Transfer learning for dependency parsing of Vedic Sanskrit

Abhiram Vinjamuri and Weiwei Sun
Department of Computer Science and Technology
University of Cambridge

{av646@cantab.ac.uk, ws390@cam.ac.uk}

Abstract

This paper focuses on data-driven dependency parsing for Vedic Sanskrit. We propose and evaluate a transfer learning approach that benefits from syntactic analysis of typologically related languages, including Ancient Greek and Latin, and a descendant language - Classical Sanskrit. Experiments on the Vedic TreeBank demonstrate the effectiveness of cross-lingual transfer, demonstrating improvements from the biaffine baseline as well as outperforming the current state of the art benchmark, the deep contextualised self-training algorithm, across a wide range of experimental setups.

1 Introduction

There is a pressing need for high-quality linguistic analysis in the study of ancient languages; this work is critically hindered by a scarcity of annotated data (Sommerschield et al., 2023). This challenge is particularly acute for Vedic Sanskrit, a low-resource language whose free word order and rich morphology create complex, non-projective dependency structures that make automatic parsing a formidable task (Ponti et al., 2019). This combination of structural complexity and data scarcity establishes Vedic Sanskrit as a critical test case for the robustness and scalability of modern data-driven parsing methods.

Vedic Sanskrit exemplifies the need for methods that can operate effectively in low-resource settings. Two dominant paradigms address this directly: transfer learning and self-learning (Alyafeai et al., 2020). The former involves transferring a model trained on a high-resource language, often syntactically related to the target language, using the model's predictions to create a large corpus of 'silver-standard' data. The latter, exemplified by Deep Contextualized Self-Training (DCST) (Rotman and Reichart, 2019), utilises a semi-supervised loop where a model is iteratively retrained on its most confident predictions over a large unlabelled corpus. The application of these techniques to Vedic Sanskrit is compelling; transfer learning could exploit structural similarities with other ancient Indo-European languages, while self-training could leverage the unannotated Vedic corpus itself to refine a parser's accuracy.

In this paper, we make the following primary contributions. First, we establish a new state-of-the-art for Vedic Sanskrit dependency parsing by proposing a cross-lingual transfer learning framework that achieves a Labelled Attachment Score (LAS) of 82.5%. This result outperforms the previous state-of-the-art, the Deep Contextualized Self-Training (DCST) method, by 2.3 points. Second, we demonstrate the remarkable data efficiency of this framework; in a rigorous few-shot setting using only 80 annotated sentences, our model achieves a LAS of 17.33%, more than doubling the performance of a randomly initialised baseline. Finally, our direct empirical comparison reveals that our transfer learning approach is a more effective and robust strategy than the complex self-training paradigm for this task.

2 Related Work

Our research evaluates competing strategies for low-resource neural dependency parsing by enhancing the foundational deep biaffine attention parser (Dozat and Manning, 2017), a powerful architecture well-suited to the non-projective structures found in free-word-order languages. Specifically, we adapt the modern trend of replacing traditional LSTM encoders with the more powerful Transformer architecture (Vaswani et al., 2017), a combination that has been successfully demonstrated (Li et al., 2019). Applying this enhanced parser to Vedic Sanskrit, we use the Vedic Treebank (Hellwig et al., 2020) to conduct two primary investigations: we first measure the impact of the Transformer encoder, and then we empirically compare the state-of-the-art DCST paradigm (Hellwig et al., 2023) against our proposed cross-lingual transfer learning framework.

The primary challenge for parsing Vedic Sanskrit is data scarcity. We address this by comparing two paradigms. The first is self-training, where a model learns from its own predictions on unlabelled data. The state-of-the-art for Vedic Sanskrit is an advanced implementation of this called DCST, pioneered by (Rotman and Reichart, 2019) and applied to Vedic Sanskrit by (Hellwig et al., 2023), which serves as our primary baseline. The second, competing paradigm is cross-lingual transfer learning, where syntactic knowledge is leveraged from related, higher-resource languages like Ancient Greek and Latin (Ammar et al., 2016). We explore this through full fine-tuning and a particularly data-efficient few-shot learning approach (Hu et al., 2022), which is crucial for extremely low-resource settings. This targeted transfer complements the broader trend of building large monolingual foundation models like SanskritT5 (Bhatt et al., 2024).

While both self-training and cross-lingual transfer are established techniques, a direct empirical comparison of their efficacy for a morphologically rich and free-word-order language like Vedic Sanskrit has been absent. This project fills that critical gap. We augment the standard biaffine parser with a more powerful Transformer encoder and use it to systematically evaluate its performance within both the complex DCST framework and a simpler, direct transfer learning framework. By testing this rigorously in full-resource and few-shot learning scenarios, our comparative framework isolates the impact of architectural choices and training paradigms, ultimately demonstrating that a simpler transfer-based method is more effective than the current state-of-the-art.

3 Methodology

We formulate dependency parsing as the task of finding the maximum-weight spanning tree in a graph of all possible head–dependent relations, following (McDonald et al., 2005). Our methodology systematically evaluates architectural enhancements and compares learning paradigms to improve upon the state-of-the-art for this graph-based approach on Vedic Sanskrit. A summary of our experimental framework is shown in Figure 1.

3.1 Baseline Parser

Our baseline is the Biaffine dependency parser (Dozat and Manning, 2017), which represents a strong foundation for this task. It consists of a contextual encoder, for which we use a standard Bidirectional LSTM (BiLSTM) (Huang et al., 2015), and a biaffine attention classifier to score all possible head–dependent arcs. To handle the free word order characteristic of Vedic Sanskrit, the decoder uses the Chu-Liu/Edmonds algorithm (Kübler et al., 2009) to efficiently extract a valid, non-projective dependency tree.

3.2 Transformer-based Parser

To better model the non-local dependencies in Vedic Sanskrit, we enhance the baseline by replacing its BiLSTM encoder with a Transformer encoder (Vaswani et al., 2017). The self-attention mechanism in this architecture is theoretically more effective at capturing long-range syntactic relationships, making it a better fit for this task. This improved Transformer-based parser serves as the foundation for our primary experiments comparing different low-resource learning strategies.

3.3 Low-Resource Learning Paradigms

Using our enhanced parser, we conduct a comparative analysis of the two prominent learning paradigms for low-resource settings:

Deep Contextualized Self-Training (DCST) First, we re-implement the existing state-of-the-art semi-supervised method for Vedic Sanskrit (Hellwig et al., 2023; Rotman and Reichart, 2019). This approach uses the parser's own output on unlabelled data to generate "pseudo-labels," which are then used to train a contextualised model that, in turn, refines the final parser.

Cross-Lingual Transfer Learning As an alternative, we propose to pre-train our parser on annotated data from related languages before fine-tuning on Vedic Sanskrit. Source languages include Ancient Greek, Latin, and Classical Sanskrit (Uni, 2020), chosen for their typological proximity. We rigorously test the quality of the transferred knowledge in a **few-shot setting**, where the pretrained encoder is frozen and only the final layers are fine-tuned on a minimal set (80 sentences) of the target data. This comparative framework allows us to isolate the impact of both architectural choices and training paradigms on final parsing performance.

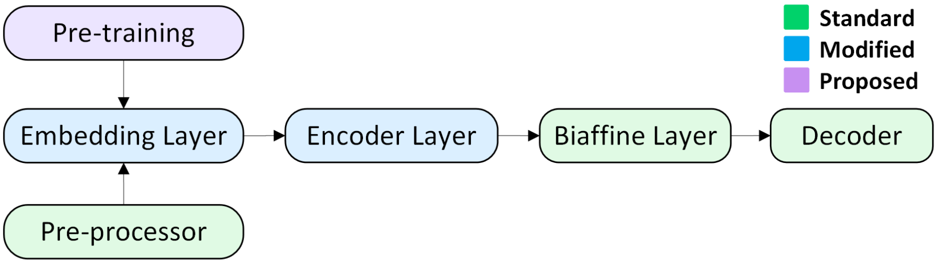

Figure 1: Overview of our experimental framework. We first enhance a baseline Biaffine parser with a Transformer encoder, then use it to compare two learning paradigms: self-training (DCST) and cross-lingual transfer learning.

4 Experiments and Analysis

We conducted a series of experiments to evaluate our proposed approach. We first establish the architectural advantage of using a Transformer encoder and then compare our cross-lingual transfer learning paradigm with the state-of-the-art self-training method (DCST). Finally, we demonstrate the data efficiency of our approach in a rigorous few-shot setting. All experiments are performed on the Vedic Treebank (Hellwig et al., 2020) and evaluated using mean UAS and LAS with 5-fold cross-validation.

4.1 Transformer vs. LSTM

Model Variant	UAS (%)	LAS (%)
Hellwig et al. (2023)	79.5	72.0
BiLSTM Encoder	82.0	73.5
Transformer Encoder	82.8	79.4

Table 1: Results on the Vedic Sanskrit test set, for baseline parser architectures with different encoders.

First, we evaluated the impact of our architectural choice, the results can be found in Table 1. Replacing the BiLSTM encoder with a Transformer encoder significantly increased the LAS parsing performance by 5.9 points with a p-value of 4.24% using paired t-tests. The gain in Unlabeled Attachment Score (UAS) was negligible. This substantial gain in Labelled Attachment Score (LAS) suggests that the Transformer's self-attention mechanism is particularly effective at capturing the complex, non-local contextual cues required for accurate dependency label assignment in a free word-order language. The negligible change in UAS indicates that while both architectures are competent at identifying basic head–dependent structures, the Transformer excels at discerning the fine-grained syntactic relationships.

4.2 Transfer Learning

Pretraining Source	UAS (%)	LAS (%)
Baseline	78.0	73.2
Classical Sanskrit	79.4	78.6
Ancient Greek	82.3	78.5
Latin	80.2	77.2

Table 2: Parsing performance on the Vedic Sanskrit validation set, for models pre-trained on different typologically related languages

Table 2 shows that pre-training on related languages provides a consistent and significant performance gain over the baseline. The choice of source language introduces important trade-offs: pre-training on Ancient Greek yields the highest UAS, suggesting its free word order and morphological richness provide a powerful inductive bias for learning syntactic structure. In contrast, pre-training on Classical Sanskrit achieves the highest LAS, likely due to a closer alignment in annotation conventions and dependency labels. This highlights that while cross-lingual transfer is broadly effective, the optimal source language may differ depending on whether the goal is to improve structural accuracy (UAS) or labelling precision (LAS).

4.3 Few-Shot Learning

Finally, to test the data efficiency of our method, we evaluated it in a challenging few-shot scenario, fine-tuning on only 80 labelled sentences. The results are shown in Table 3. The results demonstrate the remarkable efficiency of transfer learning. Pre-training on Ancient Greek yields a LAS of 17.33%, more than doubling the performance of a randomly initialised model. This success stems from our strategy of **freezing the pretrained encoder layers**. This forces the model to retain the rich, general syntactic knowledge learned from the source language, while the fine-tuning process adapts only the final classification layers to the Vedic-specific label set. This effectively separates the learning of structural representation (transferred) from label mapping (fine-tuned), confirming it as a powerful strategy for extremely low-resource settings.

Significance testing shows that while all pre-trained languages (Ancient Greek, Latin, and Classical Sanskrit) significantly outperformed the baseline on the LAS metric, only Ancient Greek did so for UAS. Crucially, there was no statistically significant performance difference found when comparing the various pre-trained languages against each other, suggesting they are all similarly effective.

Pretrain Source	UAS (%)	LAS (%)
Baseline	18.50 ± 5.68	8.50 ± 2.88
Ancient Greek	**26.17 ± 3.19**	**17.33 ± 0.52**
Latin	22.40 ± 3.36	16.80 ± 1.79
Sanskrit	23.68 ± 2.53	16.56 ± 0.61

Table 3: Effectiveness of cross-lingual transfer in a few-shot setting. All models were fine-tuned on only 80 sentences of Vedic Sanskrit. Pre-training on Ancient Greek more than doubles the Labelled Attachment Score (LAS) compared to the baseline, demonstrating a powerful inductive bias.

4.4 Transfer Learning vs. Self-training

We then compared our cross-lingual transfer learning framework against the strong DCST self-training baseline. The results are summarised in Table 4. As shown, our transfer learning approach, particularly when pre-training on Ancient Greek, establishes a new state-of-the-art, outperforming the DCST method by over 2 LAS points. This suggests that pre-training on high-quality, annotated data from a typologically similar language provides a more powerful and effective inductive bias than attempting to learn from pseudo-labels generated by the parser's own output. Our simpler, more direct pre-training approach proves to be more robust.

5 Discussion

Our experiments consistently demonstrate that a Transformer-based parser augmented with cross-lingual transfer learning is a superior approach for Vedic Sanskrit dependency parsing compared to the previous state-of-the-art. The key insight from our analysis is that pre-training on high-quality, annotated data from typologically related languages provides a more effective and robust inductive bias than the semi-supervised, pseudo-labelling approach of DCST. The model learns a strong representation of syntactic structure that requires only minimal, targeted fine-tuning. A particularly noteworthy finding is the strong performance of Ancient Greek as a source language, despite its different script not being explicitly handled by our tokeniser. This suggests that the model is capturing deep, abstract structural similarities between the languages, rather than relying on surface-level lexical overlap. This highlights the robustness of transfer learning for morphologically rich, low-resource languages.

Model	UAS (%)	LAS (%)
Biaffine	82.8	79.4
DCST	83.5	80.2
Transfer Learning	**84.6**	**82.5**

Table 4: Main parsing results on the Vedic Sanskrit test set. Our Transformer-based parser with cross-lingual transfer learning achieves the highest performance, outperforming both the baseline Biaffine parser and the DCST self-training paradigm. This result supports transfer learning as a viable alternative to more complex self-training strategies in low-resource settings.

The success of our few-shot learning experiments further underscores this point. By freezing the encoder, we showed that the core syntactic knowledge can be effectively transferred, while the fine-tuning process specialises the final layers for the target language's label set. This provides a practical and highly data-efficient roadmap for developing parsing tools for other ancient or low-resource languages where annotated data is scarce.

6 Conclusion

We establish a new state-of-the-art dependency parser for Vedic Sanskrit by demonstrating that a modern Transformer-based architecture significantly outperforms a traditional BiLSTM baseline. Our central contribution, however, is showing that a straightforward cross-lingual transfer learning framework is more effective and data-efficient than the existing, more complex self-training paradigm. We find that pre-training on typologically related ancient languages provides a powerful inductive bias that substantially improves parsing accuracy, even in rigorous few-shot settings. This work delivers a new benchmark for Vedic Sanskrit and also validates a robust methodology for tackling parsing challenges in resource-scarce linguistic contexts.

7 Limitations

Our work is subject to several limitations that suggest clear directions for future research. First, our models are constrained by the available data; the Vedic Treebank contains a notable number of unknown tokens, which introduces noise. Second, while our transfer learning approach succeeded with typologically related Indo-European languages, its effectiveness on more distant language families remains an open question. Finally, our parser operates at the sentence level, limiting its ability to capture document-level discourse phenomena such as topic chains or verse alignment, which are crucial for deeper philological analysis.

References

2020. Universal dependencies. https://universaldependencies.org/, Accessed 19 Apr. 2025.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Waleed Ammar, Gülşen Mulcaire, Yulia Tsvetkov, Benjamin Van Durme, and Chris Dyer. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Jash J. Bhatt, Kushal Pandya, Vandan Mehta, Henil Varia, and Brijesh Bhatt. 2024. Sankritt5: A t5 model for sanskrit language. *Preprint*, arXiv:2409.13920.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Oliver Hellwig, Sebastian Nehrdich, and Sven Sellmer. 2023. Data-driven dependency parsing of vedic sanskrit. *Language Resources and Evaluation*, 57:1173–1206. Accepted: 13 January 2023, Published online: 10 February 2023. Accessed: 6 Feb. 2025.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France. European Language Resources Association.

Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. *Preprint*, arXiv:2204.07305.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Cambridge University Press.

Junxian Li, Xuezhe Ma, and Eduard Hovy. 2019. Dependency parsing with partial bi-affine attention. In *Proceedings of NAACL-HLT*.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti and 1 others. 2019. Modeling language variation and universals: A survey on typological representations for cross-lingual nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4525–4549.

Guy Rotman and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3):703–747.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

A Experimental Details

A.1 Corpus Information

Our experiments utilized several corpora selected for their relevance and quality. The target language, Vedic Sanskrit (VS), was sourced from the Vedic Treebank, which contains approximately 3,700 manually annotated sentences (Hellwig et al., 2020). This corpus is considered high-quality, achieving an inter-annotator agreement of 0.75 (Uni, 2020), and was chosen as our primary data source for training and evaluation.

For cross-lingual transfer, we selected three source languages based on their typological and genealogical proximity to VS. We used high-quality, gold-standard treebanks for **Ancient Greek** and Latin (Uni, 2020), both of which share features with VS like rich inflectional morphology and free word order. We also used a silver-standard (machine-annotated) corpus for Classical Sanskrit (Uni, 2020). As the direct successor to VS, it shares key syntactic properties and serves as an effective surrogate to mitigate data sparsity.

A.2 Hyperparameter Configuration

For the pre-training phase of our transfer learning models, key hyperparameters were set as follows: models were trained for up to 120 epochs with a batch size of 16. We used the Adam optimiser with an initial learning rate of 2×10^{-5} (Hellwig et al., 2023). Rather than performing an exhaustive grid search, we employed an inference-based tuning strategy. This involved starting with established baseline values from existing literature and iteratively adjusting parameters based on gradient stability and validation metrics, which proved to be a more computationally efficient approach.

A.3 Rationale for Few-Shot Setting

The few-shot learning experiments were designed to simulate a realistic low-resource scenario where high-quality annotations are extremely scarce. We selected a sample of approximately 80 sentences from the full Vedic Treebank. This subset was chosen to capture the linguistic diversity and nuances of the complete dataset, ensuring the fine-tuning process was both efficient and effective. This small training set forces the model to rely on the inductive biases learned during pre-training, allowing for a rigorous test of knowledge transfer. The remaining data was partitioned into validation and test sets at a 1:8 ratio, providing a small but sufficient validation set for tuning and a large test set for a dependable performance estimate.

A.4 Computational Requirments

The experiments were conducted in a local environment using a standard developer laptop equipped with a modern, consumer-grade dedicated GPU. This setup proved sufficient for training and evaluating all model variants presented in this work. The software stack was built on Python with the PyTorch deep learning library.