THE VEDIC COMPOUND DATASET

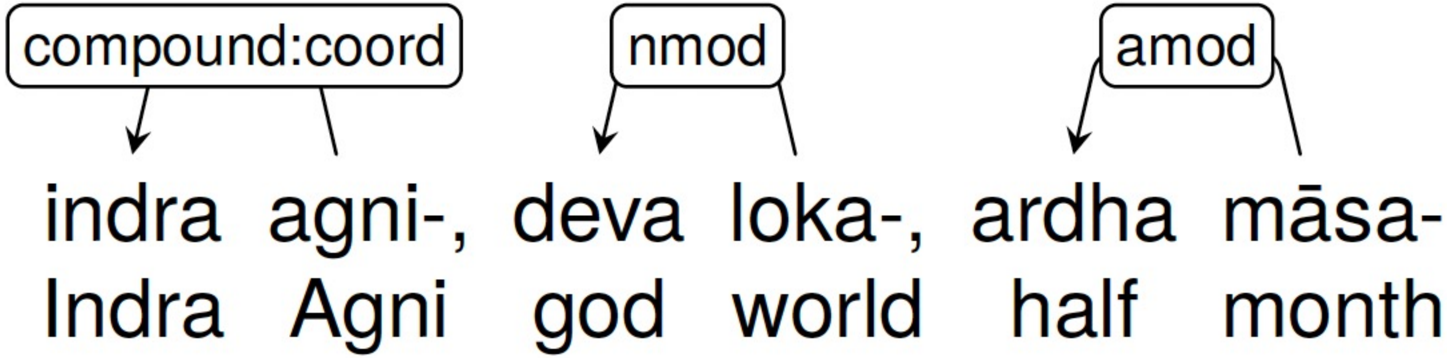
Sven Sellmer
Institute of Oriental Studies
Faculty of Modern Languages and Literatures
Adam Mickiewicz University Poznań

Oliver Hellwig
Institute of Oriental Studies
Faculty of Modern Languages and Literatures
Zurich University

Resources

- Digital Corpus of Sanskrit (http://www.sanskrit-linguistics.org/dcs/index.php)
- Vedic Treebank (https://github.com/UniversalDependencies/UD_Sanskrit-Vedic/tree/master)

Compounds in the VTB


compound:coord
nmod
amod
indra agni-, deva loka-, ardha māsa-
Indra Agni god world half month

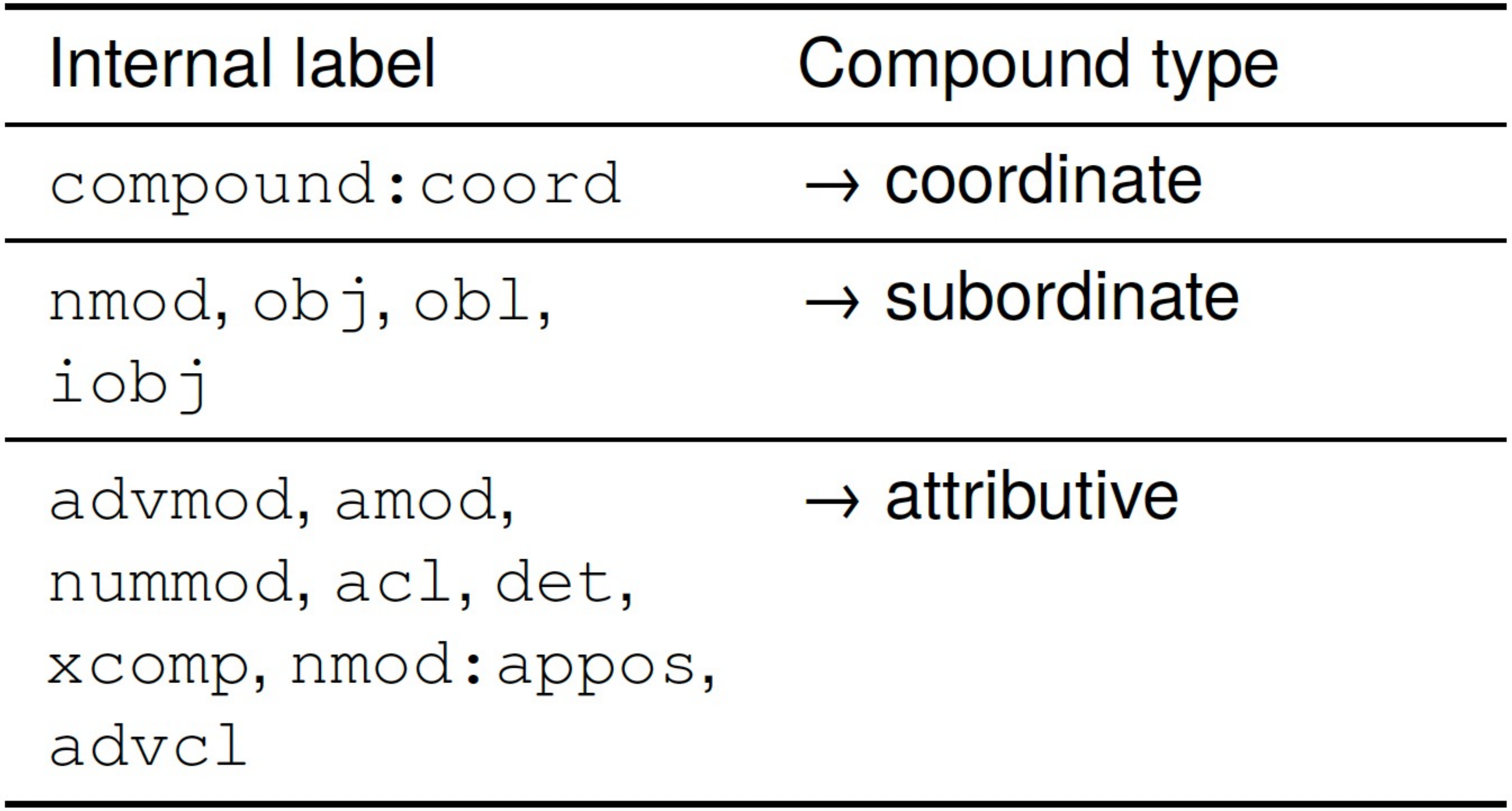
Compound classification according to Scalise & Bisetto 2005

	Endocentric	Exocentric
Coordinate	Austria-Hungary	[lacking in Engl. + Skt.]
Subordinate	horse-sacrifice	horse-faced
Attributive	blackbird	redneck

Raw information for compound classification in the VCD

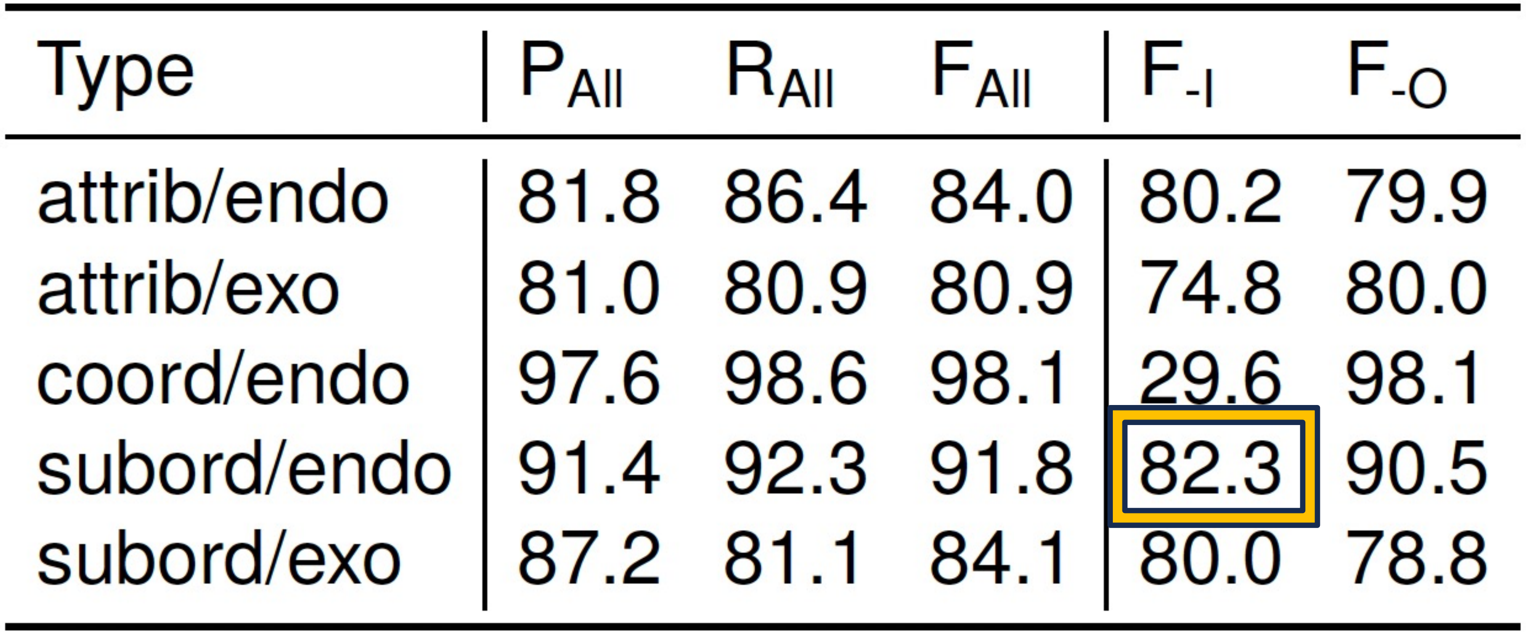
- UD label of the compound as a whole (i.e., of its final member)
- internal UD label
- POS information for both members
- case and gender of the final member

Label → Compound type



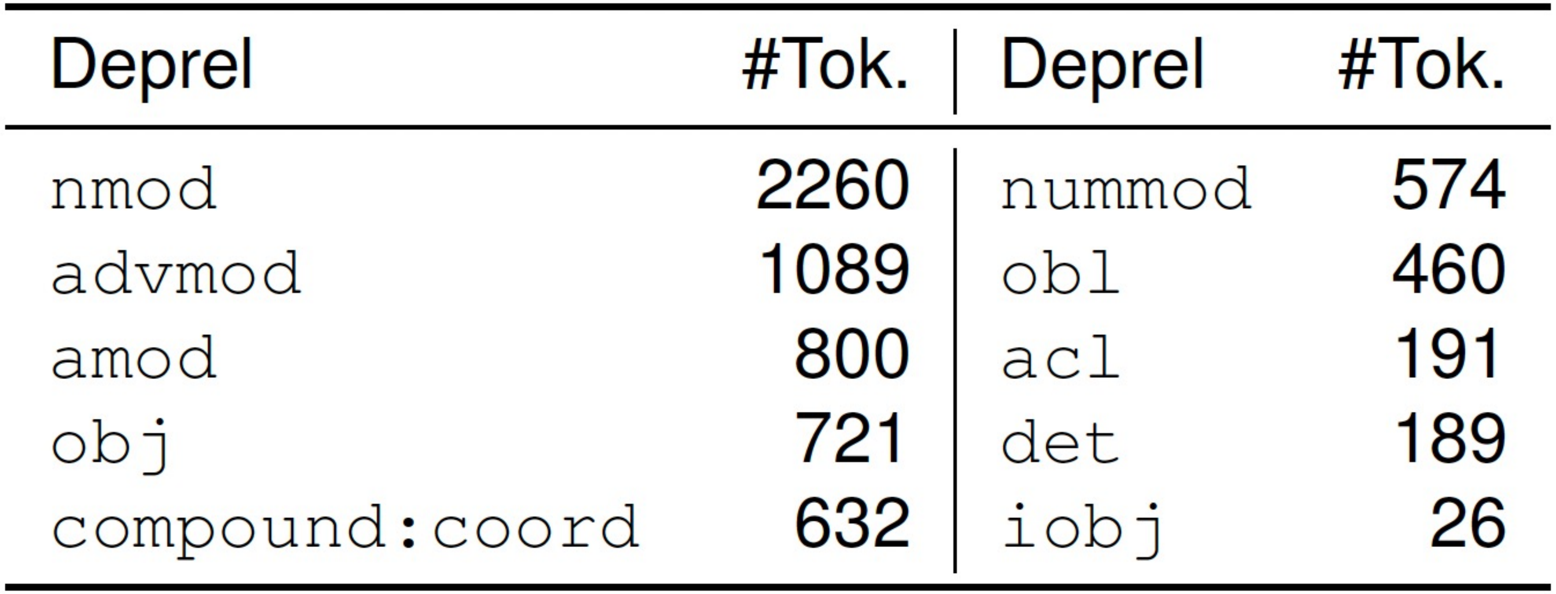
Internal label	Compound type
compound:coord	→ coordinate
nmod, obj, obl, iobj	→ subordinate
advmod, amod, nummod, acl, det, xcomp, nmod:appos, advcl	→ attributive

Results of ML classifier



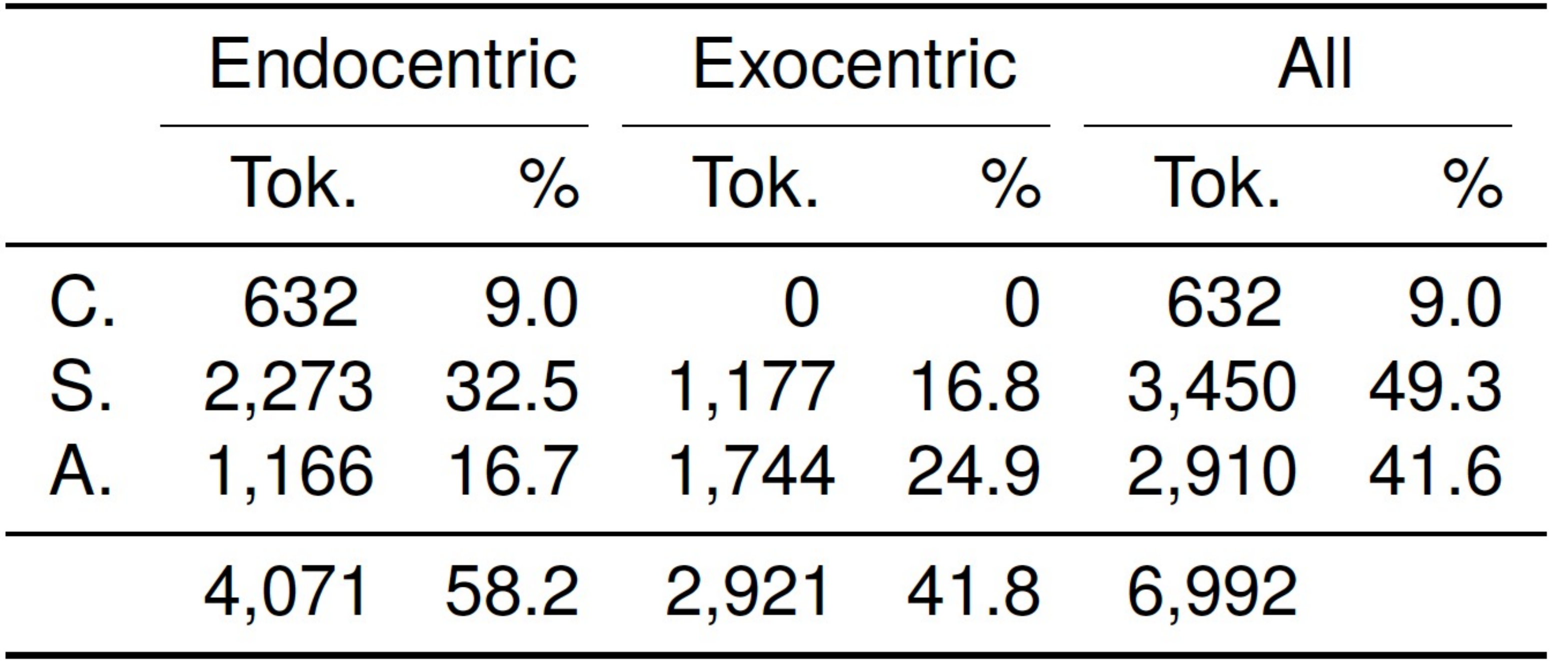
Type	P_{All}	R_{All}	F_{All}	F_{-I}	F_{-O}
attrib/endo	81.8	86.4	84.0	80.2	79.9
attrib/exo	81.0	80.9	80.9	74.8	80.0
coord/endo	97.6	98.6	98.1	29.6	98.1
subord/endo	91.4	92.3	91.8	82.3	90.5
subord/exo	87.2	81.1	84.1	80.0	78.8

Most frequent compound-internal dependency relations in the VCD



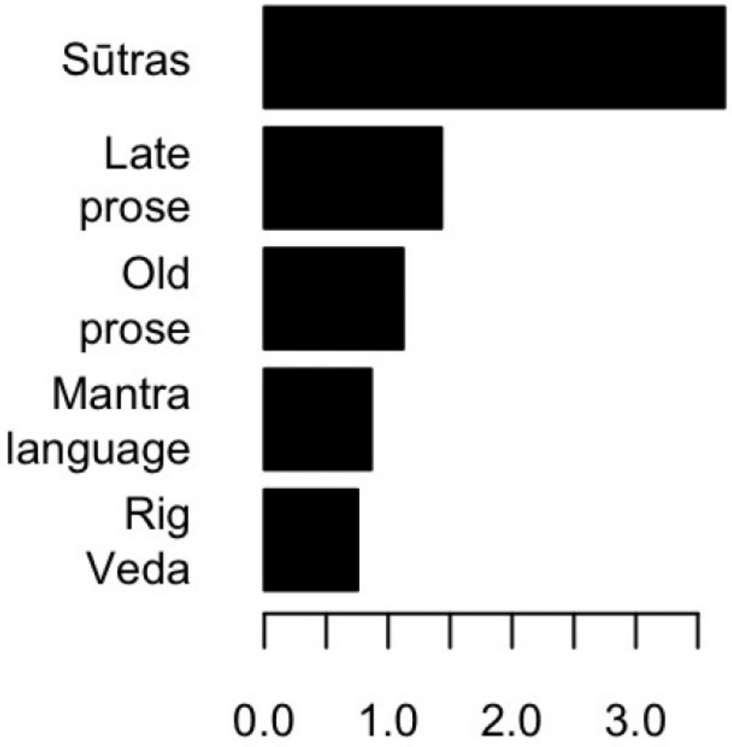
Deprel	#Tok.	Deprel	#Tok.
nmod	2260	nummod	574
advmod	1089	obl	460
amod	800	acl	191
obj	721	det	189
compound:coord	632	iobj	26

Main compound categories in the VCD

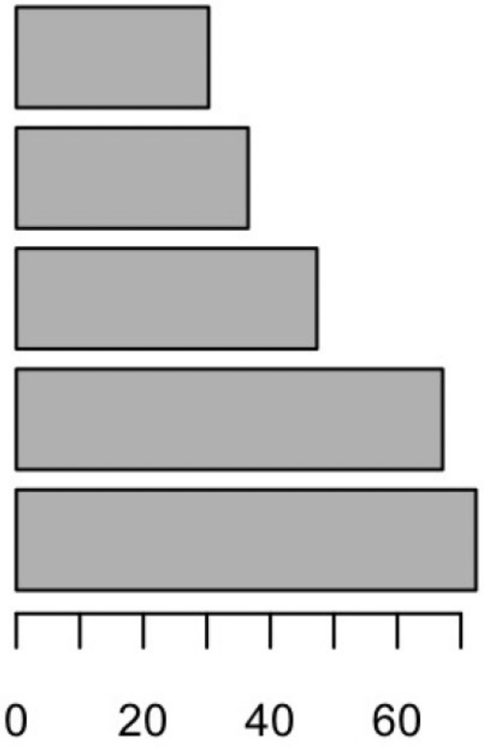


	Endocentric		Exocentric		All	
	Tok.	%	Tok.	%	Tok.	%
C.	632	9.0	0	0	632	9.0
S.	2,273	32.5	1,177	16.8	3,450	49.3
A.	1,166	16.7	1,744	24.9	2,910	41.6
	4,071	58.2	2,921	41.8	6,992	

Compound chronology



Compounds in all lemmata



Exoc. compounds in all cpds.

CONCLUSIONS

The Vedic Compound Dataset is available at:

https://github.com/SvenSellmer/VedicCompoundDataset

Acknowledgments

- 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.

THANK YOU VERY MUCH
FOR YOUR KIND ATTENTION!