TantraTagger: A Benchmark Dataset for Tantrayukti-Based Discourse Structure Labelling in Sanskrit Śāstra Texts

Tapas Khanra[1]*, **Priya Mishra**[2]*, **Malhar Kulkarni**[3], **Ganesh Ramakrishnan**[2]

Indian Institute of Technology Bombay, India

[1]tapaskhanra@iitb.ac.in

[2]{priyam, ganesh}@cse.iitb.ac.in

[3]malhar@hss.iitb.ac.in

Abstract

Classical Sanskrit śāstra texts exhibit highly structured discourse governed by Tantrayukti (TY), a formal system of rhetorical and logical devices central to Indian scholarly writing. TY has received little attention in Sanskrit Computational Linguistics, where research has largely focused on sentence-level analysis. We introduce TantraTagger, the first benchmark dataset for TY-based discourse labelling in Sanskrit śāstric and commentarial texts. The task is formulated as a multi-class classification problem: given a discourse span, a system predicts its primary TY label. The dataset consists of manually annotated discourse units with 39 fine-grained TY relations, covering core argumentative functions such as Introduction, Doubt, Reason, Example, Exception, Conclusion, and Counter-question. We sampled the *Kāśikā-vṛtti* text as a primary source and curated text spans for 39 tags. TantraTagger establishes a new benchmark for Sanskrit discourse modelling and supports computational study of classical discourse theories in low-resource settings.

1 Introduction

Discourse structure plays a crucial role in explaining how meaning is constructed and interpreted beyond the sentence level. In modern Natural Language Processing (NLP), discourse analysis has been extensively studied through frameworks such as Rhetorical Structure Theory (RST), the Penn Discourse Treebank (PDTB), and related formalisms, primarily for contemporary languages. By contrast, Sanskrit remains comparatively under-explored from a computational discourse-analytic perspective, despite its rich textual traditions and well-developed indigenous theories of textual organisation. This gap is particularly striking given the centrality of discourse-level reasoning in Sanskrit scholastic literature. Sanskrit śāstra texts—spanning grammar, philosophy, medicine, logic, and ritual theory—exhibit dense argumentative structure and highly formalized modes of exposition. Traditional authors and commentators systematically employed Tantrayukti (TY), a set of explicitly codified methodological devices such as *anumata, anāgatāvekṣaṇa, pūrvapakṣa, uttarapakṣa, prayojana*, to organise topics, raise doubts, introduce arguments, cite examples, handle exceptions, and arrive at conclusions. Unlike modern discourse theories developed for contemporary languages, TY is theory-driven, text-internal, and grounded in long-standing scholastic practice, making it a particularly suitable framework for modelling Sanskrit discourse computationally. Although TY devices are well documented in classical sources such as Kauṭilya's *Arthaśāstra*, *Carakasaṃhitā*, and *Suśrutasaṃhitā* they have not yet been systematically leveraged in computational linguistics for discourse modelling.

Annotated discourse-level corpora are essential for enabling statistical and computational analysis beyond sentence boundaries, supporting tasks such as discourse parsing, anaphora resolution, and ellipsis detection (Goyal et al., 2012). While recent Sanskrit NLP research has produced robust tools for segmentation, morphological analysis, and dependency parsing, discourse-level annotation frameworks comparable to RST (Mann and Thompson, 1987) or PDTB (Prasad

*Equal contribution.

The prompt file, and dataset are available at https://github.com/SDRMp/TantraTagger.

et al., 2008) are still lacking. Existing computational studies of Sanskrit discourse have typically focused on individual texts—most notably commentaries related to Pāṇini—and have relied on small, closed tagsets tailored to specific works (Kulkarni and Das, 2012; Das, 2016). There is currently no open benchmark that treats TY identification as a general discourse classification task across śāstric texts.

In this work, we address this gap by introducing TantraTagger, a benchmark dataset for TY-based discourse labelling in Sanskrit śāstric texts. Our contributions are threefold: (i) we formalize TY identification as a discourse-level span classification task; (ii) we release a curated corpus annotated with 39 fine-grained TY labels,[1]; and (iii) we establish a suite of baseline models, ranging from simple heuristics to pretrained transformer-based systems. Together, this work establishes the first open evaluation benchmark for TY-based discourse modelling and provides a foundation for future research on Sanskrit discourse and low-resource language understanding.

1.1 Identifying the Gap

Through frameworks like RST and PDTB, discourse analysis has been thoroughly studied in contemporary NLP; however, these approaches are primarily designed for modern European languages and fail to capture the discourse logic of classical Sanskrit texts. TY, an indigenous and explicitly theorized system that specifies how topics are introduced, doubts are raised, reasons are given, examples are provided, exceptions are stated, and conclusions are drawn, governs the highly formalized argumentative structure of Sanskrit scholarly writing. TY has not received much attention in computational linguistics, despite its crucial role in fields like grammar, philosophy, medicine, and logic. As a result, there isn't a benchmark dataset that defines TY identification as a general discourse-level NLP task, which makes it impossible to evaluate contemporary models systematically and compare approaches in a meaningful way. By creating a benchmark dataset for TY-based discourse labelling, this research will close this gap, facilitate repeatable experimentation, and advance discourse modelling for Sanskrit texts.

1.2 Evaluating the Current State of Language Models (LMs) in Local Contexts

We evaluated a variety of baselines, such as majority-class and random predictors, heuristic rule-based systems, pretrained transformer models, and large language models (LLMs) under zero-shot and few-shot settings, to determine the suitability of current language models for TY-based discourse labelling. In every setting, the results show serious limitations. Naive baselines demonstrate that the task cannot be solved by label frequency or random guessing, as they perform close to chance. Although heuristic rules based on conventional lexical markers produce slight gains, their low coverage and high fall-back rates show that surface cues are insufficient on their own. Although pretrained transformer models, such as google-muril-base-cased, perform best overall, macro-F1 scores are still low, indicating challenges in capturing fine-grained discourse functions. Notably, general-purpose LLMs show a lack of sensitivity to localized, theory-driven discourse distinctions, failing to generalize effectively despite prompt engineering, context injection, and few-shot examples. The need for benchmark datasets and task-specific modelling in this area is highlighted by these findings, which show a significant gap between the apparent linguistic competence of contemporary LMs and their capacity to model structured, indigenous discourse systems like TY.

2 Background and Motivation

2.1 Discourse Analysis in Sanskrit

Sentences and clauses in naturally occurring discourse exhibit systematic relations with their surrounding context. In recent years, discourse analysis has gained considerable prominence in

[1]These TY tags were presented as discourse relations at the 16th International Conference on the History of the Language Sciences (ICoHLS XVI), held in 2024 at Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia; a detailed account is forthcoming.

NLP, where it is broadly understood as the study of linguistic organization beyond the level of individual sentences (Vaze and Kulkarni, 2024, p. 67). Within the domain of computational Sanskrit studies, Kulkarni and Das (2012) employ the *saṅgati* device as an analytical framework for discourse-level interpretation. Further, in the thesis (Das, 2016), discourse annotation is extended to the level of topics and subtopics through *saṅgati*-based tags. At the level of Sanskrit discourse analysis, topic- and subtopic-based annotation has been explored, while more recent work by Vaze and Kulkarni (2024) concentrates specifically on inter-sentential discourse relations. From the perspective of the Pāṇinian grammatical tradition, the question arises as to whether classical Indian schools of thought articulated systematic discourse-relation frameworks to account for textual cohesion. Evidence from the broader Indian scientific and scholastic tradition suggests an affirmative answer. The Mīmāṃsā school, for instance, develops the concept of *saṅgati* as a device for discourse organization, particularly with reference to the Vedic corpus.

Beyond this, TY emerges as a comprehensive set of devices explicitly designed to encode scientific discourse relations. These devices, together with their definitions and modes of application, are attested across a wide range of classical texts, including the *Arthaśāstra*, *Carakasaṃhitā*, *Suśrutasaṃhitā*, and *Aṣṭāṅgahṛdaya*. Dr. W. K. Lele, in his work *Doctrine of the Tantra Yuktis* (Lele, 1981, p. 4), underscores the role of TYs within Pāṇinian Grammar (PG), emphasizing their importance in structuring discourse and maintaining textual coherence.

> "Thus, PA[2] (450 B.C.) has made use of twenty-eight devices of a scientific composition and therefore, it is obligatory upon us to make a thorough study of at least those twenty-eight devices to be able to understand fully the PAS.[3]"

2.2 Related Work and Motivation

Previous computational work on Sanskrit has primarily concentrated on sentence-internal phenomena, including morphological analysis, syntactic dependency parsing, and semantic role labelling. While these efforts have significantly advanced low- and mid-level linguistic processing, higher-level discourse organization remains largely unexplored. In contrast, discourse annotation frameworks for modern languages rely on manually curated datasets that encode rhetorical or pragmatic relations between textual units. Modern linguistics introduces some discourse theories like RST, PDTB, etc. For Indian languages, Subalalitha and Parthasarathi (2012) propose two parsers combining Sanskrit's "*Saṅgati*" with RST for richer discourse parsing. Tested on Tamil and English datasets, the approach achieved high precision and recall, showcasing its language-independent and nuanced capabilities.

Very recently, Dangarikar et al. (2024) introduced a set of sentence-level discourse tags for Indian languages in the volume Samanvaya: An Interlingua for Unity of Indian Languages[4]. In a complementary direction, the USR Bank for Indian languages employs Universal Semantic Representation (USR) to capture sentence meaning at the discourse level by encoding concept- and sentence-level properties such as verbal concepts with TAM (tense–aspect–modality) specifications, semantic categories of nouns, GNP (gender, number, person) features, dependency relations, anaphora, speaker viewpoints, and sentence types. Acting as an interlingua in the translation pipeline, USR representations are generated using heuristics derived from the Pāṇinian Sanskrit grammatical framework (Garg et al., 2023). For Sanskrit, Kulkarni and Das (2012) propose a computational approach to discourse analysis in the *Mahābhāṣya* based on Finite State Automata. Their framework draws upon the relational tag set marking sub-topic connections in the edition of (Kudala, 1912; Joshi, 1968), and further builds on inter-sentential annotation schemes proposed by Ramkrishnamacharyulu (2009), which identify discourse relations typically signalled by explicit connectives, often realized as indeclinable particles. Building on this line of work, the present task focuses on examining relations between two consecutive text spans as

[2] Pāṇini.
[3] *Aṣṭādhyāyī*–Pāṇini.
[4] `https://sanskrit.iitk.ac.in/interlingua/samanvaya/`

marked by such explicit markers and providing a semantic interpretation of the corresponding relationship.

The recent emergence of LLMs has significantly advanced discourse analysis in modern NLP, including investigations into LLM-based discourse modelling within frameworks such as RST (Maekawa et al., 2024). However, these frameworks are largely tailored to modern European languages and do not capture the highly formalized discourse logic of classical Sanskrit śāstra texts, which are governed by the indigenous and explicitly theorized system of TY. Sanskrit currently lacks a discourse-annotated corpus reflecting this argumentative structure, limiting systematic study of reasoning patterns and explanatory strategies central to śāstra literature. Addressing this gap, the present work initiates a discourse-annotated dataset for Sanskrit using TY devices as discourse tags, enabling computational analysis and evaluation using modern language models.[5]

2.3 Importance of Language Models in Modern Workflows

A vast array of applications, including machine translation, text generation, summarization, sentiment analysis, and question answering, are powered by language models, which are now the fundamental part of contemporary natural language processing workflows. They enable systems to interpret and produce human language in previously unattainable ways by learning linguistic patterns and contextual relationships from vast amounts of data. This has significantly broadened NLP's application, enabling tasks like semantic understanding and automated dialogue systems in a variety of fields. They are essential to both academic research and real-world AI systems across industries due to their adaptability through methods like pre-training and fine-tuning (Sajjadi Mohammadabadi et al., 2025).

2.4 Critical Challenges Faced by Existing Language Models in Handling Diversity

The ability of current language models to handle linguistic diversity is severely limited, despite their wide range of capabilities. The majority of models struggle with underrepresented and low-resource languages or language varieties because they were primarily trained on high-resource languages, especially English. This undermines equitable access to NLP technology by creating performance disparities where models are significantly less accurate for tasks involving non-English or less digitally resourced languages. With more than 7,000 languages spoken worldwide, this underrepresentation means that many languages are still underserved by contemporary models and that the advantages of NLP are not equally distributed (Qin et al., 2025). Recent work has shown that large language models perform well on several Sanskrit NLP tasks, including poetry generation (Jagadeeshan et al., 2025), text transformation (Das et al., 2025), and cross-lingual generalization (Akavarapu et al., 2025; Nehrdich et al., 2024; Sandhan et al., 2023). However, these tasks are largely sentence- or form-driven and do not require modelling discourse-level organization. Consequently, the ability of existing language models to capture the conceptual and rhetorical structure underlying TY-based discourse remains largely unexplored.

2.5 Existing Benchmarks

The majority of current work in Sanskrit NLP concentrates on low-level linguistic and lexical resources rather than discourse-level understanding, despite the fact that Sanskrit is one of the oldest documented languages and has garnered significant computational interest. The Indic-NLP catalog,[6] which gathers monolingual and parallel corpora for Indian languages, including Sanskrit, and the Digital Corpus of Sanskrit (DCS),[7] a sizable lemmatized and POS-tagged corpus spanning classical literature and utilized for numerous linguistic tasks, are two exam-

[5] Although *Mīmāṃsā* is traditionally referred to as *Vākyāśāstra* and offers a sophisticated framework for discourse-level analysis—particularly for Vedic interpretation (Subrahmanyam, 2012)—the present study does not engage with this tradition.

[6] `https://github.com/AI4Bharat/indicnlp_catalog`

[7] `http://www.sanskrit-linguistics.org/dcs/`

ples of such resources. There are initiatives like SAHAAYAK 2023 for Sanskrit–Hindi machine translation (Bakrola and Nasariwala, 2023) and the Treebank of Vedic Sanskrit (Hellwig et al., 2020), which offers syntactic annotations under Universal Dependencies for a specific set of Vedic sentences. Advanced word segmentation, morphological tagging, dependency parsing, and OCR correction tasks are among the advanced Sanskrit-specific pretrained models (e.g., ByT5-Sanskrit) that show improvement on structured linguistic tasks. Additional resources include lexicographic databases like IndoWordNet,[8] classical lexicons like the Cologne Digital Sanskrit Lexicon, intrinsic evaluation toolkits for Sanskrit embeddings, and larger text repositories like GRETIL[9] that offer machine-readable Sanskrit texts. There is currently no benchmark for discourse-functional interpretation based on traditional rhetorical systems like TY, particularly for pure Sanskrit discourse spans, despite the abundance of lexical, syntactic, and translation datasets and tools. This motivates the development of our TY annotation dataset and draws attention to a significant gap in the Sanskrit NLP benchmarks currently in use.

3 TY as an Indigenous Discourse Framework

TY represents an epistemic toolkit applied by traditional scholars to organize reasoning, foreground arguments, distinguish claims from examples, and regulate the progression of ideas across discourse. Unlike modern discourse theories, TY devices are not abstract rhetorical postulates but pragmatically motivated strategies embedded in śāstra argumentation.

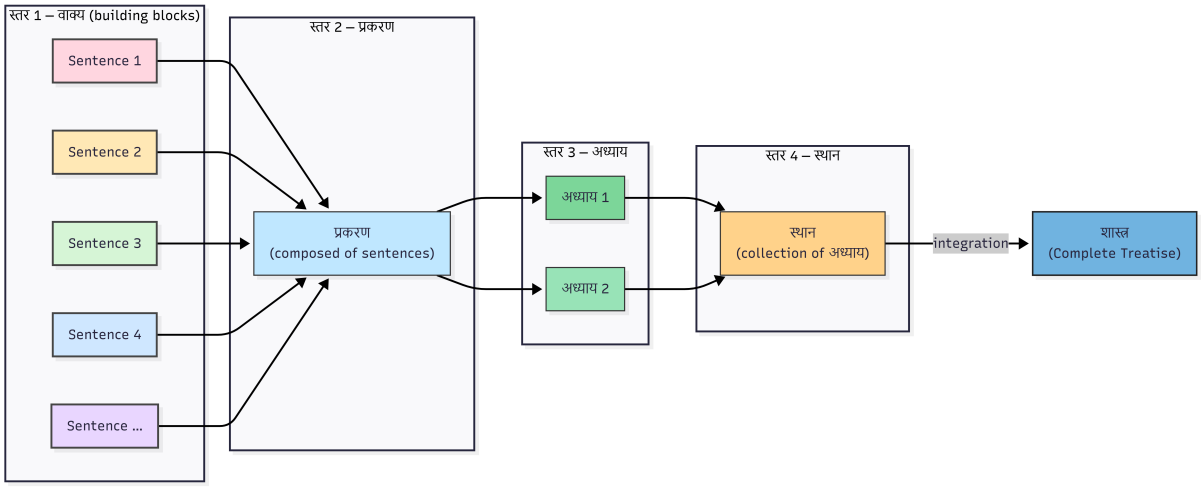

Figure 1: TY as an indigenous discourse annotation framework.

The figure[10] 1 illustrates a hierarchical discourse model for Sanskrit śāstra texts, in which meaning is composed incrementally across multiple textual levels focusing on the textual mapping for book- and chapter-level discourse. At the lowest level, individual sentences **(वाक्य)** function as basic semantic units. These are grouped into coherent discourse segments **(प्रकरण)**, which in turn form larger structural units such as sections and chapters **(अध्याय)**. Multiple chapters are organized into higher divisions **(स्थान)**, and their systematic integration yields the complete treatise **(शास्त्र)**. This layered representation makes explicit how local sentence-level meanings are progressively aggregated into chapter- and book-level discourse structures, providing a principled framework for modelling textual organization and semantic cohesion beyond the sentence boundary. These relations operate across both intra-sentence and inter-sentence levels, making them particularly suitable for discourse-level annotation. Their systematic identification enables computational modelling of reasoning flow in Sanskrit texts.

[8] https://www.cfilt.iitb.ac.in/indowordnet/

[9] https://gretil.sub.uni-goettingen.de/gretil.html#Sanskrit

[10] Figure 1 illustrates the hierarchical organization of a *śāstra* as stated in the textbook **तन्त्रयुक्तिः** (Śaṅkara Śarmā, 1949, 5) under the entry **अथ युक्तयः**, expressed by the following verse: **शास्त्रं स्थानं तथाध्यायं वाक्यं प्रकरणं च वा।**

3.1 Approaches on TY

It may be noted that the primary task of TY has been largely renounced in the context of modern scientific writing (Lele, 1981). Nevertheless, its methodological principles have found renewed application in structured academic practices, particularly in the drafting and organization of theses and dissertations, as demonstrated in recent work (Jayaraman, 2025), among others. Among recent approaches to TY, building on the foundational ideas articulated by Susarla et al. (2023), Lakkundi et al. (2025) introduce IKML (Indic Knowledge Markup Language) along with its associated collaborative web platform, e-Bhāṣya.

In this study, TY devices are modelled as discourse-level tags anchored to explicit textual spans, with each tag encoding the relationship between the annotated segment and its broader discourse context. Rather than treating TY as merely an exegetical apparatus, we formalize it as a structured annotation layer in which each tag represents a distinct discourse function. While many of these functions are associated with identifiable linguistic cues, their correct interpretation cannot be reduced to cue detection alone; they require contextual and inferential reasoning. TY thus offers a fine-grained inventory of discourse functions—such as **प्रदेश** for structuring commentary and **एकान्त** and **नैकान्त** for marking invariability and non-universality, respectively. This level of granularity enables the representation of semantic and pragmatic distinctions that fall outside the scope of *saṅgati*, aligning TY with contemporary discourse theories while retaining its indigenous conceptual grounding.

3.2 TY as an Ideal Testbed for Discourse Analysis

TY is a methodical, theory-based approach to writing and analyzing classical Sanskrit texts. It serves as a discourse-binding framework that controls the organization of arguments, the elaboration of meaning, and the disclosure of implicit content in Śāstra literature. It is made up of a wide range of logical and rhetorical strategies that have been employed for centuries in Indian scientific, philosophical, and medical writings to guarantee the interpretability, coherence, and clarity of intricate knowledge systems. TY is a great stress test for language models' capacity to capture higher-level discourse phenomena beyond word-level or sentence-level features because it encodes deep discourse functions and structural reasoning that go beyond individual lexical cues. It is a rigorous benchmark for assessing context-sensitivity, reasoning, and discourse understanding in LMs because it is frequently used in classical Sanskrit texts to reveal hidden or implied meanings and relationships between concepts. As a result, it presents significant challenges for models trained primarily on modern, high-resource languages and datasets.

4 Dataset Preparation and Annotation

4.1 Overview

The dataset uses 39 TY labels, derived from classical sources and traditional commentarial practice. Each label captures a distinct rhetorical or logical function. The label distribution can be found in Table 1. A complete list with brief meanings and cues is provided in Appendix A.

4.2 Data Collection and Contextual Span Construction

We selected the *Kāśikā-vṛtti*[11] as the primary dataset for sampling of our tag-set for the following reasons:

- It exhibits dense explanatory prose with explicit argumentative progression, characteristic of *śāstra*-style reasoning.
- Its *sūtra*–commentary structure naturally aligns with TY devices such as *nirdeśa*, *vyākhyāna*, and *atikrāntāvekṣaṇa*.

[11]Source of the e-text: Unicode e-text downloaded from GRETIL (`https://gretil.sub.uni-goettingen.de/gretil/1_sanskr/6_sastra/1_gram/jvkasipu.htm`), Göttingen Register of Electronic Texts in Indian Languages.

Given the discourse-oriented nature of TY, the unit of annotation was defined not at the sentence level but at the level of discourse spans.

Each instance in the dataset consists of three components:

- the target discourse span (text),
- the preceding span (prev_text), and
- the following span (next_text).

The inclusion of surrounding spans was designed to provide minimal but sufficient discourse context for interpreting the rhetorical function of the target span. These contextual spans were heuristically selected: for each annotated segment, we extracted the immediately preceding and immediately following textual units from the source text. This strategy was adopted to preserve local argumentative continuity while maintaining manageable input length for computational modelling.

A representative example from the dataset is shown below:


```
"paragraph_id": "JKv_1,1.3",
"text": "vrddhi-guṇau svasañjñayā śiṣyamāṇau ikaḥ eva sthāne veditavyau ",
"prev_text": "paribhāṣā iyaṃ sthāni-niyama-arthā / aniyama-prasaṅge niyamo vidhīyate ",
"next_text": "vakṣyati -- sārvadhātuka-ardhadhātukayoḥ aṅgasya guṇa iti ",
"primary_label": "upadeśa"
```


In this example, the target span ("text") provides an explicit grammatical instruction. It clarifies how the technical categories *vṛddhi* and *guṇa* (types of vowel modifications in Pāṇinian grammar) are to be interpreted in a specific rule environment. The statement restricts their application to a particular phonological position (*ik*), thereby preventing over-generalization. Although it follows a broader meta-rule concerning restriction, the sentence itself functions as a prescriptive directive guiding correct rule interpretation. Accordingly, the primary TY label assigned is *upadeśa*, since the passage delivers an authoritative grammatical instruction.

By incorporating both "prev_text" and "next_text", the dataset preserves local discourse coherence and reflects the fact that TY functions are inherently relational and context-sensitive. While the contextual spans were selected heuristically, this approach provides a consistent and reproducible method for contextual enrichment across the corpus.

4.3 Annotation Procedure

Each discourse span was manually annotated with:

- a unique paragraph identifier (paragraph_id),
- the target span (text),
- contextual spans (prev_text, next_text),
- a single label.

The decision to include the immediately preceding and following spans was motivated by the discourse structure of Śāstric texts. In classical commentarial prose, rhetorical functions often emerge through local argumentative progression rather than isolated sentences. For example:

- an objection (*pūrvapakṣa*) may only be identifiable through its contrast with a preceding claim,
- a conclusion (*nirṇaya*) often depends on resolving earlier doubts,
- an example (*dṛṣṭānta*) may be signaled only through its relation to a prior general rule.

Therefore, context was heuristically incorporated to approximate the minimal discourse window necessary for functional interpretation.

Annotation was performed by a Sanskrit domain expert with formal training in śāstric discourse traditions; this expert[12] is one of the authors of the present work. The annotation process followed classical definitions of TY categories drawn from traditional sources and was applied consistently across spans.

4.4 Dataset Distribution

To ensure representative coverage of core TY devices, the dataset was constructed with controlled sampling across labels. Table 1 summarizes the distribution of instances for selected TY categories.

TY Label	Count	TY Label	Count	TY Label	Count
प्रयोजन	26	अपवर्ग	25	उत्तरपक्ष	25
उपदेश	25	दृष्टान्त	25	पदार्थ	25
पूर्वपक्ष	25	प्रत्युत्सार	25	प्रदेश	25
स्वसंज्ञा	25	योग	24	अनुमत	23
व्याख्यान	23	संशय	23	उद्देश	22
निदर्शन	22	निर्णय	22	निर्देश	22
नैकान्त	22	प्रसङ्ग	22	वाक्यशेष	22
प्रतिप्रश्न	22	विकल्प	22	विपर्यय	22
समुच्चय	22	हेत्वर्थ	22	अनागतावेक्षण	21
नियोग	21	एकान्त	21	निर्वचन	21
अतिक्रान्तावेक्षण	20	अपदेश	20	उपमान	20
सम्भव	20	उद्धार	20	अधिकरण	20
अर्थापत्ति	20	ऊह्य	20	अतिदेश	19

Table 1: Distribution of instances across Tantrayukti categories.

5 Experimental Setup

This section details the experimental configuration used to assess the ability of language models to perform TY-based discourse labelling multi-class classification task in which a given text span is assigned one primary discourse-function label from a fixed inventory of 39 TY labels.

5.1 Dataset and Preprocessing

The evaluation is conducted on a dataset of 870 manually annotated text spans extracted from the *Kāśikā-vṛtti*. Each span is assigned exactly one primary TY discourse label, following the traditional taxonomy (e.g., *atikrāntāvekṣaṇa* 'retrospection', *arthāpatti* 'implication', *pūrvapakṣa* 'prima facie or opponent's view', etc.; see Appendix A for the complete label inventory and definitions).

The dataset is randomly split into training (80%, 696 instances) and validation (20%, 174 instances) sets using stratified sampling to preserve the original label distribution. As illustrated in Table 1 (label frequency distribution), the dataset exhibits a near-balanced distribution, with each of the 39 labels represented by approximately 20–25 instances. This relatively uniform distribution justifies the use of Accuracy and Macro-F1 as the primary evaluation metrics, as both provide interpretable and balanced views of performance.

5.2 Baselines

To rigorously evaluate our proposed task and situate them within the broader landscape of discourse tagging, we implement a comprehensive set of baselines covering the major methodological categories recommended for multi-class discourse-function classification tasks (see Table 2 for a overview). These baselines serve multiple critical purposes:

[12]The author[1] of this present work.

- they establish strict empirical lower bounds on performance,
- they provide linguistic sanity checks by testing whether simple frequency-based, surface-level, or non-neural methods suffice, and
- they offer strong classical reference points against which the gains from large language models can be meaningfully compared.

Category	**Representative baseline(s)**	**Why it matters**
Random / frequency	Majority class; uniform random; distribution-aware random	Establishes strict lower bound; reveals dataset difficulty and label-prior effects
Rule / heuristic	Cue-based heuristic with *sandhi* normalization	Linguistic sanity check; tests sufficiency of traditional TY surface markers
Encoder-only PLM	Fine-tuned `google-muril-base-cased`	Strong neural baseline; reflects contextual embedding approaches in low-resource settings

Table 2: Baseline categories and their purpose.

- Random/Frequency Baselines:
  - Baseline 0 (Majority Class): Always predicts the most frequent training label . Serves as the most conservative non-trivial lower bound.
  - Baseline 1A (Uniform Random): Samples uniformly from the 39-label set.
  - Baseline 1B (Distribution-Aware Random): Samples according to the empirical training distribution.
- Rule/Heuristic Baseline (Baseline 2):
  - A cue-based system using hand-crafted lexical patterns derived from traditional TY markers (e.g., *sah*, *tad*, *yat…tad* for atikrāntāvekṣaṇa; kim nu for *pratipraśnaḥ*; *na/neti* for *apavargaḥ*; full list in Appendix A). Unmatched spans fall back to the majority label. This baseline tests whether surface-level cues suffice for reliable tagging.
- Encoder-Only Pretrained Language Model (Baseline 4):
  - The multilingual Indic-focused BERT variant google-muril-base-cased (Khanuja et al., 2021) is fine-tuned end-to-end on the training set using cross-entropy loss. This provides the strongest supervised non-LLM reference point, capturing contextual embeddings and implicit argumentative structure.
  - This encoder-only baseline operates solely on the target discourse span without explicit neighbouring context. This controlled input formulation isolates span-level representation learning and ensures clear alignment between the input and its annotated discourse label, independent of surrounding argumentative structure.

5.3 Large Language Models

We evaluate five decoder-only large language models to investigate whether reasoning-oriented LLMs can recognize TY discourse functions — a task requiring deep understanding of ancient Indian argumentative and rhetorical structures. The models are:

- Llama 3.3 (70B)
- Qwen 2.5 (72B)
- Mixtral (8×22B)
- DeepSeek-R1 (70B)

- Qwen 3 (32B)

Prompts are designed to elicit step-by-step reasoning aligned with TY definitions.[13] No task-specific fine-tuning is applied; the evaluation focuses exclusively on in-context (zero-shot, few-shot) and constrained (MCQ) prompting capabilities.

5.4 Prompting Paradigms and Evaluation Protocol

Models are assessed under three prompting regimes:

1. **Zero-Shot (ZS)**: The model receives the discourse span, and the full list of 39 TY labels with brief definitions, and must predict the single most appropriate label without any examples.

2. **Few-Shot (FS)**: The prompt includes three representative annotated examples (span + correct label + short explanation) before the test instance.

3. **Multiple-Choice Question (MCQ)**: The task is reformulated as a 4-option multiple-choice question. Each instance includes:

   - 1 correct label
   - 2 randomly sampled distractors
   - 1 semantically difficult distractor deliberately selected from the same TY Discourse Phase.

TY Discourse Phases[14] are six sequential functional categories that group the 39 TY discourse labels according to their role in the structured argumentative flow of classical Sanskrit commentaries: from establishing the topic and context, through expansion and technical reasoning, to managing counter-arguments and reaching conclusions. These phases reflect the natural flow of reasoning in technical Sanskrit treatises—from setting up the topic, through elaboration and technical analysis, to managing objections and reaching conclusions. The phases are:

- **Phase 1 – Foundation & Topic Establishment** introduces the subject, refers back to prior statements for context, or lays the groundwork for the current discussion. Examples: *atikrāntāvekṣaṇa* (reference to past statement), *adhikaraṇa* (topic/subject), *uddeśaḥ* (introduction).

- **Phase 2 – Expansion & Elaboration** expands on the topic by providing reasons, purposes, examples, or illustrations to clarify or support the main point. Examples: *apadeśaḥ* (reason/cause), *dṛṣṭāntaḥ* (example), *prayojana* (purpose/objective).

- **Phase 3 – Technical & Logical Reasoning** involves advanced inference, logical implication, exceptions, or technical manipulation of grammatical or doctrinal rules. Examples: *ūhya* (inference), *arthāpatti* (implication/presumption), *ekānta* (exclusive view), *apavargaḥ* (exception).

- **Phase 4 – Feasibility & Linguistic Analysis** focuses on word meanings, etymology, grammatical arrangement, commentary, or practical linguistic feasibility. Examples: *nirvacana* (etymology), *vyākhyāna* (commentary), yoga (grammatical joining), *padasvarūpa* (word meaning).

[13]The prompt file used in this work is available at `https://github.com/SDRMp/TantraTagger`.

[14]The five prompt-level groupings are designed as a cognitive reasoning scaffold to help models systematically compare discourse functions across dimensions (argumentative role, temporal reference, linguistic function, dialectical positioning, and instructional function). These groupings do not correspond directly to the six Discourse Phases used for MCQ construction, which instead reflect macro-level argumentative progression. The two systems operate at different methodological levels and are not interchangeable.

- **Phase 5 – Counter-Argument & View Management** addresses objections, presents opposing views, offers alternatives, or manages the dialectical exchange (*pūrvapakṣa–uttarapakṣa* structure). Examples: *pūrvapakṣa* ('prima facie or opponent's view'), *uttarapakṣa* (refutation), *pratipraśna* (counter-question), *vikalpa* (alternative).
- **Phase 6 – Conclusion & Directive** resolves the discussion, issues authoritative instructions, extends applications, or concludes the reasoning. Examples: *nirṇaya* (decision/conclusion), *niyoga* (command/directive), *atideśa* (extended application), *anumata* (approval).

The six discourse phases were introduced in this work as an evaluation-oriented grouping of the 39 TY labels. They are not directly derived from classical literature, but were designed to organize related discourse functions and to construct semantically challenging distractors in the MCQ setting. One distractor was intentionally sampled from the same phase as the correct label to increase task difficulty and prevent coarse-grained guessing.

This MCQ formulation constrains the output space, reduces spurious predictions, and forces models to perform fine-grained semantic discrimination — closely mirroring the subtle, context-sensitive distinctions that are central to TY analysis in classical Sanskrit commentaries.

6 Results and Discussion

6.1 Baseline Performance

We first examine three simple frequency-aware and random baselines to quantify the extent to which label distribution alone can explain performance and to provide a clear reference against which non-trivial learning can be measured. Baseline 0, Baseline 1A, and Baseline 1B, as reported in Table 3. These baselines achieve near-chance performance (Macro-F1 ranging from 0.0014 for Majority Class to 0.0172 for Uniform Random), quantitatively indistinguishable from a random guess in terms of macro-averaged metrics. The results are crucial for two reasons:

1. they establish a rigorous empirical lower bound of approximately 0.02 Macro-F1; and
2. they serve as a clear reference point demonstrating that any substantial improvement over this threshold cannot be attributed to trivial effects or random behaviour.

Instead, such gains provide strong evidence that the model is capturing genuine discourse-relevant linguistic and contextual patterns characteristic of TY usage in classical Sanskrit commentaries.

6.1.1 Rule/Heuristic Baseline

Inspired from Kulkarni and Das (2012), we tried to assess whether traditional surface-level linguistic cues—as documented in Appendix A—are sufficient for reliable tagging, we implement a rule-based heuristic baseline (Baseline 2). This system uses a carefully expert curated set of lexical patterns.

The heuristic achieves only modest improvement over random baselines (Accuracy: 0.0632, Macro-F1: 0.0405). Critically, it resorts to fallback prediction (majority label) for 44.3% of instances due to the absence of explicit markers in many discourse spans. This high fallback rate underscores a fundamental characteristic of TY usage in classical Sanskrit commentaries: many functional relations are expressed implicitly through context, argumentative structure, or subtle syntactic arrangement rather than overt lexical signals—a challenge repeatedly noted in prior computational work on Sanskrit discourse (Das, 2016).

6.1.2 Encoder-only Pre-trained Language Model Baseline

As a strong modern neural reference point, we fine-tune google-muril-base-cased (Khanuja et al., 2021), a multilingual BERT variant explicitly pre-trained on Indic scripts, including Sanskrit. This model represents the encoder-only paradigm, contrasting with decoder-only large

language models by relying on bidirectional contextual encoding and supervised fine-tuning for classification.

After standard fine-tuning, muril-base-cased significantly outperforms all previous baselines, achieving Accuracy = 0.2069 and Macro-F1 = 0.1230. The improvement demonstrates the value of contextualized embeddings in capturing nuanced implications, long-range dependencies, and implicit argumentative moves that are characteristic of Pāṇinian commentaries like *Kāśikā-vṛtti*.

Baseline	**Accuracy**	**Macro F1**
0: Majority Class	0.0287	0.0014
1A: Uniform Random	0.0172	0.0172
1B: Distribution-Aware Random	0.0115	0.0145
2: Heuristic (cue-based)	0.0632	0.0405
Encoder-only PLM: `google-muril-base-cased`	0.2069	0.1230

Table 3: Performance of baseline models on the TY discourse labelling task.

These baseline results collectively highlight the non-trivial nature of TY discourse tagging: surface cues provide limited coverage, frequency priors are insufficient, and even strong contextual encoders leave substantial room for improvement—setting the stage for our investigation of large language models in Section 6.2.

6.2 LLM Performance

We evaluate five decoder-only large language models — Llama 3.3 (70B), Qwen 2.5 (72B), Mixtral (8×22B), DeepSeek-R1 (70B), and Qwen 3 (32B) — under three prompting regimes: **Zero-Shot (ZS)**, **Few-Shot (FS)**, and **Multiple-Choice Question (MCQ)**. Table 4 presents the complete performance results across all LLM models and settings. Following the near-balanced nature of our dataset, we emphasize **Accuracy** and **Macro-F1** as the primary evaluation metrics.

Model	**Setting**	**Accuracy**	**Macro F1**
Qwen 2.5 (72B)	MCQ	0.2471	0.2263
Llama 3.3 (70B)	MCQ	0.2414	0.2271
Mixtral (8×22B)	MCQ	0.2126	0.1997
Qwen 3 (32B)	MCQ	0.1494	0.1974
DeepSeek-R1 (70B)	MCQ	0.1321	0.1353
Llama 3.3 (70B)	FS	0.1609	0.1311
Qwen 2.5 (72B)	FS	0.1322	0.1307
Qwen 2.5 (72B)	ZS	0.1667	0.1230
Mixtral (8×22B)	ZS	0.1379	0.1006
Qwen 3 (32B)	ZS	0.1400	0.0823
Llama 3.3 (70B)	ZS	0.1034	0.0663
DeepSeek-R1 (70B)	FS	0.1034	0.0561
DeepSeek-R1 (70B)	ZS	0.0575	0.0528
Qwen 3 (32B)	FS	0.0690	0.0676
Mixtral (8×22B)	FS	0.1111	0.0729

Table 4: Performance of large language models on TY discourse labelling task.

The most striking pattern is the clear superiority of MCQs. In this setting, the two strongest models — Qwen 2.5 (72B) and Llama 3.3 (70B) — achieve Macro-F1 scores of 0.2263 and 0.2271, respectively, surpassing the best supervised encoder-only baseline (muril-base-cased, Macro-F1 = 0.1230) . This gain is particularly notable given the fine-grained nature of TY distinctions and the deliberate inclusion of semantically difficult distractors, which force models to perform subtle comparative reasoning (e.g., distinguishing retrospection from future reference, or doubt from counter-question). Model scale also emerges as an important factor: the 70B–72B class models consistently outperform the smaller Qwen 3 (32B) across all prompting regimes, with the gap widening most noticeably under MCQ.

In contrast, zero-shot and few-shot chain of thoughts prompting produce disappointing results, with Macro-F1 scores ranging from 0.0528 to 0.1311 — often below or only marginally above the supervised muril baseline. Despite the models' demonstrated ability to handle Sanskrit morphology and generate fluent text (refer to section 2.4), their poor performance in open-ended settings indicates that the subtle, culturally and textually specific discourse functions of TY are not intrinsically encoded in current decoder-only LLMs without additional scaffolding. The MCQ along with FS, ZS results reveal considerable headroom for future improvement, particularly through parameter-efficient fine-tuning (e.g., LoRA) or integration of domain-specific knowledge.

To further contextualize these findings, we also evaluated the proprietary GPT- 5 model (via the ChatGPT interface) on a subset of 50 samples from our validation set, focusing on the MCQ and FS settings.The results are shown in Table 5. This preliminary assessment was conducted to gauge the potential of closed-source models, which often benefit from larger-scale training and proprietary optimizations not available in open-source alternatives. In the MCQ setting, GPT-5 achieved an Accuracy of 0.5769 and a Macro-F1 of 0.3691, substantially outperforming our best open-source LLM results (e.g., Qwen 2.5 at 0.2471 Accuracy and 0.2263 Macro-F1 on the full set). For the FS setting, GPT- 5 obtained an Accuracy of 0.2979 and a Macro-F1 of 0.2295, which is competitive with or slightly better than our top open-source FS performers but still highlights the challenges. While these results are promising and suggest that proprietary models may capture more nuanced TY discourse patterns—possibly due to broader exposure to multilingual and classical texts during pre-training—they are limited to a small subset and warrant full-scale evaluation in future work. At the same time, even the best MCQ performance (Macro-F1 0.37 on the subset with GPT-5) remains substantially below what would be considered satisfactory for practical fine-grained discourse tagging in a domain as specialized as TY discourse analysis. There is therefore still a huge headroom for improvement. These directions motivate research explorations closing the gap toward robust, production-ready performance on Sanskrit discourse labelling tasks.

Setting	Accuracy	Macro F1
MCQ	0.5769	0.3691
FS	0.2979	0.2295

Table 5: GPT-5 results

6.3 Error Analysis: Linguistic Insights into LLM Failures in TY Labelling

We perform a qualitative error analysis on a subset of LLM generations from the validation set in order to comprehend the consistently poor performance of LLMs on TY discourse labelling (e.g., Macro F1 scores ranging from 0.05 in zero-shot to 0.23 in MCQ settings). This analysis emphasizes the linguistic and task-specific difficulties present in classical Sanskrit discourse by drawing on the examples given, which show both successes and failures. We classify errors according to common patterns found in all models (e.g., Qwen 2.5 72B and Llama 3.3 70B).

1. **Confusion Between Semantically Overlapping Labels**: Many TY labels have close conceptual boundaries, requiring fine pragmatic distinctions that LLMs struggle to make reliably. For instance: Text: "*apare tvāhuḥ, yady apy ekasya ākhyātasya samīpe kiṃśabdaḥ śrūyate, tathāpi sarvasya saṃśayaviṣayasya tena yogaḥ iti ubhayatra pratiṣedhena bhavitavyam iti /*"

    - True Label: **संशय** (Doubt) — The span raises conditional uncertainty about rule application in both contexts ("*ubhayatra*"), using markers like "*yady apy...tathāpi*" to express unresolved inquiry, a preparatory move in classical discourse (e.g., setting up for refutation).

- LLM Prediction: **अर्थापत्ति** (Implication/Presumption) — The model focuses on logical connection ("*tena yogaḥ*") and presumes a general principle applies despite conditions.
- Why this error? Both **संशय** and **अर्थापत्ति** involve inference from context, but the former emphasizes doubt/exploration, while the latter deduces unstated conclusions. LLMs, trained on modern multilingual data, over generalize implication for conditional structures and miss the rhetorical intent (doubt as a distinct stage in TY). Such overlaps affect 30–40% of errors in our samples.

2. **Surface-Level Detection vs. Pragmatic Instability in Reasoning**: From the reasoning traces produced by the LLMs, we consistently observed that models reliably detect overt surface-level interrogative structures (e.g., the phrase "*kiṃ nu*" in "*kiṃ nu yaṇā bhavati iha na siddhaṃ yvāvidutoryadayaṃ vidadhāti*"). In many runs, this detection led to the correct prediction of **प्रतिप्रश्न** (Counter-Question), with the model explicitly noting the interrogative form as a marker of inquiry. However, the same reasoning traces also revealed frequent instability: the LLMs often wavered between **प्रतिप्रश्न** and **संशय** (Doubt), debating internally whether the utterance represented "seeking clarification by challenging prior assumptions" (the pragmatic intent of a counter-question) or merely "general doubting." This oscillation occurred even when the final extracted label was correct, indicating shallow confidence in the decision. This pattern highlights a fundamental limitation: LLMs demonstrate strong sensitivity to explicit lexical and syntactic cues (such as interrogatives or clear negations, as seen in more reliable predictions for **अपवर्ग** (Exception) with markers like "*na siddhaṃ*" or "*varjayitvā"*). In contrast, they struggle to resolve the deeper pragmatic and context-dependent nuances required when discourse functions rely on implicit presuppositions, elliptical constructions, or *sandhi*-fused compounds—features pervasive in *Mahābhāṣya*-style argumentation. Such implicit cases, which lack strong surface signals, account for approximately 40% of the failures across all prompting regimes.

3. **Broader Task Difficulties Contributing to Low Results**:
   - **Fine-Grained Taxonomy**: 39 labels grouped into TY phases (see section 5.4) demand nuanced discrimination (e.g., **पूर्वपक्ष** vs. **उत्तरपक्ष**), akin to RST parsing where LLMs struggle with subtle relations. MCQ helps by constraining choices but exposes hard distractor confusions.
   - **Low-Resource Archaic Domain**: Even balanced data lacks the scale/volume of modern corpora; LLMs generalize poorly without adaptation.
   - **Instability in reasoning**: Step-by-step reasoning induces hallucinations or wavering (e.g., debating doubt vs. implication), unlike supervised encoders that capture context more stably.

7 Conclusion and Future Work

This work introduces the first benchmark dataset for discourse-level analysis of Sanskrit texts grounded in TY devices. By annotating excerpts from the *Kāśikā-vṛtti*, we provide a principled resource for modelling discourse organization in Sanskrit śāstric literature. Experiments across multiple model families and prompting paradigms show that large language models can partially capture TY-based discourse functions, while also highlighting persistent challenges posed by fine-grained label distinctions and the limited availability of supervised training data.

The TantraTagger benchmark lays a strong foundation for discourse-level modelling in classical Sanskrit, but several promising directions remain to extend its scope and impact. Immediate next steps include parameter-efficient fine-tuning (e.g., LoRA) of the strongest prompting models on the full training set to further improve performance on rare and pragmatically subtle TY labels. Building on this, we plan to advance toward end-to-end span detection + classification, enabling automatic identification of discourse units and label assignment over continuous

commentary passages. Longer-term goals encompass hierarchical discourse parsing to construct full rhetorical trees from *Kāśikā-vṛtti*-style texts, cross-domain rhetorical transfer to other Indic śāstra traditions (e.g., *Nyāya*, *Mīmāṃsā*), and extraction of structured argument graphs that represent dialectical relations between TY-labelled spans. We also intend to incorporate the *Saṃgraha* theory (Kulkarni, 2021) into the analytical framework, examining how principles of condensation and thematic integration interact with TY-based discourse structure.

Due to the highly specialized and theory-driven nature of TY categories, recruiting multiple expert annotators is resource-intensive. We plan to extend the dataset by involving multiple trained Sanskrit scholars to annotate additional discourse spans. This will enable the calculation of inter-annotator agreement.

These extensions will bridge TY with modern discourse theories and support advanced applications such as automated argument mining, visual argumentation maps, and knowledge-graph integration for classical Indian reasoning systems.

References

V.S.D.S. Mahesh Akavarapu, Hrishikesh Terdalkar, Pramit Bhattacharyya, Shubhangi Agarwal, Dr. Vishakha Deulgaonkar, Chaitali Dangarikar, Pralay Manna, and Arnab Bhattacharya. 2025. A Case Study of Cross-Lingual Zero-Shot Generalization for Classical Languages in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2745–2761, Vienna, Austria, July. Association for Computational Linguistics.

Vishvajitsinh Bakrola and Jitendra Nasariwala. 2023. SAHAAYAK 2023: The Multi-Domain bilingual parallel corpus of Sanskrit–Hindi for machine translation. arXiv:2307.00021 [cs.CL].

Chaitali Dangarikar, Arnab Bhattacharya, Karthika N J, Hrishikesh Terdalkar, Pramit Bhattacharyya, Annarao Kulkarni, Chaitanya S Lakkundi, Ganesh Ramakrishnan, and Shivani V. 2024. *Samanvaya: An Interlingua for Unity of Indian Languages*. Central Sanskrit University, India.

Kunal Kingkar Das, Manoj Balaji Jagadeeshan, Nallani Chakravartula Sahith, Jivnesh Sandhan, and Pawan Goyal. 2025. Still Not There: Can LLMs Outperform Smaller Task-Specific Seq2Seq Models on the Poetry-to-Prose Conversion Task? arXiv:2511.08145 [cs.CL].

Monali Das. 2016. *Discourse Analysis of Sanskrit Texts: First Attempt towards Computational Processing*. Ph.D. thesis, University of Hyderabad, Hyderabad, India.

Kirti Garg, Soma Paul, Sukhada, Riya Kumari, and Fatema Bawahir. 2023. Evaluation of universal semantic representation (USR). In Julia Bonn and Nianwen Xue, editors, *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 13–22, Nancy, France, June. Association for Computational Linguistics.

Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, and Ralph Bunker. 2012. A distributed platform for sanskrit processing. In *Proceedings of COLING 2012*, pages 1011–1028.

Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2020. The treebank of vedic Sanskrit. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5137–5146, Marseille, France, May. European Language Resources Association.

Manoj Balaji Jagadeeshan, Samarth Bhatia, Pretam Ray, Harshul Surana, P AkhilRajeev, Priya Mishra, Annarao Kulkarni, Ganesh Ramakrishnan, AP Prathosh, and Pawan Goyal. 2025. Chandomitra: Towards Generating Structured Sanskrit Poetry from Natural Language Inputs. arXiv:2506.00815 [cs.CL].

M. Jayaraman. 2025. *Tantrayukti: IKS-Based Handbook for Thesis Construction*. INDICA.

S. D. Joshi. 1968. *Patañjali's Vyākaraṇa Mahābhāṣya: Samarthāhnika (P2.1.1)*. Center of Advanced Study in Sanskrit, Poona, 1st edition.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual representations for Indian Languages. arXiv:2103.10730 [cs.CL].

Sivadatta D. Kudala. 1912. *Patañjali's Vyākaraṇa Mahābhāṣya with Kaiyaṭa's Pradīpa and Nāgeśa's Uddyota*, volume 2. Nirṇaya Sāgara Press, Bombay.

Amba Kulkarni and Monali Das. 2012. Discourse analysis of Sanskrit texts. In Eva Hajičová, Lucie Poláková, and Jiří Mírovský, editors, *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 1–16, Mumbai, India, December. The COLING 2012 Organizing Committee.

Malhar Kulkarni. 2021. Introducing fresh terminology to cognitively explain sentence meaning in the paninian grammatical tradition. *Studi Classici e Orientali*, 67(1):487–495.

Chaitanya S Lakkundi, Gopalakrishnan Rajaraman, and Sai Rama Krishna Susarla. 2025. IKML: A Markup Language for Collaborative Semantic Annotation of Indic Texts. In *Computational Sanskrit and Digital Humanities-World Sanskrit Conference 2025*, pages 109–130.

V.K. Lele. 1981. *The Doctrine of the Tantrayukti-s: Methodology of Theoretico-scientific Treatises in Sanskrit*. Chaukhamba Surabharati studies. Chaukhamba Surabharati Prakashan.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta, March. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics*, pages 85–95. Springer.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. arXiv:2409.13920 [cs.CL].

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

K.V. Ramkrishnamacharyulu. 2009. Annotating Sanskrit texts based on Śābdabodha systems. In *International Sanskrit Computational Linguistics Symposium*, pages 26–39. Springer.

Seyed Mahmoud Sajjadi Mohammadabadi, Burak Cem Kara, Can Eyupoglu, Can Uzay, Mehmet Serkan Tosun, and Oktay Karakuş. 2025. A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, 14(18).

Jivnesh Sandhan, Anshul Agarwal, Laxmidhar Behera, Tushar Sandhan, and Pawan Goyal. 2023. SanskritShala: A Neural Sanskrit NLP Toolkit with Web-Based Interface for Pedagogical and Annotation Purposes. arXiv:2302.09527 [cs.CL].

CN Subalalitha and Ranjani Parthasarathi. 2012. An approach to discourse parsing using sangati and Rhetorical Structure Theory. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 73–82.

Korada Subrahmanyam. 2012. *Mahāvākyavicāraḥ*. Śrī lakṣaṇārya bhavyasmṛti Granthamālā - dvitīya puṣpam, Bhāgyanagaram.

Sai Susarla, Suryanarayana Jammalamadaka, Vaishnavi Nishankar, Siva Panuganti, Anupama Ryali, and S Sushrutha. 2023. Shaastra maps: Enabling conceptual exploration of indic shaastra texts. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 174–187.

Saee Vaze and Amba Kulkarni. 2024. Inter Sentential Discourse Relations. In Arnab Bhattacharya, editor, *Proceedings of the 7th International Sanskrit Computational Linguistics Symposium*, pages 67–83, Auroville, Puducherry, India, February. Association for Computational Linguistics.

Śaṅkara Śarmā. 1949. *Tantrayuktiḥ*. Vaidyasārathi Mudraṇālaya, Kottayam.

A List of TY devices with concepts & cues

No.	TYs	Description	Cues
1	अतिक्रान्तावेक्षण	Reference to a statement stated previously.	यद्-तद्, तत्र, तद्, प्राक्, प्रागुक्तम्, पूर्वस्मिन्, पूर्वत्र
2	अतिदेश	Extension of application to analogous topics.	वत्, प्राग्वत्, तथा, तथैव, अतिदेश, इत्यतिदेश
3	अधिकरण	Topic of a chapter, section, or sentence.	अथ, आरभते..., इति...प्रकरणम्, इति...अध्यायः
4	अनागतावेक्षण	Reference to a statement to be stated later.	यद्-तद्, वक्ष्यामि-मः, प्रवक्ष्यामि, वक्ष्यति-ते, निरूपयिष्यति-ते
5	अनुमत	Acceptance or approval of the opinion of others.	इत्यन्ये, मते, मतेन, तन्मते, इत्येके, इत्यपरे, इत्याह
6	अपदेश	Suggestion of a cause–effect relationship.	यदि-तर्हि, यदि-ततः, यदि-तदा, यतः-ततः, इति हेतोः
7	अपवर्ग	Statement of an exception to a general rule.	न, नहि, न वा, न च, न तु, नापि, अपवाद
8	अर्थापत्ति	Implied meaning inferred from what is explicitly stated.	इति...गम्यते
9	उत्तरपक्ष	Reply refuting the objection.	तदुक्तम्, तदुच्यते, चेत्...उच्यते
10	उद्देश	Brief mention of a topic.	इति, निर्देशः, इत्यादि, इत्यादयः
11	उद्धार	Extraction of relevant information from a statement.	इति ज्ञापयति, ज्ञापितम्, इत्युक्तं भवति
12	उपदेश	Advice given by a trustworthy authority.	लिङ्, लोट्, कृत्यप्रत्यय
13	उपमान	Establishing an unknown fact through comparison.	एवम्, तथैव, तथा सति
14	ऊह्य	Inference of unstated information by reasoning.	विपरिणम्यते, योगः विभज्यते, कल्प्यते
15	एकान्त	Statement admitting no exception.	नित्यम्, नियमसूत्र
16	दृष्टान्त	An example illustrating a statement.	उदाहरणम्, यथा, तद्यथा, इत्युदाहरणम्
17	निदर्शन	Illustrated statement supported by an example.	यथा... इत्यत्र, यथात्र
18	नियोग	An injunction that must be followed.	यजेत, जुहुयात्, मा प्रमदः
19	निर्णय	A definite conclusion or determination.	इति सिद्धम्, उपसंहरति, उपसंहरन्, अत्र निर्णयः, इति भावः, अयं भावः, मन्मते, परे तु
20	निर्देश	Elaboration of a previously mentioned topic.	इति, निर्देशः, इति निर्देशात्
21	निर्वचन	Etymological explanation for better understanding.	रूपम्, इति विग्रहः, निर्वचनम्, व्युत्पत्तिः, निष्पन्नम्
22	नैकान्त	A statement that is not universally valid.	बहुलम्, अनित्य
23	पदार्थ	Meaning conveyed by an individual word.	इत्यर्थः, अयमर्थः, पदार्थः, शब्दार्थः, तस्यार्थः, अस्यार्थः
24	पूर्वपक्ष	Objection raised by an opponent.	ननु, आक्षेपः, प्रश्नः, किम्
25	प्रत्युत्सार	Supplying omitted or elliptical content.	अनुवर्तते, अनुवृत्तिः, अध्याहारः, अनुवर्तन्ते
26	प्रदेश	Initial indicator of a sequence or list.	इति, इत्यादयः, इत्यादि, इत्यादीनि, इत्यादौ चेयम्
27	प्रयोजन	Purpose or objective of an action.	अर्थ, प्रयोजनम्, तदर्थम्, इत्येतदर्थम्
28	प्रसङ्ग	Introduction of an appropriate context.	प्रसङ्गे, स्थाने, विषये, इत्यधिकारे, इति परतः
29	योग	Logical or grammatical arrangement of words.	अन्वय, योजन, इत्यन्वयः
30	वाक्यशेष	Unexpressed part of a sentence to be inferred.	इति शेषः, अयं शेषः, इति भावः, अनुगन्तव्य, अवगन्तव्य, इति वक्तव्यम्
31	विकल्प	Presentation of alternatives.	विभाषा, वा, अथ वा, अन्यतरस्याम्, विकल्प, विकल्प्यते
32	विपर्यय	Acceptance of an opposite meaning.	यदि-न, असति भवति, वर्जयित्वा, नञ्समास
33	व्याख्यान	Detailed explanation covering all aspects.	...व्याख्या, ...भाष्ये, ...वृत्तौ
34	संशय	Doubt between two opposing views.	संदेहः, संशयः, आशङ्का, शङ्का, शङ्कते, संशय, संदेह, आशङ्कोत, शङ्कते, शङ्का स्यात्
35	समुच्चय	Combination of multiple entities in one statement.	च, अपि च, किञ्च, अथ च, वा, अपि च, च
36	सम्भव	Indication of possibility.	सम्भाव्यते, सम्भवति, सम्भावना, लिङ् (सम्भावनायाम्)
37	स्वसंज्ञा	Domain-specific technical term.	संज्ञा, ...संज्ञः
38	हेत्वर्थ	Statement expressing a cause.	तस्मात्, अनेन, अस्मात्, अनेन कारणेन, यतो हि, हि, अत एव, अतः, कारणात्, अस्मात् कारणात्, हेतोः, इत्यनेन
39	प्रतिप्रश्न	Counter-question in an argumentative exchange.	ननु, ननु च, ननु वद, किमुत, वा, पुनः आक्षिपति, प्रतिप्रश्नः, किम्, पुनराक्षेपः

Table 6: Tantrayukti devices with descriptions and cues