ANNOTATION AND ANALYSIS OF DISCOURSE STRUCTURE IN THE TAMIL LANGUAGE

Thesis submitted in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING
BY RESEARCH

by

RAVI TEJA RACHAKONDA
200807015
teja.rr@gmail.com

International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2017
CERTIFICATE

It is certified that the work contained in this thesis, titled “ANNOTATION AND ANALYSIS OF DISCOURSE STRUCTURE IN THE TAMIL LANGUAGE” by RAVI TEJA RACHAKONDA, has been carried out under my supervision and is not submitted elsewhere for a degree.

08-8-2017

Date

Adviser: Dr. Rajeev Sangal
To Ray Rachakonda. The love of my life.

And Jotheeshwari Gopinath, Subha Rachakonda, Sarma Rachakonda, Aditya Rachakonda, and Amrita Lakshmi. Good people.

And Joseph Friedman, inventor of the bendy straw. Smart man.
Acknowledgments

Sincere appreciation and gratitude to Dr. Dipti Misra Sharma and Dr. Rajeev Sangal for inspiration, guidance, and assistance during my research. Special thanks to the late Prof. Lakshmi Bai for affection and trust. Warm regards to the gang at IIT for friendship.
Abstract

The behavior of discourse in a morphologically rich language such as Tamil has not been studied from a Computational Linguistics standpoint. A study of discourse in Tamil using a systematic framework such as PDTB would help identify the manner in which discourse features are realized in Tamil morphology. Added benefits of such a study include establishing the role of discourse in its relationship with other linguistic aspects, and facilitating cross-linguistic discourse analyses between language groups of differing morphological richness.
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Tamil Language</td>
</tr>
<tr>
<td>1.1</td>
<td>Morphology</td>
</tr>
<tr>
<td>1.2</td>
<td>Structure and Syntax</td>
</tr>
<tr>
<td>2</td>
<td>Discourse</td>
</tr>
<tr>
<td>2.1</td>
<td>Discourse from the Semantic standpoint</td>
</tr>
<tr>
<td>2.2</td>
<td>Approaches to Discourse Structure</td>
</tr>
<tr>
<td>2.3</td>
<td>Discourse Connectives</td>
</tr>
<tr>
<td>3</td>
<td>The Penn Discourse TreeBank</td>
</tr>
<tr>
<td>3.1</td>
<td>Discourse Relations</td>
</tr>
<tr>
<td>3.2</td>
<td>Hierarchy of Senses</td>
</tr>
<tr>
<td>4</td>
<td>A Tamil Discourse Relation Bank</td>
</tr>
<tr>
<td>4.1</td>
<td>Pilot Experiment</td>
</tr>
<tr>
<td>4.2</td>
<td>Discourse Connectives in Tamil</td>
</tr>
<tr>
<td>4.3</td>
<td>Annotation Process</td>
</tr>
<tr>
<td>4.4</td>
<td>The PDTB Approach</td>
</tr>
<tr>
<td>5</td>
<td>Modifications to PDTB</td>
</tr>
<tr>
<td>5.1</td>
<td>Changes to the Sense Hierarchy</td>
</tr>
<tr>
<td>5.2</td>
<td>Modifications to tree-structures</td>
</tr>
<tr>
<td>5.3</td>
<td>Sense annotation ambiguity</td>
</tr>
<tr>
<td>5.4</td>
<td>Miscellaneous</td>
</tr>
<tr>
<td>6</td>
<td>Corpus, Results and Unique Cases</td>
</tr>
<tr>
<td>6.1</td>
<td>The Wikipedia Tamil Corpus</td>
</tr>
<tr>
<td>6.2</td>
<td>Results</td>
</tr>
<tr>
<td>6.3</td>
<td>Combined Connectives</td>
</tr>
<tr>
<td>6.4</td>
<td>Redundant Connectives</td>
</tr>
<tr>
<td>6.5</td>
<td>Ambiguous Connectives</td>
</tr>
</tbody>
</table>
7 Discussion And Future Work .................................................................52
7.1 Are all Tamil discourse connectives suffixes?.................................53
7.2 Extending Dependency relations to Discourse .................................55

Bibliography ...............................................................................................58
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 The Dravidian Language Family</td>
<td>12</td>
</tr>
<tr>
<td>1.2 Major Dravidian languages in the Indian subcontinent</td>
<td>13</td>
</tr>
<tr>
<td>2.1 RST annotation</td>
<td>23</td>
</tr>
<tr>
<td>3.1 The PDTB Sense Hierarchy</td>
<td>29</td>
</tr>
<tr>
<td>5.1 The HDRB Sense Hierarchy</td>
<td>45</td>
</tr>
<tr>
<td>5.2 Annotation preserving Tree structure</td>
<td>47</td>
</tr>
<tr>
<td>5.3 Annotation violating the Tree paradigm</td>
<td>47</td>
</tr>
<tr>
<td>6.1 Combined Connectives</td>
<td>56</td>
</tr>
<tr>
<td>6.2 Redundant Connectives</td>
<td>58</td>
</tr>
<tr>
<td>7.1 Levels of linguistic analysis</td>
<td>65</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Results of the pilot experiment</td>
<td>32</td>
</tr>
<tr>
<td>6.1 Results of discourse annotation experiment</td>
<td>55</td>
</tr>
<tr>
<td>7.1 Possible morphological analysis of isolated connectives</td>
<td>64</td>
</tr>
</tbody>
</table>
Chapter 1

The Tamil Language

Tamil is the official language of the Indian state of Tamil Nadu and the union territory of Pondicherry, and is spoken by the Tamil people of South India, Sri Lanka, Singapore, and in emigrant Tamil communities around the world.

Tamil is one of the oldest of the Dravidian languages and one of the longest surviving classical languages in the world, with extant literature of more than 2000 years old. In addition, Tamil's standard metalinguistic terminology and scholarly vocabulary is itself Tamil, as opposed to the Sanskrit that is standard for most other Dravidian languages. In 2004, Tamil became the first language to be declared as a classical language by the Government of India\(^1\).

Tamil belongs to the South Dravidian group of the Dravidian language family [1] that comprises of around 26 other languages, of which Malayalam is the closest in relation to it (see Figure 1.1 and 1.2).

\(^1\) [http://en.wikipedia.org/Tamil_language](http://en.wikipedia.org/Tamil_language)
Figure 1.1 The Dravidian Language Family. (Languages marked with a * are officially recognized by the Government of India).
Figure 1.2 Major Dravidian languages in the Indian subcontinent.
1.1 Morphology

Tamil is an agglutinative language – also a trait that is found in other Dravidian languages like Telugu, Kannada, Tulu and other languages such as Japanese, Korean and Turkish. As a result, suffixes are used to mark noun class, number, and case, verb tense and other grammatical categories.

Most Tamil words consist of a lexical root to which are attached one or more affixes – most of them being suffixes, though there are also some valid prefixes. This process results in the fusion or alteration of sounds, and is governed by certain morphophonological rules called puNarchi, commonly known as Sandhi. Also, agglutination can occur between morphemes, called internal sandhi, or at word boundaries, where it is referred to as an external sandhi.

Suffixes can be classified into either derivational or inflectional suffixes. The former are those that change the part-of-speech of the word; the latter kind are the ones which mark categories such as person, number, mood, tense, etc., but do not alter the part-of-speech category of the base form. Suffixes are used to perform the functions of cases or postpositions. The various suffixes that nouns take can also be grouped into those corresponding to eight different grammatical cases, viz., nominative, accusative, dative, vocative, genitive, instrumental, locative, and ablative. In addition, the suffixes also can serve as markers performing other functions such as those expressing dubitation, interrogation, emphasis, etc. [2]

As there is no absolute limit on the extent of agglutination, this can lead to long words with a large number of suffixes, as in (1.1.0)

என்றும் சொல்லப்பட்டும் கொள்ளப்படும்.
Because of the nuisance of the three-wheeled vehicles (i.e., autorickshaws).”
1.2 Structure and Syntax

In terms of word order, Tamil, like other Dravidian languages, is consistently head-final. This means that in most clauses, the verb occurs at the end of it, with a very common word order being one of Subject–Object–Verb. However, word order in Tamil is relatively flexible (when compared to, say, English); hence word orders such as OSV, VSO and others are also possible, each of which may have some effects on the topicality and emphasis in the sentences.

As mentioned above, like all other Dravidian languages, declarative sentences in Tamil usually show a verb-final sentence structure. However, sentences can be expressed with a relatively free word-order due to its richness of morphological information. Thus, an S-O-V sentence like (1.2.1) can also be mentioned in an O-S-V (1.2.2), V-S-O (1.2.3), or V-O-S (1.2.4) word ordering.

రామన్ సితాయి కాదలిత్తాన్.
rAmaN sItayai kAdaliththAN.
ram.M sita.ACC love.PST.M
‘Rama loved Sita’.
(1.2.1)

సితాయి రామన్ కాదలిత్తాన్.
sItayai rAmaN kAdaliththAN.
sita.ACC ram.M love.PST.M
‘Rama loved Sita.’
(1.2.2)

каதாலித்தான் ராமன் சித்தலய்.

kAdaliththAN rAmaN sItyai.

love.PST.M ram.M sita.ACC

‘Rama loved Sita.’

(1.2.3)

kaதாலித்தான் சித்தலய் ராமன்.

kAdaliththAN sItyai rAmaN.

love.PST.M sita.ACC ram.M

‘Rama loved Sita.’

(1.2.4)

The same is the case even with other sentence types such as interrogatives, imperatives, and exclamatory sentences. An interesting feature is that a declarative sentence (1.2.5) can be converted into an interrogative (1.2.6) by in-situ replacement, i.e., by swapping the question term with the appropriate question word. A similar conversion in English, by contrast, would require additional syntactic modifications such as change in word-order, inversion of subject and auxiliary verb, etc.

(1.2.5)

ராமன் ராவணலன் தகான் றான்.

rAmaN rAvaNaNai konDrAN.

ram.M ravan.M.ACC kill.PST.M

‘Rama killed Ravana.’

(1.2.6)

The same is the case even with other sentence types such as interrogatives, imperatives, and exclamatory sentences. An interesting feature is that a declarative sentence (1.2.5) can be converted into an interrogative (1.2.6) by in-situ replacement, i.e., by swapping the question term with the appropriate question word. A similar conversion in English, by contrast, would require additional syntactic modifications such as change in word-order, inversion of subject and auxiliary verb, etc.

(1.2.5)

రாமன் ராவணலன் தகான் றான்.

rAmaN rAvaNaNai konDrAN.

ram.M ravan.M.ACC kill.PST.M

‘Rama killed Ravana.’
Whom did Rama kill?

Also, not all Tamil sentences have subjects, verbs, and objects and hence many sentences which lack one or more of the three can be considered grammatical. These aspects can be seen respectively in (1.2.7) where the entire sentence is just a verb with its agglutinations and in (1.2.8) where the sentence does not explicitly show a verb, and is called a zero copula construction.

We have been coming.

‘He is a good student.’
Chapter 2

Discourse

Discourse is a term that has differing meanings from the standpoints of social science and linguistic semantics.

In social sciences such as linguistics, sociology, anthropology, social work, cognitive psychology, social psychology, international relations, human geography, communication studies, and translation studies – each subject to its own assumptions, dimensions of analysis, and methodologies – the definition of discourse pertains to the various usages of a language attached to different social practices. The type of discourse provides the vocabulary, expressions and the style required for communication. For example, works that pertain to the topic of atheism differ widely in language, connotations, and style in conservative texts when compared to those from more liberal sources. Similarly, articles on Naxalites can refer to them as either terrorists or revolutionaries and their actions condemned or condoned accordingly, depending on the viewpoints of the respective writers.

Linguistic semantics, on the other hand, is the study that primarily pertains to the relations between signifiers, like words, phrases, etc., and signified, the meanings that they stand for. More simply, it is the study of relations between text and context. These relations between words (or groups of words) and their meanings exist at the word, phrase, clause, sentence, paragraph, document and even higher levels. Many models have been proposed that describe and explain these relations and inter-relations. A detailed view of discourse in linguistic semantics is provided in the following section.
2.1 Discourse from the Semantic standpoint

In the study of discourse in linguistic semantics, a given piece of work is segmented into individual units of discourse, called as events or abstract objects. These units usually contain a single piece of information or a fact, like for example, a syntactic clause. These events are usually related to each other by means of explicit lexical items or contextual cues.

These relations can sometimes be in the form clear syntactic constructions, as in (2.1.1) or pragmatic ideas that require contextual real-world knowledge, like in (2.1.2).

(2.1.1)

[As the temperatures began to drop in the capital city] [people are forced to stay indoors for warmth and comfort].

(2.1.2)

[I’m using an old driver]. [The game runs terribly choppy].

In (2.1.1), the events (represented in square brackets) are linked by a syntactic lexical item ‘as’ that shows a causal relationship between them.

However, no such item exists for the two events in (2.1.2) and the causal relationship is obtained from the pragmatic information about the ‘driver’ being a hardware device driver which, being out of date, is causing a game to be ‘choppy’ i.e., run with a less than optimal display rate of frames per second.

This way, discourse structure deals with all such relationships that are identified to hold between individual events of a text. The process of arriving at the framework that
is used to show the underlying discourse structure is called discourse modeling, and the process of identifying the relations between events – either manually or in an automated manner – as defined by a discourse model is called as discourse annotation.
2.2 Approaches to Discourse Modelling

One way of neatly classifying the major approaches to that are proposed to model discourse is based on the reasons behind the declaration of discourse relations between components. In other words, the triggers that signify the existence of discourse relations have led to different annotation schemes. These can be enumerated based on lexical elements, discourse structure, or both.

2.2.1 Lexical Elements

Halliday and Hasan [3] suggest that the primary reason for the existence of discourse relations is to bring about cohesion between discourse elements. These relations can be identified by the presence of certain lexical elements such as coordinating and subordinating conjunctions (and, as) or adjuncts (therefore, additionally, on the other hand, etc.).

This way, discourse relations necessitate a matrix element, a presupposed predicate element, and a lexical element that links the first two and signifies a discourse relation that helps in making the text coherent. In addition, they argue that all such linkages are binary and summarily reject the notion of a structure that exists between the discourse elements.

2.2.2 Discourse Structure

There are many other approaches that take an opposing view to that of Halliday and Hasan and emphasize the existence and importance of discourse structure and argue that discourse annotation should be performed only to associate the discourse relations with that underlying discourse structure.
Rhetorical Structure Theory (RST) postulated by Mann and Thompson [4] is one such approach that supports a discourse structure that can be realized through context-free rules, which when applied to a text yield a tree structure. In this structure, the root node shows the overall discourse structure of the text, individual discourse units are the leaves and discourse relations hold only between daughters of non-leaf nodes. Also, in each relation, one member is taken as the dominant one, the nucleus, and the other is considered as providing additional information, called as the satellite. Figure 2.1 shows the RST annotation of the (2.2.1), with each discourse unit numbered.

(1) The housing prices in Cincinnati have steadily declined over the last few years, (2) leading investors to be wary of venturing into these troubled waters. (3) As a result, the housing market has taken a major hit in these areas.

(2.2.1)

![Figure 2.1 RST annotation.](image)

Discourse GraphBank by Wolf and Gibson [5] is another approach that emphasizes discourse structure as one that triggers relations but rather asserts that it need not be a tree structure, but one in where discourse units can be arguments to more than one discourse relation, resulting in what is known as a chain graph.

**2.2.3 Lexical Elements and Discourse Structure**
Hybrid approaches like Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG) by Webber and Joshi [6] argue that discourse relations can be triggered by both lexical elements as well as the underlying structure between adjacent elements that are unmarked (by lexical elements). Thus, the discourse meaning of a text can be arrived in manner akin to that of sentence meaning.

Hence, just as sentence relations involve verbs as predicates that require nouns as arguments, at the discourse level, discourse relations exist where discourse predicates require discourse units as their arguments. Similarly, in the absence of these predicates, relations are also conveyed by adjacency, similar to how noun modifiers act at the sentence level.

As a result, the approach combines lexical elements and the underlying discourse structure where the lexicalized elements (called that we call discourse connectives) are associated with a set of trees that show their discourse configuration, again similar to how, at the sentence level, each verb is associated with frames that show its different configurations.

The Penn Discourse TreeBank (PDTB) [7] is based on the D-LTAG approach, but designed to be theory-independent, such that all relations and their arguments are annotated the same way, irrespective of the annotation approach. This helps in comparing and testing the claims of different discourse annotation schemes. A detailed description of the PDTB schema is given in Chapter III.
2.3 Discourse Connectives

A discourse connective can be considered as the predicate of a binary relation which takes two text spans (mainly clauses or sentences) as its arguments. These arguments, which are usually prepositions, events or states, are called abstract objects.

Essentially, a discourse connective connects these units to larger ones while signaling a semantic relation between them at the same time. In many languages, discourse connectives, like modality markers, are above or outside of the proposition and hence are morphologically inflexible and do not act as grammatical constituents of a sentence.

In general, they are represented by coordinating conjunctions (and, but), some subordinating conjunctions (because, if, while), some particles (also, only) and sentence adverbials (afterwards), and marginally also by some other parts-of-speech – mainly in cases of fixed compound connectives like in other words or on the contrary.

There have been attempts where discourse connectives have been classified into two groups: Structural and Anaphoric. [6] Here, structural connectives are usually subordinating conjunctions, coordinating conjunctions and other subordinators like in order to, on the other hand, etc. that are related to their arguments, either explicitly or implicitly, by means of an underlying (usually tree) structure of discourse.

Anaphoric connectives are mostly those that are discourse adverbials, such as in addition, furthermore, etc. which connected structurally to one argument and only by a later reference (anaphora) to the other.
Chapter 3

The Penn Discourse Tree Bank

The PDTB [7] is a resource built on discourse structure in Webber and Joshi [6] where discourse connectives are treated as discourse-level predicates that always take exactly two abstract objects such as events, states and propositions as their arguments. We now describe the types of connectives and their senses from the PDTB framework and provide examples from the annotated PDTB corpus.
3.1 Discourse Relations

In PDTB, discourse relations are primarily realized by the means of discourse connectives that can be explicit, implicit, or alternate lexicalizations.

The city’s Campaign Finance Board has refused to pay Mr. Dinkins $95,142 in matching funds because his campaign records are incomplete. (PDTB 0041)

The relationship between the two events described in (3.1.1) is brought about by the use of the connective because, which is hence marked as an explicit discourse connective in PDTB. In contrast, (3.1.2) shows two events that are related by a context that is not explicitly realized in the form of a connective. These are called implicit connectives, which, when inserted in the appropriate location in text, better indicate the relation that is being realized between the events.

Motorola is fighting back against junk mail. So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it. (Implicit = so) Now, thousands of mailers, catalogs and sales pitches go straight into the trash. (PDTB 0989)

Sometimes, discourse relations are not realized explicitly or implicitly in the form of words or phrases that can be termed discourse connectives, but rather as sentences or clauses that perform the same function. These are called alternate lexicalizations.

Ms. Bartlett’s previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as
its nominal subject. AltLex [Mayhap this metaphorical connection made] the BPC Fine Arts Committee think she had a literal green thumb. (PDTB 0984)

(3.1.3)

In (3.1.3), there is neither an explicit connective, nor can an implicit one be inserted to make the relationship clear; however, we see that the clause mayhap this metaphorical connection made functions like a discourse connective that links the discourse units surrounding it.

In addition, it may be the case that there is an entity-based coherence between the events, or it could be that no relation could be identified between them, as seen in (3.1.4) and (3.1.5) respectively.

Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. EntRel Mr. Milgrim succeeds David Berman, who resigned last month. (PDTB 0945)

(3.1.4)

Jacobs Engineering Group Inc.’s Jacobs International unit was selected to design and build a microcomputer-systems manufacturing plant in County Kildare, Ireland, for Intel Corp. Jacobs is an international engineering and construction concern. NoRel Total capital investment at the site could be as much as $400 million, according to Intel. (PDTB 1081)

(3.1.5)
3.2 Hierarchy of Senses

Each time a relationship is identified by means of a discourse connective, it is then coupled with a *sense* that categorizes the type of relationship expressed by that connective. These senses are classified hierarchically using four top-level *classes* ‘Comparison’, ‘Contingency’, ‘Expansion’ and ‘Temporal’. Each class is refined by its component *types* and these, in turn, are further refined by the *subtype* level. Figure 3.1 shows the sense hierarchy developed by the PDTB group originally for annotating English discourse relations.

![Figure 3.1 The PDTB Sense Hierarchy](image-url)
Chapter 4

A Tamil Discourse Relation Bank

In this Chapter, we explore the agglutinative nature of the Tamil language, as it necessitated a deeper analysis to look into suffixes that act as discourse connectives in addition to those that occur as unbounded lexical items.

We also show how we have incorporated the types of connectives and their senses from the PDTB framework in our attempts to create a Tamil Discourse Relation Bank [8], providing examples from sentences in the Tamil.
4.1 Pilot Experiment

To obtain a preliminary insight into the types of discourse connectives present in Tamil, a POS-tagged corpus of information released by the Tamil Nadu Tourism Department provided by AU-KBC Research Centre\(^2\) was analyzed manually for discourse information.

From 7508 sentences comprising of 130610 words that comprise the corpus, a total of 148 discourse connectives were identified. Further analysis revealed that there were only 61 distinct connectives among them, and that only 3 connectives occurred with a frequency of more than 5 times [8]. These most frequent connectives and their counts are provided in Table 4.1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Tamil</th>
<th>Translit.</th>
<th>Meaning</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>எனவே</td>
<td>enavE</td>
<td>‘so’</td>
<td>26</td>
</tr>
<tr>
<td>2</td>
<td>ஆனால்</td>
<td>AnAl</td>
<td>‘but’</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>மற்றும்</td>
<td>mattRum</td>
<td>‘and’</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>அல்லது</td>
<td>alladu</td>
<td>‘or’</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>என்றாலும்</td>
<td>enrAlum</td>
<td>‘though’</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>செய்து</td>
<td>ceydu</td>
<td>‘having-done’</td>
<td>2</td>
</tr>
</tbody>
</table>

\(^2\) http://www.au-kbc.org/nlp.htm
The reason for this paucity in the number of connectives that could be identified during this experiment is primarily because of the agglutinative nature of the Tamil language, which causes discourse connectives to appear as affixes instead of separate words that are identified and POS tagged as such.

As a result, a study of discourse structure in Tamil would require an in-depth analysis of the agglutinated connectives that are usually suffixed to verb forms, in addition to the connectives that occur as separate words altogether.
4.2 Discourse Connectives in Tamil

Many connectives in Tamil have both separate-word and agglutinated forms. Consider the sentence in (4.2.1) where *matRum*, the separate-word form of the connective *and*, is used:

\[\text{kumaran matRum pAri viLayADa senRanar.} \]

kumar.M and paari play.INF go.PST.PL

‘Kumaran and Paari went to play’

(4.2.1)

Now, compare it to the sentence in (4.2.2) where *-um*, the equivalent agglutinated form of *and*, is seen affixed to both the nouns *kumaran* and *pAri*:

\[\text{kumar.aum pAryaum viLayADa senRanar.} \]

kumar.M=and paari=and play.INF go.PST.PL

‘Kumaran and Paari went to play’

(4.2.2)

Although both these forms are grammatically valid, the latter is used in preference to the former by native speakers of the language. [2] While the connective in both these
cases strictly conveys a non-discourse usage, we have observed (in the following section) that this is true even with discourse connectives.
4.3 Annotation Process

The process of discourse annotation involves identifying discourse connectives in raw text and then annotating their arguments and semantics. Discourse connectives are identified as being explicit, implicit, AltLex, EntRel, or NoRel by Prasad et al [7]. By convention, annotated explicit connectives are underlined and implicit connectives are shown by the marker, “(Implicit=)”. As can be seen in (4.3.1), one of the arguments is shown enclosed between {} and the other argument is shown in []. The AltLex, EntRel or NoRel relations are shown by underlining, i.e., as “(AltLex=)” “(EntRel)” and “(NoRel)”, respectively.

\text{\{eN kAl uDaindadaN\}Al [eNNAl viLayAAdA muDiyaDdu\].}

\text{my leg break.PRF.CAUS my.INST play.INF can.FUT.NEG}

\text{‘\{My leg broke\}, hence [I cannot play].’}

(4.3.1)
4.4 The PDTB Approach

4.4.1 Explicit Discourse Connectives

Explicit discourse connectives are lexical items present in text that are used to anchor the discourse relations portrayed by them. In Tamil, they are found as affixes to the verb, as in (4.4.1.1) where the affix -Al conveys the meaning ‘so’. This is in a way similar to the simplex subordinators in Turkish, as described in Zeyrek and Webber [9]. However, like in English, explicit discourse connectives are also realized as unbound lexical items, as can be seen in (4.4.1.2) where the word eNavE means ‘hence’.

அங்கு அல்லாறு சிரிப்பிக்கவும் வலங்கு மாநிலங்களை அண்மை விளையாடும் விளையாடை அவ்வாறு.

avaradu uDalnalam sariyillAmaiyaLAl [nAngu he.HON.POSS body-wellness right.NEG.INF.CAUS four mAdangal avarAl viLayADa iyalavillai]. month.PL he.HON.INST play.INF can-no

‘{He was suffering from ill health} so [he could not play for four months].’

(4.4.1.1)
Thirukkural has been written in such a way that people from all religions can benefit from it. Hence, this book is praised by many.

(4.4.1.2)

Syntactically, explicit connectives can be coordinating conjunctions e.g., alladu (‘or’), subordinating conjunctions e.g., -Al (‘so’), sentential relatives e.g., -adaNAl (‘because of which’), particles e.g., -um (‘also’) or adverbials e.g., -pOdu (‘just then’).

Explicit connectives also occur as conjoined connectives where two or more instances of connectives share the same two arguments. Such connectives are annotated as distinct types and are annotated discontinuously, as seen in (4.4.1.3) where the connectives -um and -um are paired together to share the same arguments.
Government schools need to help in {providing nutritious food to the students} and [making sure they do exercise].

(4.4.1.3)

4.4.2 Implicit Discourse Connectives

Implicit discourse connectives are inserted between adjacent sentence pairs that are not related explicitly by any of the syntactically defined set of explicit connectives. In such a case, we attempted to infer a discourse relation between the sentences and a connective expression that best conveys the inferred relation is inserted. In (4.4.2.1), the implicit expression uthAraNamAga ('for example') has been inserted as an inferred discourse relation between the two sentences.
eora clan people.POSS language.ABL today

Angilattil vazangum sorkaL uLLaNa}. (uthAraNamAga)

english.LOC present word.PL exist.PST example-make

[dingO, vUmErA, vAlabi pONra sorkaL IyOravilirindu
dingo  womera wallaby like word.PL eora.ABL
tONriya sorkaLdAN].

originate.PST word.PL.EMP

‘{There are words that are present in English that originated from
the language of the Eora people}. (Implicit= For example) [Dingo,
Woomera and Wallaby are words with their origins in Eora].’

(4.4.2.1)

4.4.3 AltLex, EntRel and NoRel

In cases where no implicit connective was appropriately found to be placed between
adjacent sentence-pairs, we now look at three distinct classes. AltLex relations, as seen
in (4.4.3.1) are discourse relations where the insertion of an implicit connective leads
to a redundancy in its expression as the relation is already alternatively lexicalized by
some other expression that cannot be labeled as an explicit connective. (4.4.3.2) shows
an EntRel relation where no discourse relation can be inferred and the second sentence
provides further description of an entity realized in the first sentence. When neither a
discourse relation nor entity-based coherence can be inferred between the two adjacent
sentences, it is described as a NoRel, shown in example (4.4.3.3).
(mudalAvadAga mAgim, jOgEshwari, pUrivilla rayil
first-like.ADV mahim jogeshwari poorivilla rail
nilayangaLil guNDu vedittadu) (idai toDarndu)
station.PL.LOC bomb explode.PST this.ACC follow.CNT
[mErku rayilvEyiN aNaittu rayilgaLum
west railway.POSS all rail.PL=and
niruttappaTTaNa].

stop.INF-make.PST.PL

'{Initially, bombs exploded in Mahim, Jogeshwari and Poorivilla).
(AltLex=following this) [all the trains from the western railway were halted].'

(4.4.3.1)

இன்னும் பெருமளவான பேர்மன்றங்களை விளக்கினால் பார்க்கும் புரிமறை விளக்கும் பேர்மன்றங்களைச் சொன்னாலும் வேளை கேன்று.

{ivvANDu kirikket ulagakkOppai mErkindiya

30
This year's Cricket World Cup was held in West Indies from the thirteenth of March to the twenty-eighth of April. (EntRel) [In this competition, the teams representing the sixteen nations were grouped into four groups with four teams in each group].

(4.4.3.2)
Sachin Tendulkar is considered the best batsman in the world. (NoRel) [Indian bowlers are not being given proper training].

(4.4.3.3)
Chapter 5

Modifications to PDTB

In this chapter, we show cases where we needed to modify the existing PDTB schema. While we have incorporated many changes to PDTB proposed by existing approaches, we illustrate certain interesting examples that were distinct from those observed in such approaches, thereby arguing that such changes will be useful in describing the same phenomena in other Dravidian languages.
5.1 Changes to the Sense Hierarchy

The Hindi Discourse Relation Bank (HDRB) [10] is an approach that extends the PDTB to Hindi as necessitated by the features of the Hindi language. During this effort, several modifications were proposed to PDTB to make it even more language-agnostic. Here, we describe the major deviations that have been taken from PDTB that we have incorporated into our efforts.

5.1.1 Eliminating argument-specific labels

Since the order of arguments labels don’t directly pertain to the meaning of discourse relations, these labels are removed from the subtype levels. For example, the subtypes Reason and Result depending upon the argument order have been merged into a single parent type Cause. All levels in the sense hierarchy thus have the purpose of specifying the semantics of the relation to different degrees of granularity. The relative ordering of the arguments is instead specified in the definition of the type-level senses, and is inherited by the more refined senses at the subtype level.

5.1.2 Uniform treatment of pragmatic relations

The PDTB pragmatic senses are replaced with a uniform three-way classification. Each pragmatic sense at the type level is further distinguished into three subtypes: epistemic, speech-act, and propositional. While the former are described in Sweester [11], the propositional subtype involves the inference of a complete proposition. The relation is then taken to hold between this inferred proposition and the propositional content of one of the arguments.

5.1.3 The Goal sense
Under the Contingency class, a new type Goal has been added, which applies to relations where the situation described in one of the arguments is the goal of the situation described in the other argument that enables its achievement (see Figure 5.1).

Figure 5.1 The HDRB Sense Hierarchy
5.2 Modifications to tree-structures

The PDTB schema stipulates that discourse connectives essentially form logical Tree structures where the connective itself is the parent node and the arguments ARG1 and ARG2 become its child nodes.

While requirement of discourse annotation to form neat Tree structures has been debated before by Lee et al [12], we have found several example where departures from Tree structures are indeed necessary to preserve the integrity of other aspects of annotation.

Consider (5.2.1) where two connectives, the Conjunction marker -um (‘and’) and the Cause marker –Al (‘because’), appear to have combined to become -Alum (‘and-because’).

Because I played cricket and because I did exercise I am tired.

(5.2.1)
Here, we can either (a) treat the *Alum ... -Alum* as a separate connective with a unique *because-and* sense or (b) take the combined elements of *-um ... -um* and *-Al* separately with their own Conjunction and Cause senses (see Figure 5.2 and Figure 5.3, respectively).

![Figure 5.2 Annotation preserving Tree structure.](image)

In case (a) is that the ARG1 and ARG2 spans neatly separate into a Tree structure whereas the major issue is that there is no equivalent *because-and* connective sense in the PDTB guidelines.

![Figure 5.3 Annotation violating the Tree paradigm.](image)

In case (b), it is just the opposite where the each part of the combination can be expressed using a sense from the PDTB sense hierarchy but the downside is that the
argument spans overlap between the two connectives, which effectively translates to certain child nodes having more than one parent node, thereby invalidating the Tree paradigm of PDTB discourse modelling.
5.3 Sense annotation ambiguity

There are certain discourse connectives that look like they are expression a specific type of relation between their arguments at the surface level, only to see that they also show another relation in a pragmatic sense.

The connective *piNNar* (‘*after*’) usually holds a relation of a sense type *Temporal* between its arguments. However, in sentences like (5.3.1) it may be found that in addition to the temporal relation, a stronger relation of sense type *Contingency: Cause* exists between the two events as it was because of ARG1 that ARG2 happened in addition to the fact that ARG1 and ARG2 happened one after the other.

{**Bangladesh’s 7-wicket victory against Bermuda**, [it got promoted to the Super-8 competition].}
(5.3.1)

As a result, our choices are currently to either annotate only the stronger sense among the two (or maybe more) competing sense relations, or have an option in the guidelines to include multiple senses (in ranked or unranked order) for a given connective.

We chose the latter, disregarding an identified relation in the text during annotation would mean losing valuable information, and a conservative approach will help in identifying similar connectives that show multiple senses for purposes of further study.
5.4 Miscellaneous

According to the PDTB definition of AltLex, it is usually the case where a phrase performs the function of a discourse connective in expressing a relation. However, we have identified instances where an entire sentence (as opposed a clause) performs the function of a single discourse connective.

This would require a broadening in the PDTB definition of AltLex to include larger syntactic structures like sentences as well. Whether multiple sentences can be ascribed as connectives requires a deeper study into the nature and functions of discourse connectives in general, however.

The sentence shown in (5.4.1) can be seen as an AltLex that holds a relation of type *Contingency: Cause* between the abstract events XYZ and ABC.

\[
\text{XYZ. இதற்கான காரணங்கள் பின்வருமாறு. ABC.}
\]

\[
\{XYZ\}. \text{this.ACC.PST reason.PL back-come.INF}
\]

\[
\{XYZ\}. \text{The reasons for this are as follows. [ABC].}
\]

(5.4.1)

Our efforts in annotating the agglutinative morphology in Tamil required another minor change worth mentioning, where the annotation software needed to be modified to enable the annotation of lexical units smaller than words, like characters, syllables, vowel markers, etc., to accurately represent the argument and connective spans.
At the time of writing, the latest version of the PDTB annotation tool can be used for highlighting sub-word elements, and also supports agglutinative Indic languages.
Corpus, Results, and Unique Cases

Since there has been no previous work in the annotation of discourse connectives in Tamil, the theory-independent scheme of the PDTB was an important reason for us to choose it over other approaches. This way, the annotated corpus could be used to possibly test, verify, or modify existing theories, or probably postulate new ones altogether.

Since the PDTB approach has already been carried out in many diverse languages such as English, Chinese, Finnish, Turkish, Hindi, etc., it would be interesting to perform a cross-linguistic study of discourse features and connectives; one that helps provide a wider and perhaps better understanding of discourse phenomena.

Also, the original PDTB work was performed on corpora on non-agglutinative languages like English where discourse connectives occurred as separate words that occurred in text and were clearly demarcated by spaces. Using the approach on an agglutinative language like Tamil required a study to observe the morphological construction of discourse connectives in it.

For the sake of completeness, we have decided to annotate both structural and anaphoric connectives with their arguments.
6.1 The Wikipedia Tamil Corpus

We collected Tamil encyclopedia articles from the June 2008 edition of the Wikipedia static HTML dumps\(^3\). Elements such as HTML metadata, navigational links, etc. were then removed until only the text of the articles remained. A corpus was then built by collecting the texts from all the articles in the dump. The corpus thus created consists of about 2.2 million words from approximately 200,000 sentences.

Since the texts used in building the corpus were all encyclopedia articles featured in the Tamil language version of Wikipedia, the corpus covers a wide variety of topics including arts, culture, biographies, geography, society, history, etc., written and edited by volunteers from around the world.

We now describe our observations in annotating this Wikipedia corpus using the PDTB methodology in the following sections.

\(^3\) http://static.wikipedia.org/
6.2 Results

We looked at 511 sentences from the corpus and annotated a total of 323 connectives [8]. Table 6.1 shows the distribution of the annotated connectives across the different types such as Explicit, Implicit, EntRel, AltLex and NoRel.

<table>
<thead>
<tr>
<th>Connective Type</th>
<th>Count (unique)</th>
<th>Count (%)</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit</td>
<td>269</td>
<td>96</td>
<td>83.3</td>
</tr>
<tr>
<td>Implicit</td>
<td>28</td>
<td>16</td>
<td>8.6</td>
</tr>
<tr>
<td>EntRel</td>
<td>16</td>
<td>-</td>
<td>5.0</td>
</tr>
<tr>
<td>AltLex</td>
<td>8</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>NoRel</td>
<td>2</td>
<td>-</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Table 6.1 Results of discourse annotation experiment

While a higher percentage of the connectives annotated are those of the Explicit type, it can also be seen that there is a higher proportion of unique connectives in the Implicit and AltLex types. Note that since EntRel and NoRel connectives are not associated with a sense relation or a lexical item, their counts are left blank.
6.3 Combined connectives

There is a paired connective -um ... -um (...) that sometimes expresses an Expansion: Conjunction relation between the events where each -um is suffixed to the verb that describes each event. Also, there is a connective -Al which usually never occurs more than once and sometimes expresses a Contingency: Cause relation between two events.

It is interesting to see that in sentences like (5.2.1) reprinted below, the -Al combines with the -um ... -um to express something like a new type of relation. In the process, the -um ... -um causes the -Al, which is usually not doubled, to become doubled, thereby forming an -Alum ... -Alum. We call this special type of connectives as combined connectives.

![Combined Connectives Diagram](image)

**Figure 6.1** Combined connectives

{kirikket viLayADiyad}Alum {uDarpayirci}
cricket play.ADJ-that.PST.CAUSE=and body-training

seidad\textit{Alum} \[\texttt{sOrvaDaindEN}].

do.ADJ-that.PST.CAUSE=and tiredness-get.PST

‘Because \{I played cricket\} and \textit{because} \{I did exercise\} \{I am tired\}.’

(5.2.1)
6.4 Redundant connectives

The connective -O ... -O (...) that conveys a dubitative relation also combines with the -Al connective in a way similar to what was shown in Section 5.1 to form the combined connective -AlO ... -AlO (...).

However, in example (6.4.1), alladu, an equivalent of the -O ... -O connective has also occurred in addition to the combined -AlO ... -AlO connective.

![Figure 6.2 Redundant connectives](image)

{pOtti samappatt}AlO alladu {muDivu perAmal

contest equal-get.CAUSE=or or end get.NEG.COND.

pON}AlO [piNvarum muraigal mUlam aNigaL

go.CAUSE=or following rule.PL using team.PL

tarappaDuttapaDum].
‘If {a game is tied} or if {there is no result}, [the qualified teams are chosen using the following rules].’

(6.4.1)

Although we have termed these connectives as redundant, it is actually unclear whether these may be purely redundant, or could serve a purpose to emphasize the dubitative relation expressed by both alladu and -O ... -O.
6.5 Ambiguous Connectives

It is interesting to note that some connectives have multiple senses. In (6.5.1) the affixed –um connective carries the sense of typeExpansion: Conjunction ‘also’ whereas in (6.5.2) the same affix carries the sense of the subtypeContingency: Concession ‘however’.

{(idam)Mulan avar orunAL pOttiyil oNbadAyiram OttanLaI
this-using he.HON one-day contest.LOC nine-thousand run.PL.ACC
kaDanda pattAvadu vIrar eNra perumaiyai pettrAr}. [inda
cross.ADJ ten.ADJ player that greatness.ACC get.PST.HON this
OttanLaI kaDanda mudal teNNAppirikka vIrar eNra
run.PL.ACC cross.ADJ first south-africa.ADJ player that
sAdaNaiyaiｙum [nigaztiNAr].

achievement=and happen.PST.HON

‘(By this, he became the tenth player to cross nine thousand runs in one-day internationals). [He] also [attained the record of becoming the first South African player to cross these many runs].’
Though {two more teams participated when compared to last Cricket World Cup}, [the total matches played during this time were fewer].

(6.5.2)
Chapter 7

**Discussion and Future Work**

During the course of our annotation process, in addition the observations detailed in Chapter 6, we have come across some interesting phenomena that we feel raise important questions regarding the nature of discourse in the Tamil language.

We also propose a line of further inquiry into this our discourse analysis framework that may help provide valuable insight into whether the discourse information identified at higher levels is indeed an extension of the information available at the sentence and probably word level.
7.1 Are all Tamil discourse connectives suffixes?

In one of our initial experiments, we have observed that Tamil contains discourse connectives as both isolated word-forms and suffixes that are agglutinated to other base-forms. We now examine a possibility that all discourse connectives are originally suffixes, i.e., that even the isolated connectives can be split into a base and a suffix.

To expand on this line of inquiry, we can consider these isolated connectives as *fused forms* in the language, i.e., words that were originally constructed by joining two or more words, but over the course of time have become into a single word that cannot be further split into component morphemes. For example, words like *herewith, therefore*, etc., can no longer be analyzed into *here*+*with* and *there*+*fore* anymore.

The results of the pilot experiment (see Section 4.1) showing the most frequent isolated connectives is now partially reprinted below with possible morphological analyses that we have speculated to correspond to each of them.

<table>
<thead>
<tr>
<th>Tamil</th>
<th>Translit.</th>
<th>Meaning</th>
<th>Analysis</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>எனவே</td>
<td>enavE</td>
<td>‘so’</td>
<td>*ena + E</td>
<td>*‘thus’ + ‘(emph.)’</td>
</tr>
<tr>
<td>ஆனால்</td>
<td>AnAl</td>
<td>‘but’</td>
<td>*Ana + Al</td>
<td>*‘made’ + ‘and’</td>
</tr>
<tr>
<td>மற்றும்</td>
<td>mattRum</td>
<td>‘and’</td>
<td>*maRu + um</td>
<td>*‘other’ + ‘and’</td>
</tr>
<tr>
<td>அல்லது</td>
<td>alladu</td>
<td>‘or’</td>
<td>*alla + adu</td>
<td>*‘or’ + ‘that’</td>
</tr>
</tbody>
</table>
Table 7.1 Possible morphological analysis of isolated connectives

<table>
<thead>
<tr>
<th>சொற்கள்</th>
<th>enRALum</th>
<th>‘though’</th>
<th>*enRu + Al + um</th>
<th>*‘that’ + ‘cause’ + ‘and’</th>
</tr>
</thead>
<tbody>
<tr>
<td>குறிப்பிட்டு</td>
<td>ceydu</td>
<td>‘having-done’</td>
<td>ceydu</td>
<td>*‘do’ (conjunctive participle)</td>
</tr>
</tbody>
</table>

Similar to the English examples given above, the isolated word matRum (see Table 7.1) can be seen as a frozen form which could originally be separated into an anaphoric maRu (meaning other) and the conjunctive suffix -um, thereby providing an explanation for the existence of both the forms.

A historical study on such a possibility needs to be performed on perhaps larger datasets from different periods of time will shed some light into the validity of our claims, and pave the way for better understanding of the encoding of discourse information in agglutinative languages.
7.2 Extending Dependency relations to Discourse

The approaches outlined in Chapter 2 show attempts to provide an explanation for discourse relations that hold beyond the sentence level. Similarly, there are also theories such as the Panini’s Karaka Framework [13], Theta/Thematic Roles [14], and other dependency grammars [15] (for example) that variously describe the syntactic and semantic relationships that exist between elements within the sentence level.

An example of the different levels of analyzing text is shown in Figure 7.1. Thus, in an abstract sense, the discourse relations can be seen as a ‘zoomed-out’ view of dependency relations, with the former applicable above a sentence level using *discourse markers* and the latter within it employing the use of *case markers*. Also, it has been analyzed in the English language [6] that these levels are well stratified in the sense that all the relationships occurred within those levels and elements at a
sentence level did not contribute much at higher levels and hence discourse structure was less complex than structures at lower levels.

However, we have observed in the course of our work in Tamil (see Chapter 4) that discourse connectives were rarely occurring as separate word forms having clear discourse usage. As a result, elements which occurred as suffixes provided information not only at the morphological and dependency levels, but had certain functions at the discourse level as well.

இவ்வணியின் பாகிஸ்தான் வெற்றியால் அடுத்த நிலவுக்கு அதிக ஹ்ஹர்ஹக்கற்ற

{ivvaNiyiN pAkistAnukku edirANa vettriyAl} [aDuththa this-team.POSS pakistan.DAT against victory.INST/.CAUS next nilaikku thErchi pettradu].

stage.DAT qualification get-that.PST

‘Because of {this team’s victory against Pakistan}, [it got promoted to the next stage].’

(7.2.1)

In (7.2.1), we see that the morphological gloss of the verb vettriyAl, we see that the suffix –Al works both as an instrumental case marker (at dependency level) and as a connective showing cause relation (at a discourse level). Whether this is a simple case of homophony between the instrumental marker in dependency and the cause marker in discourse, or if there are markers that show meaning at different levels needs to be explored, and if it is indeed the case, their behavior studied and analyzed.
As a result, we feel that there might exist many ways in which the information at the dependency level could be useful in identifying, disambiguating, or enriching the higher level discourse relations.
Bibliography