The Forest Lion and the Bull: Morphosyntactic Annotation of the Panchatantra

Puneet Dwivedi, Daniel Zeman

Faculty of Mathematics and Physics, Charles University, Praha
Czechia

puneet.iitkgp1094@gmail.com, zeman@ufal.mff.cuni.cz

Abstract. We present the first freely available dependency treebank of Sanskrit. It is based on text from Panchatantra, an ancient Indian collection of fables. The annotation scheme we chose is that of Universal Dependencies, a current de-facto standard for cross-linguistically comparable morphological and syntactic annotation. In the present paper, we discuss word segmentation issues, morphological inventory and certain interesting syntactic constructions in the light of the Universal Dependencies guidelines. We also present an initial parsing experiment.

Keywords. Dependency syntax, morphology, word segmentation, tokenization, treebank, Sanskrit.

1 Introduction

Universal Dependencies (UD)[1] [8] is a project that defines a common annotation of part-of-speech tags, morphology and dependency syntax, applicable to many languages. It also takes care of collecting and releasing treebank data adhering to the UD standard. In terms of number of languages, UD has probably become the largest collection of freely available treebanks in the world: the latest release, UD 2.1 [7], contains 102 treebanks in 60 different languages (the first release in January 2015 consisted of 10 languages). The set already includes some classical languages of Europe (Ancient Greek, Latin, Gothic, Old Church Slavonic), as well as five modern Indian languages: Hindi, Urdu, Marathi, Tamil and Telugu. The present work is the first step towards extending UD with one of the oldest attested Indo-European languages, Sanskrit.

Sanskrit is the classical language of India and the liturgical language of Hinduism, Buddhism, and Jainism. It is also one of the official languages of India, despite the fact that it is rarely (if at all) used in everyday communication.

Sanskrit does not have a treebank of reasonable size so that data-driven approaches to parsing could be used. [4] mentions a Sanskrit treebank of around 3000 sentences (mostly modern short stories), reportedly developed under a Government of India sponsored project in 2008–2012. However, we have no knowledge about this corpus being publicly available. Our aim is to lay foundations of a corpus that will be available to everyone under a free license. The annotated part is small at present, but we are extending it and, more importantly, the resource is open for everyone to contribute. The history of the UD project has shown that presence of a language, even if incomplete, motivates people to get involved.

One peculiarity of Sanskrit processing is the non-trivial word segmentation [5]. For a long time, oral transmission played a dominant role in preserving and spreading Sanskrit stories; if they were eventually written down, the writing system closely followed pronunciation. Unlike Chinese or Japanese, Sanskrit texts do have spaces between words—just not always. Word sequences that are pronounced together are written together, too. Some of them are long compounds and can be processed as single words, but in general it is not necessary that the words within a

[1] http://universaldependencies.org/

ISSN 1405-5546
doi: 10.13053/CyS-22-4-3076

segment are syntactically or semantically related. Furthermore, a typical segment is not just a pure concatenation of words. Euphonic changes (called *sandhi*) take place on word boundaries and these transformations must be reversed before a word form can be isolated and morphologically analyzed.

2 Data

Our corpus is based on *Pañcatantra*, an ancient Indian collection of interrelated fables by Vishnu Sharma.[2] The Sanskrit text is also available from Wikisource[3] and from the Sanskrit Documents website;[4] note however that the exact wording at these sources sometimes differs.

The *Pañcatantra* corpus amounts to 20K tokens with morphological annotation converted to the UD annotation style. A small part (the preface about creation of Pañcatantra, and the beginning of the first book called *Mitrabheda*) has been manually checked, syntactically annotated and released in UD; then the rest was parsed by a parser trained on the core portion, to facilitate manual annotation and growth of the treebank.

3 Preprocessing

We used Gérard Huet's *Sanskrit Reader Companion*[5] [2, 3] to obtain possible word segmentation and morphological features for each sentence. The segmenter provides multiple hypotheses where applicable (Fig. 1); these were manually disambiguated. In some cases we even re-combined compounds that were separated in our input data but the segmentation did not make much sense (mostly proper names like *Visnuśarmā*).

The lemma and morphological information (gender, number and case for nominals; mood, tense and number for verbs) was obtained from the Sanskrit Reader along with the correct segmentation. One of the 17 universal part-of-speech tags defined in UD was also manually assigned to each word. Finally, the data was converted to the CoNLL-U file format. The format includes a mechanism to store the mapping between the surface tokens and their segmentation to syntactic words; it is thus possible to reconstruct the original text.

The word forms and lemmas are encoded in the Devanagari script (UTF-8). Roman transliteration is also available in separate attributes.

4 Word Segmentation

The UD guidelines do not recommend splitting of compounds that are written as one orthographic word. Indeed, compounds in languages like German or Swedish are not split.

In contrast, the Sanskrit Reader shows each stem as a separate segment. Sometimes it marks the segment as a part of a compound (the yellow segments in Fig. 1). This is possible because non-final parts of compounds lack the typical case-number suffixes. In some situations the compound part can be confused with a vocative form of a noun (see Table 1); fortunately, genuine vocatives are rather rare in the data and can be identified from the syntactic context.

Table 1. Declension of the masculine noun सिंह / *simha* "lion". The citation form, which is also used in non-final components of compounds, is identical to the singular vocative form.

	Sing	Dual	Plur
Nom	*simha*	*simhau*	*simhā*
Voc	*simha*	*simhau*	*simhā*
Acc	*simham*	*simhau*	*simhān*
Ins	*simhena*	*simhābhyām*	*simhai*
Dat	*simhāya*	*simhābhyām*	*simhebhya*
Abl	*simhāt*	*simhābhyām*	*simhebhya*
Gen	*simhasya*	*simhayo*	*simhānām*
Loc	*simhe*	*simhayo*	*simhe u*

There are examples of multi-word tokens that consist solely of nominal components but they do not constitute a single compound: they contain inflectional morphemes indicating internal structure. For instance, in स्नेहः सिंहगोवृषयोर्वने /

[2] More precisely, we draw upon the Sanskrit part of a bilingual edition by Jwala Prasad Mishra from 1910, our text starts at page 2 (http://archive.org/stream/PanchatantraSanskritHindi-JpMishra1910/).

[3] https://sa.wikisource.org/s/12x

[4] http://sanskritdocuments.org/all_sa/panchatantra1_unic.html

[5] http://sanskrit.inria.fr/DICO/reader.fr.html

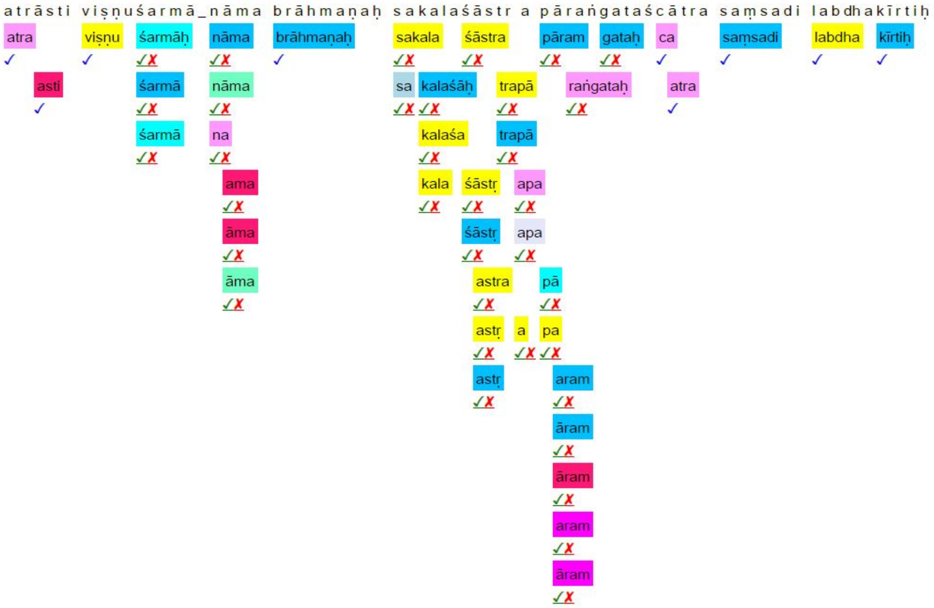

Fig. 1. An example of multiple segmentation hypotheses, as provided by the Sanskrit Reader Companion. Colors correspond to different parts of speech. Morphological analysis is also available, although not visible in this screenshot. The input string contained 7 space-delimited tokens: *atrāsti visnuśarmā nāma brāhmana sakalaśāstrapārangataścātra samsadi labdhakīrtih*. During manual disambiguation, we picked the segmentation that mostly corresponds to the top hypothesis, but we also re-combined several compounds and the result comprises 12 words: *atra asti visnuśarmā nāma brāhmana sakala śāstra pārangata ca atra samsadi labdhakīrtih*.

snehah simhagovrsayorvane "friendship of a lion and a bull in a forest" (Fig. 2), we treat *vane* separately because it is the locative modifier "in the forest", and we analyze it as modifying *snehah* "friendship". *Simhagovrsayoh* (replacing final *-r* by *-h* when reversing the sandhi euphonic changes) is a compound noun meaning "lion-bull" and sharing the genitive suffix *-yoh*. (Interestingly, the suffix also specifies the dual number because the compound *simhagovrsa* refers to two animals.) *Simha* does not have its own case morpheme, which is another indicator that *simhagovrsa* should be treated as a compound.

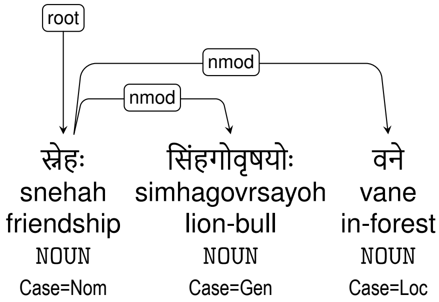

Fig. 2. "Friendship of a lion and a bull in a forest"

Sometimes we allow compound splitting because one part of the compound enters syntactic relations with words outside the compound, e.g.: धर्मोपार्जितभूरिविभवो / *dharmopārjitabhūrivibhavo* "who possessed numerous virtues and wealth" (Fig. 3). The first part, *dharmopārjita*, could be treated as a compound meaning "virtue-acquired". However, the following part is in coordination with just *dharma* "virtue", therefore we split even the compound to individual nodes.[6]

Occasionally we allow splitting of long compounds even if the external relations do not

[6] From the semantic point of view, *dharma* is a patient of *upārjita*, hence it would also seem plausible to analyze *dharma* as an object.

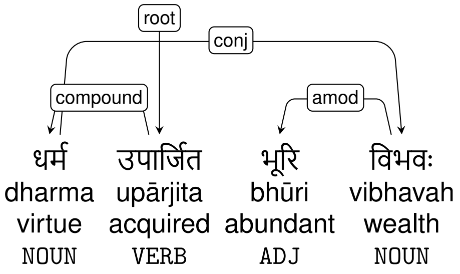

Fig. 3. "Acquired virtues and a lot of money"

provide clues shown in the previous example. In तस्य वृषभौ सञ्जीवकनन्दकनामानौ / *tasya vrsabhau sañjīvakanandakanāmānau* "his two bulls named Sanjivaka and Nandaka" (Fig. 4), the third token is a compound since both *sañjīvaka* and *nandaka* lack the *-h* suffix of the nominative.

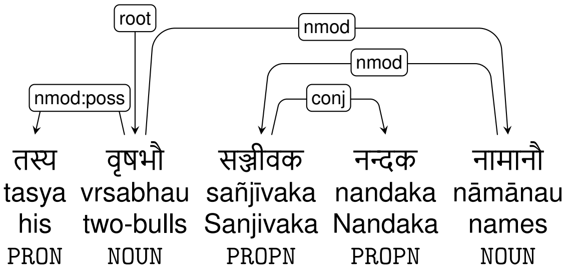

Fig. 4. "His two bulls named Sanjivaka and Nandaka"

There is only one external relation connecting the compound; however, there is internal syntactic structure between the parts and, importantly, it involves other UD relations than just `compound`. Therefore we split the token into multiple syntactic words and show the structure. Being able to recognize the named entities "Sanjivaka" and "Nandaka" as independent nodes (and match them against other occurrences of either one in the corpus) is an additional bonus.

Finally, there are multi-word tokens consisting of non-nominals and differing from patterns that are typically considered compounds. The token *sakalaśāstrapārangataścātra* in Fig. 1 is an example. Its components are *sakala śāstra pārangata ca atra* "whole science acquired and here"; they do not even form a contiguous subtree (see Fig. 5).

5 Morphology

The corpus contains 16 out of 17 "universal" part-of-speech tags defined in UD; the missing tag is `SYM` for symbols. There are two types of particles: negative (न / *na*, नहि / *nahi*, मा / *mā*) and interrogative for polar questions (किम् / *kim*). The only auxiliary verb is अस् / *as* "to be", and it is only used as copula.

We use 16 universal features: gender, number, case, degree, polarity, prontype, numtype, possessivity, reflexivity, person, politeness, verbform, mood, aspect, tense and voice. There are three genders, three numbers and eight cases (see Table 1). Besides finite verbs, there are also participles, converbs ("absolutives") and infinitives; especially the various participial forms are used frequently and do not require a finite auxiliary to accompany them. Sanskrit finite verbs can form up to six moods: indicative, conditional, optative, imperative, injunctive and benedictive. UD has defined feature values for the first four but not for injunctive and benedictive, which are specific to Sanskrit. The benedictive is extremely rare; it does not occur in our data, so we do not have to propose its annotation now.

The injunctive may have an imperative or subjunctive meaning; in particular, it is used for negative commands. There are several possibilities how to annotate the injunctive in UD:

— Define a Sanskrit-specific value Mood=Inj.

— Use Mood=Imp for both imperative and injunctive. Distinguish them by a Sanskrit-specific feature ImpType=Imp and ImpType=Inj, respectively.

— Use Mood=Jus (jussive), already defined in UD, whose meaning is also command-like, and it is also used (in Arabic) for negative commands, i.e., prohibitions.

— Use another existing UD mood with partially similar meaning, such as desiderative (Mood=Des), necessitative (Mood=Nec) or subjunctive (Mood=Sub).

We try to avoid proliferation of language-specific features, leaning against the first two options. We decided to conflate the Sanskrit injunctive with

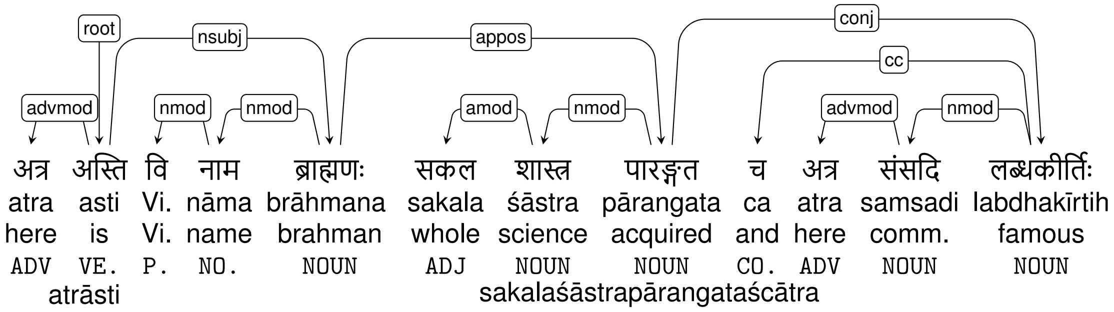
root
advmod
nsubj
nmod
nmod
appos
amod
nmod
conj
cc
advmod
nmod
अत्र अस्ति वि नाम ब्राह्मणः सकल शास्त्र पारङ्गत च अत्र संसदि लब्धकीर्तिः
atra asti Vi. nāma brāhmana sakala śāstra pārangata ca atra samsadi labdhakīrtih
here is Vi. name brahman whole science acquired and here comm. famous
ADV VE. P. NO. NOUN ADJ NOUN NOUN CO. ADV NOUN NOUN
atrāsti
sakalaśāstrapārangataścātra

Fig. 5. Dependency tree of the sentence from Fig. 1. Arthur Ryder's English translation: *There is a Brahman here named Vishnusharman, with a reputation for competence in numerous sciences.*

jussive. There is only one instance in our data: a negative command भद्र मैवं **वोचः** / *bhadra maivam* ***vocahÄ*** "Dear fellow, do not speak this way." The second word, मैवं / *maivam* is a contraction of मा एवम् / *mā evam*, where *mā* is a negative/prohibitive particle and *evam* means "this way".

The present and future tenses are tagged Tense=Pres and Tense=Fut, respectively. The perfect tense is tagged using the aspect feature (Aspect=Perf) but for the imperfect UD has two possibilities, Aspect=Imp and Tense=Imp (that is because both of them are needed in some languages); we use Tense here and we leave Aspect empty. The aorist is encoded simply as Tense=Past. Different types of participles are also distinguished by the Tense and Voice features.

6 Syntax

To estimate inter-annotator agreement and identify phenomena for which more detailed guidelines are needed, an initial portion of the dataset was annotated by two annotators and then a single annotator decided the conflicts. The two annotators agreed on 86% UPOS tags, 79% unlabeled head references, and 67% relations with labels. For short and simple sentences, the shallow Sanskrit parser[7] [4] was of some help, but unfortunately it cannot parse more complex sentences.

Being an Indo-European language, Sanskrit does not introduce phenomena that the current UD framework could not deal with. Yet we present a few examples to illustrate how certain less obvious situations are solved.

The verb अस्ति *asti* (lemma अस् *as*) is equivalent to है *hai* in Hindi and to *is* in English. It may function as copula; in accord with the UD guidelines, copulas are attached as functional modifiers of the non-verbal predicate, e.g.: कः अर्थः पुत्रेण जातेन यः न विद्वान्न न भक्तिमान् अस्ति / *kah arthah putrena jātena yah na vidvānna na bhaktimān asti* "What use having a son who is neither smart nor obedient." Here the adjective *vidvānna* "smart" is the root of the relative clause and the verb *asti* is attached to it using the relation cop (Fig. 6).

In contrast, the same verb in existential meaning takes the root position (this involves pure existentials only; locative constructions are, since the UD guidelines version 2, treated as nonverbal predicates): अत्रास्ति विष्णुशर्मा नाम ब्राह्मणः / *atrāsti visnuśarmā nāma brāhmana* "There is a Brahman here named Vishnusharman."

Infinitives are attached to the verbs that control them via the relation xcomp, which is used in UD whenever a complement clause inherits its subject from a superordinate clause, e.g.: एतस्मिन्नन्तरे ते वानराः यथेच्छया क्रीडितुम् आरब्धम् / *etasminnantare te vānarāh yathecchayā krīditum ārabdham* lit. *in-this-moment the monkeys as-with-desire to-play began*, "At the moment the monkeys began their playful frolics." The infinitive *krīditum* is attached to the past participle *ārabdham* as its controlled complement, xcomp (Fig. 7).

Occasionally it is not clear whether a sequence of clauses should be analyzed as coordination or

[7] http://sanskrit.uohyd.ac.in/scl/SHMT/index.html

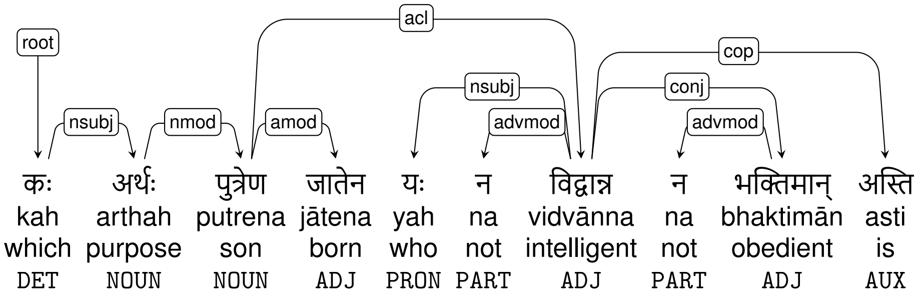

Fig. 6. Ryder's English translation: "Or why beget a son who proves a dunce and disobedient?"

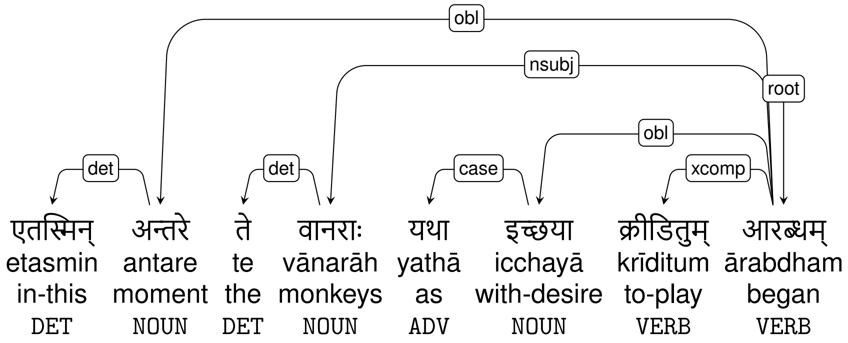
obl
nsubj
root
obl
det
det
case
xcomp
एतस्मिन् etasmin in-this DET
अन्तरे antare moment NOUN
ते te the DET
वानराः vānarāh monkeys NOUN
यथा yathā as ADV
इच्छया icchayā with-desire NOUN
क्रीडितुम् krīditum to-play VERB
आरब्धम् ārabdham began VERB

Fig. 7. Ryder's English translation: "There the monkeys began their playful frolics."

subordination. We preferred the syntactic over semantic criteria. Thus the sentence *The king listened and then spoke* is analyzed as coordination, while in *Having listened, the king spoke*, the first clause is attached as `advcl` (adverbial clause), modifying the predicate of the second clause *(spoke)*. Non-finite verb forms co-occurring with finites are indicators of subordination. The latter strategy is rather prevailing in Sanskrit (see Fig. 8 for a corpus example). Their English translation would typically use coordinate clauses but we annotate surface syntax in Sanskrit, where the first clause is subordinate. This is perfectly in the spirit of UD, which strives to preserve cross-linguistically parallel morphology and syntax, while it highlights cross-language differences when the meaning is parallel but its surface coding is not.

As in other Indo-Aryan languages, Sanskrit employs relative-correlative constructions where the relative pronoun (typically beginning in *y*-) introduces a subordinate clause, which can be analyzed as modifying a corresponding correlative / demonstrative pronoun (typically beginning in *t*-) in the matrix clause. See Fig. 9 for an example.

Some sentences have no verb; this happens mostly in *ślokas* (verse), e.g.: यस्यार्थास्तस्य मित्राणि यस्यार्थास्तस्य बान्धवाः / *yasyārthāstasya mitrāni yasyārthāstasya bāndhavāh* lit. *whose wealth his friends whose wealth his family*, meaning "One who has money, has friends; one who has money, has family." We parse this sentence as two ordinate clauses each having an embedded relative clause. Here, *yasya arthah* "whose wealth" is an adnominal clause (`acl`) modifying the demonstrative pronoun *tasya*. Both *yasya* and *tasya* are genitive forms meaning possession (Fig. 10).

7 Parsing Experiment

We have performed preliminary parsing experiments with two parsers, the Malt Parser (*stack-lazy* algorithm) [6], and UDPipe 1.2 (default settings)

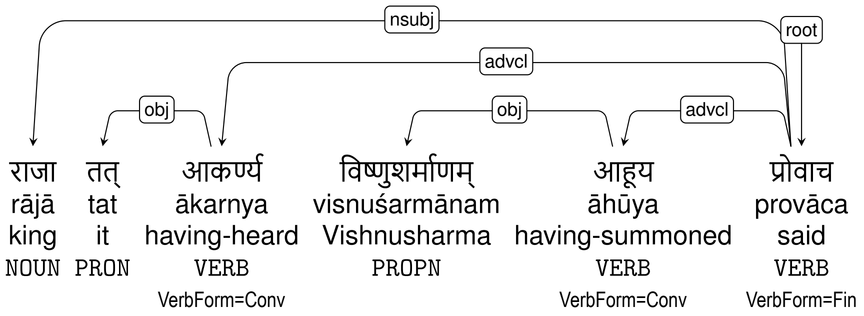
nsubj
root
advcl
obj
obj
advcl
राजा तत् आकर्ण्य विष्णुशर्माणम् आहूय प्रोवाच
rājā tat ākarnya visnuśarmānam āhūya provāca
king it having-heard Vishnusharma having-summoned said
NOUN PRON VERB PROPN VERB VERB
VerbForm=Conv VerbForm=Conv VerbForm=Fin

Fig. 8. "When the king had listened to this, he summoned Vishnusharman and said…" In the English translation, clause 1 is subordinate while clauses 2 and 3 are coordinate. In Sanskrit however, both clauses 1 and 2 are subordinate.

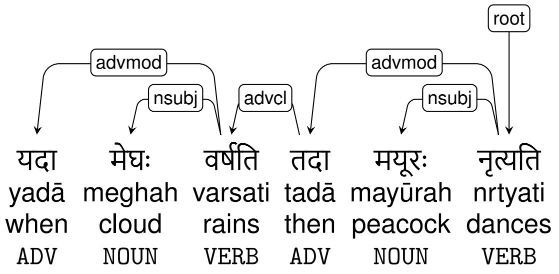
root
advmod
advmod
nsubj
advcl
nsubj
यदा मेघः वर्षति तदा मयूरः नृत्यति
yadā meghah varsati tadā mayūrah nrtyati
when cloud rains then peacock dances
ADV NOUN VERB ADV NOUN VERB

Fig. 9. A relative-correlative construction: "When it rains, peacocks dance."

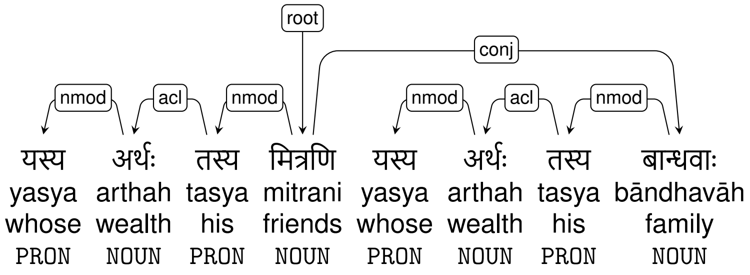
root
conj
nmod
acl
nmod
nmod
acl
nmod
यस्य अर्थः तस्य मित्राणि यस्य अर्थः तस्य बान्धवाः
yasya arthah tasya mitrani yasya arthah tasya bāndhavāh
whose wealth his friends whose wealth his family
PRON NOUN PRON NOUN PRON NOUN PRON NOUN

Fig. 10. Verbless example: "One who has wealth has friends; one who has wealth has family."

[9]. Since the corpus is so small, one has to train the parsers in a 10-fold cross-validation style; our average unlabeled attachment score reaches 61% and labeled attachment score is 51% (parsing only; this figure does not include the accuracy of word segmentation, as it is measured on gold-standard segmentation.) It is difficult to compare these numbers to previously reported work in Sanskrit parsing. [1] notes that "test data for Sanskrit syntax are not available;" his unsupervised parser is restricted to projective trees. [4] reports LAS=63% and UAS=80% on her test data (1316 sentences that are not publicly available and thus the results are not directly comparable to ours). However, we did compare our results with delexicalized parsers [10] trained on 2000 sentences from various groups of languages; the best-performing delexicalized parser was trained on Slavic languages and achieved UAS=54.67%, resp. LAS=38.99%, which is significantly lower than the lexicalized parser trained on the treebank presented in this paper. We therefore conclude

that even very small data, obtained in a cheap and fast way, can provide a better parsing model than unsupervised and semi-supervised methods.

8 Conclusions

We presented a new seed treebank for Sanskrit, a classical language of India. To our knowledge this is the first syntactically annotated data set for this language that is publicly available. Thanks to the annotation standard of Universal Dependencies, the morphological and syntactic annotation is easily understandable and comparable to existing corpora in many other languages, and existing tools can be easily employed to process the corpus and utilize the annotation. While the corpus is currently small, it can be used to train a statistical parser. Moreover, the underlying text is rather large, providing a good base for future growth of the treebank.

Acknowledgments

The work was supported by the grant 15-10472S of the Czech Science Foundation. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2015071).

References

1. **Hellwig, O. (2009).** Extracting dependency trees from Sanskrit texts. **Kulkarni, A. & Huet, G.**, editors, *Sanskrit Computational Linguistics 3, LNCS 5406*, Springer Verlag, Hyderabad, India, pp. 106–115.
2. **Huet, G. (2007).** Shallow syntax analysis in Sanskrit guided by semantic nets constraints. *Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries*, ACM, New York, NY, USA, pp. 1–10.
3. **Huet, G. (2009).** Formal structure of Sanskrit text: Requirements analysis for a mechanical Sanskrit processor. **Huet, G., Kulkarni, A., & Scharf, P.**, editors, *Sanskrit Computational Linguistics 1 & 2, LNAI 5402*, Springer-Verlag, pp. 162–199.
4. **Kulkarni, A. (2013).** A deterministic dependency parser with dynamic programming for Sanskrit. *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, Praha, Czechia, pp. 157–166.
5. **Mittal, V. (2010).** Automatic Sanskrit segmentizer using finite state transducers. *Proceedings of the ACL 2010 Student Research Workshop*, Uppsala, Sweden, pp. 85–90.
6. **Nivre, J. (2009).** Non-projective dependency parsing in expected linear time. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, pp. 351–359.
7. **Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., & et al. (2017).** Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
8. **Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016).** Universal dependencies v1: A multilingual treebank collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 1659–1666.
9. **Straka, M., Hajič, J., & Straková, J. (2016).** UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association, Portorož, Slovenia, pp. 4290–4297.
10. **Zeman, D. & Resnik, P. (2008).** Cross-language parser adaptation between related languages. *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, International Institute of Information Technology, Hyderabad, India, pp. 35–42.

Article received on 05/01/2018; accepted on 08/02/2018.
Corresponding author is Daniel Zeman.