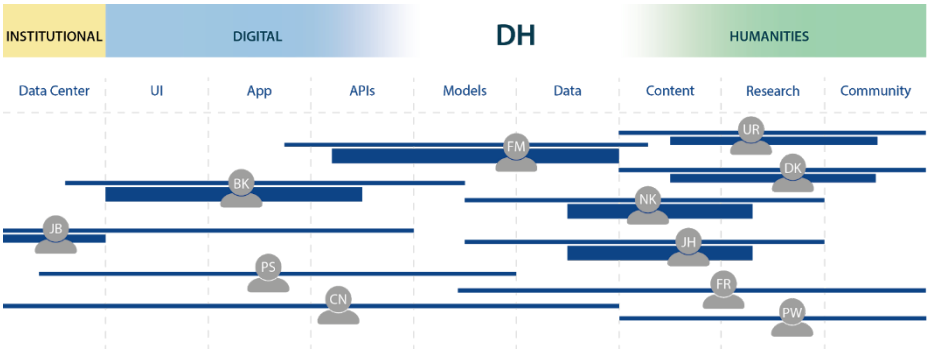

Fig. 1. Roles and coverages. Family constellation for the collaboration in the *VedaWeb* project.

As illustrated by the diagram, which we informally call our “family constellation”, there is no simple bifurcation between digital humanities and linguistics. Rather, each team member contributes to a certain range of domains, several ones of which are not clearly or not purely either technical or linguistic. This differentiation and, crucially, the overlap of competence areas among team members makes this project possible in the first place, as no single team member could be competent in or responsible for the whole range of domains.

In addition to differentiation and overlap, regular communication and close feedback loops ensure productive, agile and goal-oriented work dynamics, the success of which is mirrored by the fact that, half of the funding phase having elapsed, the project is exactly in time with its planned schedule. Most online communication, collaboration and data exchange takes place in GitLab (https://about.gitlab.com/), where several of the linguistic bodies of data are stored, too. Using GitLab's issue tracking system, team members are able to quickly provide feedback on each other’s work and effectively react to changing requirements in the project.

Team meetings in person take place on a regular basis (roughly once a month on average). In this way, developments can be steered into the right direction, mis-communications can be detected early, and there is room for creativity and adaptation. In fact, since the project is on time and running smoothly, the team is already addressing certain steps that were originally believed to be beyond the first funding period, such as the inclusion of further text versions and translations. The overlapping structure of the competence areas among the team members, paired with the regular meetings, which are supplemented by highly regular (often daily) online communication and collaboration, give rise to a sense of plasticity, i.e. to a dynamic, adaptable development of the project. Given the outlined project dynamics, the intermediate targets planned in the original application are reached so far and, at the same time, the project development can be flexibly adapted and expanded to find ideal solutions.

4.1 Simple and Challenging Issues

“What was easy and what was difficult – and why?” In general, team members at the more technical and at the more linguistic pole were frequently surprised at what exactly it was that was either easy or difficult for the others. To give some examples, implementing multiple, combinable full-text searches was easily realisable, which to a linguist who does not normally work with the support of elaborate digital search functions was surprising. At the same time, the fact that it would be technically quite difficult to have search functions run over diversely structured sets of data was not expected. Conversely, from the point of view of the DH researchers, the significant complexity of the Rigveda into different sub-structures including books, hymns and verses of different length and partially different type turned out to be a surprisingly complex and challenging data basis.

In the case of these and other examples, it was only possible to identify what was easy and what was difficult, and act accordingly, through the outlined constant and intensive communication among all project members. Furthermore, teasing out

exactly the level of complexity and time requirements for individual tasks was enabled through the structure of differentiation and overlap with regard to competence areas and responsibilities of the different team members.

4.2 Mutual Change of Views

"How did researcher and technician change each other's way of looking at things?" The linguistic researchers gained considerable new insights into the opportunities afforded by a digital research platform endowed with powerful search mechanisms and an integration of textual with lexicographic data. In developing the *VedaWeb* platform, the linguists gained a much deeper understanding of the affordances of building both a web-based application and a TEI model for the purpose of ensuring data longevity. On the other hand, the technical researchers gained new insights into the complexities of ancient texts and further linguistic resources, both with regard to their respective internal structures as well as the linguistic needs of accessing those resources. Numerous ones of these insights gained into each other's fields are likely to benefit the project members also above and beyond the *VedaWeb* project, as they are representative of DH work and historical, corpus-linguistic work respectively.

4.3 Disciplinary Blind Spots

"Did they, for instance, make each other aware of blind spots they had?" From day one of the *VedaWeb* project, the researchers and DH experts were in a constant dialogue that frequently uncovered blind spots in various domains and on all sides. Detecting these blind spots and addressing them is considered by the project team as central to the hitherto significant advances in building the *VedaWeb* platform. In particular, the repeated discussion of certain issues in much detail and in both professional and laymen's terms to include all team members in the best possible way proved essential to reaching the goals of the first project phase.

For all team members, it proved essential to clarify the technical and data-oriented terminology and the needs of both DH specialists and humanities scholars to establish data models capable of reflecting the complexity of linguistic data. For instance, the complexity and variety of the textual layers (including different versions of the original text according to different orthographic conventions and sub-structures), translations as well as levels of linguistic annotations (e.g. morpho-syntactic, lexical, metric, lexicographic) was a major blind spot for the DH specialists. One challenge in this regard is that the linguists often possess tacit knowledge of certain structural facts and it takes an extended and detailed communication process to make these explicit, or to even realise the need for explicitness. In some ways, the linguists have only gained a deepened understanding of the crucial need for explicitness for the purposes of modelling in the course of the project. There have also been various further domains in which blind spots became apparent, such as the need to address the philological genesis of the texts or the need to clarify copyright and other legal issues. In this regard, the linguists have been on a learning curve given the differences in

requirements of a web-based resource in comparison to their normal academic outlets (i.e. publication of books or journal papers).

4.4 The Whole Is More than the Sum of Its Parts

"Did the combination of thinking from a DH research question and thinking from a technical solution lead to new insights?" The integration of textual and lexicographic data, as well as various types of linguistic annotation layers into a digitized and standardized format modelled in TEI revealed manifold inconsistencies and gaps in the input. In order to address these challenges, a detailed correction of, e.g., mappings between textual data and lexical entries has been ongoing since the beginning of the *VedaWeb* project. In this way, the transfer to digital formats has led to a significant and necessary increase in quality of the linguistic data.

Besides the new possibilities with regard to, e.g., combined searches across multiple layers of text and annotations, the digitization of the Rigveda also opens up fields that have hitherto been hardly accessible to specialists of Old Indo-Aryan texts. This is true, for instance, of semantic searches that will enable a novel type of computerized research by linguists, philologists, philosophers, and other specialists of Old Indo-Aryan texts.

4.5 Towards Improved Collaboration

"How could better training or education of scholars and digital experts make collaboration easier, more effective and more efficient?" DH researchers collaborate with researchers from manifold disciplines, so that it seems unrealistic to include in their training specifics of any one humanities discipline. However, DH specialists need a general understanding of the objects of study in the humanities, of research questions and research methods, knowledge of which should certainly be included in their training. As the digital experts are all graduated in Digital Humanities (instead of, e.g., computer science), the project team members of *VedaWeb* have such training and in part previous experience with specifically linguistic projects, and were accordingly ideally prepared. Any specific insights that go beyond this general background was then gained in the course of the project through the outlined processes of regular and detailed communication and collaboration.

Similarly, detailed training in DH methods, approaches and resources may not be possible or necessary to include in a linguistics training. However, a general understanding of the different fields of DH work is certainly highly desirable, especially given the intense, ongoing digitization of research in the humanities disciplines. For instance, the fact that there is a significant difference between building a web application, on the one hand, in contrast to ensuring sustainability of data by modelling them in TEI, on the other hand, is an insight into the nature of DH projects that is likely to benefit collaboration from early project stages. The insufficient insights of the humanities researchers into these components of technical DH work was a hindering factor in the early stages of *VedaWeb* and a training for humanities researchers that includes a general understanding of these and other

general structures and processes is likely to be beneficial for collaborative work across disciplines. In particular, this would involve developing a deepened understanding for the need of modelling and formalizing knowledge structures, making them explicit and thus usable for computational implementation. Moreover, there is the need for a basic understanding of the interaction of different technical components including data formats, databases, search engines, program logic, web interfaces and usability.

5 Conclusion

We have described an instance of the "humanities problem" for linguists of handling large textual corpora, in this case the corpus of Old Indo-Aryan texts. In this project, the "technical solution" consists of developing a formal data model for the given knowledge domain and the construction of a web-based research platform that integrates multiple layers of textual data (different versions of the original text and various translations) with linguistic annotations (e.g. morpho-syntactic and metrical annotations) as well as with lexicographic data via APIs to specialized dictionaries. Half of the funding time having elapsed, it has become clear that it is not only helpful to collaborate closely among the different team members, but that a continuous and intense process of communication, exchange and replanning is crucial for the realization of the project. In particular, regular discussion to clarify terminology, on the one hand, and the uncovering of blind spots among team members, on the other hand, have proved indispensable. By engaging in such close collaboration, the project has been able to stay in time with its planned schedule despite a long initial phase of finding a common understanding and language, which is in part ongoing. Above and beyond the communicative practices adopted, the project benefits significantly from a diverse team with a variety of backgrounds and competences. A further advantage is that the DH specialists in the team are already experienced in handling linguistic data. However, we believe that interdisciplinary project teams in general can reach a comparable level of mutual understanding when embracing particularly close collaboration [1, 5]. This involves the continuous discussion and clarification of terminology, of data formats and of models, as well as short feedback loops to create a shared understanding of humanities research questions and to reach tailored DH solutions.

References

1. Deegan, M., McCarty W. (eds.): Collaborative Research in the Digital Humanities. Routledge, Farnham (2011).
2. Geldner, K.F.: Der Rig-Veda. Aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen von Karl Friedrich Geldner. Harvard University Press, Cambridge (Mass.) (2003)[1951–57].
3. Grassmann, H.: Wörterbuch zum Rig-veda. O. Harrassowitz, Wiesbaden (1873).
4. Grassmann, H.: Rig-veda. Übersetzt und mit kritischen und erläuternden Anmerkungen versehen von Hermann Grassmann. F.A. Brockhaus, Leipzig (1876).

5. Griffin, G., Hayler, M.S: Collaboration in Digital Humanities Research – Persisting Silences. Digital Humanities Quarterly 12(1) (2018). http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html
6. Griffith, R.T.H.: The Hymns of the Rigveda. Lazarus, Benares (1896).
7. Haig, G., Schnell, S.: Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual Version 7.0 (2014), https://www.academia.edu/10328418/Haig_Geoff_and_Stefan_Schnell._2014._Annotations_using_GRAID_Grammatical_Relations_and_Animacy_in_Discourse_._Manual_Version_7.0, last accessed 2019/02/11.
8. Hellwig, O., Hettrich H., Modi A., Pinkal M.: Multi-layer annotation of the R̥gveda. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7.-12.5. 2018. European Language Resources Association (ELRA), Paris (2018).
9. Hintze, A.: „Lohn“ im Indoiranischen: eine semantische Studie des Rigveda und Avesta. Beiträge zur Iranistik 20. Reichert, Wiesbaden (2000).
10. Kazzazi, K.: „Mann“ und „Frau“ im R̥gveda: mit einem Exkurs über Wörter für „Frau“ im Atharvaveda. Innsbrucker Beiträge zur Sprachwissenschaft 98. Inst. für Sprachwissenschaft, Innsbruck (2001).
11. Keydana, G., Luraghi, S.: Definite referential null objects in Vedic Sanskrit and Ancient Greek. Acta Linguistica Hafniensia 44(2), 116–28 (2012). doi: 10.1080/03740463.2013.776245
12. Kölligan, D.: Vedisch *nigut-* ‘Feind, Übeltäter’. Historische Sprachforschung 120, 134–141 (2007). www.jstor.org/stable/40849294
13. Kulikov, L.: The Vedic -ya-Presents: Passives and Intransitivity in Old Indo-Aryan. Rodopi, Amsterdam (2012).
14. Oldenberg, H.: R̥gveda. Textkritische und exegetische Noten. Weidmann, Berlin (1909/1912).
15. Reinöhl, U.: Grammaticalization and the Rise of Configurationality in Indo-Aryan. Oxford University Press, Oxford (2016).
16. Renou, L.: Études védiques et paninéennes. Boccard, Paris (1956–1969).
17. Roesler, U.: Licht und Leuchten im Rgveda. Untersuchungen zum Wortfeld des Leuchtens und zur Bedeutung des Lichts. Indica et Tibetica Verlag, Swisttal-Odendorf (1997).
18. Ryan, K.: Onsets contribute to syllable weight: statistical evidence from stress and meter. Language 90(2), 309–341 (2014).
19. Scarlata, S., Widmer P.: Vedische exozentrische Komposita mit drei Relationen. Indo-Iranian Journal 58(1), 26–47 (2015).
20. Schmid, W.P.: Die Wurzel *VI* im Rgveda. In: Mélanges d’indianisme. à la mémoire de Louis Renou, pp. 613–624. de Boccard, Paris (1968).
21. Widmer, P.: Der altindische vrkī́-Typus und hethitisch nakki-: Der indogermanische Instrumental zwischen Syntax und Morphologie. Die Sprache, Zeitschrift für Sprachwissenschaft 45(1–2), 190–208 (2005).
22. Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., Stegmann, J.: SusTEInability of linguisitc resources through feature structures. Literary and Linguistic Computing 24(3), 363–372 (2009).