Discourse Level Tagger for Mahābhāṣya - a Sanskrit Commentary on Pāṇini's Grammar

Monali Das

University of Hyderabad
monactc.85@gmail.com

Amba Kulkarni

University of Hyderabad
ambapradeep@gmail.com

Abstract

Mahābhāṣya is an important commentary on Pāṇini's grammar for Sanskrit and is highly structured. The traditional scholars have tagged it manually showing its underlying discourse structure. The traditional grammar also discusses clues for discourse level annotations. Taking into account these clues we have developed an automatic tagger for tagging the *Mahābhāṣya*. This tagger is described in this paper, along with its performance evaluation. We have also extended this tag-set to on another important text Śābarabhāṣya.

1 Introduction

Discourse level analysis is an important module in NLP which takes us beyond sentence level analysis. While grammar is basically about how words combine to form sentences, text and discourse analysis is about how sentences combine to form texts (Salike, 1995).

Four types of discourse structures are discussed in the modern linguistics.

1. Topic Structure: The topic structure gives a broad outline of the topics in a given text.

2. Functional Structure: It identifies various sections within a topic serving different functions.

3. Event Structure: This identifies various events in the discourse and show them on the time-line.

4. Coherence Relations: Based on the linguistic clues and from the functional and event structure various coherence relations are identified.

(Mann and Thompson, 1988) proposed a Rhetorical Structure Theory (RST) for discourse analysis. The intentional and informational relations between adjacent texts are represented in the form of a binary tree structure. (Wolf and Gibson, 2005) suggested a Dependency Graph Structure (DG) to show the relations between the topic and sub-topic segments. The segments here unlike RST need not be adjacent. (Polanyi, 1988) suggested a Linguistic Discourse Model (LDM) which resembles RST. In this model each node inherits the properties of its parent node and a parent node is an interpretation of its children nodes and the relations between them.

There have been several efforts in the recent past in computational linguistics in the field of discourse analysis. The major efforts is the proposal of discourse tag-set by (Joshi et al., 2007) in the form of Penn Dependency TreeBank (PDTB). The annotation scheme originally developed for English was further extended to Arabic (Alsaif, 2012), Chinese (Xue, 2005), Turkish (Zeyrek and Webber, 2008) and Hindi (Prasad et al., 2009). (Mladova et al., 2008) suggested a scheme for Czech and English (Praguian Discourse TreeBank) which is consistent with the dependency tree bank of Prague.

The automatic annotation and evaluation help us in understanding the discourse structures further. (Jurafsky et al., 2000) presented a framework for modelling and automatic classification of dialogue acts with an accuracy of 74.7%. (Jinova et al., 2012) experimented with semi-automatic annotation of intra-sentential discourse relations in PDT. The automatic annotation of discourse structure of various user forums for thread solvedness classification accuracy is an important application of the automatic discourse structure analysis being ex-

plored by (Baldwin et al., 2012), (Xi et al., 2004), (Fortuna et al., 2007), (Kim et al., 2010).

All these discourse theories have been centered around English and other European languages, though PDT and RST have been recently being applied to Arabic, Hindi and Tamil as well. These theories may be tried on Sanskrit corpus as well. But there are two important considerations. The first and foremost is that Sanskrit comes with a rich heritage of linguistic theory of its own. The rich grammatical tradition of India has a vast literature on almost all aspects of language analysis ranging from phonology to discourse analysis. Thus it is natural to look at these theories which have been discussed over centuries, with a focus on Sanskrit literature. The second reason is Sanskrit literature comes with annotations, though not by original authors, but by the later commentators. These annotations are done following the grammatical theories. Hence it is natural to use these theories for computational models related to Sanskrit.

In this paper in the next section we give a brief survey of Indian theories of discourse analysis. In section 3 we take up one particular Sanskrit text viz. *Mahābhāṣya* of *Patañjali* which is available as a tagged text and discuss its internal structure providing tagging guidelines for the tag-set. In section 4 we look at the clues suggested in the Sanskrit literature and built a FSA for automatic tagging of the discourse relations. In section 5 we discuss the extension of this tagger to handle another important text - *Śābarabhāṣya* - a commentary on *Mīmāṁsā sūtra*.

2 Indian Theories of Discourse Analysis

Indian Grammatical Theories evolved in order to understand the Vedic literature and protect them. The theories developed in three directions. The first one, dealing with word and sentence formation and analysis. This lead to the establishment of *Vyākaraṇa* (grammar) school. The second one dealing with the semantics, logic and inference lead to the formation of *Nyāya* (logic) school of thought and the third one dealing with discourse analysis resulted in *Mīmāṁsā* (exegesis) school. As

we glance the Indian literature, we notice discussions on various aspects of textual analysis which basically deal with the coherence of texts. The coherence is judged at different levels right from the relations between two adjoining sentences to the coherence at the level of texts with respect to the discipline. The coherence levels discussed in the literature are as follows -

1. Śāstra saṅgati (Bhaṭṭācārya, 1989): This is a subject level coherence.

2. Adhyāya saṅgati (Bhaṭṭācārya, 1989): This is a chapter or a book level coherence. Among the scientific literature in Sanskrit, we find three different trends under the adhyāya saṅgati. One is known as *Bhāṣya paramparā*. Here the original text is in the form of sūtras (compact aphorisms). This is followed by a commentary explaining the sūtras, optionally followed by an explanation (ṭīkā), a note (ṭippaṇi) etc. The commentaries may be nested, i.e. there is a commentary on the original sūtras and then sub-commentary on this commentary, and further sub-commentary on the sub-commentaries and so on. At each stage the number of commentaries may be more than one. The sūtras (compact aphorism) as well as the commentaries and sub-commentaries follow a certain discourse structure.

Another trend is where the original text establishes a theory, and the later scholars write criticisms on it attacking the original view and proposing a new view. This trend is known as *Khaṇḍana-maṇḍana paramparā*. And there can be a series of such texts criticizing the previous theory in the series and proposing a new theory. The structure of these texts then leads to a tree structure, where the siblings indicate different criticisms of the same text leading to different view points.

The third trend is to write *Prakaraṇa granthas* (books dealing with a specific important topic among several topics discussed in the texts in sūtra form). These books are thus related to the original

sūtra texts, but also have their own nested commentaries.

3. Pāda saṅgati (Bhaṭṭācārya, 1989): This is a section level coherence.

4. Adhikaraṇa saṅgati (Bhaṭṭācārya, 1989): This is a topic level coherence. *Mīmāṁsakas* (exegesists) discussed about this saṅgati. Further *Naiyāyikas* (logicians) discussed about six topic level relations (Śāstri, 1916). They are

   (a) Prasaṅga - Corollary.
   (b) Upodghāta - Pre-requisite.
   (c) Hetutā - Causal dependence.
   (d) Avasara - Provide an opportunity for further inquiry.
   (e) Nirvāhakaikya - The adjacent sections have a common end.
   (f) Kāryaikya - The adjacent sections are joint causal factors of a common effect.

5. Sub-topic level analysis: Under this level we see two different tag sets one found in exegesist's analysis and the other in grammarian's analysis.

   A. Exegesist's (Mīmāṁsakas) tag-set (Bhaṭṭācārya, 1989):
      (a) Ākṣepa - Objection
      (b) Dṛṣṭānta - Example
      (c) Pratyudāharaṇa - Counter-example
      (d) Prāsaṅgika - Corollary
      (e) Upodghāta - Pre-requisite
      (f) Apavāda - Exception
   B. Grammarians (Vaiyākaraṇas) tag-set (Joshi, 1968):
      (a) Praśna - Question
      (b) Ākṣepa - Objection
      (c) Samādhāna - Justification
      (d) Uttara - Answer
      (e) Vyākhyā - Elaboration

6. Inter-sentential relations: In (Ramakrishnamacaryulu, 2009), inter-sentential relations are classified into 9 sub-headings. These are given below along with the lexical clues which mark these relations.

   (a) Hetuhetumadbhāvaḥ (cause effect relationship): yataḥ (because), tataḥ (hence), yasmāt (because), tasmāt (hence), ataḥ (therefore).
   (b) Asāphalyam (failure): kintu (but).
   (c) Anantarakālīnatvam (following action): atha (then).
   (d) Kāraṇasatve'api kāryābhāvaḥ / kāraṇābhāve'api kāryotpattiḥ (non-productive effort or product without cause): yadyapi (even-though), tathāpi (still), athāpi (hence).
   (e) Pratibandhaḥ (conditional): yadi (if), tarhi (then), cet (if/provided), tarhyeva (then only).
   (f) Samuccayaḥ (conjunction): ca, apica (and also).
   (g) Pūrvakālīkatvam: The non-finite verb form ending with suffix *ktvā* 'adverbial participial'.
   (h) Prayojanam (Purpose of the main activity): The non-finite verb form ending with suffix *tumun* 'to-infinitive'.
   (i) Samānakālīkatvam (Simultaneity): The non-finite verb form ending with suffix *Śatṛ and Śānac* 'present participle'.

Thus we notice that the Indian grammatical theories have, to a large extent, guidelines for analysis of a text at different levels of discourse.

3 Structure of Mahābhāṣya

Pāṇini's *Aṣṭādhyāyi* (circa 500 BC) is an important milestone in the developmental history of Indian theories of language analysis. *Aṣṭādhyāyi* is in sūtra[1] (compact aphorism) form and hence to understand it one needs an explanation. Patañjali wrote a detailed commentary on *Aṣṭādhyāyi* known as *Mahābhāṣya* (the great/large commentary). It provides explanation of these sūtras and also throws light on important aspects of linguistic analysis. It is the second important milestone of the Indian grammatical tradition. The text

---

[1]alpākṣaraṁ asandigdhaṁ sāravat viśvato mukhaṁ. astobham anavadyaṁ ca sūtram sūtravido viduḥ.
A sūtra contains minimum number of words, is unambiguous, contains the essence of the topic, has universal validity and is devoid of any faults.

of *Mahābhāṣya* follows a very well defined discourse structure.

Mahābhāṣya contains commentaries on various sūtras and the supplementary sūtras called as *Vārttika*-s. The relevant level of discourse analysis for *Mahābhāṣya* is then the adhikaraṇa (topic) level analysis and all the lower level analysis viz. sub-topic level and inter-sentential. The sūtra/vārttika sets up the new topic and all the discussions under this follow a well defined structure. The topic level analysis of a commentary on one sūtra was taken up ealier (Kulkarni and Das, 2012). Figure 1 shows the structure of a commentary on the sūtra P2.1.1 (*Samarthaḥ Padavidhiḥ*).
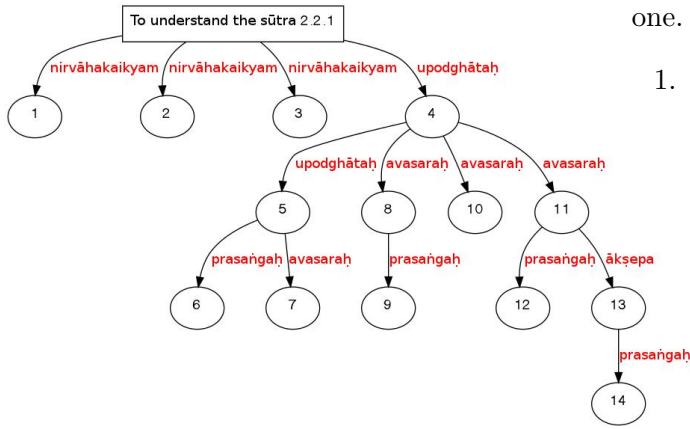

Figure 1: Structure of a commentary '*samarthaḥ padavidhiḥ*'

There are 14 topics under the main headings of the sūtra P2.1.1 (*samarthaḥ padavidhiḥ*). These 14 topics are related to each other by a set of relations, which show the coherence of the discussion under this sūtra. These relations are the topic level or *adhikaraṇa saṅgati*-s. The tagging at this level involves semantic analysis of the text.

The original Mahābhāṣya does not have any of the mark-ups. The first version of marked up texts is published by Nirṇaya Sāgara Press, Bombay in early 1900 in 6 volumes.

We find the text marked up at sub-topic level analysis. No text contains the description or explanation of the tags used, nor is there any prologue mentioning the purpose of this tagging, as is the case with any typical Sanskrit texts centuries old. So the first task we took up is to provide a manual listing the tags used, the semantics associated with these tags

and at least an example from the *Mahābhāṣya* illustrating the tag.

3.1 Mahābhāṣya's Sub-Topic Level Tag-set

The annotated *Mahābhāṣya* has 9 major tags and several sub-tags under each major tag. These sub-tags are different in different volumes. But the major tags are the same in all the books. We give below a brief description of these 9 tags, with an example each from the actual tagged Mahābhāṣya. The example in most cases consists of a pair of Sanskrit sentences/paragraphs. The first one sets the context under which the next set of sentence(s) is uttered. The label is attached to the second one.

1. **Praśna (Question)** An independent question about some topic or an argument is marked with a question tag. For example,

   - **Skt:** atha vidhiśabdārthanirupaṇādhikaraṇam.
     Eng: Now starts the section in which the meaning of the word *Vidhi* is examined.[2]
   - **Skt: (praśnabhāṣyam)**: vidhiḥ iti kaḥ ayam śabdaḥ. [1.1][3]
     Eng: What is this word *Vidhi*?

2. **Uttara (Answer)** An answer to a question is tagged with this tag. For example,

   - **Skt: (praśnabhāṣyam)**: vidhiḥ iti kaḥ ayam śabdaḥ.
     Eng: What is this word *vidhi*?
   - **Skt: (uttarabhāṣyam)**: vipūrvādghajñaḥ karmasādhana ikāraḥ. vidhīyate vidhiriti. kim punarvidhīyate. samāso vibhaktividhānam parāṅgavadbhāvaśca. [1.2]
     Eng: The letter *i* denoting the passive sense (is added) after (the root) *ghajñ* preceded by the pre-verb *vi*.

---

[2] *English translations are taken from Patañjali's Vyākaraṇa Mahābhāṣya Samarthāhnika (P 2.1.1) Edited with Translation and Explanatory Notes by S D Joshi.*

[3] *The first number denotes to the topic number and the second number denotes to the sub-topic number of the sūtra samarthaḥ padavidhiḥ P2.1.1.*

What is prescribed by Pāṇini's rules is *vidhi*: 'operation'. But what could that be which is prescribed? 'compounding', 'prescription of case-ending' and 'treatment as a part of the following word'.

3. **Ākṣepa (Objection)** An objection to an answer or resolution is marked as an *ākṣepa*. For example,

   - **Skt: (uttarabhāṣyam)**: vākye pṛthagarthāni. rājñaḥ puruṣaḥ iti. samāse punaḥ ekārthāni rājapuruṣaḥ iti.
     Eng: In the compounded word-group words have separate meanings of their own, like rājñaḥ puruṣaḥ: king's man. But in a compound, words have a single meaning, like in rājapuruṣaḥ: king-man.
   - **Skt: (ākṣepabhāṣyam)**: kimucyate pṛthagarthāni iti yāvatā rājñaḥ puruṣa ānīyatāmityukte rājapuruṣa ānīyate rājapuruṣa iti ca sa eva. [4.44-45]
     Eng: Why do you say 'words have separate meanings of their own'? Because when we say 'let the king's man be brought', the king-man is brought. And when we say 'let the king-man be brought', the same man is brought.

4. **Samādhāna (Resolution)** This tag is used to mark an answer to an objection. For example,

   - **Skt: (ākṣepabhāṣyam)**: yadi sāpekṣamasamartham bhavati iti ucyate rājapuruṣo'bhirūpaḥ rājapuruṣo darśanīyaḥ atra vṛttirna prāpnoti.
     Eng: If we accept the statement, 'what requires an outside word is treated as semantically unconnected' then the word-composition rājapuruṣa: king-man in the expressions rājapuruṣaḥ abhirūpaḥ: handsome king-man, rājapuruṣaḥ darśanīyaḥ: goodlooking king-man would not result from the uncompounded word-groups abhirūpaḥ

rājñaḥ puruṣaḥ and darśanīyaḥ rājñaḥ puruṣaḥ.

   - **Skt: (samādhānabhāṣyam)**: na eṣaḥ doṣaḥ. pradhānam atra sāpekṣam bhavati ca pradhānasya sāpekṣasya api samāsaḥ. [3.27-28]
     Eng: Nothing wrong here. Because it is here the main member which requires an outside word. And compounding does take place, even if the main member requires an outside word.

5. **Bādhaka (Rejection)**: This tag is used to mark the rejection of the arguments such as, objection, answer of an objection, refutation, criticism etc. For example,

   - **Skt: (samādhānabhāṣyam)**: na eṣaḥ doṣaḥ. samudāyāpekṣā atra ṣaṣṭhī sarvaṁ gurukulam apekṣate.
     Eng: Nothing wrong here. Here the genitive qualifies the whole word gurukulam.
   - **Skt: (samādhānabādhakabhāṣyam)**: yatra tarhi na samudāya apekṣā ṣaṣṭhī tatra vṛttiḥ na prāpnoti. [3.30-31]
     Eng: Then when a word in genitive does not qualify the whole, it should not result in a compound formation.

6. **Sādhaka (Reaffirmation)**: This tag is used to mark the reaffirmation of an argument which has been earlier rejected. For example,

   - **Skt: (ākṣepabādhakabhāṣyam)**: nanu ca gamyate tatra sāmarthyam. kumbhakāraḥ nagarakāraḥ iti.
     Eng: But is it not so, that, when we say *kumbhakāraḥ*: 'pot-maker' or *nagarakāraḥ*: 'city-maker', we do apprehend semantic connection between pot and maker.
   - **Skt: (ākṣepasādhakabhāṣyam)**: satyaṁ gamyate utpanne tu pratyaye. sa eva tāvat samarthādutpādyaḥ. [2.12-13]
     Eng: Yes, that is true. It is apprehended once a suffix has been added. But that same suffix must first be

generated after the semantically connected word.

7. **Udāharaṇa (Example)**: This tags the example.

   - **Skt:** **(subalo-podāharaṇabhāṣyam)**: supaḥ alopaḥ bhavati vākye. rājñaḥ puruṣaḥ iti. samāse punaḥ na bhavati. rājapuruṣa iti. [5.49]
     Eng: Non-disappearance of case-ending occurs in an un-compounded word-group, like rājñaḥ puruṣaḥ: king's man. But in a compound it does not occur, as in rājapuruṣaḥ: king-man.

8. **Dūṣaṇa (Criticism)**: This tag is to mark criticism. For example,

   - **Skt:** **(vyākhyābhāṣyam)**: samānavākya iti prakṛtya nighātayuṣmadasmadādeśā vaktavyāḥ.
     Eng: Under the heading of 'within the same sentence' the accents and substitutions for yuṣmad and asmad are to be stated.
   - **Skt:** **(dūṣaṇavārttikam)**: yoge pratiṣedhaścādibhiḥ. [9.115]
     Eng: When there is connection with and the prohibition should also be stated.

9. **Vyākhyā (Explanation)**: Explanation of either an objection, answer or alternative view is marked with this tag. For example,

   - **Skt:** **(samādhānavārttikam)**: pṛthagarthānāmekārthībhāvaḥ samarthavacanam. **Eng:** The word *samartha* means single integrated meaning of the separate meanings.
   - **Skt:** **(vyākhyābhāṣyam)**: pṛthagarthānāṁ padānāmekārthībhāvaḥ samarthamityucyate. [4.42]
     Eng: The single integrated meaning of the words which have separate meaning is called *samartha*.

Out of these 9 tags, 3 tags viz. *sādhaka, bādhaka and dūṣaṇa* are rare. The four tags

Tags	Frequency
Praśna (Question)	10
Pratipraśna (Counter question)	1
Pratipraśnottara (Answer to a counter question)	1
Uttara (Answer)	6
Ākṣepa (Objection)	47
Pratyākṣepa (Counter objection)	6
Pratyākṣepasamādhāna (Answer to a counter objection)	2
Samādhāna (Answer to an objection)	40
Vyākhyā (Explanation)	34
Udāharaṇa (Example)	8

Table 1: Tags with their frequencies of *Samarthāhnikam*

praśna, uttara, ākṣepa and samādhāna also have sub-tags viz. *pratipraśna (a question to a question), pratipraśnottara (answer to a question to a question), pratyākṣepa (counter objection) and pratyākṣepasamādhāna (answer to a counter objection)*, which are more frequent. The distribution of these 10 tags - 6 main tags and 4 sub-tags, in one book (*samarthāhnikam*) among 87 books of Mahābhāṣya is shown in Table 1.

4 Mahābhāṣya Tagger

The tags used in Mahābhāṣya are general and are found to be used in the texts from other disciplines such as philosophy etc. Traditional texts also discuss clues to mark these tags. For example, Bhoja Rājā's Śṛṅgāra Prakāṣa (Dvivedī and Dvivedī, 2007), Śābdabodhamīmāṁsā (Tātāchārya, 2005) and Avyaya Kośa (Srivatsāṅkācārya, 2004) do provide such clues for some tags. The clues for two tags *praśna (question) and ākṣepa (objection)*, and their sub-tags viz. *pratipraśna (counter question) and pratyākṣepa (counter objection)* are in the form of possible Lexical word(s).

A. The list of words that classify a paragraph as a *praśna* (question) are the wh-words viz. *kim, kaḥ, kimartham, katara, kutaḥ, kva, kāni, katham, kayā, kena* etc.

B. The list of words that classify a paragraph as a *ākṣepa* (objection) are *evam api, katham, kvacit, yadi, kasmāt na, nanu ca, yatra tarhi, kim punaḥ* etc.

We looked at one book (*āhnika*) of the tagged text for the clues for other two tags viz. *pratipraśna and pratyākṣepa*. The lexical clues are,

Pratipraśna: *kaḥ punaḥ, kaḥ ca, kim ca.*

Pratyākṣepa: *na vā, kasmāt na, kasya punaḥ, kaḥ vā, kva ca, kiñca.*

From the semantics of the tags, it is clear that an *uttara* (answer) tag should follow a *praśna* (question). The *pratipraśna* (counter question) follows a *praśna* (question) and *pratipraśnottara* (answer to a counter question) follows a *pratipraśna* (counter question). An *ākṣepa* (objection) arises only after an answer *(uttara)*. The *samādhāna* (answer to an objection) is for an *ākṣepa* (objection) and *pratyākṣepa* (counter objection) is to a *ākṣepa* (objection). The *pratyākṣepa samādhāna* (answer to the counter objection) will follow the *pratyākṣepa* (counter objection). Any *samādhāna* (answer to an objection) can be followed by a *ākṣepa* (objection) only. *Varttika*-s are the supplementary rules, which can occur at the very beginning or it can upon one *praśna* (question) or *ākṣepa* (objection). This is represented as a finite state automata in Figure 2.

The digital version of the tagged text of complete *Mahābhāṣya* is not available. Only *4* books out of *87* books were available at the time of testing. Of these we used one book for framing the rules and gathering cues and tested our automata on the rest of the three books. Once digitized versions of other books are available the automata tagger will be tested on them. The precision and recall for the three books is given in Table 2. Table 3 gives the precision and recall for each tag in all the three books together. In these three books, we did not find any instance of *pratyākṣepa* and *pratyākṣepasamādhāna.*
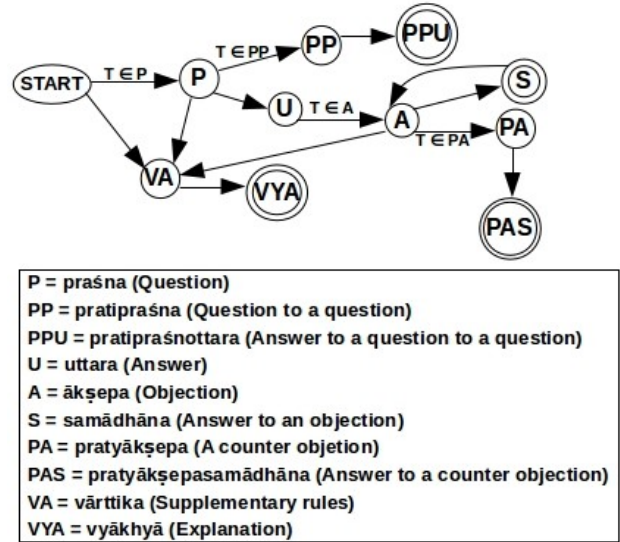

P = praśna (Question)
PP = pratipraśna (Question to a question)
PPU = pratipraśnottara (Answer to a question to a question)
U = uttara (Answer)
A = ākṣepa (Objection)
S = samādhāna (Answer to an objection)
PA = pratyākṣepa (A counter objetion)
PAS = pratyākṣepasamādhāna (Answer to a counter objection)
VA = vārttika (Supplementary rules)
VYA = vyākhyā (Explanation)

Figure 2: Finite State Automata of Tags

5 Extending the tagger

The tagger performed satisfactorily on a part of *Mahābhāṣya* text. Till the digital copies of other parts of the *Mahābhāṣya* text become available, we thought of evaluating the tagger on other commentary. Since the commentaries are to establish the necessity of a *sūtra*, the assumption is that the basic structure of commentary should be the same. We chose a commentary on the *Mīmāṃsā sūtra*-s. The *Mīmāṃsā sūtra*-s are written by *Jaiminī* around the end of 2nd century AD. It consists of 12 adhyāya-s (chapters) and 60 pāda-s (sections). This text provides rules for the interpretation of the *Veda*-s. Earlier scholars wrote commentaries on *Mīmāṃsā sūtra* but unfortunately they are lost. A major commentary is composed by *Śabarasvāmī* around 5th century and is well known as *Śābarabhāṣya*. This is the only extant and authoritative commentary on full 12 chapters of the *Mīmāṃsāsūtra of Jaiminī*.

The commentary used for our study is the Hindi translation of *Śābarabhāṣyam* by *Yuddhiṣṭhira Mīmāṃsaka*. This Hindi translation is marked up with only two tags viz. *Ākṣepa* and *Samādhāna*. From the tags in Hindi translation, we constructed the digitalized version of the tagged Sanskrit commentary. Since the tagged text had only two tags, we modified our automaton to suit this structure removing the nodes corresponding to

Books (Āhnika-s)	Precision	Recall	F Score
Kārakāhnikam	90.04%	86.45%	88.20%
Paspaśāhnikam	76.47%	79.13%	77.77%
Prātipadikārthaśeṣāhnikam	85.37%	77.21%	81.08%

Table 2: Precision Recall Table

Tags	Total Tags	Precision	Recall	F Score
P	78	80.00%	76.92%	78.42%
PP	4	100.00%	75.00%	85.71%
PPU	1	50.00%	100.00%	66.66%
U	64	82.76%	75.00	78.68%
A	65	70.49%	66.15%	58.25%
S	70	67.24%	55.71%	60.93%
VA	111	98.23%	100.00%	99.10%
VYA	109	95.58%	99.08%	97.29%

Table 3: Precision Recall Table for Each Tags

other tags. This led us to modify the lexical cues for the tags as well. So the cues for *Ākṣepa* in *Śābarabhāṣya* included the clues from both *Praśna* as well as *Ākṣepa* of the *Mahābhāṣya*. In addition, we found some stylistic variation in the cues. While many of the cues from *Mahābhāṣya* did not find any place in the *Śābarabhāṣya* it included one new phrase *na brūmaḥ* (literally do not say this), as a marker for *Ākṣepa*. Modifying the tagger accordingly, we tested it on the 3rd chapter 1st section 6th adhikaraṇa's 12th sūtra of *Śābarabhāṣya* viz. *Aruṇādhikāra* and the precision and recall were found to be 96.00% and 82.76% respectively.

6 Conclusion

Sanskrit has a rich grammatical tradition and offers theoretical insights for discourse analysis as well. The principles for analysis being language independent, these insights should be applicable to other languages as well. In this paper we have shown the applicability of this analysis in the context of scientific text in Sanskrit. The implementation as a FSA is language independent with the clue set as a language dependent component. It will be interesting to use this further on various forums on internet.

References

Amal AlSaif and Katja Markert. 2010. The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2046–2053, Valletta, Malta. European Language Resources Association (ELRA).

Amal Alsaif. 2012. *Human and Automatic Annotation of Discourse Relations for Arabic*. Ph.D. thesis, School of Computing, The University of Leeds.

Timothy Baldwin, Li Wang, and Su Nam Kim. 2012. The utility of discourse structure in identifying resolved threads in technical user forums. In *Proceedings of COLING 2012: Technical Papers*, pages 2739–2756, Mumbai. COLING 2012 Organizing Commitee.

Śri Jīvānanda Vidyāsāgar Bhaṭṭācārya. 1989. *Śrī Mādhavācāryaviracita Jaiminīya Nyāyamālāvistaraḥ*. Kṛṣṇadās Academy, Varanasi.

Revāprasād Dvivedī and Sadāśivakumār Dvivedī. 2007. *Sarasvatī Kaṇṭhābharaṇasya Bhojarājasya Sāhityaprakāśa Nāma Śṛṅgāraprakāśah, Prathamabhāgaḥ 1 - 14 Prakāśāḥ*. Indiragandhi Kala Kendram, New Delhi and Kalidasa Samsthan, Varanasi.

B. Fortuna, E.M. Rodrigues, and N. Milic-Frayling. 2007. Improving the classification of newsgroup messages through social network analysis. In *In Proceedings of the 16th*

ACM Conference on Information and Knowledge Management (CIKM 2007), page 877880, Lisbon, Portugal.

Pavlina Jinova, Jiri Mirovsky, and Lucie Polakova. 2012. Semi-automatic annotation of intra-sentential discourse relations in pdt. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA), COLING*, pages 43–58, Mumbai. COLING 2012 Organizing Commitee.

S D Joshi and J.A.F. Roodbergen. 1975. *Patañjali's Vyākaraṇa Mahābhāṣya Kārakāhnikam (P 1.4.23–1.4.55)*. Center of Advanced Study in Sanskrit, Pune.

S D Joshi and J.A.F. Roodbergen. 1981. *Patañjali's Vyākaraṇa Mahābhāṣya Prātipadikārthaśeṣāhnikam (P2.3.46–2.3.71)*. Center of Advanced Study in Sanskrit, Pune.

S D Joshi and J.A.F. Roodbergen. 1986. *Patañjali's Vyākaraṇa Mahābhāṣya Paspaśāhnikam*. Center of Advanced Study in Sanskrit, Pune, first edition.

Aravind Joshi, Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Livio Robaldo, and Bonnie Webber, 2007. *The Penn Discourse Treebank 2.0 Annotation Manual. The PDTB Research Group.*

S D Joshi. 1968. *Patañjali's Vyākaraṇa Mahābhāṣya Samarthāhnika (P 2.1.1) Edited with Translation and Explanatory Notes*. Center of Advanced Study in Sanskrit, Poona, first edition.

Daniel Jurafsky, Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Association for Computational Linguistics*, volume 26:3, pages 339–373. Association for Computational Linguistic.

Su Nam Kim, Li Wang, and Timothy Baldwin. 2010. Tagging and linking web forum posts. In *In Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL-2010)*, pages 192–202, Uppsala, Sweden.

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 Lucie Polakova Eva Hajicova and Jiri Mirovsky, editors, *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA), COLING*, pages 1–16, Mumbai. COLING 2012 Organizing Commitee.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Yudhiṣṭira Mīmāṃsaka. 1990. *Mīmāṃsāśābarabhāṣyam*, volume 2. Ramlal Kapur Trust, Sonipat, Haryana.

Lucie Mladova, Sarka Zikanova, and Eva Hajicova. 2008. From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 2564–2570, Marakes, Maroko. European Language Resources Association (ELRA).

Livia Polanyi. 1988. A formal model of discourse structure. *Journal of Pragmatics*, 12(5–6):601–638.

Rashmi Prasad, Umangi Oza, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP*, pages 158–161, Suntec, Singapore. Association for Computational Linguistics.

K V Ramakrishnamacaryulu. 2009. Annotating sanskrit texts based on Śābdabodha systems. In Amba Kulkarni and Gérard Huet, editors, *Proceedings Third International Sanskrit Computational Linguistics Symposium*, pages 26–39, Hyderabad, India. Springer-Verlag, LNAI 5406.

Raphael Salike. 1995. *Text and Discourse Analysis, Language Workbooks series, series editor Richard Hudson*. Routledge, London.

Ananta Śāstri. 1916. *Kārikāvalī with the commentaries Muktāvalī, Dinakarī, Rāmarudrī*. Nirṇaya Sāgara Press, Bombay.

Sri V Srivatsāṅkācārya. 2004. *Avyaya Kośa*. Samskrit Education Society, Chennai 4.

N S R Tātāchārya. 2005. *Śābdabodhamīmāṃsā: The sentence and its significance Part-I to V*. Institute of Pondicherry and Rashtriya Sanskrit Sansthan, New Delhi.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Journal of Computational Linguistics*, 31(2):249–287.

W. Xi, J. Lind, and E. Brill. 2004. Learning effective ranking functions for newsgroup search.

In *In Proceedings of 27th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 394–401, Sheffield, UK.

Nianwen Xue. 2005. Annotating discourse connectives in the chinese treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 84–91, Ann Arbor, Michigan. Association for Computational Linguistics.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In *The 6th Workshop on Asian Language Resources (ALR 6), IJCNLP*, pages 65–71, Hyderabad, India. Asian Federation of Natural Language Processing.

Yuping Zhou and Nianwen Xue. 2012. Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 69–77, Jeju Island, Korea. The Association for Computational Linguistics.