Inter Sentential Discourse Relations

Saee Vaze
Department of Sanskrit Studies,
University of Hyderabad
20hsph03@uohyd.ac.in

Amba Kulkarni
Department of Sanskrit Studies,
University of Hyderabad
ambakulkarni@uohyd.ac.in

Abstract

In this paper, we present a tagging scheme for inter-sentential discourse relations that is developed, based on the insights from Indian Grammatical Tradition. We rely on the three factors - *ākāṅkṣā*, *yogyatā* and *sannidhi* to decide the connectivity between two consecutive sentences. Various clues are identified that bind the two consecutive sentences. A tag set is presented based on the explicit discourse markers. Implementation of discourse level analysis based on the explicit discourse markers is tested on the *Śrimadbhagvadgītā* corpus. It is observed that some discourse markers are ambiguous and it is not trivial to develop a disambiguation module for such markers.

1 Introduction

The term discourse analysis has gained a lot of attention in the recent past. It typically refers to a linguistic unit that goes beyond a sentence.[1] Thus, the discourse analysis goes beyond the scope of sentence boundaries and looks at the text as a unit of language. In understanding the meaning of a discourse, both the linguistic and non-linguistic factors contribute. The linguistic factors include coherence markers while non-linguistic background includes, speaker-listener dynamics, situationality etc. In Natural Language Processing, the core interest is in producing computer-processable models of discourse at different levels such as sentence, paragraph, text, etc. Varied work has been done on the topic already on different levels and languages.

From the theoretical point of view the work by Indologists and Sanskrit scholars in the field of discourse analysis is very rich and valuable. Scharf and Hock (2015) provides an exhaustive bibliography of works in the field of general discourse and formal syntax. However, there is very little work from the perspective of computational linguistics with regards to Sanskrit language. There are several efforts in the West especially in the field of computational linguistics. Treatment of cohesion by Halliday and Hasan (1976) attempts to look at the text as a linguistic phenomenon. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) was the first effort towards establishing the discourse structure in the form of a graph, by connecting two adjacent units by a discourse relation. Another seminal effort was made by the team lead by Arvind Joshi in the project Penn Discourse Tree-Bank (Prasad et al., 2006) which focuses on the structure of arguments and how a connective enables a certain discourse relation, implicit or explicit. The discourse tree-banks were created from a huge data from Wall Street Journal (Mann and Thompson, 1988) following the RST framework. The other discourse databanks include Linguistic Discourse Model (Polanyi, 2008), and the Discourse Graphbank (Wolf and Gibson, 2005). The Discourse-Lexicalized Tree Adjoining Grammar (Webber and Joshi, 1998) was developed following the Penn Discourse Treebank guidelines. With the emergence of such computational guidelines and resources for several languages discourse tagged datasets were developed for languages other than English such as Czech (Mladová et al., 2008), Chinese (Jiang et al., 2018) and Turkish (Zeyrek et al., 2010) to name a few. Similar efforts were made

[1] https://www.merriam-webster.com/dictionary/discourse

for some Indian Languages resulting in Hindi Discourse Relation Bank (Umangi et al., 2009), Bangla RST Discourse Treebank (Das and Stede, 2018), Annotated Tamil Corpus (Rachakonda and Sharma, 2011), Annotations of Connectives and Arguments in Malayalam (Kumari and Devi, 2016) etc.

Regarding Sanskrit, in the recent past Kulkarni and Das (2012) presented a brief summary of the various sets of discourse relations found in the Indian grammatical tradition(IGT). They have also shown the usefulness of these relations by developing a Finite State Automaton to tag the texts in *Mahābhāṣya* following the cues available. Recently Terdalkar and Bhattacharya (2019) developed a Question Answering system for special domains. Apart from these there is not much work in the area of discourse analysis in Sanskrit.

In what follows we brief our approach to discourse analysis in Sanskrit following IGT followed by a review of earlier work on Discourse analysis in Sanskrit. In Section 3 we present various clues that mark the *ākāṅkṣā* between the two consecutive sentences. We identify the explicit discourse markers in Sanskrit that connect two consecutive sentences. This is followed by a discussion on the implementation, challenges and evaluation.

2 Discourse Analysis in Sanskrit

The computational models such as RST and Penn Discourse may be tried for Sanskrit as well. However there are three considerations why we decided to follow the IGT. The first and foremost concerns with the rich linguistic tradition of India. The theories of *śābdabodha* that deal with the process of understanding texts are almost as old or albeit a little older than *Pāṇini*'s grammar. *Jaimini* in his composition of *Mīmāṁsāsūtra* not only provided his interpretations of the vedas, but also provided a glimpse of what principles he followed in interpreting the texts. Further *Śabara* elaborates these principles including the ones which *Jaimini* merely indicated. The seeds sown by these *Māmāṁsaka*s further grew into various guidelines to decide the coherence between the textual segments. The *Naiyāyika*s and the *Vaiyākaraṇa*s also followed the *Mīmāṁsaka*s resulting into various sets of coherence relations proposed by them for describing the coherence between the various segments of the texts. These relations cover a wide range of units starting from the sentences to paragraphs to chapters to texts to discipline. Depending on the style of the texts, and the type of unit the text belongs to, different annotation schemes were proposed by different schools. A detailed description of this is available in Kulkarni and Das (2012).

The second consideration is that these discourse relations are also used by the commentators while commenting upon important texts, or editors who used them as subtitles providing some hints towards understanding the cohesion, and sources for coherence markers. For example the Nirṇaya-sāgara edition of the *Mahābhāṣya* has subtitles which show the logical structure of the discourse.

The third consideration is the following. *Mīmāṁsaka*s discuss three factors viz. *ākāṅkṣā*, *yogyatā* and *sannidhi* as important factors for the verbal cognition. These factors play an important role throughout the process of verbal cognition - not limited just to the sentential analysis - but extending to the understanding of the complete text. Thus these three factors can be considered to be guidelines for identifying the clues and connecting the segments of the texts accordingly. Having developed a sentential parser (Kulkarni, 2019) based on the theories of *śābdabodha*, where all these factors were used for sentential analysis, it gave us a confidence that these factors can be further extended for discourse analysis as well.

Hence we decided to base our approach to discourse analysis following the IGT.

2.1 Earlier work and its limitations

The relations in the tag-set proposed by Krishnamacharyulu (2009) contain inter-sentential relations as well. These inter-sentential relations are marked by some connectives which are indeclinables. Some of these connectives occur in pairs. Kulkarni and Das (2012) had proposed a tagging scheme for them. Each of these connectives takes two arguments. Following logicians convention, these arguments are named by the general terms *anuyogika*[2] (combining) and *pratiyogī* (having a counter part). So, if C is the connective connecting two sentences S1 and S2 then the general structure is represented as in Figure 1.

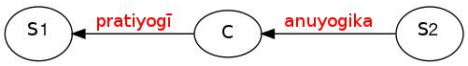

Figure 1: Discourse structure with single connective

When there are two parallel connectives C1 and C2 connecting S1 and S2 then the relation between them is represented as in Figure 2.

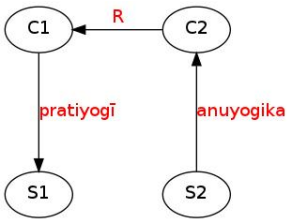

Figure 2: Discourse structure with paired connectives

Here R binds C1 and C2. The relation of the connectives with the sentence is through the main verbs. The sentences are further parsed as dependency trees. In case of paired connectives, the usage allows using either of them or both of them in a sentence. When only one of them is used in a sentence then the structure in Figure 2 collapses to Figure 1.

There were two problems with this work. The first one was related to the names of the relations. The two terms *anuyogika* and *pratiyogī* are very general that they do not convey the semantics of the relation between the two sentences. Secondly, we also noticed that there were several other indeclinables which the authors had missed. So in the next section, we look at the relations between two consecutive sentences, marked by explicit markers, and provide a semantic interpretation of that relationship. We also enlist all possible indeclinables that can mark the relations between the consecutive sentences.

3 Inter-Sentential Discourse Relations

The inter-sentential relations are identified with the help of *ākāṅkṣā*, *yogyatā* and *sannidhi*. Sometimes the fourth factor *tātparya* is also considered to an essential factor in the process of *śābdabodha*. This factor is more relevant from the word sense disambiguation point of view, and also for choosing the level of signification of the word. Hence we focus only on the first three factors. We define the basic elementary units between which we establish the relations. Then we look at the clues that guide us in proposing a relation. Then comes the mutual compatibility between the elementary units which confirms the proposed relation. We assume that the elementary units are consecutive ones, ensuring that the third factor sannidhi is taken care of.

[2]S2 is the anuyogi. So if the arrowhead is pointing towards S2 the name of the relation would have been anuyogi. In this diagram, the arrowhead is pointing towards C, and hence the name of the relation is inverse of anuyogi, i.e. anuyogika.

3.1 Elementary Unit

Elementary Unit refers to the basic elements between which the relation is to marked. We take *vākya* (sentence) as a unit, where *vākya* is defined as '*eka tiṅ vākyam*'. That is a group of related words with one finite verbal form is termed as a sentence. Further, participles, especially those with *kta*, *ktavatu* and the *kṛtya* suffixes such as *anīyar* etc. are typically used as if they are finite verbs (Speijer, 1886) (in section 9). Hence group of related words with such forms, without any finite verbal form, are also considered to be a sentence. (Others not specially mentioned in the list of sentences with non-finite exception, such as *satī-saptamī*, *tumun* etc. would be considered as a single unit and would not fall under the domain of inter-sentential discourse relations.)

With this definition, the following group of words

prātaḥkāle rāmaḥ śālāṁ gacchati. tatra pāṭhaṁ paṭhati. krīḍati ca. sāyaṁkāle gṛhaṁ āgacchati.

consists of four sentences, as delimited with the full stops. Now consider the following sentence:

Sanskrit : *yadi tvam icchasi tarhi ahaṁ tava gṛhaṁ āgacchāmi iti rāmaḥ śāmaṁ vadati.*
Gloss : if you wish{2p,sg,pres}, then I your house{loc} come{1p,sg,fut} so Rama{nom} Shyama{acc} say{3p,sg,pres}
Eng: "If you wish I will come to your house" says Rama to Shyama.

Following the definition of *eka-tiṅ vākyaṁ*, here there are three sentences viz.

1. *tvam icchasi*,
2. *ahaṁ tava gṛhaṁ āgacchāmi*, and
3. *rāmaḥ śyāmaṁ vadati.*

connected by three connectives *yadi*, *tarhi* and *iti*. The words *yadi* and *tarhi* are the pair connectives, and both these connectives have an expectancy of two sentences. The third connector *iti* is a marker for the *karman* (*vākya-karma-dyotakaḥ*) which is in sentential form. Thus now the complex sentence formed by joining the two sentences with the pair of connectives *yadi-tarhi* acts as a *karma* for the verb *vad*.

3.2 *Ākāṁkṣā*

Literally *ākāṅkṣā* is the desire on the part of a listener to know (*jñātum icchā*). In the case of understanding a sentence, the desire is to know how the words in a sentence are connected to each other producing a unified meaning. This *ākāṁkṣā* is expressed in language through various means. As is mentioned in Kulkarni (2019), there are different linguistic clues that mark the expectancies in a sentence, such as

- suffix, as in the case of '*vanaṁ gacchati*', [forest{nom} go{3p,sg,pres}] the suffix '*am*' marks the *karmatva* and thus has an expectancy of a transitive verb to connect with,
- position, as in the case of a sentence starting with the word '*api*', there is an expectancy of a sentence such as '*tvam gacchasi*',[you{nom} go{3p,sg,pres}] so that complete expression expresses a question,
- indeclinables such as '*na*' which have an expectancy of a verbal form to connect to, and finally
- the underlying verbal root in a verbal form has an expectancy for various *kāraka*s.

Similarly the sentence level connections are expressed through various means such as indeclinables, relative position of sentences, semantics associated with the verbal roots, and so on.

- The indeclinables such as *yadi*, *tarhi*, *yataḥ*, *tataḥ*, *atha*, *yathā-tathā*, *yadyapi*, etc. provide a cue that the consecutive sentences (or group of sentences) are related. For example, the two sentences

  Sanskrit: *rāmaḥ paṭhati. tathāpi parīkṣāyām uttīrṇaḥ na bhavati.*
  Gloss: Rama{nom} study{3p,sg,pres} then exam{loc} pass{nom} neg be{3p,sg,pres}
  English: Rama studies. (Even) then he does not pass the exam.

  are connected to each other showing the failure to get the desired results even after performing the necessary task.

- The relative position of the sentences in a conservation also provides us a clue about the temporal sequence between the events associated with the verbal forms.

  Sanskrit: *rāmaḥ prātaḥkāle uttiṣhati. snānam karoti. dugham pītvā śālām gacchati.*
  Gloss: Rama{nom} morning{loc} wake{3p,sg,pres}. Bath{acc} do{3p,sg,pres}. Milk{nom} drink{geund} school{acc} go{3p,sg,pres}
  English: Rama wakes up in the morning. Takes a bath. Goes to school, after drinking milk.

  Here we notice that there is a temporal sequence, and thus the order in which the activities happened is marked in the position of these sentences. There is no lexical unit which marks such relation.

- The use of pronouns connect the sentences when the anaphora resolution is made.

  Sanskrit: *rāmaḥ śālām gacchati. saḥ tatra pāṭham paṭhati.*
  Gloss: Rama{nom} school{acc} go{3p,sg,pres}. He{nom} there lesson{acc} study{3p,sg,pres}
  English: Rama goes to school. There he studies a lesson.

  Here the use of the pronoun '*saḥ*' for Rāma by the speaker, needs to be resolved by the listener. Only then the listener can understand the conversation.

- The semantics associated with verbs also raise certain expectancies. For example look at the two *śloka*s the first one and the seventh one from *Saṅkṣepa-rāmāyaṇam*. The first one viz.

  tapassvādhyāyaniratam tapasvī vāgvidām varam
  nāradam paripapraccha vālmīkirmunipuṅgavam

  has the verbal form *paripapraccha*(asked) which has an expectancy of an answer. This expectancy is fulfilled by the verbal form *abravīt* (said) from the seventh *śloka*, viz.

  ... śrūyatām iti āmantrya prahṛṣṭaḥ vākyam abravīt.

In this paper we focus only on the lexical units that express the sentential expectancies.

3.3 *Yogyatā*

Some indeclinables such as '*atha*', can be used to denote conjunction as well as succeeding action. Similarly words such as *yasmāt-tasmāt* or *yena-tena* can represent the *kāraka* relations such as *apādāna* or instrument, alternately these words may also denote a *hetuḥ* - a cause-effect relation. In order to decide the appropriate role in the context, we need to look at the context - both linguistic as well as non-linguistic, identify the linguistic and ontological factors that can help in the disambiguation, and so on. We look at one particle '*hi*' (See Sec 5.1) which is ambiguous between a causal marker and a definiteness marker, and show how difficult it is to address the problem of ambiguity.

4 Discourse Relation tagging and clues

Below we present the list of discourse markers we have come across so far (and also implemented), along with the relation(s) they express and the tagging with an example sentence.

Relation	**Markers**
Succeeding (*anantarakālaḥ*)	*tataḥ, anantara, atha*
Simultaneity (*samānakālaḥ*)	*yadā-tadā*
Co-location (*samānādhikaraṇaḥ*)	*yatra-tatra*
Similarity (*sādṛśyam*)	*yathā-tathā*
Cause-Effect (*kārya-kāraṇam*)	*yataḥ-tataḥ, ataḥ, yasmāt, tasmāt, yena, tena, hi*
Possiblity (*āvaśyakatā-pariṇāmaḥ*)	*yadi-tarhi, iti, cet*
Hindrance in cause-effect (*vyabhicāraḥ*)	*yadyapi-tathapi, cedapi, athāpi, tarhyapi*
Antithesis (*virodhaḥ*)	*parantu, kintu*
Conjunction (*samucccayaḥ*)	*ca, api, cāpi, athaca, athāpi, evañca*
Disjunction (*anyataraḥ*)	*vā, uta, yadvā, athavā, utāpi, utasvit*

Table 1: List of discourse relations and markers

1. **Succeeding (*anantarakālaḥ*) :**
   Here the relation of succeding activity to the preceeding is marked. The presence of indeclinables such as *atha*, *tataḥ*, etc. trigger the relation of the current sentence with the previous one. The activity denoted by the current sentence is marked as the succeding activity for the activity denoted by the previous sentence. See Figure 3.

   Sanskrit : *ahaṃ śṛṇomi atha likhāmi.*
   Gloss : I{nom} listen {3p,sg,pres} then write{3p,sg,pres}
   English : I listen, then I write.

   Before moving to the next relation, we highlight the salient features of the discourse graph representation.

   - The relations between two sentences is through a link between the head (*mukhya viśeṣya*) of the two sentences.
   - The arrow head is with the node that satisfies the property named by the edge label.
   - The direction of the arrow, unlike in dependency trees, does not denote the dependency, or the head and the sub-ordinate.
   - In order to distinguish the discourse relations from the intra-sentential relations, discourse relations are marked with double line.

   Thus in Figure 3 the verbs from the two sentences viz. *śṛṇomi* and *likhāmi* are related by the relation of *anantarakālaḥ*, and the marker for this relation is the word *atha*.

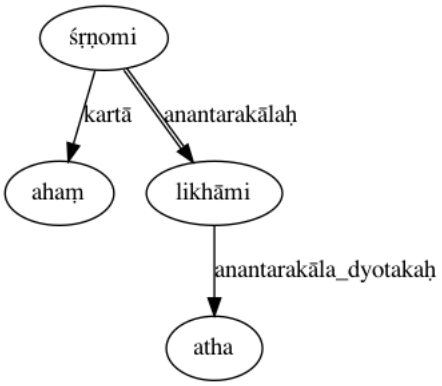

Figure 3: anantarakālaḥ relation

2. **Simultaneity (*samānakālaḥ*) :**
In Sanskrit, there are two ways of expressing the simultaneity. One is the use of present participles (*kṛt* suffixes - *śatṛ* and *śānac*) which are part of intra-sentential relations. The second is the use of indeclinable pair *yadā-tadā*. In this case, the verbs in finite form from both the sentences are connected with a relation *samānakālaḥ*. The words *yadā* and *tadā* mark a relation of *kālādhikaraṇam* (time locative) (See Figure 4).
Here is an example:

Sanskrit : *yadā bharataḥ mārge gacchati tadā saḥ devālayam paśyati.*
Gloss: when Bharata{nom} path{loc} go{3p,sg,pres} then he{nom} temple{acc} see{3p,sg,pres}
English : On his way Bharata sees a temple.

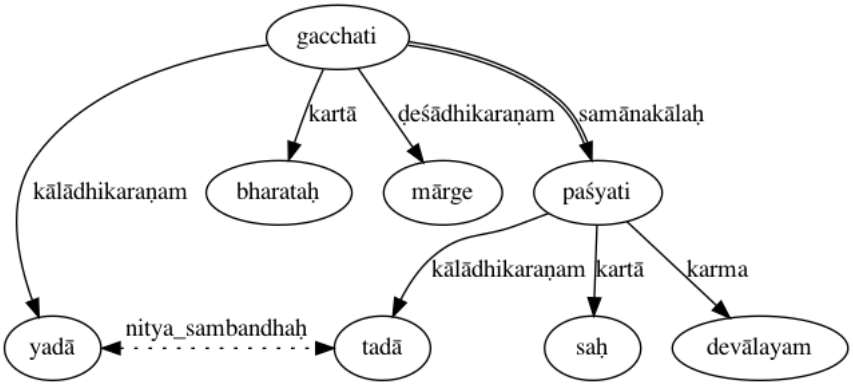

Figure 4: samānakālaḥ relation

3. **Co-location (*samānādhikaraṇaḥ*) :**
This relation indicates that the activities indicated by the two consecutive sentences are performed at the same location. This relation is marked by the pair of indeclinables *yatra-tatra*. The consecutive sentences use these two words denoting the *deśādhikaraṇam* (place locative), or only one of them is used in one sentence (See Figure 5).

Sanskrit : *yatra nāryaḥ tu pūjyante tatra devatāḥ ramante.*
Gloss : where women{nom} emph_marker worship{3p,pl,pres} there Gods{nom} reside{3p,pl,pres}

English : Where women are worshiped there reside the Gods.

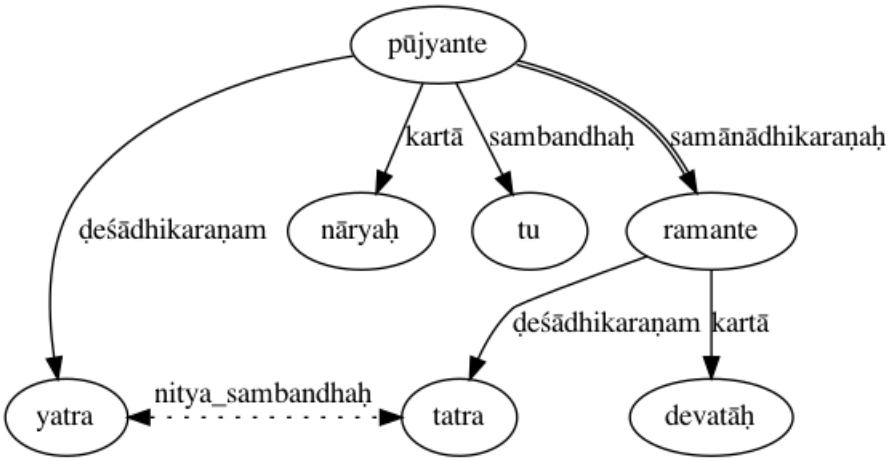

Figure 5: samānādhikaraṇaḥ relation

4. **Similarity (*sādṛśyam*) :**
This relation is of similarity. The similarity is between the two activities expressed through two consecutive sentences (See Figure 6). The example is :

Sanskrit : *janāni karmāṇi yathā kurvanti tathā te phalam prāpnuvanti.*
Gloss : people{nom} deeds{acc} as do{3p,pl,pres} so they{nom} fruit{acc} reap{3p,pl,pres}
English: People reap the fruits as per their deeds.

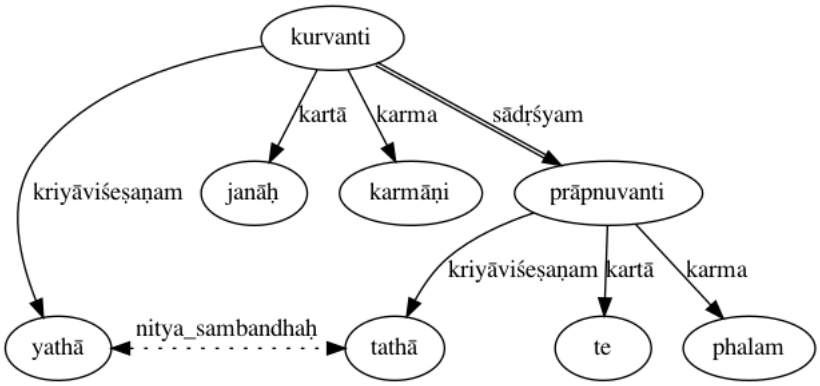

Figure 6: sādṛśyam relation

5. **Cause-Effect (*kārya-kāraṇa*) :**
When there is certainty about the cause, or the event expressing the cause has already taken place or there is a certainty that a certain event expressing the cause is going to happen, to express the certainty of the result following the cause, such constructions are used. This is a dichotomous relation where the sentence expressing the cause is marked with *yataḥ* indicating the reason/cause (*kāraṇa-dyotakaḥ*) and the sentence expressing the result is marked with the connective *tataḥ* which is an indicator of the result (*kārya-dyotakaḥ*). It is possible that only one connective among the two is used. Still it gives the same

meaning viz. cause-effect relation (See Figure 7). An example of this type of construction is:

Sanskrit : *yataḥ avarṣat tataḥ mayūraḥ nṛtyati.*
Gloss : because rain{3p,sg,pp} therefore peacock{nom} dance{3p,sg,pres}
English : Because it has rained, (therefore) peacock is dancing.

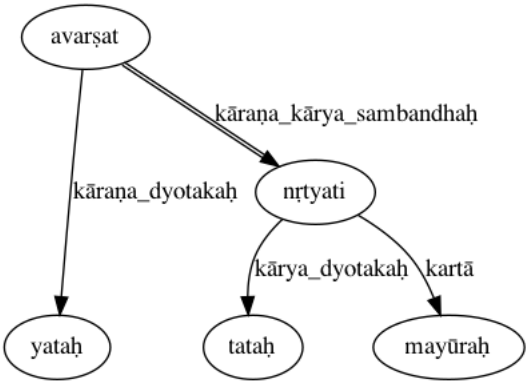

Figure 7: kārya-kāraṇa relation

6. **Conditional (*āvaśyakatā-pariṇāma*)** :
In slight contrast with the previous one, there are conditional sentences where there is no certainty of the event indicating the cause. To indicate the possibility of the resulting event provided the event corresponding to the cause takes place, such constructions are used. These sentences are marked with *āvaśyakatā-pariṇāma-sambandhaḥ*, which is a dichotomous relation where, the marker *yadi* indicates the necessity (*āvaśyakatā-dyotakaḥ*) and the marker *tarhi* indicates the result (*pariṇāma-dyotakaḥ*). The markers are used either in pair or individually as well (See Figure 8). An example of this type is:

Sanksrit : *yadi paṭhasi tarhi uttīrṇaḥ bhaviṣyasi.*
Gloss : if read{2p,sg,pres} then pass{nom} be{2p,sg,fut}
English : If you study (then) you will pass.

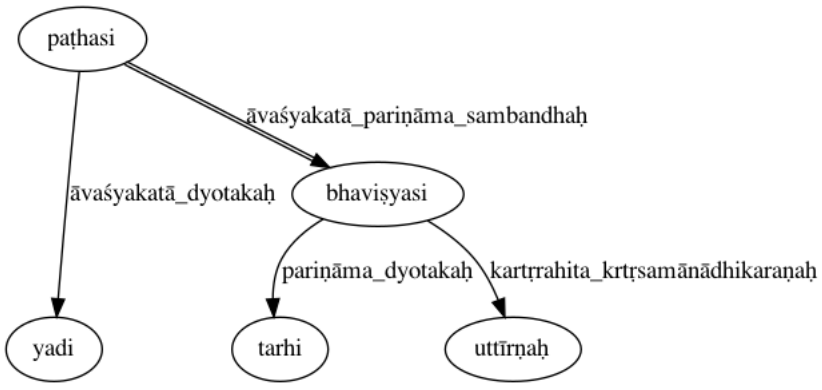

Figure 8: āvaśyakatā-pariṇāma relation

7. **Anomaly (*vyabhicāra*) :**
This is an exception or violation in naturally occurring cause-effect relationship. The pair of words *yadyapi* (even though) and *tathāpi* (even then) are the markers that trigger such

relations. We mark *vyabhicāra-sambandhaḥ* between the two finite verbs indicating the actions. The marker *yadyapi* is tagged as *kāraṇa dyotakaḥ* and *tathāpi* as *kārya dyotakaḥ*. The exceptions may be of two types. In one case even though the cause is present, the expected result is absent, and in the second case the result is present even though the desired cause is missing, thus violating the concomitance between the cause and effect (See Figures 10 and 9). The two types of examples are:

Sanskrit : *yadyapi varṣā bhavati tathāpi mayūraḥ na nṛtyati.*
Gloss: even-if rain{3p,sg,pres} happen even-then peacock{nom} neg dance{3p,sg,pres}
English: Even if it rains, even-then peacock does not dance.

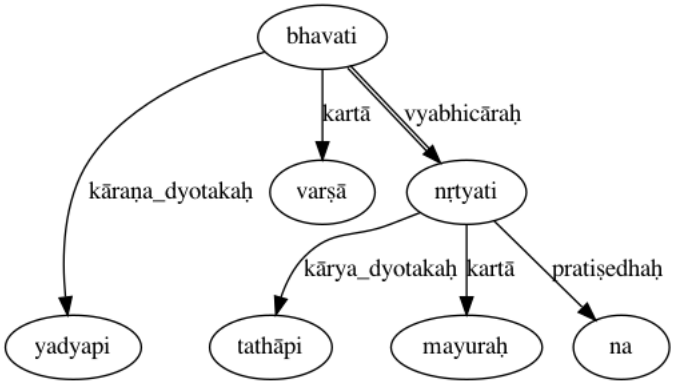

Figure 9: vyabhicāra relation type 1

Sanskrit: *yadyapi saḥ vaidyaḥ na asti tathāpi saḥ cikītsāṃ jānāti.*
Gloss: Even-if he{nom} doctor{nom} neg be{3p,sg,pres} even-then he{nom} cure{acc} know{3p,sg,pres}
Eng: Even if he is not the doctor, even-then he knows the cure.

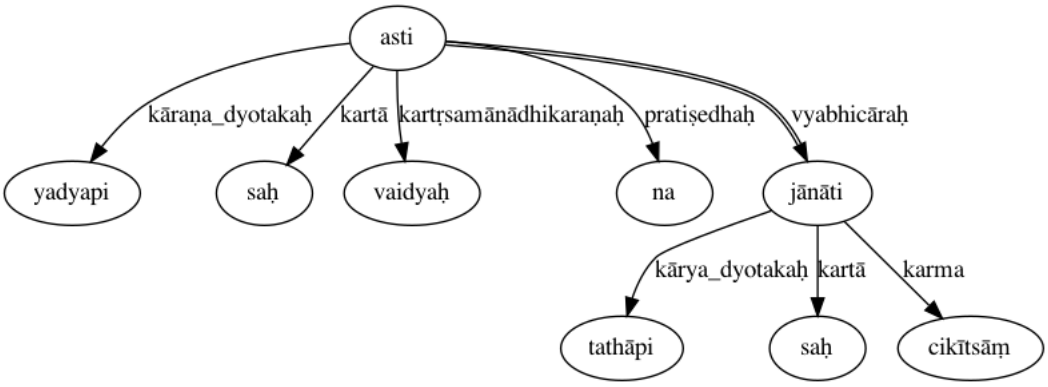

Figure 10: vyabhicāra relation type 2

8. **Antithesis (*virodhaḥ*) :**
Antithesis shows contradiction or opposition. It is typically marked by particles such as *parantu* and *kintu* (See Figure : 11). For example :

Sanskrit : *gajendraḥ tīvram prayatnam akarot parantu nakra-grahāt na muktaḥ.*
Gloss : Elephant{nom} hard effort{acc} do{3p,sg,past} but crocodile-grip{abl} no free{1p,sg,ppp}

English : Elephant tried hard but couldn't escape from the crocodile-grip.

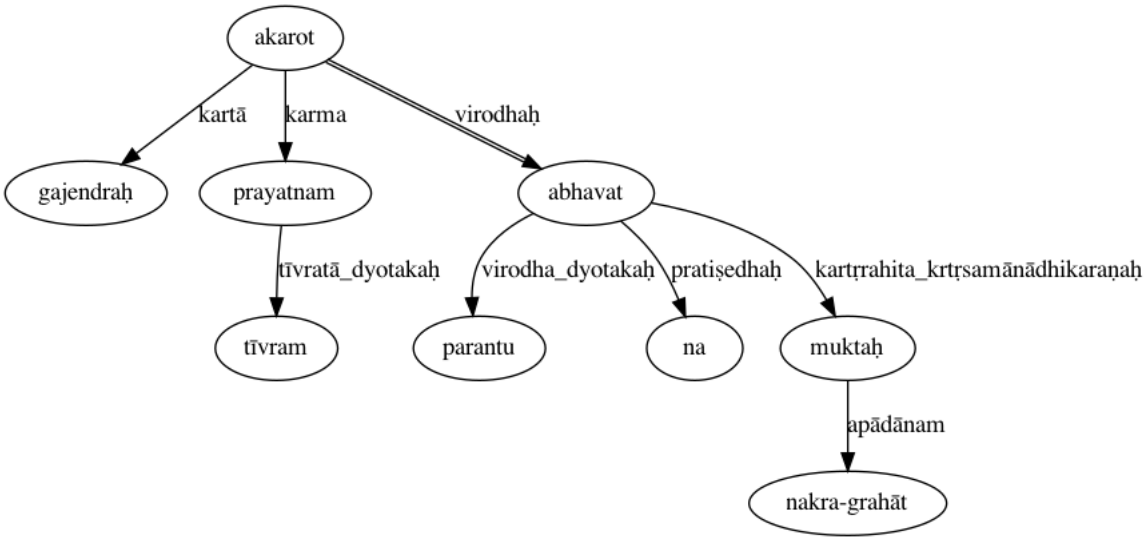

Figure 11: virodhaḥ relation

9. **Conjunction (*samuccayaḥ*):**
The conjuncts conjoined by the conjunctions are marked by this relation (See Fig 12). The example is :

Sanskrit : *Bhikṣām aṭa api ca gāṃ ānaya.*
Gloss : alms{dat} roam{2p,sg,imp} also and cow{acc} bring{2p,sg,imp}
English : Roam around for alms and bring the cow.

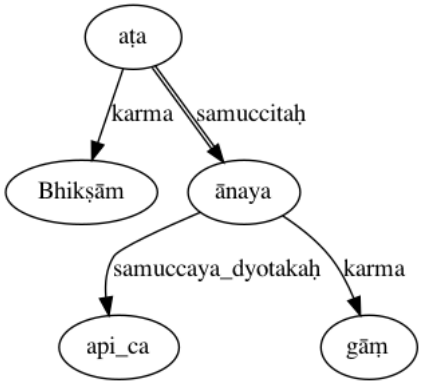

Figure 12: samuccayaḥ relation

The detailed discussion on the representatoon of conjuncts with various categories from computational point of view is presented in (Kulkarni and Panchal, 2019).

10. **Disjunction (*anyataraḥ*) :**
The disjuncts conjoined with disjunctive markers are marked by this relation (See Fig 13). The Example is :

Sanskrit : *sītā śvaḥ kāryakrame gāsyati athavā nartsyati.*
Gloss : Sita{nom} tomorrow program{loc} sing{3p,sg,fut} or dance{3p,sg,fut}

English : Sita will sing in tomorrow's program or will dance.

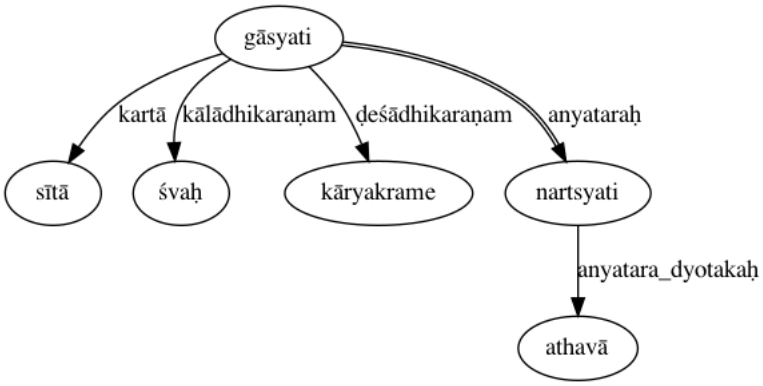

Figure 13: anyataraḥ relation

5 Implementation, Challenges and Evaluation

We selected *ŚrimadBhagvadGītā*(SBG) as a corpus for testing. There are 700 verses where each verse may consist of more than one sentence or more than one verse may consitute one sentence. The criterion for deciding the boundary of a sentence is '*eka ṭiṇ vākyam*'[3] and '*sākāṇkṣam cet vibhāge syāt*'.[4] Thus any group of words having one and only one finite verb and where every word is related to at least one other word from the group is termed as a sentence. Detecting sentence boundary has been earlier addressed by Hellwig (2016). It has been observed that a considerable number of errors are produced by the sentence boundary algorithm when sentences are smaller in length and especially without the use of copula. Since we were interested in the performance of the inter-sentential discourse analyser, to avoid cascading effect on the errors, we decided to manually annotate the sentence boundaries.

All the sentences having the inter-sentential markers were extracted from SBG. The distribution of various markers and the relations they mark is shown in Table 2. In the case of unambiguous markers, it was trivial to mark the relations automatically. The ambiguous markers fall under two categories. The first one has pronouns that are ambiguous due to the ambiguity of the case markers such as ablative and instrumental case suffixes which can mark a *kāraka* relation such as *apādānam* or *karaṇam* and a non-*kāraka* relation like *hetuḥ*. In this case unless the pronoun reference is identified, it is difficult to decide what relation is marks. The second category has ambiguous indeclinables such as *hi* and *atha*. In order to understand the problems in disambiguation, as a case study we looked at all the instances of *hi* in SBG. We describe below our observations.

5.1 Disambiguation of '*hi*'

The Sankrit-Hindi Apte's dictionary has the following four different senses of the word *hi*.

1. isaliye ki, kyoṁki
2. nissandeha, niścaya hi
3. udāharaṇasvarupa

[3]a group of words with one finite verb is a sentence.

[4]When a group of words is split into two parts and a word from one group has an expectancy for the word from the other group, all the words together from one sentence.

Markers	**Relation**	Frequency
tasmāt	*kārya-kāraṇa-sambandhaḥ*	21
tasmāt	*apādānam*	2
yasmāt	*kārya-kāraṇa-sambandhaḥ*	1
yasmāt	*apādānam*	1
tataḥ	*kārya-kāraṇa-sambandhaḥ*	8
tataḥ	*apādānam*	8
tataḥ	*anantarakālaḥ*	7
yataḥ	*kārya-kāraṇa-sambandhaḥ*	2
yataḥ	*apādānam*	2
ataḥ	*kārya-kāraṇa-sambandhaḥ*	2
ataḥ	*apādānam*	2
tena	*karaṇam*	7
yena	*karaṇam*	8
hi	*kārya-kāraṇa-sambandhaḥ*	49
hi	*sambandhaḥ*	17
tadā	*samānakālaḥ*	12
yadā	*samānakālaḥ*	11
tatra	*deśādhikaraṇam*	13
yatra	*samānādhikaraṇaḥ*	5
tathā	*sādṛśyam*	13
tathā	*kriyāviśeṣaṇām*	8
yathā	*sādṛśyam*	13
yathā	*kriyāviśeṣaṇām*	4
yadi	*āvaśyakatā-pariṇāmaḥ*	4
cet	*āvaśyakatā-pariṇāmaḥ*	6
ṭathāpi	*vyabhicāraḥ*	1
yadyapi	*vyabhicāraḥ*	1
anantaram	*anantarakālaḥ*	2
atha	*anantarakālaḥ*	1
atha	*samuccayaḥ*	4
atha	*praśnārthaḥ*	5
ca	*samuccayaḥ*	49
api	*praśnārthaḥ*	1
api	*sambandhaḥ*	52
vā	*anyataraḥ*	6
vā	*praśnārthaḥ*	1
athavā	*anyataraḥ*	2

Table 2: List of discourse relations and frequency occured in Śrimadbhagvadgītā

4. kevala, akelā

The Sanskrit-English Monnier William's dictionary has the following 3 different senses.

1. for, because, on account of

2. just, pray, do

3. indeed, assuredly, surely, of course, certainly

Speijer (1886) (§429) while commenting on it observes "*hi* was at the outset an emphatic, a weak 'indeed', but generally it is a causal particle, at least in prose.". Further in §443 Speijer states "... it has rather a general employment when annexing sentences which contain some motive, reason, cause or even an illustration of that which preceeds."

For the purpose of annotation we do not distinguish between the two usages marking emphasis (sense 2 of Sanskrit-Hindi) and marking exclusiveness (sense 4 of the Sanskrit-Hindi). We treat them under a generic term *sambandaḥ*, but we distinguish these usages from the usage of one marking the cause. The reason for not distinguishing between the emphasis and exclusiveness is that for their disambiguation just a sentence level information is not sufficient. One has to look at the context that may involve extra-linguistic information. There were total 66 shlokas that have '*hi*'. *Śaṅkarācārya* has commented on all the *śloka*s from 10th verse of the second chapter. There were 5 instances of '*hi*' till the 9th verse of the second chapter. Excluding these 5, among the remaining 61, *Śaṅkara* has marked 46 instance of '*hi*' as causal indicator, and 15 fall under the second category.

Both the authors classified '*hi*' in two categories independently without reffering to the *Śaṅkarabhāśya*. The classification is represented in Table 3.

	kārya-kāraṇam	**sambandaḥ**	total
Annotator 1	33	33	66
Annotator 2	49	17	66

Table 3: Inter-annotator agreement

The inter-annotator confusion matrix is shown in the Table 4. The comparison of the annotations of both the authors with that of *Śaṅkara* is shown in the Tables 5 and Table 6.

Thus we notice that, if we consider the *Śaṅkarabhāśya* as the gold data, the performance of the annotators measured against the gold data is not very satisfactory. Annotator 1 could

Annotator1↓	Annotator2 →		
	kārya-kāraṇam	**sambandaḥ**	Total
kārya-kāraṇam	27	6	33
sambandaḥ	22	11	33
Total	49	17	66

Table 4: confusion matrix: Annotator 1 and 2

Annotator1↓	Śaṅkarbhāṣya →		
	kārya-kāraṇam	**sambandaḥ**	Total
kārya-kāraṇam	25	4	29
sambandaḥ	21	11	32
Total	46	15	61

Table 5: confusion matrix : Annotator 1 and Śaṅkarabhāśya

Annotator2↓	Śaṅkarbhāṣya →		
	kārya-kāraṇam	**sambandaḥ**	Total
kārya-kāraṇam	39	6	45
sambandaḥ	7	9	16
Total	46	15	61

Table 6: confusion matrix : Annotator 2 and Śaṅkarabhāśya

mark only 60% of the cases correctly and the annotator 2 marked around 79% of the cases correctly. The disagreement between two annotators is also high, around 42%. If we look at the reason behind these differences we notice that the use of '*hi*' as an emphatic marker or exclusiveness marker sometimes also imply *kārya-kāraṇam*. This is observed in the commentary on '*vyākhyānato viśeṣapratipattiḥ na hi sandehāt alakṣaṇam*'. Here, we notice that almost all commentators before *Nāgeṣa* consider the use of '*hi*' as an emphatic marker, but *Nāgeṣa* categorically calls it *kāraṇa-dyotakaḥ*. The point we would like to drive here is that it is not trivial to disambiguate and many-a-times the whole context, the purpose etc. need to be taken into account.

Based on the data available we came up with some heuristics that uses the information of position of '*hi*' and the presence of pronouns or negative particle before it to classify '*hi*' into two categories. The results of the heuristics are shown in table 7. We note that the machine has provided correct results in 83% of the cases, outperforming both the annotators! It, of course, remains doubtful, if the heuristics developed for SBG will hold good across various genre of texts. The simple heuristic used is: if the word '*hi*' is at the second or the third position in a sentence, then it is marked as a *kāry-kāraṇa-bhāva*, with some special rules when the pronouns and indeclinables occur before the word *hi*. In all other cases it is marked as a *sambandhaḥ*.

Gold Data↓	Machine Results →		
	kārya-kāraṇam	**sambandaḥ**	Total
kārya-kāraṇam	41	5	46
sambandaḥ	5	10	15
Total	46	15	61

Table 7: machine produced results with comparison to gold data

6 Conclusion

The task of identifying analysing and implementing inter-sentential discourse relations with IGT perspective is still at an initial stage. We have identified various inter-sentential relations based on the explicit markers. Our next task is to identify the pair of verbs showing the expectancies such as to ask and to answer, to buy and to sell, etc. Another important task is to develop a module for anaphora resolution. We also noticed that the disambiguation is not an easy task. While some heuristics helped us in disambiguating the word *hi* it is not yet clear how much domain dependenct are these heuristic rules. The most important task ahead is therefore modeling *yogyatā*.

References

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In *International Conference on Language Resources and Evaluation*.

G V Devasthali. 1959. *Mīmāṃsā: The vākya śāstra of Ancient India*. Booksellers' Publishing Co., Bombay.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. English Language Series, London.

Dhanurdhar Jha. 2002. *Vākyārtha vivecanam*. Naag Publishers, Jaipur.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A macro-level Chinese discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

K V R Krishnamacharyulu. 2009. Annotating the Sanskrit texts based on the śābdabodha systems. In *3rd International Sanskrit Computational Symposium*. LNAI Springer Verlag.

Amba Kulkarni and Monali Das. 2012. Discourse analysis of Sanskrit texts. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 1–16, Mumbai, India, dec. The COLING 2012 Organizing Committee.

Amba Kulkarni and Sanjiv Panchal. 2019. Co-ordination in Sanskrit. In *Indian Linguistics, 80(1-2)*, pages 59–76.

Amba Kulkarni. 2019. *Sanskrit parsing based on the theories of Śābdabodha*. Indian Institute of Advanced Study, Shimla and D K Publishers (P) Ltd.

Amba Kulkarni. 2021. Sanskrit parsing following Indian theories of verbal cognition. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(2), Apr.

S Sheeja Kumari and Sobha Lalitha Devi. 2016. Annotations of connectives and arguments in malayalam language. volume 25, pages 280–285. 1st Global Colloquium on Recent Advancements and Effectual Researches in Engineering, Science and Technology - RAEREST 2016 on April 22nd & 23rd April 2016.

William C. Mann and Sandra A. Thompson. 1988. *Rhetorical structure theory: Towards a functional theory of text organization.*

Lucie Mladová, Šárka Zikánová, and Eva Hajicová. 2008. From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank. In *Proc. of LREC*.

Madhusudan Penna. 2021. *Pūrva Mīmāṃsā śāstra*, volume 2. Booksellers' Publishing Co., Bombay.

Livia Polanyi. 2008. The linguistic structure of discourse. In *The Handbook of Discourse Analysis*, pages 265–281.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2006. *The Penn Discourse TreeBank - Annotation Mannual 1.0*. University of Pennsylvania.

Ravi Teja Rachakonda and Dipti Misra Sharma. 2011. Creating an annotated Tamil corpus as a discourse resource. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 119–123, Portland, Oregon, USA, June. Association for Computational Linguistics.

Kevalanand Saraswati. 1888. *Mīmāṃsā koṣa*, volume 7. Pradnya Pathashala Mandal Granthamala.

P.M. Scharf and H.H. Hock. 2015. *Sanskrit Syntax: Selected Papers Presented at the Seminar on Sanskrit Syntax and Discourse Structures, 13-15 June, 2013, Université Paris Diderot*. Sanskrit Library.

J. S Speijer. 1886. *Sanskrit Syntax*. Leyden : E.J. Brill, University of Cornell.

Hrishikesh Terdalkar and Arnab Bhattacharya. 2019. Framework for question-answering in sanskrit through automated construction of knowledge. In *6th International Sanskrit Computational Linguistics Symposium (ISCLS)*, pages 98–117.

Oza Umangi, Prasad Rashmi, Kolachina Sudheer, Misra Sharma Dipti, and Joshi Aravind. 2009. The Hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161, Suntec, Singapore, aug. Association for Computational Linguistics.

Bonnie Lynn Webber and Aravind K. Joshi. 1998. Anchoring a Lexicalized Tree-Adjoining Grammar for discourse. In *Discourse Relations and Discourse Markers*.

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

Deniz Zeyrek, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, İhsan Yalçinkaya, and Ümit Deniz Turan. 2010. The annotation scheme of the Turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 282–289, Uppsala, Sweden, July. Association for Computational Linguistics.