Semi-automatic Approach for Tamil Discourse Relation Annotation

Frances Yung, Enosh Peter Ponraj, and Vera Demberg
Saarland University, Saarland Informatics Campus, Germany
{frances,vera}@coli.uni-saarland.de
enpe00001@stud.uni-saarland.de

Abstract

Discourse relations (DRs) specify the logical relations between text spans and are essential for modeling extended discourse. Resources annotated with DRs can help train large language models (LLMs) to recognize and generate these relations more naturally. However, there is currently no open-source DR-annotated resource for Tamil. Annotation is particularly challenging because many Tamil discourse connectives are realized as morphologically complex suffixes rather than standalone tokens, often involving phonological alternations. In this work, we present a DR-annotated dataset for Tamil based on the PDTB framework. We adopt a semi-automatic pipeline: 1) projection of automatic English discourse annotations onto Tamil in a parallel corpus; 2) lexical normalization using a morphological analyzer; and 3) manual verification of each instance. The resulting resource contains approximately $7{,}200$ explicit DR annotations and a lexicon of 450 Tamil discourse connectives. The annotated data is available for download at https://github.com/Enosh-P/Tamil-Semi-Automatic-Discourse-Relation-Dataset/

Keywords: Tamil, discourse relations, discourse annotation, PDTB, language resources

1. Introduction

Discourse refers to a coherent and structured set of sentences, such as those found in conversations and written or spoken texts, expressed in natural language. For a discourse to be meaningful and coherent, its segments should be connected through logical relations, such as cause-effect, elaboration, contrast and exemplification. These connections are referred to as coherence relations or discourse relations (DRs; Hobbs, 1978). DRs provide the structural backbone of discourse, allowing readers to interpret the overall communicative goal by connecting individual units of texts (Sanders and Spooren, 2011). Incorporating knowledge of DRs has been shown to enhance the reasoning capabilities of large language models (LLMs), improving their performance in tasks such as text generation (Guan et al., 2021; Liu et al., 2026), summarization (Xu et al., 2020; Liu and Demberg, 2024), sentiment analysis (Choi et al., 2016) and question-answering (Verberne et al., 2007; Sovrano et al., 2025).

The largest DR-annotated resources are in English, including the Penn Discourse Treebank (PDTB; Prasad et al., 2008; Webber et al., 2019) and the RST-Discourse Treebank (Carlson et al., 2003). Beyond English, discourse resources have been developed for a growing number of languages (Stede, 2004; da Cunha et al., 2011; Zhou and Xue, 2012; Synková et al., 2024). For Indian languages, notable efforts include the Hindi Discourse Relation Bank (Oza et al., 2009), the Bangla RST Discourse Treebank (Das and Stede, 2018), and other resources for Hindi, Malayalam, and Tamil created for corpus studies or model training (Rachakonda and Sharma, 2011; Gopalan et al., 2017; Sheeja S and Lalitha Devi, 2022). However, none of Indian language discourse resources are publicly accessible.

In this work, we present a DR-annotated resource for Tamil, developed following the PDTB framework. In this framework, DR annotations are anchored to explicit discourse signals, known as discourse connectives (DCs), which lexically signal the underlying relations. Applying this annotation scheme to Tamil is particularly challenging because, unlike many other languages where DCs are typically isolated tokens, Tamil often realizes DCs as morphological suffixes attached to other content words, as shown in the example below.

Example (1)

Tamil: மழை பெய்ததால் **அவன் வீட்டில் இருந்தான்**
malzhai peithatha-aal **avan veetil irunthaan**
gloss: *rain fall-PST-CAUS* 3SG.M house-LOC be-PST-3SG.M

English translation: Because **it rained**, *he stayed home.*

DR sense: CONTINGENCY.CAUSAL.REASON

Here, the DC *"because"* in English marks the DR REASON between the clauses *"he stayed home"* (called the *Argument1*) and *"it rained"* (called the *Argument2*. On the other hand, in Tamil, instead of an isolated word, the relation is marked by the suffix *"-aal"*, which is part of the word *"peithathaaal"*. Nonetheless, not all DCs in Tamil are suffixes. There are also single-token DCs, such as ஆனால் (*aana*, but) and அதேபோல் (*athepol*, similarly). While suffixal DCs also occur in the

Proceedings of WILDRE-8, pages 14–24
12 May 2026.

Turkish language and are included in the lexicon of Turkish DCs (Zeyrek and Başıbüyük, 2019), they are were not annotated in the Turkish Discourse Bank (Zeyrek et al., 2010; Zeyrek and Kurfalı, 2017). To our knowledge, the current resource is the first effort to annotate DCs that are not individual tokens.

To create a DR-annotated resource for Tamil, we propose using **annotation projection**, in which automatic DR annotations are projected from English to Tamil via a parallel corpus. DC candidates together with their arguments are first extracted from the Tamil texts based on their alignment with the English words. To identify connective suffixes, a **morphological analyzer** is applied to the extracted Tamil DC candidates. Finally, an extended series of **manual verification and normalization steps** is carried out to ensure the quality and consistency of the annotations.

Our annotation pipeline results in a corpus of 7200 explicit DR annotation, from which we also derive a lexicon of 450 Tamil DCs, including The statistics of our corpus reveal that suffixal DCs are as frequent and ambiguous as free-token DCs in Tamil, showing that it is necessary to specifically model both types of DCs. This highlights the significance of our resource for training and evaluating models of Tamil discourse processing.

2. Related Work

2.1. Tamil Language

Tamil is a Dravidian language spoken by approximately 78 million people and is one of the oldest living languages in the world today. Its script is classified as an Abugida, a writing system that lies between an alphabet and a syllabary (Sarveswaran, 2024), comprising a total of 247 distinct characters. Tamil is an agglutinative language with a complex morphological system in which grammatical and semantic information is expressed through suffixation on root nouns, verbs, adjectives and adverbs (Caldwell, 1875; Sarveswaran, 2024). In particular, Tamil DCs are often realized as verbal suffixes, though they may also appear as nominal derivatives and discourse particles.

Tamil is considered a low-resource language due to the limited availability of annotated resources. Most related to the current work, Rao et al. (2011) introduce a manually annotated Tamil corpus of 8500 sentences (352 DRs; 13 unique DCs) focusing exclusively on CAUSAL relations. Using Conditional Random Fields (CRFs) to trained on this data, they reported that identifying causal marker spans is particularly challenging, as the position of the DCs varies due to Tamil's relatively free word order. Rachakonda and Sharma (2011) extend the annotation to other DR types, resulting in a corpus of 511 sentences with 323 DRs (of which 269 are explicit) with 96 unique DCs. Finally, Gopalan et al. (2017) conducted a corpus-based study of cross-linguistic variations across Hindi, Tamil, and Malayalam based on DR annotations of the three languages, including 1341 explicit DRs in Tamil.

Unfortunately, the manual annotation efforts described above are not publicly available. We propose a more scalable approach to high-quality DR annotation for Tamil that combines automatic preprocessing and manual verification. The resulting data is released as an open resource.

2.2. PDTB framework for discourse annotation

In this work, we decided to annotate DR in Tamil based on the PDTB framework because we also want to identify the explicit connectives that trigger the annotated DR in order to construct a lexicon of Tamil DCs. In this framework, each explicit DR annotation specifies the DC, the spans of the two arguments it connects, and a sense label, as seen in Example (1). *Arg2* (in *italics*) is the clause to which the DC is syntactically attached, and corresponds to the label name of the relation, i.e. *"it rained"* is the REASON. The other argument is defined as *Arg1* (in **bold**).

The sense labels in PDTB are arranged in a three-level hierarchy (28 Level-3, 22 Level-2 labels, and 4 Level-1 labels in PDTB3.0). Since DR is an ambiguous phenomenon that even human often disagree on (Sanders et al., 1992; Spooren and Degand, 2010; Das et al., 2017; Zikánová et al., 2025; Hewett and Stede, 2025), existing DR identification tasks typically models up to the granularity of the Level-2 labels (Knaebel, 2021; Braud et al., 2025). Following this, we the current resource is also labelled with Level-2 labels.

2.3. Annotation projection

Since DR resources exist for several languages, a number of previous studies have explored annotation projection from a resource-rich language, typically English, to low-resource languages (Versley, 2010; Laali and Kosseim, 2017; Sluyter-Gäthje et al., 2020; Yung et al., 2023; Bourgonje and Lin, 2024). In particular, Bourgonje and Lin (2024) combine machine translation with word alignment, allowing English DR annotations, automatically produced by PDTB-trained parsers, to be projected to a wide range of target languages.

In contrast, our work projects discourse annotations onto human-translated Tamil text in a parallel corpus, rather than relying on machine-translated output. This choice is motivated by the poor performance of current English–Tamil machine transla-

tion systems (BLEU score of 4.35; Ramesh et al., 2020).

A key limitation of word-alignment–based annotation projection is that alignments are typically available only at the word level, whereas Tamil DCs often occur as suffixes. To address this, unlike prior approaches that rely on projection for direct annotation, we employ annotation projection only as a bootstrapping step to identify potential DC candidates. To accurately identify and extract subword-level DCs, each candidate is subsequently manually analyzed and verified with the help of a Tamil morphological parser (Sarveswaran et al., 2018). The complete processing pipeline is described in Section 4.

3. Adapting PDTB scheme for Tamil

We aim to create a Tamil discourse resource following the PDTB framework. In this contribution, we annotate the spans of the explicit DC tokens or suffixes, together with their arguments and sense labels.

In the PDTB, the DCs that are annotated include conjunctions and discourse adverbials, which are single or multiple tokens. In contrast, for Tamil, we annotate free-standing conjunctions and also suffixal connectives that are attached to verbs, normalized verbs or nouns. Multi-word connectives (as in Example 2), and multi-span connectives are also annotated, but we found only a few cases in our corpus.

Example (2)

Tamil: அவர்கள் தருவதை ஏற்றுக் கொண்டு அடங்கிப்போய் ஊக்குவிக்கின்றனர், **அதன் மூலம்** சமூக கொந்தளிப்புகளை தடுப்பதற்கு உதவுகிறார்கள்.
Avarkaḷ taruvatai ettruk koṇdu aṭaṅkippōy ūkkuvikki a ar, **ata mūlam** camūka kontaḷippukaḷai taṭuppata ku utavuki ārkaḷ.
gloss: *they give*-ACC *accept*-CVB *take*-CVB *submit*-CVB *encourage*-PRS.3PL ***through that*** *social unrest*-PL.ACC *prevent*-INF.DAT *help*-PRS-3PL

English translation: Encourage the passive acceptance of handouts, and **thereby** help prevent social explosions.

DR sense: CONTINGENCY.CAUSAL

PDTB also annotates implicit DRs, but this step requires a list of explicit discourse connectives, and is usually performed by inserting a connective from the list in between two adjacent sentences.

This is problematic in the case of Tamil, as a comprehensive lexicon of Tamil DCs is not yet available, making the choice of DCs for implicit DR annotation unclear. Second, the insertion of DCs is particularly challenging in Tamil and less intuitive than in English due to its complex inflectional morphology. Lastly, the available Tamil–English parallel corpora do not consist of continuous texts, whereas implicit DRs are usually annotated between adjacent sentences. These practical constraints make the current approach unsuitable for implicit DR annotation, and we hence focus on only annotating explicit relations.

As described in Section 2.2, arguments in the PDTB framework are labeled *Argument 1* and *Argument 2* depending on the syntactic attachment of the connective, while the DC span is annotated separately. In our Tamil corpus, we represent argument and DC spans using the notation adopted in the DISRPT shared task (Zeldes et al., 2021). This representation is more flexible and facilitates comparison across discourse frameworks, while preserving all information encoded in the original PDTB format.

Specifically, *Argument 1* and *Argument 2* are ordered according to their linear position in the text, with *Argument 1* preceding *Argument 2*. An additional tag, either *1<2* or *1>2*, indicates whether the connective is syntactically associated with *Argument 1* or *Argument 2*. The DC span is included within the argument span in its original position. This notation is particularly suitable as the Tamil suffixal DCs cannot be detached from their host tokens without rendering the text ungrammatical.

4. Methodology

4.1. Data

We chose the Tamil Samanantar Dataset (Ramesh et al., 2022) as the parallel corpus for annotation projection. The corpus contains a total of 5 million web-crawled sentence pairs taken from news, education and science domains We make use of the first $200,000$ sentence pairs of the corpus.

4.2. Annotation projection from English

Our workflow for annotating DRs on the Tamil text in the parallel corpus consists of three steps. We first apply the Discopy discourse parser (Knaebel, 2021) to obtain discourse annotations on the English side of the parallel corpus. Although the parser identifies both explicit and implicit relations, we retain only explicit relations. Sentence pairs that do not contain an explicit DC on the English side are excluded[1]. We also discard cross-sentence relations, as the order of the sentences fed to the parser does not

[1] Even though the corresponding Tamil sentence may contain an explicit DC due to *explicitation* in translation.

correspond to their original discourse order[2]. After this filtering, we obtain 37,819 sentence pairs.

We then apply AWESoME-align (Dou and Neubig, 2021) to the screened sentence pairs to compute token-level alignments between English and Tamil. Using the token spans of DCs and arguments on the English side, we extract the corresponding aligned tokens on the Tamil side. The DR labels are directly projected as the annotations for the Tamil DRs. As described in Section 3, the argument whose span begins earlier in the text is labeled *Arg1*.

Among the aligned sentence pairs, only 8,225 English DCs are aligned, sometimes jointly with other English tokens, to tokens in the corresponding Tamil sentences. This suggests that in many cases, the English DC is not translated as an explicit standalone token in Tamil; but this may also result from alignment errors.[3] We focus on the aligned DC tokens to identify explicit DRs on the Tamil side. The candidate alignments that include an English DC correspond to 3,322 unique Tamil word forms. Our next step is to identify Tamil DCs, both standalone tokens and suffixes, by separating and removing the content-bearing word segments.

4.3. Suffix-level analysis

We apply the ThamizhiFST morphological analyzer (Sarveswaran et al., 2018) to the candidate alignments to separate the suffixes from the main word stems. The alignments are then grouped on the Tamil side based on the identified suffix (if any), and each alignment is manually inspected to specify valid DC-to-DC alignments, where the Tamil DC may be realized either as a token or as a suffix, while the English DC may consist of one or multiple tokens. For example, based on the alignments shown below, it can be inferred that the Tamil DC corresponding to the English DC *"when"* is the suffix -போது (bothu).

- தேவைப்படும்**போது** *thevaipadum**bothu*** (when needed)
- சென்ற**போது** *sendra**bothu*** (when ... went)
- நகர்த்தும்**போது** *nagarthum**bothu*** (when moved)

The manual alignment post-editing is carried out by one of the authors, a native Tamil speaker. During this process, morphological segmentation errors (over-/under segmentation) and word-alignment errors in the candidate alignments are corrected simultaneously. Specifically, when a candidate alignment is incorrect, i.e., when the discourse sense expressed by the English DC is not conveyed by any part of the aligned Tamil tokens, the original sentence pair is examined to identify alternative Tamil expressions that may realize the labeled DR. If such expressions are found, the alignment is updated accordingly (as in Example (3)). If not, indicating that the DR is implicit or paraphrased in Tamil, the DR instance is discarded (as in Example (4)).

Example (3)

Tamil: அந்த விளையாட்டு ஆபத்தானது, ஏனென்றால் திருட்டும் கொலையும் சர்வசாதாரணமானது என காண்பிக்கிறது.
antha vilaiyaattu aabathaanathu, **yenendraal** thiruttu kollaium sarvasaathaaranamaanathu ena kaanbikkapadukirathu.
gloss: *that game* dangerous-PRS, ***because*** theft-AND murder-AND ordinary-PRS *be* show-PRS.

English translation: The game is considered dangerous **because** it trivializes robbery and murder.

DR sense: CONTINGENCY.CAUSE

Example (4)

Tamil: அங்கே ஒரு மளிகை கடையை திறந்து முழு குடும்பமாக **மாறிமாறி** அதை கவனித்துக்கொண்டோம்.
Ange oru malligai kadai thiranthu mulu kudumbamum **maarimaari** athai kavanithukondanar.
gloss: *There one grocery store*-ACC *open*-CVB *whole family*-ADV *alternatively*-ADV *it take-care*-1PL-NOM

English translation: We opened a grocery store there **and** our whole family took turns working in it.

DR sense: EXPANSION.CONJUNCTION

In Example (3), "because" is incorretly aligned to ஆபத்தானது *aabathaanathu* (dangerous) and is therefore revised manually to ஏனென்றால் *yenendraal* (because). In Example (4), மாறிமாறி *MaariMaari* (again and again) is aligned with "and". However, in this context, மாறிமாறி *MaariMaari*, it should be aligned to "took turns". "And" is therefore inferred implicitly in the Tamil sentence and is not included in the dataset.

After this suffix-level alignment post-editing and verification, a total of 7223 explicit DRs and 1702 unique Tamil DC candidates are collected, but 1260 of these only occur once in the corpus. This is not surprising given the highly inflectional nature

[2]We nevertheless retain the predicted DC spans and project them to Tamil, while labeling the arguments and senses as *unknown*. This part of the data can still be used for DC identification

[3]James and Krishnamurthy (2025) report an AER of 68.2 using Awesome Align for English-Tamil word alignment.

of Tamil. Many of these candidates correspond to different surface forms of the same token or suffix. Therefore, additional normalization is required to extract a list of unique Tamil DCs.

4.4. Manual normalization

As a low-resource language, Tamil has limited preprocessing tools available. Consequently, the normalization of DC candidates is performed largely manually by the same author, with support from string-matching techniques. The main goal of the normalization process is to identify and group various variants of the same DC into a standardized canonical entry.

Phonetic Variation Some Tamil DCs exhibit variation in pronunciation that may also be reflected in their written form, diverging from the standard spelling. These forms therefore represent orthographic variants of the same discourse connective. In Example (5), ஆனா *aana* is an informal spoken variant of ஆனால் *aanaal*, reflected in the phonetical spelling. Both forms correspond to the same connective, which is translated as *but* in English. Here are some more examples of phonetic variants:

- "similarly"
  standard: அதேபோல் *athepol*
  informal: அதேப்போல *atheepola*
  phonetic variation: அதபோல *athapola*
- "but"
  standard: ஆனால் *aanaal*
  phonetic variation: ஆனா *aana*

Example (5)

Tamil: இது மோசமா தெரியலாம் **ஆனா** அந்த கதவுகள்... முழுவதும் மாற்றப்போகுது... அதன் கீழ்பகுதிலயிருந்து தோட்டம் வரை
Ithu mosama theriyalaam **aana** antha kadhavugal.. muluvathum maatrapokuthu... athan keelpaguthiilirrunthu thoodam varai
gloss: *this bad*-NOM *may-look* ***but*** *those door*-PL *whole change-go*-FUT *its down-part-from*-CVB *garden untill*
English translation: I know this looks bad, **but** those patio doors are going to completely revolutionise the flow from their downstairs to their garden
DR sense: COMPARISON.CONCESSION

Since variants of the same DC typically share a common prefix, we inspect the list of DC candidates sorted alphabetically and group variants accordingly. Based on these groupings, we construct a mapping from surface variants to standard DC forms. In total, 90 candidates are identified as variants and grouped into 44 standard DCs, with each standard DC having 1 to 5 variants. Using this mapping, all DC variants (1428 out of the 7233) in the corpus are assigned a normalized standard DC on top of the raw form.

Sandhi consonant linking We also account for orthographic variants arising from sandhi consonant insertion. These are phonological alternations at word boundaries that commonly involve the insertion of a consonant to ease pronunciation when a vowel-final word is followed by a vowel-initial word. Such additional consonants could also be suffixed to a DC, as in அதற்கு**ப்** *adharkku* (therefore), இவ்விதமாய்**த்** *ivvidhamaayi* (in this manner), and உதாரணமாக**க்** *udaaranamaaka*(for example). The last consonant of these DCs are the linking consonants.

Emphasis suffix Another type of variation concerns suffixal DCs that appear in emphasized forms, as shown in the examples below.

- காட்டு *kaatu* (show)
- காட்டு**வதற்கு** *kaatuvatharkku* (in order to show)
- காட்டு**வதற்கே** *kaatuvatharkke* (in order to show; with an emphasis marker)

Since both linking consonants and emphasis suffixes are governed by strict morphological rules, their normalization is straightforward. We generate the corresponding consonant-linking and emphatic forms of each DC as variants and group all such DC candidates under a single canonical entry.

Change in DR sense in translation On top of variants identification, we also manually verify the projected DRs to identify any sense mismatches. The DR sense predicted by the discourse parser on the English texts could be incorrect. [4] Also, DRs are sometimes explicitated or implicitated (Meyer and Webber, 2013; Lapshinova-Koltunski and Carl, 2022; Yung et al., 2023), which means that a DC can be inserted or omitted, or translated to a more/-less specific connective. Our focus is not the meaning shift of the DR translation, but rather whether the projected English DR sense is valid for the Tamil DC in the translation. To verify this, each DR sense projected from English is inspected against the corresponding Tamil connective in the collected lexicon. When the projected sense is intuitively incompatible with the meaning of the Tamil DC, the original sentence pair in the corpus is examined and the sense label is revised accordingly.

In Example (6), the sense of the English DC, CONJUNCTION, is projected to the Tamil DC ஆனால்

[4] The reported accuracy of the Discopy parser on PDTB explicit relations is 78% F1 (Knaebel, 2021).

Sense	unique DC counts word-type	unique DC counts suffix-type	corpus frequency word-type	corpus frequency suffix-type	Top-3 Tamil DCs	Coverage
Expansion.Conjunction	206	91	1072	1093	-உம் (491), மற்றும் (245), மேலும் (236)	45.3%
Comparison.Contrast	93	47	1152	334	ஆனால் (794), -உம் (124), எனினும் (57)	65.6%
Temporal.Asynchronous	116	45	777	373	பின்னர் (179), பிறகு (76), -இல் (66)	27.9%
Contingency.Cause	113	59	615	329	-ஆல் (151), எனவே (90), ஆகவே (70)	33.0%
Temporal.Synchrony	88	48	245	474	-போது (174), -இல் (85), போது (76)	46.7%
Contingency.Condition	56	42	130	370	-ஆல் (190), -ஆனால் (28), -இருந்தால் (18)	47.3%
Expansion.Alternative	13	17	90	30	அல்லது (75), -ஆவிட்டால் (6), -ஆல் (4)	70.8%
Comparison.Concession	19	19	27	59	-உம் (17), -ஆலும் (9), -கூட (5)	36.0%
Expansion.Instantiation	4	1	19	1	உதாரணமாக (14), இதுபோன்று (3), உதாரணமாகும் (1)	90.0%
Contingency.Purpose	1	1	1	18	-காக (18), இருக்க (1)	100.0%
Comparison.Similarity	3	1	11	1	அதேபோல் (5), இதேபோல (4), போன்று (2)	91.7%
Expansion.Equivalence	2	1	11	1	அதாவது (10), என்றால் (1), -ஆக (1)	100.0%
Total	**327**	**127**	**4150**	**3083**	—	—

Table 1: DR sense distribution across unique Tamil DC counts and corpus frequency.

anaal (but). However, this is actually an *explicitation* of the DR; the relation is more explicitly expressed with a DC specifically for CONTRAST or CONCESSION. Since the assigned sense does not match the Tamil DC intuitively, the corresponding corpus samples are inspected and the correct sense, CONCESSION, is assigned.

Example (6)

Tamil: (இவ்வாறு) அவர்கள் சூழ்ச்சி செய்தார்கள். **ஆனால்** அவர்கள் அறியாதவாறு நாமும் சூழ்ச்சி செய்தோம்.
(ivvaaru) avargal soolchi seithaargal **anaal** avargal ariyaathavaaru naamum soolchi seithoom.
gloss: *(Thus) they maneuver do*-PST-3PL ***but*** *they unknowingly we-also maneuver do*-PST-1PL-INCL

English translation: So they plotted a plot, **and** we planned a plan, while they perceived not.

DR sense: EXPANSION.CONJUNCTION →CONTINGENCY.CONCESSION

In Example (7), *"so"* is aligned to வதற்காக *vata kāka* (for that), and the English discourse parser assigns it the label CONTINGENCY.CAUSE. However, the correct connective span should be *"so as to"*, which expresses a CONTINGENCY.PURPOSE relation. Accordingly, the sense label for the Tamil DC is manually revised from CAUSE to PURPOSE.

Example (7)

Tamil: இந்த கிராமங்களில் தலித் மற்றும் பொது வீடுகளுக்கு இடையிலான ஏற்றத்தாழ்வைக் குறைக்க 50 சமூக-பொருளாதார நிலை குறிகாட்டிகளை மேம்படுத்து**வதற்காக** இப்போது இது மறுவடிவமைப்பு செய்யப்பட்டுள்ளது.
Inta kirāmaṅkaḷil talit ma um potu vīṭukaḷukku iṭaiyilā a ē attā vaik ku aikka 50 camūka-poruḷātāra nilai ku ikāṭṭikaḷai mēmpaṭuttu**vata kāka** ippōtu itu ma uvaṭivamaippu ceyyappaṭṭuḷḷatu.
gloss: *This village*-PL-LOC *Dalit and general house*-PL-DAT *between inequality*-ACC *reduce*-INF 50 *socio-economic status indicator*PL *improve*-**INF-PURPOSE** *now this redesign do*-PASS-*pst*

English translation: This has now been redesigned to include 50 socio-economic indicators that have to be improved **so as to** reduce the inequality between Dalit and general households in these villages.

DR sense: CONTINGENCY.CAUSE →PURPOSE

During this DR sense verification process, we additionally identify four sense labels that are not identified by the parser on the English side and therefore are never projected automatically. These labels are incorporated into the Tamil DC lexicon for the corresponding DCs: EXPANSION.INSTANTIATION, EXPANSION.EQUIVALENCE, CONTINGENCY.PURPOSE, and COMPARISON.SIMILARITY.

5. Results

After all verification and normalization steps, the final dataset contains 7233 explicit DR annotations involving 454 unique Tamil DCs. Our **Tamil Discourse Relation Bank** is freely downloadable from https://anonymous.4open.science/r/Tamil-Semi-Automatic-Discourse-Relation and the data structure is shown in Listing 1.

Listing 1: Tamil Discourse Relation Bank data structure

```
"tamil_ann": {
    "line": "string",
    "arg1": {
      "raw": "string",
      "span": [int, int, ...]},
    "arg2": {
      "raw": "string",
      "span": [int, int, ...]},
    "rel_direction": "1>2 or 2<1",
    "connective": {
      "raw_text": ["string", "string"],
      "canonical": ["string", "string"],
      "type": "word/suffix",
      "morphology": {
          "stem": "string",
          "suffix": "string"},
      "span": [
        [int, int, ...],
        [int, int, ...]]}
    "relation": {
      "type": "Explicit",
      "sense": "Contingency.Cause"}
}
```

The sense distributions of both the annotated DRs and the unique DCs are presented in Table 1. We observe that suffixal DCs are prevalent in Tamil discourse, accounting for 25% unique DC types and 40% of the DC occurrences in the corpus. Tamil DCs are also highly ambiguous. To quantify the ambiguity, , we compute the normalized entropy of the sense-label distribution for each unique DC. A DC that is consistently annotated with a single sense label has an entropy of 0, indicating minimal ambiguity. In contrast, a DC whose sense labels are evenly distributed will result in an entropy of 1, indicating maximal ambiguity and difficulty in predicting its sense.

Figure 1 presents the distribution of entropy values for word type and suffixal DCs. The results show that word-type DCs are generally less ambiguous than suffixal DCs, as most have entropy values close to 0, indicating that they are associated with a limited range of senses or display a strongly biased distribution toward a dominant sense. In contrast, more than half of the suffixal DCs exhibit higher entropy values, meaning that they can be interpreted in different ways with similar chance.

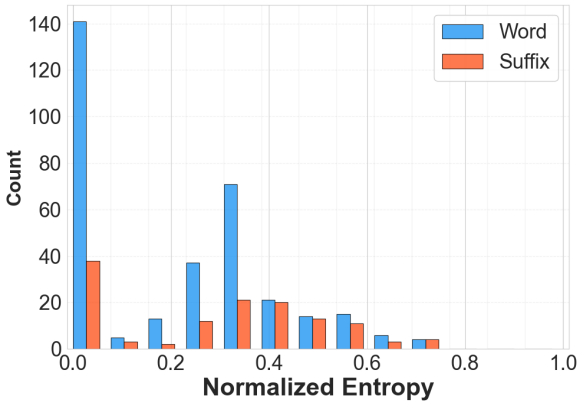

Figure 1: Distribution of sense entropy per DC

As a simple baseline, assigning each DC of the entire corpus its most frequent sense based solely on the DC lexicon yields an accuracy of 58% compared with the actual annotation. In a more realistic setting, assigning the top sense acquired from 80% of the data to the rest of the unseen data yields an accuracy of 27% only. This highlights the difficulty of explicit DR classification in Tamil. Our dataset therefore provides valuable training data for models that aim to predict the correct sense of Tamil DCs in context.

6. Conclusion

In this work, we propose a semi-automatic pipeline for creating a discourse-annotated resource for Tamil. Our approach automatically projects discourse annotations from a resource-rich language, English, and extends this projection by incorporating morphologically segmented Tamil DC suffixes, since Tamil expresses discourse mainly through verb suffixes participial constructions. To make sure these morphological elements align well with English tokens, the projection based on automatic alignment is followed by a series of manual post-editing steps. Each sample of the final 7233 explicit DRs has gone through manual verification to ensure the reliability of the DR annotations. Based on the corpus data, we also constructed the first lexicon of Tamil DCs, which are either isolated tokens or suffixes.

7. Limitation

The main limitation of the current work is its focus on intra-sentential explicit relations, due to the constraint of the parallel corpus (i.e., non-consecutive sentences). Furthermore, coherence shifts occur in manual translation (Blum-Kulka, 1986). On the one hand, the current dataset excludes explicit Tamil connectives that are implicitated or originally implicit in the English texts, as we only align the explicit connectives identified by the discourse parser. On the other hand, subtle translation nuance can unintentionally alter the coherence relations in the target language text. These limitation may affect the accuracy of the annotation projections, and lead to a biased picture of the distribution of Tamil DRs in general. As future work, we plan to develop annotation guidelines based on the DC lexicon and to conduct fully manual annotation on a subset of the current data as well as on a monolingual Tamil corpus. This manually annotated dataset will serve as an additional evaluation benchmark for Tamil shallow discourse parsing.

8. Bibliographical References

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies, 17:35.

Peter Bourgonje and Pin-Jie Lin. 2024. Projecting annotations for discourse relations: Connective identification for low-resource languages. In Proceedings of Workshop on Computational Approaches to Discourse, pages 39–49, St. Julians, Malta. Association for Computational Linguistics.

Chloé Braud, Amir Zeldes, Chuyuan Li, Yang Janet Liu, and Philippe Muller. 2025. The DISRPT 2025 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the Shared Task on Discourse Relation Parsing and Treebanking, pages 1–20, Suzhou, China. Association for Computational Linguistics.

Robert Caldwell. 1875. A comparative grammar of the Dravidian or South-Indian family of languages. Trübner.

Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. 2016. Document-level sentiment inference with social, faction, and discourse context. In Proceedings of the Meeting of the Association for Computational Linguistics, pages 333–343, Berlin, Germany. Association for Computational Linguistics.

Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Rolland Bartilotti. 2011. The RST Spanish treebank on-line interface. In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 698–703, Hissar, Bulgaria. Association for Computational Linguistics.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In Proceedings of the Workshop on Recent Advances in RST and Related Formalisms, pages 11–19, Santiago de Compostela, Spain. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, pages 2112–2128, Online. Association for Computational Linguistics.

Sindhuja Gopalan, Lakshmi S, and Sobha Lalitha Devi. 2017. Cross linguistic variations in discourse relations among Indian languages. In Proceedings of the International Conference on Natural Language Processing, pages 402–407, Kolkata, India. NLP Association of India.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, pages 6379–6393, Online. Association for Computational Linguistics.

Freya Hewett and Manfred Stede. 2025. Disagreements in analyses of rhetorical text structure: A new dataset and first analyses. In Proceedings of the Linguistic Annotation Workshop, pages 35–47, Vienna, Austria. Association for Computational Linguistics.

Jerry R Hobbs. 1978. Why is discourse coherent. Technical report.

Antony Alexander James and Parameswari Krishnamurthy. 2025. POS-aware neural approaches for word alignment in Dravidian languages. In Proceedings of the Workshop on Challenges in Processing South Asian Languages, pages

154–159, Abu Dhabi, UAE. International Committee on Computational Linguistics.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In Proceedings of the Workshop on Computational Approaches to Discourse, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Majid Laali and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In Proceedings of the International Conference Recent Advances in Natural Language Processing, pages 407–416.

Ekaterina Lapshinova-Koltunski and Michael Carl. 2022. Using translation process data to explore explicitation and implicitation through discourse connectives. In Proceedings of the Workshop on Computational Approaches to Discourse, pages 42–47, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Dongqi Liu, Hang Ding, Qiming Feng, Jian Li, Xurong Xie, Zhucun Xue, Chengjie Wang, Jiangning Zhang, and Yabiao Wang. 2026. Discorag: Discourse-aware retrieval-augmented generation. arXiv preprint arXiv:2601.04377.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In Proceedings of the Workshop on Discourse in Machine Translation, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Akshai Ramesh, Venkatesh Balavadhani parthasa, Rejwanul Haque, and Andy Way. 2020. Investigating low-resource machine translation for English-to-Tamil. In Proceedings of the Workshop on Technologies for MT of Low Resource Languages, pages 118–125, Suzhou, China. Association for Computational Linguistics.

Ted JM Sanders and Wilbert Spooren. 2011. Communicative intentions and coherence relations. In Coherence in Spoken and Written Discourse: How to create it and how to describe it., pages 235–250. John Benjamins Publishing Company.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. Discourse processes, 15(1):1–35.

Kengatharaiyer Sarveswaran. 2024. Morphology and syntax of the tamil language. arXiv preprint arXiv:2401.08367.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2018. Thamizhifst: A morphological analyser and generator for tamil verbs. In International Conference on Information Technology Research (ICITR), pages 1–6. IEEE.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In Proceedings of the Language Resources and Evaluation Conference, pages 1044–1050, Marseille, France. European Language Resources Association.

Francesco Sovrano, Monica Palmirani, Salvatore Sapienza, and Vittoria Pistone. 2025. Discolqa: zero-shot discourse-based legal question answering on european legislation. Artificial Intelligence and Law, 33(2):323–359.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. Corpus Linguistics and Linguistic Theory, 6(2):241–266.

Manfred Stede. 2004. The Potsdam commentary corpus. In Proceedings of the Workshop on Discourse Annotation, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Pavlína Synková, Jiří Mírovský, Lucie Poláková, and Magdaléna Rysová. 2024. Announcing the Prague discourse treebank 3.0. In Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pages 1270–1279, Torino, Italia. ELRA and ICCL.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval, pages 735–736.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora, pages 83–82.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 5021–5031, Online. Association for Computational Linguistics.

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Vera Demberg. 2023. Investigating explicitation of discourse connectives in translation using automatic annotations. In Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 21–30, Prague, Czechia. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Proceedings of the Shared Task on Discourse Relation Parsing and Treebanking, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek and Kezban Başıbüyük. 2019. TCL - a lexicon of Turkish discourse connectives. In Proceedings of the International Workshop on Designing Meaning Representations, pages 73–81, Florence, Italy. Association for Computational Linguistics.

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 Linguistic Annotation Workshop, pages 282–289, Uppsala, Sweden. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In Proceedings of the Linguistic Annotation Workshop, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

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 (Volume 1: Long Papers), pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.

Šárka Zikánová, Anna Nedoluzhko, Jiří Mírovskỳ, and Eva Hajičová. 2025. Gold data and multiple understanding of discourse relations. In International Conference on Text, Speech, and Dialogue, pages 250–262. Springer.

9. Language Resource References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue, pages 85–112. Springer.

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Sindhuja Gopalan, Lakshmi S, and Sobha Lalitha Devi. 2017. Cross linguistic variations in discourse relations among Indian languages. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 402–407, Kolkata, India. NLP Association of India.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. Transactions

of the Association for Computational Linguistics, 10:145–162.

Pattabhi RK Rao, Sobha Lalitha Devi, et al. 2011. Automatic identification of cause-effect relations in tamil using crfs. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 316–327. Springer.

Kumari Sheeja S and Sobha Lalitha Devi. 2022. Automatic identification of explicit connectives in Malayalam. In Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference, pages 74–79, Marseille, France. European Language Resources Association.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The PDTB 3.0 annotation manual. Philadelphia, University of Pennsylvania.