Universal Dependencies for Sanskrit: A Pilot Study

Puneet Dwivedi, Daniel Zeman
Charles University, Faculty of Mathematics and Physics
Malostranské náměstí 25, Praha, Czechia
puneet.iitkgp1094@gmail.com, zeman@ufal.mff.cuni.cz

Abstract

We present the first steps towards a treebank of Sanskrit within the Universal Dependencies framework. Our dataset is tiny at the moment, consisting of 215 sentences—a result of a summer internship project. Nevertheless, this seems to be, to the best of our knowledge, the first publicly available piece of syntactically annotated Sanskrit text. At the time of submission of this abstract, it is being extended with the final goal of around 1000 sentences. We also present a parsing experiment, with results surpassing delexicalized parsing.

Keywords: universal dependencies, sanskrit, treebank, syntactic annotation

1. Introduction

Universal Dependencies (UD)[1] (Nivre et al., 2016) 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 2.0 (Nivre et al., 2017), contains 70 treebanks in 50 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 three modern Indian languages: Tamil, Hindi and Urdu. 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. (Kulkarni, 2013) 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 (Mittal, 2010). 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 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.

We were only able to add syntactic annotation to a tiny fraction of Pañcatantra, namely to the preface about creation of Pañcatantra, and to the beginning of the first section called *Mitrabheda*, 215 sentences in total.

3. Preprocessing

We used Gérard Huet's *Sanskrit Reader Companion*[5] (Huet, 2007; Huet, 2009) to obtain possible word segmentation and morphological features for each sentence. The segmenter provides multiple hypotheses

[1] `http://universaldependencies.org/`

[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`

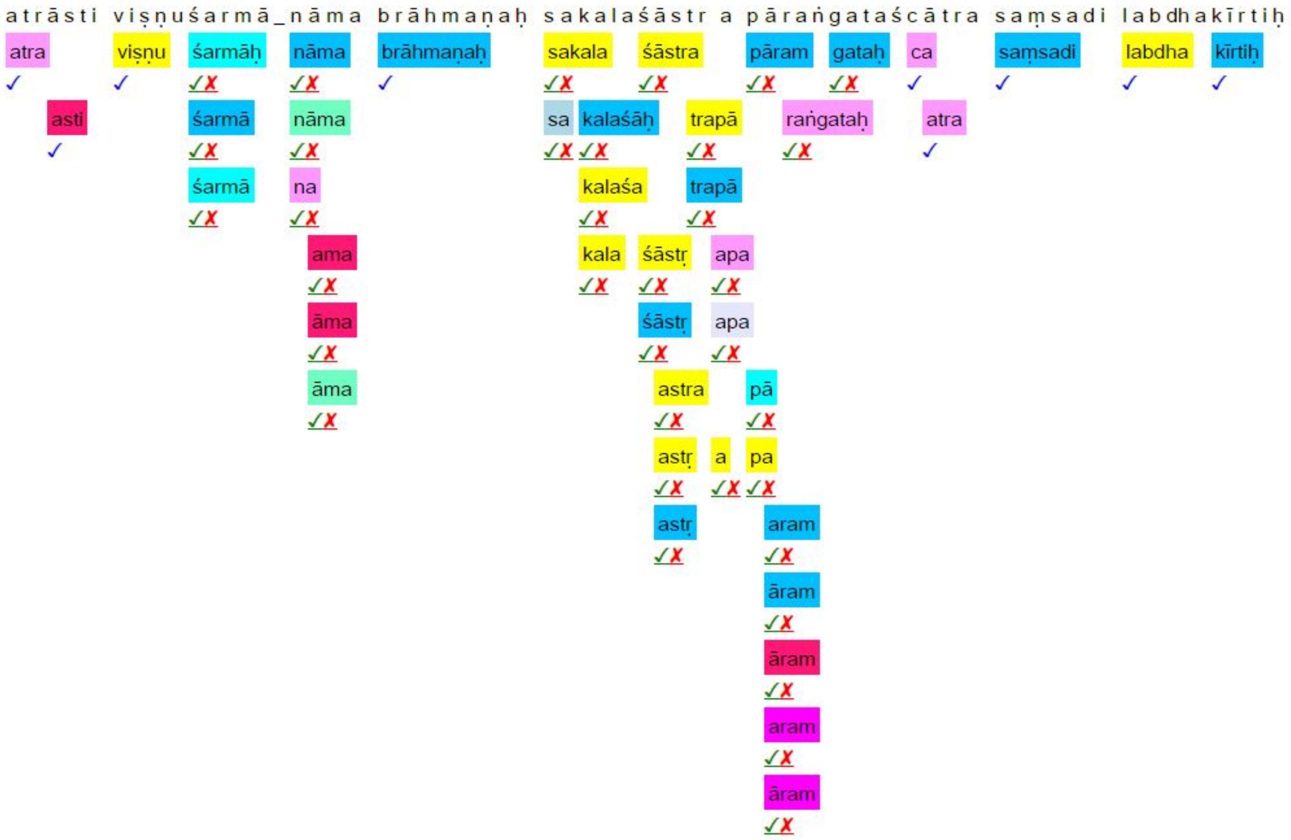

Figure 1: An example of multiple segmentation hypotheses, as provided by the Sanskrit Reader Companion. Colors correspond to different parts of speech. Morphological analysis is also available, although not visible in this screenshot. The input string contained 7 space-delimited tokens: *atrāsti viṣṇuśarmā nāma brāhmaṇaḥ sakalaśāstrapāraṅgataścātra saṁsadi labdhakīrtiḥ*. 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 viṣṇuśarmā nāma brāhmaṇaḥ sakala śāstra pāraṅgata ca atra saṁsadi labdhakīrtiḥ*.

where applicable (Figure 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 *Viṣṇuśarmā*). The lemma and morphological information (gender, number and case for nominals, and mood, tense and number for verbs) was obtained from the Sanskrit Reader together 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 dependency annotation was done manually (one annotator only). For short and simple sentences, the shallow Sanskrit parser[6] (Kulkarni, 2013) was of some help, but unfortunately it cannot parse the more complex sentences.

4. Illustrative Examples

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. Example: कः अर्थः पुत्रेण जातेन यः न विद्वान्न न भक्तिमान् अस्ति / *kaḥ arthaḥ putreṇa jātena yaḥ na vidvānna na bhaktimān asti* "What use having a son who is neither smart nor religious." Here the adjective *vidvānna* "smart" is the root of the relative clause and the verb *asti* is attached to it using the relation `cop`.

In contrast, the same verb in existential or locative meaning takes the root position: अत्रास्ति विष्णुशर्मा नाम ब्राह्मणः / *atrāsti viṣṇuśarmā nāma brāhmaṇaḥ* "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. Example: एतस्मिन्नन्तरे ते वानराः यथेच्छया क्रीडितुम् आरब्धम् / *etasminnantare te vānarāḥ yathec-*

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

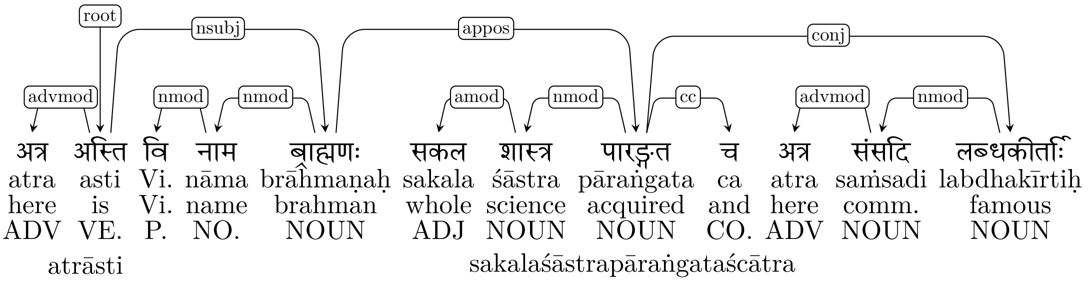

Figure 2: Dependency tree of the sentence from Figure 1. Arthur Ryder's English translation: *There is a Brahman here named Vishnusharman, with a reputation for competence in numerous sciences.*

chayā krīḍitum ārabdham lit. *in-this-moment the monkeys as-with-desire to-play began*, "At the moment the monkeys began their playful frolics." The infinitive *krīḍitum* is attached to the past participle *ārabdham* as its controlled complement, `xcomp`.

Occasionally it is not clear whether a sequence of clauses should be analyzed as coordination or subordination. We preferred the syntactic over semantic criteria. Thus the sentence *The king thought and then spoke* is analyzed as coordination, while in *Having thought, 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.

Some sentences are devoid of any verb, this happens mostly in *ślokas* (verse). Example: यस्यार्थास्तस्य मित्राणि यस्यार्थास्तस्य बान्धवाः / *yasyārthāstasya mitrāṇi yasyārthāstasya bāndhavāḥ* lit. *whose wealth his friends whose wealth his family*, meaning "One who has money, has friends; one who has money, has family." We analyze this sentence as two coordinate clauses, each comprising an embedded relative clause. The phrase *yasya arthaḥ* "whose wealth" is an adnominal clause (`acl`) modifying the demonstrative pronoun *tasya*. Both *yasya* and *tasya* are genitive forms, expressing possession. See Figure 3 for the full dependency tree.

5. Statistics

The treebank at present consists[7] of 215 sentences resp. 1254 surface tokens, which were split into 1521 syntactic words. 35 dependencies are non-projective. This makes 2.3% of all dependency relations, which is only slightly above the average of all UD treebanks.

The corpus contains 16 out of 17 "universal" part-of-speech tags defined in UD; the missing tag is `SYM` for symbols. There is only one particle, but a frequent one: न / *na* (negation). The only auxiliary verb is अस् / *as* "to be".

We use 15 universal features: gender, number, case, degree, polarity, prontype, numtype, possessivity, reflexivity, person, verbform, mood, aspect, tense and voice.

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

6. Parsing Experiment

We have performed preliminary parsing experiments with two parsers, the Malt Parser (*stack-lazy* algorithm) (Nivre, 2009), and UDPipe (Straka et al., 2016). Since the corpus is so small, one has to train the parsers in a 10-fold cross-validation style; our average labeled attachmet score reaches 61% (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. (Hellwig, 2009) notes that "test data for Sanskrit syntax are not available;" his unsupervised parser is restricted to projective trees. (Kulkarni, 2013) 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 (Zeman and Resnik, 2008) 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.

7. Conclusion

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. We opted for the annotation scheme of Universal Dependencies, which emerged as a de-facto standard and lingua franca of dependency syntax. 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.

8. Bibliographical References

Hellwig, O. (2009). Extracting dependency trees from Sanskrit texts. In Amba Kulkarni et al., editors,

[7] The numbers will be updated for the final version of the paper, as the treebank is still growing.

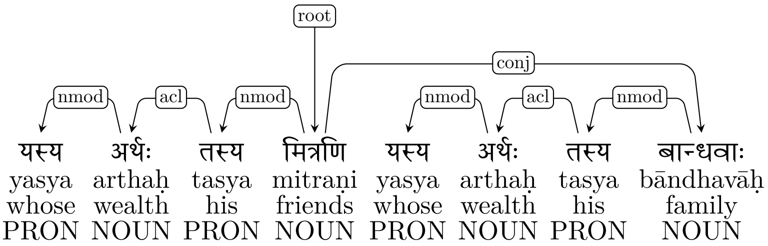

Figure 3: Verbless example: “One who has wealth has friends; one who has wealth has family.”

Sanskrit Computational Linguistics 3, LNCS 5406, pages 106–115, Hyderabad, India. Springer Verlag.

Huet, G. (2007). Shallow syntax analysis in Sanskrit guided by semantic nets constraints. In Proceedings of the 2006 International Workshop on Research Issues in Digital Libraries, New York, NY, USA. ACM.

Huet, G. (2009). Formal structure of Sanskrit text: Requirements analysis for a mechanical Sanskrit processor. In Gérard Huet, et al., editors, Sanskrit Computational Linguistics 1 & 2, LNAI 5402. Springer-Verlag.

Kulkarni, A. (2013). A deterministic dependency parser with dynamic programming for Sanskrit. In Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 157–166, Praha, Czechia.

Mittal, V. (2010). Automatic Sanskrit segmentizer using finite state transducers. In Proceedings of the ACL 2010 Student Research Workshop, pages 85–90, Uppsala, Sweden, July.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Portorož, Slovenia.

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In 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, pages 351–359, Singapore.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. European Language Resources Association.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In IJCNLP 2008 Workshop on NLP for Less Privileged Languages, pages 35–42, Hyderabad, India. International Institute of Information Technology.

9. Language Resource References

Nivre, Joakim and Agić, Željko and Ahrenberg, Lars and Aranzabe, Maria Jesus and Asahara, Masayuki and Atutxa, Aitziber and Ballesteros, Miguel and Bauer, John and Bengoetxea, Kepa and Bhat, Riyaz Ahmad and Bick, Eckhard and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Candito, Marie and Cebiroğlu Eryiğit, Gülşen and Celano, Giuseppe G. A. and Chalub, Fabricio and Choi, Jinho and Çöltekin, Çağrı and Connor, Miriam and Davidson, Elizabeth and de Marneffe, Marie-Catherine and de Paiva, Valeria and Diaz de Ilarraza, Arantza and Dobrovoljc, Kaja and Dozat, Timothy and Droganova, Kira and Dwivedi, Puneet and Eli, Marhaba and Erjavec, Tomaž and Farkas, Richárd and Foster, Jennifer and Freitas, Cláudia and Gajdošová, Katarína and Galbraith, Daniel and Garcia, Marcos and Ginter, Filip and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Gómez Guinovart, Xavier and Gonzáles Saavedra, Berta and Grioni, Matias and Grūzītis, Normunds and Guillaume, Bruno and Habash, Nizar and Hajič, Jan and Hà Mỹ, Linh and Haug, Dag and Hladká, Barbora and Hohle, Petter and Ion, Radu and Irimia, Elena and Johannsen, Anders and Jørgensen, Fredrik and Kaşıkara, Hüner and Kanayama, Hiroshi and Kanerva, Jenna and Kotsyba, Natalia and Krek, Simon and Laippala, Veronika and Lê Hồng, Phương and Lenci, Alessandro and Ljubešić, Nikola and Lyashevskaya, Olga and Lynn, Teresa and Makazhanov, Aibek and Manning, Christopher and Mărănduc, Cătălina and Mareček, David and Martínez Alonso, Héctor and Martins, André and Mašek, Jan and Matsumoto, Yuji and McDonald, Ryan and Missilä, Anna and Mititelu, Verginica and Miyao, Yusuke and Montemagni, Simonetta and More, Amir and Mori, Shunsuke and Moskalevskyi, Bohdan and Muischnek, Kadri and Mustafina, Nina and Müürisep, Kaili and Nguyễn Thị, Lương and Nguyễn Thị Minh, Huyền and Nikolaev, Vitaly and Nurmi, Hanna and Ojala, Stina and Osenova, Petya and Øvrelid, Lilja and Pascual, Elena and Passarotti, Marco and Perez, Cenel-Augusto and Perrier, Guy and Petrov, Slav and Piitulainen, Jussi and Plank, Barbara and Popel, Martin and Pretkalniņa, Lauma and Prokopidis, Prokopis and Puolakainen, Tiina and Pyysalo, Sampo and Rademaker, Alexandre and Ramasamy, Loganathan and Real, Livy and Rituma, Laura and Rosa, Rudolf and Saleh, Shadi and Sanguinetti, Manuela and Saulīte, Baiba and Schuster, Sebastian and Seddah, Djamé and Seeker, Wolfgang and Seraji, Mojgan and Shakurova, Lena and Shen, Mo and Sichinava, Dmitry and Silveira, Natalia and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Simov, Kiril and Smith, Aaron and Suhr, Alane and Sulubacak, Umut and Szántó, Zsolt and Taji, Dima and Tanaka, Takaaki and Tsarfaty, Reut and Tyers, Francis and Uematsu, Sumire and Uria, Larraitz and van Noord, Gertjan and Varga, Viktor and Vincze, Veronika and Washington, Jonathan North and Žabokrtský, Zdeněk and Zeldes, Amir and Zeman, Daniel and Zhu, Hanzhi. (2017). Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, ISLRN http://hdl.handle.net/11234/1-1983.