The Vedic Compound Dataset

Sven Sellmer, Oliver Hellwig
Institute of Oriental Studies, Adam Mickiewicz University Poznań
Department of Comparative Language Science, University of Zürich
sven@amu.edu.pl, oliver.hellwig@uzh.ch

Abstract

This paper introduces the Vedic Compound Dataset (VCD), the first resource providing annotated compounds from Vedic Sanskrit, a South Asian Indo-European language used from ca. 1500 to 500 BCE. The VCD aims at facilitating the study of language change in early Indo-Iranian and offers comparative material for quantitative cross-linguistic research on compounds. The process of annotating Vedic compounds is complex as they contain five of the six basic types of compounds defined by Scalise and Bisetto (2005), which are, however, not consistently marked in morphosyntax, making their automatic classification a significant challenge. The paper details the process of collecting and preprocessing the relevant data, with a particular focus on the question of how to distinguish exocentric from endocentric usage. It further discusses experiments with a simple ML classifier that uses compound internal syntactic relations, outlines the composition of the dataset, and sketches directions for future research.

Keywords: Compounding, Sanskrit, Vedic, Dependency annotation

1. Introduction

Since the beginnings of modern linguistics, Sanskrit compounds have played a special role in research on compounding (see, e.g., Wujastyk, 1982), which is reflected by the fact that even some terms of the Indian grammatical tradition have entered current linguistic terminology (see Tab. 1). Sanskrit – especially its oldest form, known as Vedic, which was used from ca. 1500–500 BCE – is also of fundamental importance for Indo-European and cross-linguistic studies. Up until now, there exists a substantial collection of annotated compounds for classical and Neo-Sanskrit.[1] Many of these annotations, originating from works composed in the 19th and 20th c. CE, offer, however, only limited insights for historical linguistics due to their relatively recent composition. The Vedic Compound Dataset (VCD) introduced in this paper is the first resource to provide annotated compounds from Vedic, making it particularly well-suited for studying language change in the formative period of Sanskrit.[2]

2. Previous research

Quantitative cross-linguistic research on compounds has been less intensive than in other areas of linguistics (Moyna, 2019), but Guevara and Scalise (2009) have recently produced valuable statistics, in which, however, data for Indo-Aryan as well as ancient languages are lacking. The VCD fills both of these gaps to some extent and yields relevant comparative material (see Sec. 5).

Concerning annotation, Vedic compounds constitute an interesting challenge. As will be discussed in Section 4, they contain five of the six basic types of compounds that are used in Guevara and Scalise 2009. In addition, neither the compound internal relation between the words constituting them (see the examples in Section 3) nor the relation between a compound and the rest of the sentence are consistently marked in morphosyntax, which poses a challenge to their automatic classification. Over the past decade, several attemps at automatic classification of classical Sanskrit compounds have been undertaken. While Krishna et al. (2016) obtain 74% F-score for a dataset with four coarse compound categories by applying a Random Forest classifier to a set of manually defined linguistic markers, Sandhan et al. (2019) achieved a comparable F-score of 73% using an approach that combined a recurrent architecture with static word embeddings, bypassing the need for extensive feature engineering. Most recently, Sandhan et al. (2022) argued that compound classification needs to take syntactic properties of the surrounding text into account. They therefore combined compound classification with morphosyntactic tagging and dependency parsing in a joint learning task. Using a deep learning architecture with contextualized word embeddings, they report an F-score of 85.7% for coarse compound classification.

While these contributions have significantly advanced automatic Sanskrit compound classification, the present study did not use these systems for compound annotation for several reasons. Firstly, previous studies used classical Sanskrit data, but our focus is on Vedic compounds. The significant lexical differences between Vedic and classical Sanskrit can make applying these systems

[1] https://sanskrit.uohyd.ac.in/Corpus/

[2] The VCD is available at https://github.com/SvenSellmer/VedicCompoundDataset.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 50–55. May 25, 2024.

to Vedic texts problematic. Secondly, while an F-score of 85.7% is remarkable, it does not meet the high standards required for creating a reference dataset. Thirdly, the compound categories employed by these studies do not encompass all categories proposed by Bisetto and Scalise, limiting their applicability to our research. In what follows, we will present how we collected and prepared our data (Sec. 3 and 4), devoting particular attention to the recognition of their endocentric-exocentric dimension, and discuss experiments with a simple ML classifier. We will then discuss the composition of the dataset (Sec. 5) and draw conclusions for future research (Sec. 6).

3. Data collection

Our data is derived from two closely linked resources. The Digital Corpus of Sanskrit (Hellwig, 2010–2024) offers lexical and morphosyntactic annotations for Vedic and classical Sanskrit texts. Within the DCS, compounds that have a non-lexicalized reading (see below) are divided into their constituent parts. For instance, the coordinate compound *indrāgni-* 'Indra and Agni' is separated into the words *indra-* and *agni-*, each with its own morphosyntactic information. This preprocessing of the source data makes the identification of compounds significantly easier. The Vedic Treebank (VTB, Hellwig et al. 2020), containing approximately 32,000 sentences, supplements the DCS with a layer of Universal Dependencies (UD) annotations. The syntactic annotation of the VTB was carried out by a team of experts, who employed enhanced annotation guidelines (see Hellwig et al., 2023).

The standard UD guidelines offer only limited possibilities for a differentiated treatment of compounds,[3] which is unsatisfactory in view of the versatile role of compounds as an interface between syntax and lexicon and especially of the fact that Vedic compounds – like Sanskrit compounds in general (Lowe, 2015) – contain various syntactic structures, which tend to become diachronically increasingly complex. Therefore, the team extended the annotation guidelines (Biagetti et al., 2020) with the aim of enabling the annotator to make explicit the internal syntactic structure of a compound in the same way as UD labels show the relations obtaining between the words in a sentence. For instance, the compounds *indra-agni-*[4] 'Indra and Agni' (as a pair), *deva-loka-* 'world of the gods', and *ardha-māsa-* 'half-month' are annotated as follows:

The information – not immediately obvious in the latter two examples – that a word is a non-final member of a compound was incorporated into the VTB via the "Compound" feature (to be distinguished from the label `compound`, which is only used for coordinate compounds in the VTB). Compounds can include a limited number of particles and adverbs in addition to nominal forms (e.g. *sa-ratha-*, lit. 'with-chariot', i.e. "having a chariot"). Most of these indeclinables do not exist as standalone words. Since they constitute a closed lexical set, they can be directly integrated into compound detection. Adverbs that are part of compounds but do not belong to this closed set (e.g. *su-* 'well', which also occurs independently) were addressed individually during annotation.

To detect compounds in the VTB, we conducted a scan of the VTB's conllu file for instances of the "Compound" feature and built compounds by tracing the syntactic arcs of the non-terminal members until we reached an inflected word form, which had to be the terminal member. In the example *deva-loka-* given above, *deva-* is labeled with the Compound feature in the VTB. Following the arc with the `nmod` label, we arrive at *loka-* which has an inflectional ending in a real world case and thus must constitute the terminal member of the compound.

The VCD is specifically designed to contain only two-word compounds. Apart from time restrictions, this focus is due to the fact that longer compounds of n words can typically be analyzed as multi-level mixed types consisting of $n-1$ elements. Furthermore, the oldest Vedic texts predominantly contain two-word compounds (see e.g. Macdonell, 1910, 143). By limiting our data selection to these short compounds, we ensure that our data covers the entire Vedic period. We equally did not include compounds that were identified as lexicalized by the annotator of the DCS. These compounds are typically technical terms. For instance, the term *agnihotra-* is a compound of the words *agni-* 'sacrificial fire' and *hotra-* 'sacrificial libation'. However, an *agnihotra-* is not merely a 'libation into the sacrificial fire', but a specific type of such a libation (see e.g. Renou, 1953). Despite their semantic transparency, such lexicalized compounds are annotated as single words in the DCS and are not identified as compounds in its dictionary. As a result, we lack access to information indicating that *agnihotra-* is a compound, as well as its internal syntactic relation. The integration of such lexicalized compounds into the VCD remains an open issue for future research.

[3] See https://universaldependencies.org/docs/en/dep/compound.html.

[4] For convenience, all euphonic ('sandhi') changes have been removed in this paper.

	Endocentric	Exocentric
C.	Austria-Hungary (*dvandva*)	[lacking in E. and S.]
S.	horse-sacrifice (*tatpuruṣa*)	horse-faced (*bahuvrīhi*)
A.	blackbird (*karmadhāraya*)	redneck (*bahuvrīhi*)

Table 1: Compound classification according to Scalise and Bisetto 2005 (C. = coordinate; S. = subordinate; A. = attributive); indigenous terms in brackets.

Internal label	Compound type
`compound:coord`	→ coordinate
`nmod`, `obj`, `obl`, `iobj`	→ subordinate
`advmod`, `amod`, `nummod`, `acl`, `det`, `xcomp`, `nmod:appos`, `advcl`	→ attributive

Table 2: Internal labels and the dimension coordinate/subordinate/attributive.

4. Compound classification

Among the various possible classification schemes for compounds, we adopted the version proposed by Scalise and Bisetto (2005), for three reasons: 1. it is not only well-suited for Sanskrit (as can be seen in Biagetti, 2024) but also for many other languages; 2. it has already been employed in cross-linguistic studies (Scalise and Guevara, 2006; Guevara and Scalise, 2009), so that it facilitates the reusability of our dataset in such contexts; 3. it is convenient due to its conceptual simplicity, as opposed to its refined version in Scalise and Bisetto, 2009, which was too finegrained for our time budget.

This scheme has two-dimensions, as exemplified by the rows and columns of Tab. 1. Firstly it classifies compounds as endo- vs. exocentric, where an exocentric compound is understood as one that is not a hyponym of its formal head (see Bauer, 2017, 37, e.g., a redneck is not a kind of neck; for other definitions see Bauer, 2008 and Moyna, 2019). In Sanskrit grammar, these are called *bahuvrīhi*s: compounds that, as a whole, are (sometimes secondarily nominalized) adjectives though their final member is a noun. Secondly, it encodes the relation between the first and the final member, which may either be coordinate (i.e., dvandvas in the strict sense, Ralli 2019), subordinate, or attributive.

For the actual task of compound classification, the VCD provides the following information:

- UD label of the compound as a whole (i.e., of its final member)
- internal UD label
- POS information for both members
- case and gender of the final member

For classification according to this scheme we used an algorithm that was partly rule-based, and partly relied on human expertise:

1. The distinction between coordinate, subordinate, and attributive compounds could easily be made on the basis of the internal label, as shown in Table 2.

2. Coordinate compounds being endocentric by default in Sanskrit, subordinate und attributive compounds were then classified under the aspect of their exocentricity.

2a. In about 1/5 of the cases, this can be done automatically,[5] namely, where a mismatch between the gender of the compound and the gender of its final member as an independent noun can be observed. For instance, the compound *aśva-mukha-* can be either endocentric ('face of a horse') or exocentric ('horse-faced'). Now, *mukha-* 'face' is a neutral noun, so wherever *aśva-mukha-* features a non-neuter ending it must refer as an adjective to a masculine or feminine noun (e.g., *aśva-mukhaḥ rākṣasaḥ* 'a horse-faced demon') and so be an exocentric compound.

2b. Further, a sizeable subgroup of exocentric compounds (ca. 600 tokens) could be classified on the basis of their morphology: the so-called root or synthetic compounds with a verbal root noun as final member are always exocentric (Scarlata, 1999); e.g., *prathama-ja* 'first-born', from $\sqrt{jan}$ 'to be born'. Detecting them could not be fully automatized as there is no appropriate POS tag in the DCS flagging them as verbal roots.

2c. In the remaining ca. 1,700 cases, the decision to classify a given compound as exo- or endocentric could only be made by a human expert on the basis of its use in the actual context. Dictionary information could be used in cases in which the translation indicated exclusively exo- or endocentric usage. But such hints were not available for all compounds, and in addition turned out to be not always reliable. Opposite to what one may expect, the UD label of the final compound member did not allow to decide between exo- and endocentric usage. For example, in their prototypical role as adnominal modifiers, *bahuvrīhi*s are linked by `acl` to their referents (Biagetti et al., 2020, Sec.

[5] In accented texts, also the location of the accent in a compound often is indicative of exocentricity (Wackernagel, 1905, § 113), but this information was not available to us as it is lacking in the DCS.

2.7.2). However, even this label cannot serve as a reliable indicator of exocentricity, because it also appears with endocentric compounds, for instance, in relative clauses. In addition, only about 30% of all *bahuvrīhi*s are used as adnominal modifiers, as they are, for instance, often substantivized and function as independent nouns. As a consequence, the annotation of these 1,700 compounds had to be done manually, which turned the present step into the most time-consuming one.

The description of the annotation process suggests that many decisions are rule-based, i.e., can be made based on the internal and external syntactic relations of compounds and their morphosyntactic information. We hypothesized that a simple classification algorithm with access to the syntactic gold information of the VCD could learn these rules. To test this hypothesis, we implemented a multinomial regression model. The predictors for this model include the aforementioned compound-internal and external syntactic labels, as well as the part-of-speech tags and lemmata of the two words constituting a compound. The model is trained to predict which of the five classes in the scheme of Bisetto and Scalise (Table 1) a compound belongs to. The results of a tenfold cross-validation (see Table 3) show that the system achieves F-scores above 80%, even for complicated classes that involve decisions between endo- and exocentric use. As the F-scores of the two ablation tests in columns 5 and 6 of Table 3 (-I: no compound internal syntactic labels; -O: no outer labels) indicate, this success is mainly due to the availability of compound-internal syntactic relations from the VTB. While ignoring the labels that connect compounds with the rest of the sentence and thus indicate their syntactic function (-O) keeps the F-scores largely unchanged, ignoring their inner syntactic labels (-I) leads to substantially lower F-scores for three out of five types. Specifically, the low F_{-I}-score for coordinate endocentric compounds likely results from the fact that they are not distinguished by POS information from subordinate compounds, but occur with much lower frequency (see Tab. 5.) These findings can inform future research in automatic compound classification.

Type	P_{All}	R_{All}	F_{All}	F_{-I}	F_{-O}
attrib/endo	81.8	86.4	84.0	80.2	79.9
attrib/exo	81.0	80.9	80.9	74.8	80.0
coord/endo	97.6	98.6	98.1	29.6	98.1
subord/endo	91.4	92.3	91.8	82.3	90.5
subord/exo	87.2	81.1	84.1	80.0	78.8

Table 3: Results of the multinomial classifier for compound types, 10-fold cross-validation. All: all predictors, -I: no internal syntactic labels, -O: no outer labels.

Deprel	#Tok.	Deprel	#Tok.
`nmod`	2260	`nummod`	574
`advmod`	1089	`obl`	460
`amod`	800	`acl`	191
`obj`	721	`det`	189
`compound:coord`	632	`iobj`	26

Table 4: Most frequent compound-internal dependency relations in the VCD.

5. Composition of the dataset

The VCD contains almost 7,000 two-word compounds together with information on morphology, internal and external syntactic relations, chronology, and Vedic subtraditions. A few plots and tables may serve to give an overview of the composition of our dataset. Tab. 4 lists the most frequent compound internal labels in the VCD. It thus gives insights into the syntactic processes active during compounding and so can serve as a starting point for cross-linguistic comparison and for the construction of fine-grained semantic frames. Tab. 5 shows the distribution of the tokens over Scalise and Bisetto's classification. The numbers confirm the general cross-linguistic observations in Guevara and Scalise 2009, 118–119, that in terms of frequency S. > A. > C. Regarding the endo-/exocentric distinction, exocentric compounds make up 41.8% of all compounds in the VCD. This is a remarkably high percentage compared with the statistics in Scalise et al. 2009, where this ratio ranges from 8.4% (Germanic languages) to 35.4% (Romance languages). The ratio for Vedic gets even higher when the diachronic dimension of our dataset is taken into account. As can be seen in Fig. 1, right, it drops from an extreme ratio of 72.4% in the archaic Rig Vedic period, a figure reminiscent of what Bauer (2008, 68) reports for some African and Australian languages, to 30.3% in the late Sūtras. Notably, this development runs counter to the general rise in compound usage, as shown in Fig. 1, left.

	Endocentric		Exocentric		All	
	Tok.	%	Tok.	%	Tok.	%
C.	632	9.0	0	0	632	9.0
S.	2,273	32.5	1,177	16.8	3,450	49.3
A.	1,166	16.7	1,744	24.9	2,910	41.6
	4,071	58.2	2,921	41.8	6,992	

Table 5: Counts of the main compound categories (tokens) in the VCB.

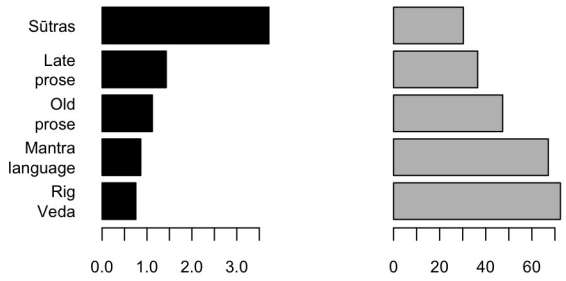

Figure 1: Percentages of compounds among all lemmata (= left) and of exocentric compounds among all compounds (= right), across Vedic literary periods; earliest in the bottom-most row.

6. Conclusion

Up until now, the diachronic, geographical and sociolinguistic development of Vedic literature remains incompletely understood (Witzel, 1997). The compounds collected in the VCD, showing clear diachronic trends regarding their endo-/exocentric dimension (see Fig. 1), thus provide valuable data for gaining deeper insights into the linguistic developments during this period as well as for time-stamping Vedic texts (Hellwig, 2024). They can further prove fruitful for comparative studies in an Indo-European and cross-linguistic framework, as they contain data about one of the earliest attested Indo-European languages.

The rule-based parts of collecting the dataset were comparatively straightforward, but to distinguish between exocentric and endocentric compounds of the attributive and subordinate types turned out to be a time-consuming process. It is important to note that this work would have been unnecessary if such a distinction could be directly established on the basis of the UD labels. It would be therefore helpful to add an appropriate UD sublabel to, e.g., `nmod` and `amod`, to indicate *bahuvrīhi*s in various languages. This would be a small extra effort, because for a human expert annotating a whole sentence it is usually evident whether a given compound is exocentric. It is to be expected that DL dependence parsers will then be able to process these annotations and to determine the exocentricity of compounds with high precision. This would be a highly desirable outcome for the research on compounds in general, as their exocentric/endocentric dimension is of fundamental importance. In addition, due to the general tendency of exocentric compounds for having a metonymic meaning (Bauer, 2008; Barcelona, 2008), such a sublabel would also be relevant for metonomy recognition.

7. Ethical considerations

We are not aware of any ethical issues arising from the composition or use of our data set.

8. Limitations

Four limitations of our dataset should be mentioned. Firstly, it must be understood that – though of considerable size for an ancient language – it is based on only about 35% of the extant Vedic literature – nevertheless, its chronologically balanced composition and the wide variety of texts it draws on make it useful for quantitative linguistic studies. Secondly, as discussed on p. 2 above, we did not consider compounds that were treated as lexicalized in the DCS. Thirdly, for the reasons explained on p. 2, we restricted ourselves to two-word compounds for the time being. We plan to overcome these limitations in future versions of the VCD. Finally, it should be noted that the POS tags taken over from the VTB are not completely reliable. In the VCD, they have been manually corrected in a number of instances, but not in the form of a systematic revision.

9. Acknowledgements

We would like to express our gratitude to three anonymous reviewers, who made a number of very helpful remarks and suggestions.

Research for this paper was funded by the German Federal Ministry of Education and Research, FKZ 01UG2121.

10. Bibliographical References

Antonio Barcelona. 2008. The interaction of metonomy and metaphor in the meaning and form of 'bahuvrīhi' compounds. *Annual Review of Cognitive Linguistics*, 6:208–281.

Laurie Bauer. 2008. Exocentric compounds. *Morphology*, 18(1):51–74.

Laurie Bauer. 2017. *Compounds and Compounding*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.

Erica Biagetti. 2024. Early Vedic compounds: A typological reappraisal. *Studies in Language*, 48(1):1–64.

Erica Biagetti, Oliver Hellwig, Salvatore Scarlata, Paul Widmer, and Elia Ackermann. 2020. Annotation guidelines for the Vedic Treebank. v2.0 – July 2020.

Emiliano Guevara and Sergio Scalise. 2009. Searching for universals in compounding. In Sergio Scalise, Elisabetta Magni, and Antonietta Bisetto, editors, *Universals of language today*, pages 101–128. Springer, Amsterdam.

Oliver Hellwig. 2010–2024. DCS - The Digital Corpus of Sanskrit.

Oliver Hellwig. 2024. To compound or not to compound? A diachronic Bayesian analysis of compounds and equivalent constructions in Vedic Sanskrit. *Indogermanische Forschungen*, 129.

Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of Vedic Sanskrit. *Language Resources and Evaluation*, 57:1173–1206.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of Vedic Sanskrit. In *Proceedings of the LREC*, pages 5139–5148.

Amrith Krishna, Pavankumar Satuluri, Shubham Sharma, Apurv Kumar, and Pawan Goyal. 2016. Compound type identification in Sanskrit: What roles do the corpus and grammar play? In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 1–10, Osaka, Japan. The COLING 2016 Organizing Committee.

John J. Lowe. 2015. The syntax of Sanskrit compounds. *Language*, 91(3):71–115.

Arthur Anthony Macdonell. 1910. *Vedic Grammar*. Trübner, Strassburg.

María Irene Moyna. 2019. Exocentricity in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Angela Ralli. 2019. Coordination in compounds. In *Oxford Research Encyclopedias – Linguistics*. Oxford University Press.

Louis Renou. 1953. *Vocabulaire du rituel védique*. Librairie C. Klincksieck, Paris.

Jivnesh Sandhan, Ashish Gupta, Hrishikesh Terdalkar, Tushar Sandhan, Suvendu Samanta, Laxmidhar Behera, and Pawan Goyal. 2022. A novel multi-task learning approach for context-sensitive compound type identification in Sanskrit. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4071–4083, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jivnesh Sandhan, Amrith Krishna, Pawan Goyal, and Laxmidhar Behera. 2019. Revisiting the role of feature engineering for compound type identification in Sanskrit. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 28–44, IIT Kharagpur, India. Association for Computational Linguistics.

Sergio Scalise and Antonietta Bisetto. 2005. The classification of compounds. *Lingue e Linguaggio*, 2:319–332.

Sergio Scalise and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford handbook of compounding*, Oxford handbooks in linguistics, pages 34–53. Oxford University Press, Oxford ; New York.

Sergio Scalise, A. Fábregas, and F. Forza. 2009. Exocentricity in compounding. *Gengo Kenkyu*, 135:49–84.

Sergio Scalise and Emiliano Guevara. 2006. Exocentric compounding in a typological framework. *Lingue e linguaggio*, 2:185–206.

Salvatore Scarlata. 1999. *Die Wurzelkomposita im Ṛg-Veda*. Reichert, Wiesbaden.

Jakob Wackernagel. 1905. *Altindische Grammatik. Band II, 1: Einleitung zur Wortlehre. Nominalkomposition*. Vandenhoeck & Ruprecht, Göttingen.

Michael Witzel. 1997. The development of the Vedic canon and its schools: The social and political milieu. In Michael Witzel, editor, *Inside the Texts, Beyond the Texts*, pages 257–345. Department of Sanskrit and Indian Studies, Harvard University, Cambridge, Mass.

Dominik Wujastyk. 1982. Bloomfield and the Sanskrit origin of the terms 'exocentric' and 'endocentric'. *Historiographia Linguistica*, 9(1):179–184.