INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue4)
Available online at www.ijariit.com

Universal Dependencies for Sanskrit

Puneet Dwivedi
Indian Institute of Technology, Kharagpur
puneet.iitkgp1094@gmail.com

Easha Guha
Indian Institute of Technology, Kharagpur
easha.guha5@gmail.com

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 less than 200 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. We also present a parsing experiment, with results surpassing delexicalized parsing.

Keywords: Natural Language Processing, Dependency Parsing, Universal Dependencies, Text Segmentation, Machine Learning

I. INTRODUCTION

Universal Dependencies (UD) [6] 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 a number of languages, UD has probably become the largest collection of freely available treebanks in the world: the latest release, UD 1.3, contains 54 treebanks in 40 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, and Old Church Slavonic), as well as two modern Indian languages: Tamil and Hindi. 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)[3] Mentions a Sanskrit treebank of around 3000 sentences (mostly modern short stories), reportedly developed under a Government of India sponsored a 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 believe that others will pick up the ball, improve and extend the annotation; 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)[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 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 analysed.

II. DATA PREPROCESSING

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

sometimes differs. We were only able to add a syntactic annotation to a ti any fraction of Pañcatantra, namely to the preface about the creation of Pañcatantra, and to the beginning of the rest section called Mitrabheda, 190 sentences in total.

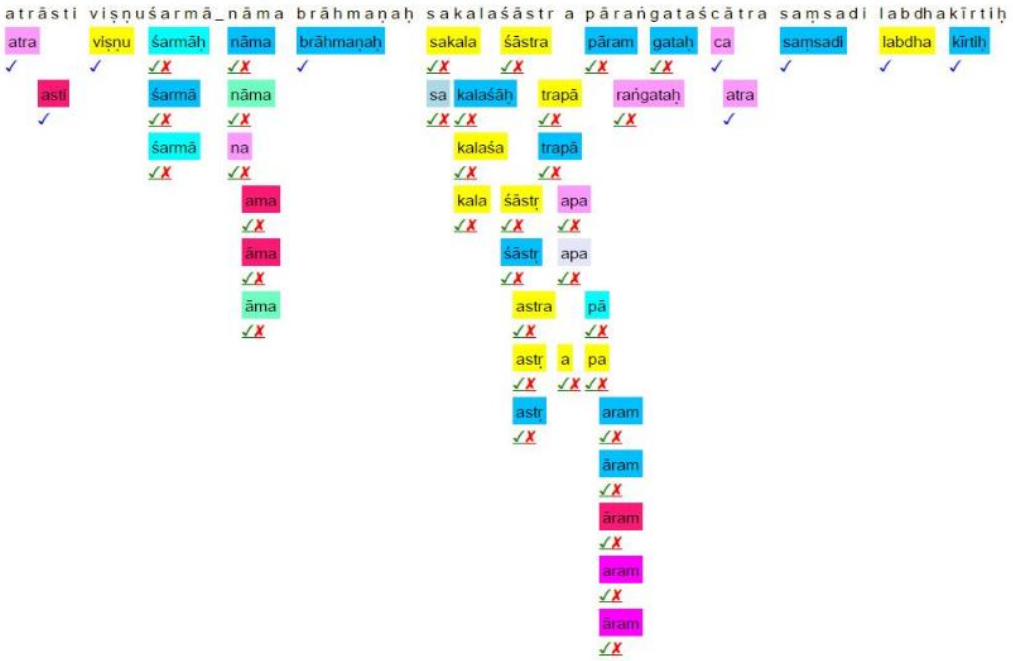

Figure 1: An example of multiple segmentation hypotheses, as provided by the Sanskrit Reader Companion. Colors correspond to different parts of speech. Morphological analysis is also available, although not visible in this screenshot. The input string contained 7 space-delimited tokens: atrāsti viṣṇuśarmā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

III.PREPROCESSING

We used Gérard Huet's Sanskrit Reader Companion (Huet, 2007; Huet, 2009) to obtain possible word segmentation and morphological features for each sentence. The segmenter provides multiple hypotheses 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) were obtained from the Sanskrit Reader together with the correct segmentation. One of the 17 universal part-of-speech tags de ned in UD was also manually assigned to each word. Finally, the data was converted to the CoNLL-U 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 (Kulkarni, 2013)[4] was of some help, but

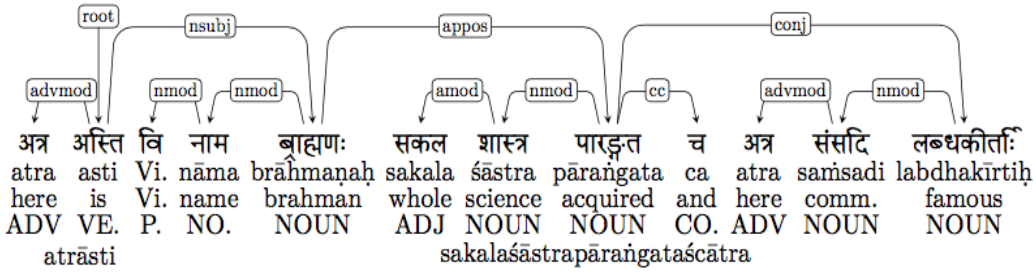

Figure 2: Dependency tree of the sentence from Figure 1. Arthur Ryder's English translation: *There is a Brahman here named Vishnusharman, with a reputation for competence in numerous sciences.*

unfortunately, it cannot parse the more complex sentences.

IV. 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 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 modi ers of the non-verbal predicate. ***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: ***atrāsti viṣṇuśarmā nāma brāhmaṇaḥ*** "*There is a Brahman here named Vishnusharman*." The infinitive is attached to the verbs that control them via the relation xcomp, which is used in UD whenever a clause inherits its subject from a superordinate clause. Example: e***tas Minnan tare te vānarāḥ yathecchayā krīḍitum ārabdham lit.*** "*In-this-moment the monkeys as-with-desire to-play began, There 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 had a think and then spoke"* is analyzed as coordination, while in Having a thought, the king spoke the first clause is attached as advcl, modifying the predicate of the second clause (spoke). Non-nite verb forms co-occurring with nites are indicators of subordination.

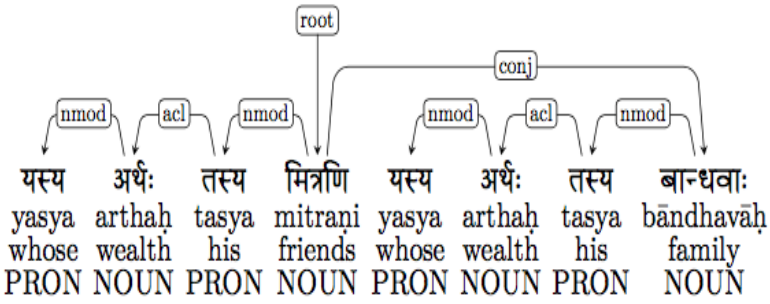

Figure 3: Verbless example: "One who has wealth has friends; one who has wealth has family."

Some sentences are devoid of any verb, this happens mostly in ślokas (verse). Example: **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 of 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.

V. STATISTICS

The treebank at present consists of 190 sentences resp. 1031 surface tokens, which were split into 1206 syntactic words. 35 dependencies are non-projective. This makes 2.9% of all dependency relations, which is only slightly above the average of all UD treebanks.

The corpus contains 13 out of 17 "universal" part-of-speech tags de ned in UD. Missing are punctuation (PUNCT), symbols (SYM), interjections (INTJ) and also subordinating conjunctions (SCONJ). There is only one particle, but a frequent one: न / na (negation). The only auxiliary verb is अस् / as "to be". We use 13 universal features: gender, number, case, prontype, numtype, possessivity, reflexivity, person, verbform, mood, and 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.

VI. RESULTS AND DISCUSSIONS

We have performed a preliminary parsing experiment using the Malt Parser's stack-lazy algorithm (Nivre, 2009)[6]. Training on 100 sentences and testing on 40, we obtained UAS=66.95% and LAS=61.45%. It is difficult to compare these numbers to previously reported work in Sanskrit parsing.

(Hellwig, 2009) [1] 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 from the unpublished Sanskrit treebank). However, we did compare our results with delexicalized parsers (Zeman and Resnik, 2008) [7] 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%. 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.

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. We plan to contribute the data to the next release of the Universal Dependencies treebanks in November 2016. 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.

REFERENCES

[1] Oliver Hellwig. 2009. Extracting dependency trees from Sanskrit texts. In Amba Kulkarni and Gérard Huet, editors, Sanskrit Computational Linguistics 3, LNCS 5406, pages 106–115, Hyderabad, India. Springer Verlag.

[2] Gérard Huet. 2007. Shallow syntax analysis in Sanskrit guided by semantic nets constraints. In Pro- seedings of the 2006 International Workshop on Research Issues in Digital Libraries, New York, NY, USA. ACM.

[3] Gérard Huet. 2009. The formal structure of Sanskrit text: Requirements analysis for a mechanical San- skirt processor. In Gérard Huet, Amba Kulkarni, and Peter Scharf, editors, Sanskrit Computational Linguistics 1 & 2, LNAI 5402. Springer-Verlag.

[4] Amba Kulkarni. 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.

[5] Vipul Mittal. 2010. Automatic Sanskrit segmentizer using nite state transducers. In Proceedings of the ACL 2010 Student Research Workshop, pages 85–90, Uppsala, Sweden, July.

[6] Joakim Nivre, Marie-Catherine de Marne e, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Man- ning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multi lingual treebank collection. In Proceedings of the 10th Inter- national Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Portorož, Slovenia.

[6] Joakim Nivre. 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.

[7] Daniel Zeman and Philip Resnik. 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.

