The Thai Discourse Treebank: Annotating and Classifying Thai Discourse Connectives

Ponrawee Prasertsom
University of Edinburgh, UK
p.prasertsom@sms.ed.ac.uk

Apiwat Jaroonpol
Department of Linguistics,
Chulalongkorn University,
Thailand
apiwat.jaroonpol@gmail.com

Attapol T. Rutherford*
Department of Linguistics,
Chulalongkorn University,
Thailand
attapol.t@chula.ac.th

Abstract

Discourse analysis is a highly applicable area of natural language processing. In English and other languages, resources for discourse-based tasks are widely available. Thai, however, has hitherto lacked such resources. We present the Thai Discourse Treebank, the first, large Thai corpus annotated in the style of the Penn Discourse Treebank. The resulting corpus has over 10,000 sentences and 18,000 instances of connectives in 33 different relations. We release the corpus alongside our list of 148 potentially polysemous discourse connectives with a total of 340 form-sense pairs and their classification criteria to facilitate future research. We also develop models for connective identification and classification tasks. Our best models achieve an F_1 of 0.96 in the identification task and 0.46 on the sense classification task. Our results serve as benchmarks for future models for Thai discourse tasks.

1 Introduction

Discourse analysis is a challenging area of Natural Language Processing (NLP) because the model must take into account context larger than the sentence. NLP has reached a high accuracy rate for tasks whose input is a sentence or a pair of sentences, but tasks involving larger linguistic units attract more attention from the research community as the discourse processing would enable us to improve many downstream tasks such as information extraction (Regneri and Wang, 2012; Tang et al., 2021), text complexity assessment (Davoodi and Kosseim, 2016), and automatic essay grading (Wang et al., 2018; Nadeem et al., 2019). While corpora with discourse annotation exist across languages, there is a lack of such a corpus for languages with little morphology and punctuation. These languages are particularly in need of a discourse corpus because fewer formal cues are available for automated discourse-related tasks such as argument identification. For example, thanks to its morphology, sentential adverbs that signal discourse relations in English often have the *-ly* suffix (e.g., *additionally*, *actually*). One such language, Thai, lacks both systematic descriptions of discourse elements and annotated corpora that enable supervised machine learning methods and analysis of discourse phenomena.

The Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004) proposes a lexically grounded method for discourse annotation. Each discourse relation has two arguments, which are two semantically related contiguous textual spans, and a *sense*, which categorizes the relationship. *although* is a canonical example of a discourse connective that signals a concession relation between the two arguments of the discourse relation. *afterwards* might signal a temporal relation. Under the PDTB paradigm, a discourse relation can be *explicit* (signalled by a connective) or *implicit* (inferred). The discourse analysis in this style results in a list of discourse relations. PDTB-style annotation has been adopted successfully in many languages such as English (Prasad et al., 2019), Hindi (Oza et al., 2009), Czech (Poláková et al., 2013), and Chinese (Li et al., 2014; Zhou et al., 2014).

Following previous endeavors in these languages, we present the Thai Discourse Treebank, a corpus of Thai discourse connectives annotated on 10,000 sentences from 384 documents in the style of PDTB. All documents are annotated for explicit discourse connectives, their senses, and connectives they are paired with (if any; see Section 3.2). Additionally, 180 out of 384 documents are annotated for connective arguments. We then construct a Thai discourse connective identification system trained on the corpus, which serves

*Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 613–629, 2024. https://doi.org/10.1162/tacl_a_00650
Action Editor: Asli Celikyilmaz. Submission batch: 3/2023; Revision batch: 10/2023; Published 5/2024.

as a benchmark for the task of shallow discourse parsing and as a proof of the utility of the corpus. In addition, we provide an extensive list of the connectives and the analysis of their linguistic properties.

At cursory glance, the task of annotating explicit discourse connectives may seem trivial or even amenable to automation. However, dictionary-based rules are demonstrably insufficient (see Section 5.2) because many connectives are ambiguous among different senses (cf. Webber et al., 2019). Furthermore, there is the added layer of complexity brought about by homonymy, where words that serve as connectives in certain contexts might serve other roles such as a coordinating conjunction or relativizer in other contexts. Finally, the distinctive nature of Thai orthography, which lacks the punctuation that demarcate clause or sentence boundaries, further elevates the difficulty of discourse argument identification, which the PDTB-style annotation requires explicit clausal units to perform. This absence means that manual linguistically informed annotation is essential to ensure accuracy.

Our corpus is an initiative to address these challenges. We believe that it represents the most complete study in the context of explicit discourse connectives in a low-to-medium resource language such as Thai. This focus was intentional, as we aim to study (explicit) discourse connectives as a structural element that signals coherence and discourse relations. By honing in on these connectives, we aim to contribute a dataset that not only tackles the issues highlighted but also offers insights into the discourse analysis in Thai language, which can transfer to other languages where the clausal units are not clearly marked such as Lao and Burmese.

Our main contributions can be summarized as follows:

1. We construct the first and large (10,000-sentence) Thai corpus annotated with explicit discourse relations.[1] The corpus is annotated following the English PDTB-3 scheme, with over 18,374 instances of connectives in 33 different relations (Section 4).
2. We compile the first extensive list and classification of 148 polysemous connectives with a total of 340 form-sense pairs. The list can be extended and new connectives can be classified following our criteria (Section 3).
3. We develop and train machine learning models for identifying connectives and the discourse relation senses they convey. Our best model achieves an F_1 score of 0.96 in the connective identification task, and 0.46 in the sense classification task (Section 5).

[1] https://github.com/nlp-chula/thai-discourse-treebank.

2 Related Work

2.1 Discourse Corpora

We are not aware of any other discourse corpora for Thai. Thus, the current work has introduced Thai into the pool of languages with PDTB-style corpora. Such corpora already include the PDTB for English (Miltsakaki et al., 2004; Prasad et al., 2006, 2008, 2019), the Prague Discourse Treebank for Czech (Poláková et al., 2013), the CUHK Discourse Treebank for Chinese (Zhou et al., 2014), the Chinese Discourse Treebank (Lu et al., 2014), the Hindi Discourse Relation Bank (Oza et al., 2009), the Leeds Arabic Discourse Treebank (Al-Saif and Markert, 2010), and the Turkish Discourse Bank (Zeyrek et al., 2013).

Some of the aforementioned corpora adopt a sense classification scheme that is different from the PDTB. For example, the CUHK treebank (Zhou et al., 2014) has its own set of Level-2 and Level-3 senses. The Hindi treebank also added the sense GOAL when there was no PURPOSE sense in PDTB2 (Oza et al., 2009). In this regard, the Thai Discourse Treebank is no different, adding EXPANSION.GENEXPANSION as needed (see Section 3.3). The modification is minimal, so strong compatibility with the PDTB is maintained.

These corpora also provide discourse annotation on an existing constituency or dependency treebank. However, the Thai Discourse Treebank annotates over tokenized data, as there was no sufficiently large hand-annotated treebank at the time of corpus construction. This issue is not worrying, however, because the full corpus focuses on annotating the connective senses and pairs only, which do not require structural information.

Apart from PDTB, a popular framework for discourse annotation is Rhethorical Structure Theory (RST; Mann and Thompson, 1988). RST provides discourse tree structures and has been applied to English (Carlson et al., 2003) and other

languages such as Spanish, Chinese (Li et al., 2014; Cao et al., 2018), and Russian (Toldova et al., 2017). Future work should aim to extend and align the current corpus with the LST20-based Blackboard Treebank[2] which has since become available, and annotate it using either the PDTB or RST scheme.

2.2 Connective Identification and Sense Classification

Previous work has been done on both connective identification and classification tasks for explicit connectives, but we are unaware of any work that has been done on Thai.

In English, models often rely on word-pair features (Rutherford and Xue, 2014), syntactic features such as constituency or dependency features. Early models including Naive Bayes (Pitler and Nenkova, 2009) and Maximum Entropy (Lin et al., 2014; Wang and Lan, 2015) already achieved good results ($F_1 > 0.90$) on both identification and classification tasks under different datasets derived from PDTB-2. The properties of discourse connectives themselves have also been exploited to help annotate implicit discourse relations (Rutherford and Xue, 2015).

More recent approaches adopt (deep) neural networks and embeddings. Scholman et al. (2021) directly compared two deep learning models (Nie et al., 2019; Knaebel and Stede, 2020) with two simpler models (Lin et al., 2014; Wang and Lan, 2015) across different domains on the identification task. They found Discopy (Knaebel and Stede, 2020), which uses BERT pretrained embeddings, performs best in the identification task in three out of four test datasets, including PDTB-2 ($F_1 \approx 0.95$). Other high-performing models for the identification task include ensembles that incorporate RNNs as well as automated and gold-standard syntactic features (Yu et al., 2019). Gessler et al. (2021) achieved SOTA results on the DISRPT2021 shared connective identification and sense classification tasks, which cover a wide variety of languages. Their model for the identification task is a CRF model that uses a vector derived from Transformer-based word embeddings and Bi-LSTM character embeddings. Similarly, their sense classification models are BERT-based, but also incorporate handcrafted syntactic features such as syntactic dependencies. The task is set up as a next sentence prediction task and the models take discourse argument pairs (sentences) as input.

Other languages have received relatively less attention. The CoNLL 2016 shared task is the first attempt to spark interest in the multilingual discourse parsing task for English and Chinese (Xue et al., 2016). Existing models use well-known architectures and similar features: German Random Forest connective identifiers (Bourgonje and Stede, 2018), German BERT-based sense classifiers (Bourgonje and Stede, 2020), Chinese SVM sense classifiers with lexical features such as length and POS (Huang and Chen, 2011), MaxEnt sense classifiers with syntactic features (Wang and Lan, 2016), and an end-to-end Chinese discourse parser that uses a BERT-based sense classifier (Hung et al., 2020).

3 Thai Discourse Connectives

Following the original PDTB (Miltsakaki et al., 2004), we define *discourse connectives* as *predicates that take two or more abstract objects as arguments*, where *abstract objects* refer to events, states, or propositions (Asher, 1993). These arguments are called *Arg1* and *Arg2*. As a convention, *Arg2* is identified as the argument the connective binds to. In Thai, the arguments are mostly clauses. In this section, we briefly discuss Thai discourse connectives and their properties, which inform our annotation guidelines for the Thai Discourse Treebank. Detailed information for each connective is provided alongside the corpus.

3.1 Syntax and Semantics of Discourse Connectives

As in English, Thai discourse connectives can be identified and categorized into three types according to their syntactic and discourse-semantic behaviors.

Co-ordinating Conjunctions (e.g., English *and* and Thai และ 'and') assert all their arguments. This means that all arguments of a co-ordinating conjunction participate in the sentence's truth value. A cross-linguistically applicable way to test this is to perform a negation test (Cristofaro, 2005). (1) is true only if 'Kim went to school' is false and/or 'Kim ate lunch' is false.

[2] https://bitbucket.org/kaamanita/blackboard-treebank.

(1) มัน ไม่ จริง ที่ คิม ไป โรงเรียน **และ** คิม
it not true that Kim go school and Kim
กิน ข้าวเที่ยง
eat lunch
'It is not true that [Kim went to school and Kim ate lunch].'

We treat all of the coordinating conjunctions as discourse connectives. The connectives of this type always appear at the beginning of an independent clause.

Subordinating Conjunctions (English *if* and Thai *แม้* 'even though') have one argument that do not participate in the truth value, namely, the subordinate clause. (2) is true when 'we still work' is false, regardless of whether 'we are sleepy' is false or not.

(2) มัน ไม่ จริง ที่ เรา ทำงาน **แม้** เรา
it not true that we work even though we
ง่วง
sleepy
'It is not true that [we still work even though we are sleepy]'

This negation test is crucial for our analysis because Thai does not mark sentence boundaries. A discourse connective that is a subordinating conjunction binds a clause expressing a proposition insensitive to the negation test. Additionally, discourse connectives of this type can occur either at the beginning or the end of a subordinate clause that they bind to, depending on the idiosyncrasy of each connective. Subordinating conjunctions that occur clause-finally typically occur in pairs (see Section 3.2). *แม้* 'even though' in (2) is an example of a connective that precedes the argument that it binds with syntactically. ... *ปุ๊บ* ... *ปั๊บ* 'as soon as' (3).

(3) คิม ทำงาน เสร็จ ปุ๊บ ก็ เกิด
Kim work finish **when** LINKER happen
เรื่อง ปั๊บ
incident **when**
'The incident happened as soon as Kim finished work.'

Adverbials also have an argument that does not participate in the truth value, but, unlike subordinating conjunctions, exhibit properties parallel to pronouns (Webber et al., 2003; Forbes-Riley, 2005). They can establish a relation from within an embedded clause to the matrix clause (4).

(4) [คน [ที่ เคย มี งาน]] **ต่อมา** ก็
person who once have job **then** LINKER
ตกงาน ได้
fall-job can
'[Someone [who is once employed]] can **then** become unemployed.'

They may also refer to inferred events. (5) shows that the connective refers to the inferred event of 'not being tired'.

(5) ถ้า เหนื่อย ให้ นั่ง **ไม่อย่างนั้น** ก็ วิ่ง
if tired IMP sit otherwise LINKER run
ต่อ
continue
'If tired, sit. Otherwise, continue running.'

Adverbials come in four classes: *clause-initial*, *clause-final*, *clause-peripheral* (initial/final), or *post-subject*, which is also called *second position* (Enfield, 2007). Importantly, post-subject connectives exclusively appear immediately after the subject, unlike English adverbs, which can also occur elsewhere. An example of this is *เลย* 'consequently', as in (6).

(6) ฉัน หิว พ่อ **เลย** ทำ
I hungry father **consequently** make
ข้าวผัด ให้ กิน
fried rice for eat
'I am hungry, so my father made fried rice for me.'

3.2 Pairedness

Some **subordinating conjunctions** occur in **pairs**. In such cases, one conjunction (the *head*) selects another conjunction that can occur with it (the *child*). (7) illustrates this for the conjunctions such as ... *ทีไร* ... *ทุกที* 'whenever'. In these cases, *Arg2* is the argument that does not participate in the truth value under negation.

(7) [ฝน ตก **ทีไร**$_{Arg2}$] [รถ ติด **ทุกที**$_{Arg1}$]
rain fall when car stuck always
'[Whenever it rains$_{Arg2}$], [there is traffic jam$_{Arg1}$].'

Co-ordinating conjunctions and **adverbials** do not have lexically specific pairs.

All three types can optionally be paired with ก็, a general post-subject adverbial that can signal a variety of discourse relations (Iwasaki and Ingkaphirom, 2005). ก็ strictly occurs in the argument that is a matrix clause and linearly comes later. Examples involving co-ordinating conjunctions and adverbials are given in (8) and (9).

(8) [คิม มี เงิน$_{Arg1}$] **และ** [คิม **ก็** มี
Kim have money **and** Kim LINKER have
บ้าน$_{Arg2}$]
house
'Kim has money and a house.'

(9) [คิม วิ่ง ขึ้น เขา$_{Arg1}$] **ดังนั้น** [เธอ ก็
Kim run up hill so Kim LINKER
เหนื่อย$_{Arg2}$]
tired
'Kim ran uphill, so she was tired.'

For subordinating conjunctions, it can only occur when the subordinate clause is fronted, as in (10a), but not (10b).

(10) a. [**เมื่อ** คิม อ่าน หนังสือ$_{Arg2}$] [เธอ
when Kim read book she
ก็ เข้าใจ$_{Arg1}$]
LINKER understand
'[When Kim had read the book$_{Arg2}$],
[she understood it.$_{Arg1}$]'

b. * [คิม **ก็** เข้าใจ$_{Arg1}$] [**เมื่อ** เธอ อ่าน หนังสือ$_{Arg2}$]
'When Kim had read the book, she understood it (intended).'

3.3 Senses

We largely follow the PDTB 3.0 (Prasad et al., 2019) for describing the senses of Thai connectives. Each PDTB sense is a result of three-level classification as the sense inventory is organized hierarchically. For example, CONTINGENCY.CAUSE.REASON consists of CONTINGENCY (Level-1 sense), CAUSE (Level-2 sense) and REASON (Level-3) sense, respectively.

The only deviation from the English PDTB is our new sense EXPANSION.GENEXPANSION. This describes the sense of ซึ่ง and other connectives with the same function. This sense applies when *Arg2* provides information that can only be fully comprehended with *Arg1* as prior knowledge, and is equivalent to BACKGROUND in Rhetorical Structure Theory (Mann and Thompson, 1988). Approximate English translations are the adverbials 'in this regard', or 'in regard to this', as in (11).

(11) เศรษฐกิจ มี แนวโน้ม ถดถอย
economy have tendency recess
ซึ่ง รัฐบาล ยัง ไม่ ตื่นตัว
in.this.regard government yet NEG alert
พอ
enough
'The economy is in recession. In this regard, the government is still not alert enough.'

Under the original PDTB guideline, ซึ่ง should fall under EXPANSION.CONJUNCTION. Our decision to propose a separate sense is because ซึ่ง and the quintessential Conjunction และ 'and' are hardly interchangeable in Thai. (12b) is unacceptable for many Thai speakers, and has a different meaning from (12a) even for those who accept it.

(12) a. ตำรวจ ได้ จับ คนร้าย แล้ว
police PERF arrest criminal PERF
ซึ่ง ความ คืบหน้า จะ ได้
in.this.regard NOM advance PRSP get
รายงาน ต่อไป
report next
'The police have arrested the criminal. In regard to this, further developements will be reported later.'

b. ? ตำรวจ ได้ จับ คนร้าย แล้ว และ ความ คืบหน้า จะ ได้ รายงาน ต่อไป
'The police have arrested the criminal, and further developments will be reported later (intended).'

4 The Thai Discourse Treebank

4.1 Corpus Construction

We first annotated discourse connectives on a sample of 50 documents from the LST20 corpus. (Boonkwan et al., 2020). The corpus consists of news articles in various genres from politics to science and technology and has been previously annotated with word boundaries, sentence boundaries, part-of-speech tags, and named entities. We

Thai Discourse Treebank (18374)		**PDTB3** (25406)		**CDTB** (11084)	
Expansion.Conjunction	21.28	Expansion.Conjunction	34.67	Conjunction	62.78
Contingency.Cause	15.96	Comparison.Concession	19.11	Expansion	14.47
Temporal.Asynchronous	13.17	Contingency.Cause	9.28	Temporal	6.52
Expansion.GenExpansion	12.89	Temporal.Asynchronous	8.31	Causation	5.59
Comparison.Concession	10.33	Temporal.Synchronous	7.68	Contrast	4.43
Contingency.Condition	8.86	Contingency.Condition	5.87	Purpose	2.68
Contingency.Purpose	7.86	Comparison.Contrast	5.33	Conditional	2.02
Comparison.Contrast	6.72	Expansion.Manner	2.15	Progression	1.30
Expansion.Disjunction	1.71	Contingency.Purpose	1.50	Alternative	0.20
Temporal.Synchronous	1.22	Expansion.Instantiation	1.28		

Table 1: The 10 most common level-2 connective senses (shown in percentages) in the Thai Discourse Treebank, PDTB3 (Prasad et al., 2006) and the CDTB data for the 2016 CoNLL discourse shared task (Xue et al., 2016). The numbers in parentheses are the total numbers of connectives.

retokenized the documents so that every connective counts as one token because some connectives consist of many tokens according to the LST20 annotation guidelines. To ease the workload of the annotators, we automatically annotated potential connectives for their connective status and senses by simple dictionary lookups using our list of Thai discourse connectives that is compiled by a native Thai linguist and classified according to the criteria in Section 3. As an initial phase, an annotator with expertise in Thai linguistics corrected senses and removed non-connectives, wrongly labeled by the automated process. They also tagged clauses and verb phrases as arguments for each discourse connective/relation. Largely following Prasad et al. (2019), *Arg2* is assigned to the argument attached to the connective, whereas *Arg1* is the other argument. As the final pass, they identified additional connectives not in our initial connective list. We now have a more complete list of discourse connectives for the next phase.

We sampled 334 documents more so that the total number of sentences is close to 10,000. The corpus was first pre-processed as above and pre-annotated using the discourse connective list. The documents were then assigned roughly equally to three linguistically trained annotators for each task (argument span and sense annotation). We conducted double annotation for sense and argument annotation to measure inter-annotator agreement, with each annotator annotatoing on the first 7 sentences of 21 randomly sampled documents. We achieve high inter-annotator agreement for connective annotation tasks (Cohen's $\kappa > 0.75$; Table 2), and decent agreement for argument annotation (Jaccard index ≈ 0.6–0.7; Table 3). Because these disagreements were not substantial, we proceeded with the full annotation without double annotation.

Ant #1	**Ant #2**	**P/κ (Status)**	**P/κ (Sense)**	**#Inst**
Ant0	Ant1	0.92 / 0.83	0.84 / 0.76	475
Ant1	Ant2	0.88 / 0.91	0.89 / 0.82	488
Ant2	Ant0	0.96 / 0.77	0.82 / 0.77	501
Ant3	Ant4	0.92 / 0.84	0.89 / 0.85	449
Ant3	Ant5	0.91 / 0.83	0.86 / 0.81	549
Ant4	Ant5	0.92 / 0.83	0.91 / 0.87	503

Table 2: Raw agreement proportions (P) and Cohen's κ for connective status and sense annotation. *Ant* = annotator. *Ant0-Ant2* work on documents with only connective annotation. *Ant3-Ant5* annotate both arguments and connectives.

Ant #1	**Ant #2**	**Agreement**	**Union size**
Ant3	Ant4	0.70	8275
Ant3	Ant5	0.63	10932
Ant4	Ant5	0.72	8446

Table 3: Jaccard similarity (intersection over union) for argument span annotation. Each annotation set is a set of 4-tuples ⟨document, token ID, tag (*Arg1*/*Arg2*), connective⟩. Only arguments of words identified as connectives by both annotators are considered.

The final corpus consists of 10,602 sentences from 384 documents, 180 of which have complete annotation of explicit discourse relations (discourse connective and its two argument spans).

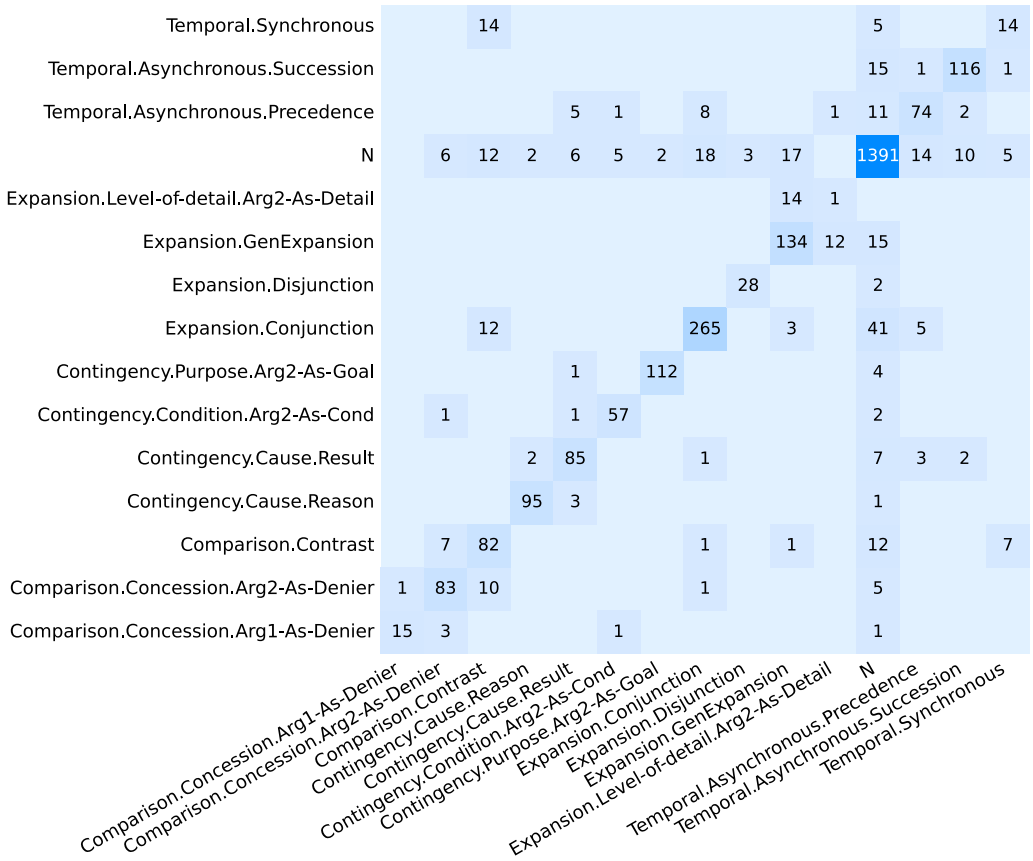

Figure 1: Contingency matrix from the double annotation performed by the annotators shows that the non-connectives are confused more often than the senses are confused.

The documents are news articles about politics (134 documents; 34.89%), crimes and accidents (60 documents; 15.63%), and economics (43 documents; 11.20%). The corpus has the total of 18,374 instances of discourse connectives from 33 different senses. We consider this corpus to be a complete first release, with potential plans to update the corpus in later versions, just as the PDTB did.

We compare the most common level-2 senses in our Thai corpus and other two PDTB-style corpora (Table 1). The exact rankings differ, but the most common senses tend to be the same, with Conjunction at the top.

4.2 Analysis of Annotation Disagreements

We conducted an analysis to determine the source of the disagreements and to highlight the points where the annotators should give special care. To streamline our analysis, we decided to merge these pairs into a single contingency matrix because the annotators exhibited similar patterns of disagreement (Figure 1). However, we filter the matrix to include only the top 15 senses, as lower-ranked senses yield an insufficient number of instances for meaningful analysis. And note that the contingency matrix is not symmetric like a confusion matrix in a classification task because the contingency matrix compares two annotators with different annotations.

Most frequent disagreements come from confusing connectives and non-connectives in multi-functional words. By *multi-functional*, we mean the connectives that have non-connective homonyms which serve a grammatical function. The most common is between EXPANSION.CONJUNCTION and N in และ, ทั้ง and พร้อม. และ behaves similarly to English *and* in that they can conjoin S, VP or NP. Thus, the annotators sometimes struggled to distinguish between the use of NP coordination, which is not a discourse connective, and VP- and S-coordination, which is. Similarly, พร้อม is a relatively rare connective and is often used as an adjective meaning 'ready.'

Other frequent disagreements involve EXPANSION.GENEXPANSION, TEMPORAL.ASYNCHRONOUS.

SUCCESSION, COMPARISON.CONTRAST, and TEMPORAL.ASYNCHRONOUS.PRECEDENCE. For each of these, (e.g., GENEXPANSION ซึ่ง), an exclusive particle (e.g., CONTRAST แต่), an auxiliary (e.g., PRECEDENCE แล้ว), or a preposition (e.g., PRECEDENCE ก่อน). (13) gives an example of a disagreement on ก่อน, where one annotator mistakes it as a non-connective.

(13) **ก่อน** เข้า จับกุม นั้น ตำรวจ สืบทราบ ...
before enter arrest TOP police discover ...
'Before arresting, the police discovered ...'

Some of these disagreements are due to genuine ambiguity, and no mistakes were made. In the common case (14), แล้ว is ambiguous between a post-clausal aspect auxiliary (non-connective) and a subordinator attached to the second clause (a PRECEDENCE connective).

(14) เจ้าหญิง คิโกะ ทรงเข้า โรงพยาบาล **แล้ว**
princess Kiko enter hospital then/PERF
ให้ แพทย์ ถวาย การดูแล
have doctor give care
'Princess Kiko entered the hospital and had the doctors take care of her.'

Whether multifunctionality leads to disagreements depends on its frequency as a connective and syntactic behavior, however. COMPARISON and CONTINGENCY connectives induce disagreements less, even though they have a non-connective counterpart because they tend to be more often connective than not. For instance, แต่ is most frequently a COMPARISON.CONCESSION.ARG2-AS-DENIER connective, not an exclusive particle. In addition, the purpose connective เพื่อ can be easily distinguished from its non-connective prepositional counterpart because the connective usually requires a complementizer (e.g., ให้) when taking a clause, whereas the preposition does not.

For these potentially tricky cases, annotators analyzed the syntax during annotation (e.g., checking whether the argument is an NP, VP, or S) and consulted a supervising linguist in hard cases.

Sense disagreements can result from closely related, confusable senses. On top of the confusability between non-connectives and connectives, even among instances of clear connectives, some senses are harder to tease apart than others. Notably, โดย is solely responsible for all disagreements involving EXPANSION.GENEXPANSION and EXPANSION.LEVEL-OF-DETAIL.ARG2-AS-DETAIL. Although the annotators distinguish its sense as a preposition (non-connective) perfectly, they sometimes encounter challenges in distinguishing between the subtle nuances between these two senses. EXPANSION.GENEXPANSION relates to a broader expansion of information or ideas, while EXPANSION.LEVEL-OF-DETAIL.ARG2-AS-DETAIL involves the inclusion of specific details or finer granularity within the discourse. For these easily confusable senses, annotators were instructed to work wth clearer paraphrases (e.g., denier แต่, but not contrastive แต่, can be paraphrased with another denier such as ถึง) and consult a supervising linguist when in doubt.

Argument span disagreements can result from attachment ambiguity. Unlike sense labelling, disagreements in argument agreement spans are more sporadic and do not recur. One discernible pattern involves attachment ambiguity where it is not clear whether a subordinate clause is attached to S, VP, or an embedded VP. In (15), for example, it is not clear whether the clause with หลัง attaches to a whole clause including the subject, only the VP of the clause, or the VP-complement of เล็ง. Such ambiguities barely affect the semantics of the sentence. Like (14), neither annotator is incorrect, but their annotations differ because they opt for different underlying structures.

(15) คมช. เล็ง เลิก
military government aim repeal
กฎอัยการศึก วันที่ 5 ธ.ค. หลัง ปชช.
martial law day 5 December after people
เข้าใจ คมช.
understand mil gov
'The military government aims to repeal the martial law on December 5 after the people have reached an understanding.'

5 Discourse Connective Identification and Classification Models

With our corpus, we develop classifiers for two basic shallow discourse parsing tasks:

1. **Connective identification**: determining whether a word is an explicit discourse connective. This task is formulated as a binary

classification problem or a binary sequence tagging problem.

2. **Sense classification**: classifying a given word as a non-connective or one of the discourse senses conveyed by that connective. In other words, this task takes care of connective identification and sense classification in one go.

We train models with all three PDTB sense levels under two conditions, i.e., given raw tokens vs. gold-standard connectives. For each task, we experiment with three families of models.

1. **Dictionary** To establish a baseline, we create a simple model that always classifies as connective a word that is more often than not a connective in the connective identification task, and selects the most frequent sense in the training set for a given token in the sense classification task.

2. **Conditional Random Fields (CRF)** We formulated the task as a sequence tagging problem and trained CRF models on both tasks, using different combinations of the following features:

   (a) Word form

   (b) Gold-standard part-of-speech tag annotated in LST20 corpus or automatic part-of-speech tag, using PyThaiNLP's perceptron engine (Phatthiyaphaibun et al., 2016)

   (c) All word and/or POS tag uni-, bi-, tri-grams within a context window of 3 words around a given word

3. **Pretrained Language Model (PLM)** We formulated the task as a token classification task. We finetuned WangchanBERTa (Lowphansirikul et al., 2021), a RoBERTa-style pretrained Thai language model, on both tasks. We experiment with feeding one sentence at a time and with feeding sentence pairs as input.

For sense classification with gold-standard connectives, we use dictionary and PLM models, but switch CRF out for Maximum Entropy (MaxEnt) models with the same features since the task operates on specific tokens (i.e., the known connectives), not a sequence of tokens.

5.1 Experiment Setup

We randomly split the data into the train-dev-test sets (80:10:10). Since the sense inventory is organized in a hierarchy, we train separate models for level-1 (5-way), level-2 (20-way), and leaf-level (35-way) senses for the sense classification task. We add an extra label for *non-connective*.

It is worth noting that the F_1 measures from PLMs are calculated on predictions postprocessed for consistency. PLMs such as WangchanBERTa come with their own tokenizers. If the tokenizer incorrectly tokenizes the text, we have no chance of being correct. Thus, we postprocess the predicted labels as follows. If a predicted connective ends with one of the 10 words that the WangchanBERTa tokenizer frequently mismerges into the previous word (จะ, ที่, มี, ให้, เป็น, ต้อง, ไม่, ว่า, ขอ, การ), we separate these words out into a distinct token. We then map each predicted label to the gold-standard token that 1) shares the same character index span, or, failing that, 2) starts at an index contained within the auto-tokenized token. This ensures an equal number of tokens across models.

5.2 Key Results

We evaluate each model using macro-averaged F_1 scores. To obtain conservative measures, we exclude the label *non-connective* from our calculations. Table 4 gives the macro-averaged F_1 scores across tasks. To summarize briefly, across all tasks except one, the best-performing models are PLMs ($F_1 = 0.96$ for the identification task, and 0.46 for the leaf-level sense classification task), followed by log-linear models and the baseline dictionary-based systems, respectively (Table 4). Our key findings are as follows.

Dictionary-based models are insufficient for sense classification tasks. The dictionary-lookup system using the majority sense performs somewhat decently in the identification task ($F_1 = 0.72$). However, the system progressively performs worse as the sense classification task becomes finer-grained. The system achieves 0.69 F_1 score in classifying level-1 sense when given a gold-standard discourse connective because the percentage of majority senses of most common discourse connectives generally ranges from 0.7 to 1.0, and is as low as 0.3 in the case of the multiply ambiguous ก็ (Table 5). This confirms

Model	F_1			
	Identification	**Sense LV1**	**Sense LV2**	**Sense**
Raw tokens				
WangchanBERTa - 1 sentence	**0.956**	**0.886**	**0.517**	**0.460**
WangchanBERTa - 2 sentences	0.906	0.882	0.508	0.413
CRF_{Word}	0.725	0.660	0.323	0.297
$\text{CRF}_{\text{Auto-POS+Word}}$	0.732	0.679	0.328	0.303
$\text{CRF}_{\text{POS+Word}}$	0.784	0.760	0.360	0.336
$\text{CRF}_{\text{Word 1,2,3-grams}}$	0.853	0.815	0.433	0.380
$\text{CRF}_{\text{Auto-POS+Word 1,2,3-grams}}$	0.858	0.811	0.438	0.385
$\text{CRF}_{\text{POS+Word 1,2,3-grams}}$	0.868	0.831	0.501	0.418
Dictionary	0.718	0.643	0.326	0.302
Gold standard connectives				
WangchanBERTa - 1 sentence	–	**0.964**	0.636	**0.610**
WangchanBERTa - 2 sentences	–	0.865	0.624	0.538
$\text{MaxEnt}_{\text{Word}}$	–	0.921	0.634	0.518
$\text{MaxEnt}_{\text{Auto-POS+Word}}$	–	0.921	0.677	0.573
$\text{MaxEnt}_{\text{POS+Word}}$	–	0.922	**0.723**	0.576
$\text{MaxEnt}_{\text{Word 1,2,3-grams}}$	–	0.941	0.596	0.538
$\text{MaxEnt}_{\text{Auto-POS+Word 1,2,3-grams}}$	–	0.930	0.592	0.531
$\text{MaxEnt}_{\text{POS+Word 1,2,3-grams}}$	–	0.937	0.607	0.538
Dictionary	–	0.689	0.352	0.326

Table 4: Macro-F_1 scores for each model. The scores are calculated after excluding the class *non-connective*. Models under *Raw tokens* are trained and tested on all of the tokens in the data. Models under *Gold standard connectives* are trained and tested on only gold-standard connectives. **Boldfaced** figures are the highest F_1 scores among the models under the same dataset (raw tokens/gold-standard connectives) and task (connective identification/Level 1–3 sense classification). All models are tested on the test set.

the premise that there is a need for a machine learning–based solution for these tasks in Thai. In line with previous results in English and other languages, the best-performing models are PLMs. These models outperform the log-linear models in all settings except when given gold-standard connectives in the level-2 sense setting (Table 7). In the level-1 sense setting, the F_1 scores are in the range of 0.82–0.94 for PLMs, which are higher than 0.72–0.90 for the best CRF models (Table 6).

Functionally ambiguous connectives are difficult for models and humans alike. As Table 9 (left) shows, most of the model misclassifications result from confusion between N (non-connective) and another sense. Interestingly, the errors that the models make are the same ones that cause disagreement among annotators (Table 9, right; see also Section 6), suggesting the natural difficulty rather than model-specific flaws.

The fact that these errors are due to multifunctionality can explain why POS information is important for CRF models (F_1 increases by 0.05 from CRF_{Word} to $\text{CRF}_{\text{Word+POS}}$), but not for the

Connective	Count	Majority sense	% Sense
และ	2105	Expansion.Conjunction	0.97
แต่	1187	Comparison.Concession.Arg2-As-Denier	0.84
เพื่อ	1069	Contingency.Purpose.Arg2-As-Goal	1.00
ก็	1012	Contingency.Condition.Arg2-As-Cond	0.29
โดย	1007	Expansion.GenExpansion	0.87
เพราะ	899	Contingency.Cause.Reason	1.00
ซึ่ง	732	Expansion.GenExpansion	1.00
จึง	556	Contingency.Cause.Result	0.86
หาก	460	Contingency.Condition.Arg2-As-Cond	1.00
เมื่อ	395	Temporal.Asynchronous.Succession	0.93
ส่วน	378	Comparison.Contrast	0.84
แล้ว	286	Temporal.Asynchronous.Precedence	0.72
ถ้า	285	Contingency.Condition.Arg2-As-Cond	0.99
เนื่องจาก	277	Contingency.Cause.Reason	1.00
ขณะ	266	Comparison.Contrast	0.74
ก่อน	235	Temporal.Asynchronous.Precedence	0.97
หรือ	219	Expansion.Disjunction	1.00
จน	204	Contingency.Cause.Result	0.87
นอกจากนี้	172	Expansion.Conjunction	1.00
หลังจาก	168	Temporal.Asynchronous.Succession	1.00

Table 5: Top 20 connective tokens vary in their percentages of majority senses they convey.

Sense LV1	F_1			Support
	PLM	CRF	Dict	
Comparison	0.91	0.88	0.81	383
Contingency	0.94	0.90	0.89	717
Expansion	0.88	0.82	0.56	731
Temporal	0.82	0.72	0.31	308
micro avg	0.89	0.85	0.71	2139
macro avg	0.89	0.83	0.64	2139
weighted avg	0.89	0.85	0.68	2139

Table 6: Macro-F_1 scores for level-1 sense classification of raw tokens. The scores in the CRF column are from the best CRF model, $\text{CRF}_{\text{POS+Word 1,2,3-grams}}$.

Sense LV2	F_1			#Inst.
	PLM	CRF	Dict	
Comparison.Concession	0.90	0.87	0.83	256
Comparison.Contrast	0.74	0.74	0.55	125
Comparison.Similarity	0.00	0.00	0.00	2
Contingency.Cause	0.93	0.94	0.92	366
Contingency.Cause+Belief	0.20	0.00	0.46	9
Contingency.Condition	0.89	0.79	0.79	192
Contingency.Condition+SA	0.00	0.00	0.00	1
Contingency.Neg-Condition	0.00	0.00	0.00	5
Contingency.Purpose	0.96	0.97	0.89	144
Expansion.Conjunction	0.87	0.84	0.27	398
Expansion.Disjunction	0.86	0.91	0.00	39
Expansion.Exception	0.00	0.00	0.00	4
Expansion.GenExpansion	0.86	0.81	0.76	265
Expansion.Instantiation	0.57	0.40	0.00	3
Expansion.Level-of-detail	0.00	0.21	0.00	16
Expansion.Substitution	0.00	0.29	0.00	6
Temporal.Asynchronous	0.81	0.74	0.34	277
Temporal.Synchronous	0.72	0.51	0.06	31
micro avg	0.86	0.83	0.68	2139
macro avg	0.52	0.50	0.33	2139
weighted avg	0.86	0.82	0.61	2139

Table 7: Macro-F_1 scores for level-2 sense classification of raw tokens. The scores in the CRF column is of the best CRF model, $\text{CRF}_{\text{POS+Word 1,2,3-grams}}$.

corresponding MaxEnt models that operate on tokens known to be connectives (F_1 increases by 0.001 from $\text{MaxEnt}_{\text{Word}}$ to $\text{MaxEnt}_{\text{POS+Word}}$). It also explains why PLMs perform best, as they have been shown to implicitly encode morphosyntactic information (Hewitt and Manning, 2019; Manning et al., 2020). Both POS and syntax could help with cases of multifunctionality such as adjectival พร้อม 'ready' vs. conjunctive พร้อม mentioned in Section 4.2 (although not always; cf. Section 6). Future developments of discourse-based models should thus focus on extracting grammatical information, and may test whether using two separate models for the joint task yield better performance, since the present models fare well in the isolated identification task (Best PLM $F_1 = 0.96$).

Connective sense	F_1			N
	PLM	CRF	Dict	
Comparison.Concession.Arg1-As-Denier	0.80	0.62	0.62	60
Comparison.Concession.Arg2-As-Denier	0.85	0.84	0.80	196
Comparison.Contrast	0.75	0.74	0.55	125
Comparison.Similarity	0.00	0.00	0.00	2
Contingency.Cause+Belief.Reason+Belief	0.00	0.00	0.67	5
Contingency.Cause+Belief.Result+Belief	0.33	0.00	0.00	4
Contingency.Cause.Reason	0.97	0.97	0.97	222
Contingency.Cause.Result	0.84	0.86	0.83	144
Contingency.Condition+SA	0.00	0.00	0.00	1
Contingency.Condition.Arg1-As-Cond	0.00	0.00	0.00	13
Contingency.Condition.Arg2-As-Cond	0.86	0.81	0.83	179
Contingency.Negative-Condition.A1-NC	0.00	0.00	0.00	1
Contingency.Negative-Condition.A2-NC	0.00	0.00	0.00	4
Contingency.Purpose.Arg1-As-Goal	0.00	0.00	0.00	3
Contingency.Purpose.Arg2-As-Goal	0.98	0.98	0.90	141
Expansion.Conjunction	0.87	0.84	0.27	398
Expansion.Disjunction	0.83	0.91	0.00	39
Expansion.Exception.Arg2-As-Excpt	0.00	0.00	0.00	4
Expansion.GenExpansion	0.85	0.80	0.76	265
Expansion.Instantiation.Arg2-As-Instance	0.86	0.40	0.00	3
Expansion.Level-of-detail.Arg2-As-Detail	0.00	0.21	0.00	16
Expansion.Substitution.Arg1-As-Subst	0.00	0.00	0.00	3
Expansion.Substitution.Arg2-As-Subst	0.00	0.00	0.00	3
Temporal.Asynchronous.Precedence	0.64	0.58	0.24	106
Temporal.Asynchronous.Succession	0.85	0.78	0.38	171
Temporal.Synchronous	0.69	0.52	0.06	31
micro avg	0.84	0.82	0.67	2139
macro avg	0.46	0.42	0.30	2139
weighted avg	0.83	0.80	0.60	2139

Table 8: Macro-F_1 scores for level-3 sense classification of raw tokens. The scores in the CRF column is of the best CRF model, $\text{CRF}_{\text{POS+Word 1,2,3-grams}}$.

CRF models and PLMs do not differ in the types of errors they make. The PLMs and CRF models appear to make the same types of errors, almost across the board. The relative performances exhibit the same trends across senses. In Level-1 sense classification, for example, Contingency being the easiest class, followed by Comparison, Expansion, and Temporal (Table 6). The exceptions to this trend appear to be due to sparse data rather than the true performance. In level-2 classification, CRF models either beat PLMs by a margin of 1% or in senses with few 40 tokens (39 tokens with Expansion.Disjunction, 16 with Expansion.Level-of-detail and 6 with Expansion.Substitution). Similarly, in level-3 sense classification, CRF models beat PLMs in only three senses, two of which (Expansion.Disjunction, Expansion.Level-of-detail.Arg2-As-Detail) have fewer than 40 tokens (Table 8).

Connective-related tasks in Thai require wide context. The tested models work with different context window size, from the dictionary-based system with essentially no context, to CRF models with local contexts (N-gram features), to

Sense A	Sense B	Count
Expansion.Conjunction	N	59
Expansion.GenExpansion	N	32
Expansion.GenExpansion	Expansion.Level-of-detail.Arg2-As-Detail	26
N	Temporal.Asynchronous.Succession	25
N	Temporal.Asynchronous.Precedence	25
Comparison.Contrast	N	24
Comparison.Contrast	Temporal.Synchronous	21
Comparison.Concession.Arg2-As-Denier	Comparison.Contrast	17
Comparison.Contrast	Expansion.Conjunction	13
Expansion.Conjunction	Temporal.Asynchronous.Precedence	13
Contingency.Cause.Result	N	13
Comparison.Concession.Arg2-As-Denier	N	11
N	Temporal.Synchronous	10
Contingency.Cause.Result	Temporal.Asynchronous.Precedence	8
Expansion.Instantiation.Arg2-As-Instance	N	7

Sense A	Sense B	Count
Expansion.Conjunction	N	81
Expansion.GenExpansion	N	66
N	Temporal.Asynchronous.Precedence	38
Comparison.Contrast	N	30
N	Temporal.Asynchronous.Succession	27
Contingency.Condition.Arg2-As-Cond	N	23
Comparison.Concession.Arg2-As-Denier	Comparison.Contrast	22
Contingency.Cause.Result	N	21
Comparison.Concession.Arg2-As-Denier	N	20
Comparison.Concession.Arg1-As-Denier	Comparison.Concession.Arg2-As-Denier	17
Expansion.Disjunction	N	13
N	Temporal.Synchronous	12
Temporal.Asynchronous.Precedence	Temporal.Asynchronous.Succession	9
Contingency.Condition.Arg2-As-Cond	Temporal.Asynchronous.Succession	9
Expansion.Conjunction	Temporal.Asynchronous.Precedence	9

Table 9: The left table shows the top 15 disgreement between annotators. The right table shows the top 15 disagreement (misclassification) between the best PLM model and the gold standard data.

Error sources	Count
Similar senses	32
Co-ordination	26
Other grammatical functions (multifunctionality)	22
Arg1-Arg2 misidentification	13
NP-S misidentification	2
No discernable pattern	20

Table 10: Common sources of errors in the results.

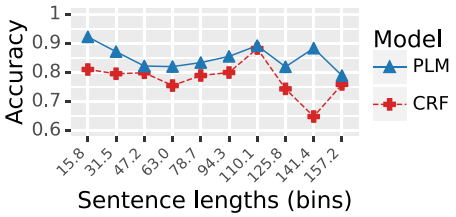

Figure 2: The best PLM model performs better than the best CRF model regardless of sentence length.

the PLMs with large context window (416 tokens during training). The trend is that more contextual information yields higher performance. One apparent anomaly is the PLMs with two-sentence inputs should perform best as they have the largest context size. However, RoBERTa-style PLMs, like WangchanBERTa, are not pretrained on the next sentence prediction task to learn about the discourse information across a sentence pairs like in BERT-style PLMs. This might explain why it is better to feed the model one sentence at a time.

6 Error Analysis of Models

We conducted a more detailed error analysis on the predictions of our best PLM model by inspecting the error rates under different sentence lengths. We also extracted 10 most common per-token-and-misclassification errors (115 instances, total) and manually annotated each instance for its error type (Table 10). In line with our key results, our error analysis reveals a striking parallel between the kinds of mistakes that the models make and the ones where annotators disagree (cf. Section 4.2).

PLMs perform better than CRF models regardless of sentence length. We explore the effect of sentence lengths on the error rates. One might expect that CRF models may perform better or similar to PLMs when the sentence is short as local context is often enough. However, we find no such difference, suggesting that wide contexts are important even in short sentences (Figure 2).

Similar senses induce misclassification. We note that 32 out of the 115 errors extracted are cases of similar senses in แต่ and โดย; 19 แต่ instances involve a misclassification between COMPARISON.CONCESSION.ARG2-AS-DENIER and COMPARISON.CONTRAST. Both senses highlight a contrast between two arguments, the difference being that only the former also involves denying an implicature. (16) gives an example.

(16) ทุกคน อาจจะ ไม่ สนิท กัน
Everyone maybe not close each.other
เท่าไหร่ แต่ หลังจาก คุ้นเคย กัน
much **but** after familiar each.other
แล้ว ก็ เริ่ม ทำงาน สนุก
then LINK begin work fun
'Everyone may not be close to each other but after getting to know each other, the work begins to be fun.'

The 13 errors involving โดย feature a misclassification between EXPANSION.GENEXPANSION and EXPANSION.LEVEL-OF-DETAIL.ARG2-AS-DETAIL.

As noted in Section 4.2, these two senses are also closely related.

Errors are primarily syntactic. Most remaining errors are syntax-related. The most interesting cases are those due to multifunctionality, recapitulating the inter-annotator disagreements. For example, in (17), ก่อน is a PRECEDENCE connective misclassified as a non-connective, plausibly because it can also be a preposition (like English *before*).

(17) รัฐบาล ต้อง เลิก การกระทำ ที่ เป็น
gov. must stop action REL COP
อุปสรรค ต่อ การสมานฉันท์ ก่อน สาย เกินไป
obstacle to unity before late too
'The government must stop actions that hinder unity before it is too late.'

Another particularly common error type involves the model misclassifying a co-ordinator as a non-connective that has the same form (or vice-versa). In (18), หรือ is a connective co-ordinating the two VPs but is misclassified as a non-connective, plausibly because it can also co-ordinate NPs in other cases (cf. English *or*).

(18) แต่เดิม ชนเผ่า เหล่านี้ ไม่ มี ความ
initially tribe these NEG have NOM
สามารถ ใน การ ปลูก ผัก หรือ เลี้ยง
able in NOM grow vegetables or raise
สัตว์
animal
'Initially, these tribes were not able to grow vegetables or to raise animals'

Finally, some errors are specific to แต่. While itself a connective, แต่ could also be selected by ถึง 'although'. When paired this way, the argument following แต่ is *Arg1* instead of *Arg2* because *Arg2* is taken to be the clause attached to ถึง, an annotation choice made to uniformly treat all subordinate clauses as *Arg2*. This and the previous two error types highlight the need for more syntactic information in NLP tasks when working with languages with few orthographical cues to syntax (e.g., spaces and punctuation).

7 Conclusion

In this paper, we present the Thai Discourse Treebank, the first Thai corpus annotated with PDTB-style discourse relations. We develop the annotation guidelines for Thai and annotate 18,374 instances of discourse connectives over 10,602 sentences on top of the LST20 corpus. We benchmark the task of connective identification and sense classification using log-linear models and PLM-based systems. We also provide extensive qualitative analyses of the data and model results. In later releases of the data, we hope to incorporate syntactic information and implicit connectives.

Acknowledgments

We would like to thank reviewers and action editors for their helpful reviews and feedback.

References

Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Springer Netherlands. Dordrecht. https://doi.org/10.1007/978-94-011-1715-9

Prachya Boonkwan, Vorapon Luantangsrisuk, Sitthaa Phaholphinyo, Kanyanat Kriengket, Dhanon Leenoi, Charun Phrombut, Monthika Boriboon, Krit Kosawat, and Thepchai Supnithi. 2020. The annotation guideline of lst20 corpus. https://doi.org/10.48550/arXiv.2008.05055

Peter Bourgonje and Manfred Stede. 2018. Identifying explicit discourse connectives in German. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5037

Peter Bourgonje and Manfred Stede. 2020. Exploiting a lexical resource for discourse connective disambiguation in German. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 5737–5748, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.505

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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. https://doi.org/10.1007/978-94-010-0019-2_5

Sonia Cristofaro. 2005. *Subordination*, Oxford Studies in Typology and Linguistic Theory, New York. Oxford University Press, Oxford.

Elnaz Davoodi and Leila Kosseim. 2016. On the contribution of discourse structure on text complexity assessment. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 166–174, Los Angeles. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-3620

N. J. Enfield. 2007. *A Grammar of Lao*. Number 38 in Mouton Grammar Library, New York. Mouton de Gruyter, Berlin. https://doi.org/10.1515/9783110207538

K. Forbes-Riley. 2005. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23(1):55–106. https://doi.org/10.1093/jos/ffh032

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.disrpt-1.6

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1419

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Shyh-Shiun Hung, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. A complete shift-reduce Chinese discourse parser with robust dynamic oracle. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.13

Shoichi Iwasaki and Preeya Ingkaphirom. 2005. *A Reference Grammar of Thai*. Cambridge University Press. Google-Books-ID: YE29njS4qSUC.

René Knaebel and Manfred Stede. 2020. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.codi-1.7

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2114, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1224

Ziheng Lin, Hwee Tou Ng, and Min-Yan Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*,

20(2):151–184. https://doi.org/10.1017/S1351324912000307

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. https://doi.org/10.48550/ARXIV.2101.09635

Jill Lu, Jennifer Zhang, Nianwen Xue, and Yuping Zhou. 2014. Chinese Discourse Treebank 0.5. https://doi.org/10.35111/NJB6-WB02

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3). https://doi.org/10.1515/text.1.1988.8.3.243

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. In *Proceedings of the National Academy of Sciences*, 117(48):30046–30054. https://doi.org/10.1073/pnas.1907367117, PubMed: 32493748

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal. European Language Resources Association (ELRA).

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4450

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1442

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. https://doi.org/10.3115/1698381.1698410

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, and Pattarawat Chormai. 2016. PyThaiNLP: Thai natural language processing in Python. https://doi.org/10.5281/zenodo.3519354

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics. https://doi.org/10.3115/1667583.1667589

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the Prague discourse treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing.

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).

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. https://doi.org/10.35111/QEBF-GK47

Michaela Regneri and Rui Wang. 2012. Using discourse information for paraphrase extraction. In *Proceedings of the 2012 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pages 916–927, USA. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1068

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1081

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.codi-main.9

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.60

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3604

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K15-2002

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-2004

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1090

Bonnie Webber, Rashmi Prasad, and Alan Lee. 2019. Ambiguity in explicit discourse connectives. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 134–141, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-0411

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–587. https://doi.org/10.1162/089120103322753347

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the*

CoNLL-16 shared task, pages 1–19, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-2001

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143, Minneapolis, MN. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2717

Deniz Zeyrek, Işın Demirşahin, AB Sevdik-Çallı, and Ruket Çakıcı. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *D&D*, 4(2):174–184.

Lanjun Zhou, Binyang Li, Zhongyu Wei, and Kam-Fai Wong. 2014. The CUHK discourse TreeBank for Chinese: Annotating explicit discourse connectives for the Chinese TreeBank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 942–949, Reykjavik, Iceland. European Language Resources Association (ELRA).