

Discourse Annotation in the PDTB: The Next Generation

Rashmi Prasad¹, Bonnie Webber², Alan Lee³

¹Interactions LLC

rprasad@interactions.com

²School of Informatics, University of Edinburgh

bonnie@inf.ed.ac.uk

³University of Pennsylvania

alanlee@lexiconree.org

Abstract

We present highlights from our work on enriching the Penn Discourse Treebank (PDTB), to be released to the community in Fall 2018 as the PDTB-3. We have increased its coverage of discourse relations (from ~40K to ~53k), the majority in intra-sentential contexts. Our work on these new relations has led us to augment and/or modify aspects of the annotation guidelines, including the sense hierarchy, and all changes have been propagated through the rest of the corpus.

1 Introduction

The last decade has seen growing interest in enabling language technology and psycholinguistics to move beyond the sentence, to what can be derived from larger units of text. This has led to greater interest in the properties of discourse. One such property is the coherence between clauses and sentences arising from low-level discourse relations. This level of meaning has been made overt through manual annotation in the Penn Discourse TreeBank (PDTB), developed with NSF support.¹ Version 2.0. of the PDTB (Prasad et al., 2008), released in 2008, contains over 40K tokens of annotated relations, making it the largest such corpus available to date. Largely because the PDTB was based on the simple idea that discourse relations are grounded in an identifiable set of explicit words or phrases (discourse connectives) or simply in sentence adjacency, it has been taken up and used by many researchers in the NLP community and more recently, by researchers in psycholinguistics as well. It has also stimulated the development of similar resources in other languages (Chinese (Zhou and Xue, 2015), Czech (Poláková et al., 2013), Hindi (Oza et al., 2009), Modern Standard Arabic (Al-Saif and Markert, 2010), Turkish (Zeyrek and Webber, 2008) and French (Danlos et al., 2012)) and domains (biomedical texts (Prasad et al., 2011), conversational dialogues (Tonelli et al., 2010)), the organization of community-level shared tasks on *shallow discourse parsing* (Xue et al., 2015; Xue et al., 2016), and a cross-lingual discourse annotation of parallel texts, the TED-MDB corpus (Zeyrek et al., 2018), to support both linguistic understanding of coherence in different languages and improvements in machine translation of discourse connectives.

Given only three years in which to develop guidelines, and annotate and release the PDTB, we knew that it would be incomplete (Prasad et al., 2014). With additional support from the NSF, we have now addressed many of the gaps in the corpus, adding over 17K new discourse relations. Most of the new relations occur intra-sententially, but there are also ~300 inter-sentential implicit relations between adjacent sentences whose annotation is missing from the PDTB-2.² This paper focuses on the new intra-sentential relations annotated in the PDTB-3. We also discuss major modifications and extensions to the PDTB guidelines, including the sense hierarchy, which have resulted from our study of the new relations, and which have been propagated throughout the corpus. PDTB-3, which we plan to release to the community in Fall 2018, will contain over 53K tokens of discourse relations, and as with PDTB-2, will

¹<http://www.seas.upenn.edu/~pdtb>

²Separate from the PDTB-3, Prasad et al. (2017) address the annotation of cross-paragraph implicit relations that are not annotated in either PDTB-2 or PDTB-3. These annotations are provided for 145 texts from Sections 01, 06, and 23 of the Wall Street Journal corpus, producing a full-text annotated sub-corpus merged with the PDTB-3 annotations for the same texts. However, because the annotation guidelines developed for the cross-paragraph annotation depart in some respects from the PDTB guidelines in ways not incorporated in PDTB-3, these annotations will be released to the community separately, via *github* (<https://github.com/pdtb-upenn/full-text>).

be distributed through the Linguistic Data Consortium (LDC), along with a detailed annotation manual (which will also be available from the PDTB website).

Section 2 describes the range of new constructions annotated in the corpus. Section 3 describes changes to the senses and relation types, and Section 4 describes some other modifications to the guidelines. We close with a discussion of mapping the new PDTB-3 senses to the ISO-DR-Core set of discourse relations (Section 5), and the conclusion (Section 6).

2 New Relations

While the PDTB-2 contains over 40K tokens of discourse relations, there are some syntactic and textual contexts where discourse relations were not annotated. In particular, PDTB-2 guidelines (PDTB-Group, 2008) limited annotation to (a) explicit relations lexicalized by discourse connectives, and (b) implicit relations between paragraph-internal adjacent sentences and between (semi-)colon separated clauses within sentences. Further, discourse connectives were drawn from the pre-defined syntactic classes of subordinating conjunctions, coordinating conjunctions, and discourse adverbials. And strict constraints were placed on the syntactic realization of relation arguments: with a few well-defined exceptions, arguments had to be realized as one or more clauses or sentences.

Defining the scope of the annotation in this way, however, precluded consideration of a wider set of discourse relations.

First, the general restriction to explicit connectives precluded subordinate clauses — in particular **free adjuncts** and **free TO-infinitives**, that can occur without lexical subordinators while bearing an implicit relation to their matrix clause. Ex. (1) shows a free adjunct related via an implicit REASON sense to its matrix clause, explaining why treasurys opened lower. The free TO-infinitive in Ex. (2) is related via CONDITION to its matrix clause, specifying the hypothetical purpose from which the competitive edge would have to follow.

- (1) *Treasurys opened lower*, Implicit=as a result of reacting negatively to news that the producer price index – a measure of inflation on the wholesale level – accelerated in September. (CONTINGENCY.CAUSE.REASON)[wsj_2428]
- (2) *Banks need a competitive edge* Implicit=if (they are) to sell their products. (CONTINGENCY.CONDITION.ARG2-AS-CONDITION)[wsj_0238]

Second, the restriction to explicit connectives from the limited set of syntactic classes precluded relations triggered by **prepositional subordinators** like *for*, *by*, *in*, *with*, *instead of*, etc., that can complementize for clauses, as in Exs. (3-6).

- (3) *But with foreign companies snapping up U.S. movie studios, the networks are pressing their fight harder than ever.* (CONTINGENCY.CAUSE.REASON) [wsj_2451]
- (4) Wall Street analysts have criticized Bethlehem *for not following its major competitors in linking with a foreign company to share costs and provide technology to modernize old facilities or build new ones.* (EXPANSION.MANNER.ARG2-AS-MANNER) [wsj_0782]
- (5) James Cleveland, a courier *who earned a Bravo Zulu for figuring out how to get a major customer’s 1,100-parcel-a-week load to its doorstep by 8 a.m.*, considers himself far more than a courier. (CONTINGENCY.CAUSE.REASON) [wsj_1394]
- (6) But on reflection, Mr. Oka says, he concluded that Nissan is being prudent *in following its slow-startup strategy instead of simply copying Lexus.* (EXPANSION.SUBSTITUTION.ARG1-AS-SUBST) [wsj_0286]

Third, the restriction on arguments to clauses (with a small set of specific exceptions) precluded relations between **conjoined verb phrases**. The PDTB-2 exceptions to clausal realization did allow verb phrases to be valid arguments, but not of the VP conjunction itself. Thus, in Ex. (7), while *because* was annotated, the VP conjunction *and* was not. Conjoined VPs have now been annotated in the PDTB-3 (Webber et al., 2016), as in Ex. (8) and Ex. (9).

- (7) She became an abortionist accidentally, *and continued because it enabled her to buy jam, cocoa and other war-rationed goodies.* (CONTINGENCY.CAUSE.REASON) [wsj_0039]
- (8) She *became an abortionist accidentally, and continued because it enabled her to buy jam, cocoa and other war-rationed goodies.* (EXPANSION.CONJUNCTION) [wsj_0039]

Intra-S Context	~Num
Free Adjuncts	~2200
Free TO-infinitives	~1500
Prep. Clausal Subordination	~1600
Conjoined VPs	~5800
S Conjunction Implicits	~1800
Total	~13000

Table 1: Approximate distribution of new intra-sentential relations in PDTB-3. Exact distributions will be announced with the release of the corpus.

(9) Stocks *closed higher in Hong Kong, Manila, Singapore, Sydney and Wellington*, **but were lower in Seoul**. (COMPARISON.CONTRAST) [wsj_0231]

As discussed in Webber et al. (2016), in order to maintain alignment with the shared subject of VP conjunctions, we excluded the shared subject from both the left and right conjunct arguments of the conjunction, as in Ex. (8-9). Webber et al. also discuss the fact that these arguments can also be linked by an additional implicit relation, as in Ex. (8), where an implicit temporal PRECEDENCE sense is inferred between the arguments in addition to the explicit CONJUNCTION. Rather than associating the additional sense inference with the conjunction itself, these implicit relations have been annotated as separate tokens, as in Ex. (10). However, the explicit and implicit relations are *linked* in the underlying representation, with a linking mechanism to indicate that two relations hold between the same two arguments.

(10) She *became an abortionist accidentally*, Implicit=then **and continued because it enabled her to buy jam, cocoa and other war-rationed goodies**. (TEMPORAL.ASYNCHRONOUS.PRECEDENCE) [wsj_0039]

The distribution of new intra-S relations is given in Table 1, showing VP conjunctions accounting for about half of the total. However, about 20% of these tokens are implicit relations inferred in addition to those associated with the explicit conjunction, as with the PRECEDENCE sense in Ex. (10). The “S Conjunction Implicits” category in the table is in fact a consequence of our finding that additional implicit inferences can be associated not just with VP conjunctions but with intra-sentential S conjunctions as well. In PDTB-2, these additional inferences were either not annotated, or associated with the explicit connective, alone or in addition to the default CONJUNCTION sense. For PDTB-3, therefore, all S conjunction relations in PDTB-2 were revisited and reconsidered for these additional inferences, with new implicit tokens added to the corpus where needed. The number of the additional implicit inferences shown in the table account for 32% of the discourse relations associated with S conjunctions in the corpus.

For the annotation task, the intra-sentential contexts discussed above were automatically identified using the Penn Treebank (Marcus et al., 1993) and Propbank (Palmer et al., 2005) annotations. As described in Prasad et al. (2015), subordinated clause contexts were identified by first searching for adjunct (ARGM) arguments of verbs in the Propbank, and then filtered to select clausal arguments, by aligning the Propbank ARGM arguments with the Penn Treebank (PTB). The resulting tokens were then divided into separate sets using further heuristics applied to the PTB clausal structures, thus creating distinct well-defined subtasks corresponding to free adjuncts, TO-infinitives, and preposition-subordinated clauses. Guidelines were then created separately for each of these subtasks after a study of a development sample within each subset. VP conjunction and S conjunction contexts were identified with heuristics applied solely to the PTB, and were also annotated as a separate subtask, with its own set of guidelines. Detailed guidelines for these different contexts will be included in the PDTB-3 manual accompanying the release of the corpus.

Temporal	Synchronous	--
	Asynchronous	Precedence Succession

Contingency	Cause $+/-\beta, +/\zeta$	Reason
		Result
		Negative-result*
	Condition $+/\zeta$	Arg1-as-cond
		Arg2-as-cond
	Negative condition $+/\zeta$	Arg1-as-negcond
		Arg2-as-negcond
	Purpose	Arg1-as-goal
Arg2-as-goal		
Arg2-as-negGoal		

Comparison	Contrast	--
	Similarity	--
	Concession $+/\zeta$	Arg1-as-denier*
		Arg2-as-denier

Expansion	Conjunction	--
	Disjunction	--
	Equivalence	--
	Instantiation	Arg1-as-instance
		Arg2-as-instance
	Level-of-detail	Arg1-as-detail
		Arg2-as-detail
	Substitution	Arg1-as-subst
		Arg2-as-subst
	Exception	Arg1-as-excpt
Arg2-as-excpt		
Manner	Arg1-as-manner	
	Arg2-as-manner	

Figure 1: PDTB-3 Sense Hierarchy. Only asymmetric senses are specified further at Level-3, to capture the directionality of the arguments. Superscript symbols on Level-2 senses indicate features for implicit beliefs ($+/-\beta$) and speech-acts ($+/\zeta$) that may or may not be associated with one of the defined arguments of the relation. Senses marked with the asterisk (*) are the subtypes that did not occur as belief and speech-act senses in the corpus. Features are shown on the sense only for clarity, but should not be seen as a property of the relation, rather of the arguments.

3 Changes to Senses and Relation Types

Figure 1 shows the most recent PDTB-3 sense hierarchy, which simplifies and extends the PDTB-2 sense hierarchy. Simplifications include (a) restricting Level-3 sense to differences in directionality, (b) eliminating rare and/or difficult-to-annotate senses, and (c) replacing separate senses with features that can be added to a given sense (Section 3.1). Extensions mainly involve senses needed for annotating some of the new intra-sentential relations (Section 3.2).

3.1 Simplifying the relation hierarchy

Although the hierarchy retains the same four Level-1 senses, senses at Level-3 now only encode *directionality* of the arguments, and so only appear with asymmetric Level-2 senses.³ Those Level-3 senses in the PDTB-2 that did not convey directionality were either moved to Level-2 — SUBSTITUTION (renamed from the PDTB-2 CHOSEN ALTERNATIVE) and EQUIVALENCE — or eliminated due to their rarity or the difficulty they posed for annotators — in particular, those under the Level-2 senses of CONTRAST, CONDITION and ALTERNATIVE (now renamed DISJUNCTION).

With respect to directionality, annotating intra-sentential discourse relations revealed asymmetric Level-2 senses for which the relation’s arguments occur in either order (rather than the single order assumed in the PDTB-2). In particular, the argument conveying the condition in CONDITION relations can be either **Arg2** (as was the case throughout the PDTB-2) or *Arg1* as in Ex. 11, while the argument conveying the “chosen alternative” (now called “substitute”) in SUBSTITUTION relations can be either **Arg2** (as was the case throughout the PDTB-2) or *Arg1*, as in Ex. 12. In the case of the rare sense called EXCEPTION, it was not previously noticed that in some of the tokens so annotated, the exception appeared in **Arg2**, while in the rest, the exception appeared in *Arg1* (Ex. 13). Finally, while all cases of the INSTANTIATION sense in PDTB-2 were annotated with the assumption that it was always **Arg2** that provided the instance, we have now found (rare) evidence that the instance can be realized as *Arg1* as well (Ex. 14).

³A sense relation \mathfrak{R} is *symmetric* iff $\mathfrak{R}(Arg1, Arg2)$ and $\mathfrak{R}(Arg2, Arg1)$ are semantically equivalent. If a relation is not symmetric, it is *asymmetric*.

- (11) ARG1-AS-COND: *Call Jim Wright's office in downtown Fort Worth, Texas, these days and the receptionist still answers the phone, "Speaker Wright's office.* [wsj_0909]
- (12) ARG1-AS-SUBST: "The primary purpose of a railing is to *contain a vehicle and not to provide a scenic view,*" [wsj_0102]
- (13) ARG1-AS-EXCPT: *Twenty-five years ago the poet Richard Wilbur modernized this 17th-century comedy merely by avoiding "the zounds sort of thing," as he wrote in his introduction. Otherwise, the scene remained Celimene's house in 1666.* [wsj_1936]
- (14) ARG1-AS-INSTANCE: *In a country where a bribe is needed to get a phone, a job, and even into a school, the name Bofors has become a potent rallying cry against the government. **That illustrates the kind of disappointment many Indians feel toward Mr. Gandhi, whom they zestfully elected and enthusiastically supported in his first two years in power.*** [wsj_2041]

Level-2 pragmatic senses have been removed from the hierarchy and replaced with features that can be attached to a relation token to indicate an inference of *implicit* belief (epistemic knowledge) or of a *speech act* associated with arguments, rather than with the relation itself. Figure 1 shows the senses for which these features have so far been found to be warranted, based on the empirical evidence found during annotation. Ex. 15 shows an implicit CAUSE.RESULT relation but one where the result **Arg2** argument is the (speaker's/writer's) *belief* that the deadline could be extended. **Arg2** is therefore annotated with a +belief feature because the belief is implicit. Similarly, Ex. 16 shows a CONCESSION.ARG2-AS-DENIER relation, but what's being denied (or cancelled) is the speech act associated with **Arg2**, and this is annotated as a feature on **Arg2** because it is implicit.

- (15) RESULT+BELIEF: That deadline *has been extended once* Implicit=so **and could be extended again.** [wsj_2032]
- (16) ARG2-AS-DENIER+SPEECH-ACT: He spends his days *sketching passers-by, or* **trying to.** [wsj_0039]

Also simplifying the PDTB2 hierarchy is removal of the LIST sense, which turned out not to be distinguishable from CONJUNCTION. And the names of two asymmetric PDTB-2 senses have been changed to bring out commonalities. In particular, RESTATEMENT has been renamed LEVEL-OF-DETAIL, with its SPECIFICATION and GENERALIZATION subtypes in the PDTB-2 now just taken to be directional variants renamed ARG2-AS-DETAIL and ARG1-AS-DETAIL, respectively. Similarly, the sub-types of CONCESSION, opaquely called CONTRA-EXPECTATION and EXPECTATION, have been renamed to reflect simply a difference in directionality: ARG1-AS-DENIER and ARG2-AS-DENIER, respectively.

3.2 Augmenting the Sense Hierarchy

New senses have been introduced into the hierarchy on an "as needed" basis. These include the asymmetric Level-2 senses of MANNER under EXPANSION, and PURPOSE and NEGATIVE CONDITION under CONTINGENCY, along with their Level-3 directional variants. Parallel to the negative counterpart of CONDITION (NEGATIVE CONDITION), we also found evidence for negative counterparts of RESULT (NEGRESULT) and ARG2-AS-GOAL (ARG2-AS-NEGGOAL). The symmetric Level-2 sense SIMILARITY was added under COMPARISON because of its obvious omission from the PDTB-2 as the complement of the symmetric sense CONTRAST. The definitions and examples for these new senses are given in Table 2.

The entire PDTB2 has been updated to reflect the sense modifications. Most often, the mapping is simply 1:1 and has been done automatically. Where the mapping is 1:N or M:N, manual review has been required, with further adjudication to ensure both agreement and consistency.

3.3 Hypophora: A New Relation Type

Among the inter-sentential relations missing from PDTB-2, we found many pairs such as Exs. (17-18), where the first sentence (*Arg1*) expresses a question seeking some information, and the second (**Arg2**) provides a response to fulfil that need. As with the EntRel relations in the PDTB, these relations cannot be instantiated with connectives, explicitly or implicitly.

- (17) *If not now, when?* "**When the fruit is ripe, it falls from the tree by itself,**" he says." [wsj_0300]
- (18) *Of all the ethnic tensions in America, which is the most troublesome right now?* **A good bet would be the tension between blacks and Jews in New York City.** [wsj_2369]

<p>SIMILARITY: One or more similarities between <i>Arg1</i> and Arg2 are highlighted with respect to what each argument predicates as a whole or to some entities it mentions.</p> <p>... , the Straits Times index is up 24% this year, so investors who bailed out generally did so profitably. <u>Similarly</u>, Kuala Lumpur’s composite index yesterday ended 27.5% above its 1988 close. [wsj_2230]</p>
<p>CAUSE:NEGATIVE RESULT: <i>Arg1</i> gives the reason/explanation/justification for why Arg2 does not result.</p> <p>A search party soon found the unscathed aircraft in a forest clearing <u>much too small</u> to have allowed a conventional landing.</p>
<p>NEGATIVE CONDITION: One argument describes a situation presented as unrealized (the antecedent or condition), which if it doesn’t occur, would lead to the situation described by the other argument (the consequent).</p> <p>ARG1-AS-NEGCOND: In Singapore, a new law requires smokers to <i>put out their cigarettes before entering restaurants, department stores and sports centers</i> <u>or</u> face a \$250 fine. [wsj_0037]</p> <p>ARG2-AS-NEGCOND: <u>Unless</u> the Federal Reserve eases interest rates soon to stimulate the economy, profits could remain disappointing. [wsj_0322]</p>
<p>PURPOSE: One argument presents an action that an agent undertakes with the purpose (intention) of achieving the goal conveyed by the other argument.</p> <p>ARG1-AS-GOAL: She ordered <i>the foyer done in a different plaid planting</i>, <u>Implicit=for that purpose</u> and made the landscape architects study a book on tartans. [wsj_0984]</p> <p>ARG2-AS-GOAL: <i>Skilled ringers use their wrists to advance or retard the next swing</i>, <u>so that</u> one bell can swap places with another in the following change. [wsj_0089]</p> <p>ARG2-AS-NEGGOAL: We can applaud Mr. Pryor’s moment of epiphany, even as we understand <i>that he and his confreres need restraint</i> <u>lest</u> they kill again. [wsj_1698]</p>
<p>MANNER: The situation described by one argument presents <i>how</i> (i.e., the manner in which) the situation described by other argument has happened or is done.</p> <p>ARG1-AS-MANNER: He argued that program-trading by roughly 15 big institutions is <i>pushing around the markets</i> <u>Implicit=thereby</u> and scaring individual investors. [wsj_0987]</p> <p>ARG2-AS-MANNER: A native of the area, he is back now after riding the oil-field boom to the top, <i>then surviving the bust</i> <u>Implicit=by</u> running an Oklahoma City convenience store. [wsj_0725]</p>

Table 2: New senses in PDTB-3

The response to the question can answer the information need explicitly, as in Exs. (17-18), or implicitly (Ex. 19). And the answer can also indicate that the information need cannot be fulfilled (Ex. 20).

(19) *So can a magazine survive by downright thumbing its nose at major advertisers?* **Garbage magazine, billed as "The Practical Journal for the Environment," is about to find out.** [wsj_0062]

(20) *With all this, can stock prices hold their own?* **"The question is unanswerable at this point"** she says. [wsj_0681]

Because these relations involve dialogue acts (Bunt et al., 2017), which we treat as distinct from discourse relations, and because they are uninstantiable as connectives, we have added a new coherence relation type for them — called HYPOPHORA.

Of course, not all questions in a discourse are dialogue acts. HYPOPHORA does not apply when the subsequent text relates to a question in other ways – for example, with rhetorical questions that are posed for dramatic effect or to make an assertion, rather than to elicit an answer, as in Ex. (21), or if the subsequent text provides an explanation for why the question has been asked, as in Ex. (22). In such cases, an implicit connective can be asserted and a discourse relation can be inferred to hold, as shown.

(21) *What’s wrong with asking for more money?*

Implicit=because **Money is not everything, but it is necessary, and business is not volunteer work.** (CONTINGENCY.CAUSE.REASON+BELIEF) [wsj_0094]

(22) *"What sector is stepping forward to pick up the slack?"* he asked.

Implicit=because **"I draw a blank."** (CONTINGENCY.CAUSE.REASON+SPEECH-ACT) [wsj_0036]

4 Modifications to Other Guidelines

In this section, we present other major modifications to the PDTB guidelines.

4.1 Argument Labeling Convention

The first modification, described earlier in Webber et al. (2016), relates to the argument labeling two-part convention in PDTB-2, where

- For spans linked by an explicit discourse connective, **Arg2** was the argument to which the connective was attached syntactically, and the other was *Arg1*. This allowed the arguments to subordinating conjunctions to be labeled consistently, independent of the order in which the arguments appeared. The same was true for coordinating conjunctions, whose argument order is always the same, and for discourse adverbials, whose *Arg1* always precedes the adverbial, even when *Arg1* is embedded in **Arg2**.
- For implicit discourse relations, *Arg1* was always the first (lefthand) span and **Arg2**, the adjacent (righthand) span.

Blindly applying this convention in annotating *intra-sentential* discourse relations can produce inconsistent labeling because of (1) variability in where an explicit connectives can attach within a sentence; and (2) the ability of marked syntax to replace explicit connectives.

The first problem can be illustrated with paired connectives like *not only ...but also*. Here, both members of the pair may be present (Ex. 23), or just one or the other (Ex. 24 and Ex. 25):

- (23) Japan not only outstrips the U.S. in investment flows but also outranks it in trade with most Southeast Asian countries ... [wsj_0043]
- (24) The hacker was pawing over the Berkeley files but also using Berkeley and other easily accessible computers as stepping stones ... [wsj_0257]
- (25) Not only did Mr. Ortega's comments come in the midst of what was intended as a showcase for the region, it came as Nicaragua is under special international scrutiny ... [wsj_0655]

A labeling convention that requires **Arg2** to be the argument to which the explicit connective attaches will choose a different argument for **Arg2** in Ex. (24) than in Ex. (25), and an arbitrary argument in the case of Ex (23), when semantically, the lefthand argument is playing the same role in all three cases, as is the righthand argument.

The second problem can be illustrated with preposed auxiliaries, which signal a CONDITION sense between the clause with the preposed auxiliary (as *antecedent*) and the other clause (as *consequent*). As with subordinating clauses, the two clauses can appear in either order:

- (26) Had the contest gone a full seven games, ABC could have reaped an extra \$10 million in ad sales ... [wsj_0443]
- (27) ...they probably would have gotten away with it, had they not felt compelled to add Ms. Collins's signature tune, "Amazing Grace," ... [wsj_0207]

Since there is no explicit connective in either clause, if position is used to label *Arg1* and **Arg2**, the result can again be inconsistent.

To avoid inconsistency, while not requiring any change to existing labels in the PDTB-2, we have adopted the following new convention:

- The arguments to inter-sentential discourse relations remain labeled by *position*: *Arg1* is first (lefthand) argument and **Arg2**, the second (righthand) argument.
- With intra-sentential coordinating structures, the arguments are also labeled by *position*: *Arg1* is first argument and **Arg2**, the second one, independent of which argument(s) have attached coordinating conjunction(s).
- With intra-sentential subordinating structures, *Arg1* and **Arg2** are determined syntactically. The subordinate structure is always labeled **Arg2**, and the structure to which it is subordinate is labeled *Arg1*.

4.2 AltLex Identification and Annotation

The convention for identifying instances of *Alternative Lexicalizations* (or AltLex) in the PDTB-2 was that, in the absence of an explicit connective, if annotators inferred a relation between the sentences but felt that the insertion of an implicit connective would be redundant, they were encouraged to identify the non-connective expression in **Arg2** that they took as the source of the perceived redundancy as the AltLex,

Annotating intra-sentential discourse relations in the PDTB-3 has led to modifying the above convention in two ways — what is annotated as AltLex and where AltLex can be annotated.

With respect to what is annotated as AltLex, reliably identifiable AltLex expressions in the PDTB-2 included one part that conveyed the relation and one part that referred anaphorically or elliptically to *Arg1*, as in “after that” or “a likely reason for the disparity is” (Prasad et al., 2010). To allow for AltLex expressions in the context of intra-sentential discourse relations, we have allowed expressions of any form or syntactic class to be labeled as AltLex, including adjectives and adjective-modifiers such as *additional*, *next*, *further*, and *earlier*. While these expressions continue to suggest the relation, unlike AltLex expressions in PDTB-2, the reference to *Arg1* may be implicit. That is, while *next* implies next to something, that something may be implicit.

One consequence of this new convention is that words such as *further* and *next*, that can appear as discourse adverbials, or as adverbials modifying verbs or adjectives, or as adjectives themselves, will be annotated as Explicit connectives when they are discourse adverbials, as in Exs. (28-29), and otherwise as AltLex phrases, as in Ex. (30), where *further* modifies *fractioning*.

- (28) Stephen G. Jerritts, president and chief executive officer, said *customers weren't willing to commit to an expensive NBI hardware systems because of the company's financial troubles*. Further, he said, *argii*[the company doesn't have the capital needed to build the business over the next year or two]. (EXPANSION.CONJUNCTION) [wsj_0092]
- (29) *Inspired by imports, Mr. Rosen now makes fur muffs, hats and flings. This year he produced a men's line and offers dyed furs in red, cherry red, violet, royal blue and forest green.... From Asia, he has mink jackets with floral patterns made by using different colored furs.* Next he will be testing pictured embroidery (called kalega) made in the Far East. (TEMPORAL.ASYNCHRONOUS.PRECEDENCE) [wsj_1586]
- (30) *The show, despite a promising start, has slipped badly in the weekly ratings as compiled by A.C. Nielsen Co., finishing far below “Tonight” on NBC, a unit of General Electric Co., and “Nightline” on ABC-TV, a unit of Capital Cities/ABC Inc.* Further fractioning the late-night audience is the addition of the “Arsenio Hall Show,” syndicated by Paramount Communications Inc (EXPANSION.CONJUNCTION) [wsj_2395]

With respect to where *AltLex* can be annotated, PDTB-3 annotators have been permitted to include material the AltLex expression from both *Arg1* and *Arg2*. This is motivated by examples like Exs. (31-33), where the underlined segmented in *Arg1* and **Arg2** together signal the sense of the relation.

- (31) *Marni Rice plays the maid with so much edge **as to steal her two scenes**.* (CONTINGENCY.CAUSE.RESULT) [wsj_1163]
- (32) *some of the proposals are so close **that non-financial issues such as timing may play a more important role**.* (CONTINGENCY.CAUSE.RESULT) [wsj_0953]
- (33) *Things have gone too far **for the government to stop them now**.* (CONTINGENCY.CAUSE.RESULT) [wsj_2454]

We have also allowed AltLex to span an entire argument, which would typically be **Arg2**, to adequately represent the expression of discourse relations with *syntactic constructions*. For example, in Ex. (34), it is the *syntactic inversion* of the predicate that signals the CONCESSION sense. And in Ex. (35), it is the AUX-inversion that signals the CONDITION sense. In both these cases, as in others like these, the entire **Arg2** is selected as the AltLex span, which is a unique indication that it is the syntactic construction that serves as the AltLex.

- (34) **Crude as they were**, *these early PCs triggered explosive product development in desktop models for the home and office.* (COMPARISON.CONCESSION.ARG1-AS-DENIER) [wsj_0022]
- (35) **Had the contest gone a full seven games**, *ABC could have reaped an extra \$10 million in ad sales on the seventh game alone, compared with the ad take it would have received for regular prime-time shows.* (CONTINGENCY.CONDITION.ARG2-AS-CONDITION) [wsj_0443]

Since researchers may be interested in analyzing these constructional AltLex's further, we have assigned them the relation type ALTLEXC, to indicate that they are a sub-type of Altlex. Tokens of this type have all the same fields as an AltLex. They are just marked for easy identification and review.

5 Mapping to ISO-DR-Core

Existing annotation frameworks (which, apart from the PDTB, have led to the creation of several other corpora of coherence relations, including Afantenos et al. (2012), Carlson et al. (2003), Reese et al. (2007), Sanders and Scholman (2012), and Wolf and Gibson (2005)) exhibit some major differences in their underlying assumptions, but there are also strong compatibilities. ISO DR-Core (ISO 247617-8: 2016) forms part of an effort to develop an international standard for the annotation of discourse relations.⁴ One of the outcomes of this effort (Bunt and Prasad, 2016) was to provide clear and mutually consistent definitions of a set of *core* discourse relations (senses) – ISO-DR-Core – many of which have similar definitions in different frameworks, and provide mappings from ISO-DR-Core relations to relations in different frameworks, including the PDTB.

With the extensions to the sense hierarchy in PDTB, it is therefore of interest to ask if the new senses in PDTB are mappable to the ISO-DR-Core relations. We find that while the new senses PURPOSE, NEGATIVE CONDITION, SIMILARITY and MANNER have a 1:1 mapping with the relations in ISO-DR-Core, ARG2-AS-NEGGOAL (under Level-2 PURPOSE) and NEGATIVE RESULT (under Level-2 CAUSE) do not. What this suggests is that like the negative counterpart of condition, ISO-DR-Core should be extended to include the negative counterpart for CAUSE, as NEGATIVE CAUSE, and for PURPOSE, as NEGATIVE PURPOSE. However, it remains an open question whether these relations should be defined in a way that captures both argument directionalities. In the PDTB, we have as yet found no evidence for the reverse directionality for either of these senses.

6 Conclusion

We have presented highlights from our work on enriching the PDTB with new relations, which has also led to modifications and extensions to the PDTB guidelines. Annotating a further ~13K discourse relations and reviewing existing PDTB-2 annotation to bring it in line with the new guidelines has highlighted the importance of assessing *consistency* across the corpus — that similar tokens are annotated in a similar way, no matter when they were annotated. Such *semantic consistency* (Hollenstein et al., 2016) is meant to facilitate improvement in all future applications of the PDTB-3. Consistency checks are described in the detailed annotation manual that will accompany the corpus in its LDC distribution, as well as being available at the PDTB website.

Acknowledgements

Missing from the list of authors is the name of our colleague and dear friend, Aravind K Joshi. Aravind conceived of the Penn Discourse TreeBank and was instrumental in its development from its beginnings in 2005 through the new PDTB-3. Aravind died peacefully on 31 December 2017, and we miss him more than we can say. An obituary will appear in *Computational Linguistics* 44(3), September 2018. We would also like to thank Robin Malamud, James Reid and Samuel Gibbon for their hard work and valuable contribution to annotating the PDTB-3. This work was supported by the National Science Foundation (NSF Grants 1422186 and 1421067).

References

- Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, et al. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).
- Amal Al-Saif and Katja Markert. 2010. The Leeds Arabic discourse treebank: Annotating discourse connectives for Arabic. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, Valletta, Malta.

⁴<https://www.iso.org/standard/60780.html?browse=tc>

- Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.
- Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. In *Multimodal Interaction with W3C Standards*, pages 109–135. Springer.
- 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.
- Laurence Danlos, Diégo Antolin-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le fdtb: French discourse tree bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 471–478. ATALA/AFCP.
- Nora Hollenstein, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, 19(2):313–330.
- Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Mishra Sharma, and Aravind Joshi. 2009. The Hindi Discourse Relation Bank. In *Proceedings of the ACL 2009 Linguistic Annotation Workshop III (LAW-III)*, Singapore.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.
- 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 (IJCNLP)*, pages 91–99.
- 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 6th International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech, Morocco.
- Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1023–1031. Association for Computational Linguistics.
- Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188.
- Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora and complementary annotation. *Computational Linguistics*, 40(4):921–950.
- Rashmi Prasad, Bonnie Webber, Alan Lee, Sameer Pradhan, and Aravind Joshi. 2015. Bridging sentential and discourse-level semantics through clausal adjuncts. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 64–69.
- Rashmi Prasad, Katherine Forbes-Riley, and Alan Lee. 2017. Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16.
- Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis, and Jason Baldridge, 2007. *Reference Manual for the Analysis and Annotation of Rhetorical Structure*. University of Texas at Austin.
- Ted J. Sanders and Merel Scholman. 2012. Categories of coherence relations in discourse annotation. Presented at the International Workshop on Discourse Annotation. Utrecht Institute of Linguistics, Universiteit Utrecht.
- PDTB-Group. 2008. The Penn Discourse TreeBank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.
- Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

- Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind K. Joshi. 2016. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop*, Berlin, Germany.
- Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational linguistics*, 31(2):249–287.
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the CoNLL-15 shared task*, pages 1–16, Beijing, China.
- Nianwen Xue, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany.
- Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources, The Third International Joint Conference on Natural Language Processing, (IJCNLP-2008)*, pages 65–71, Hyderabad, India.
- Deniz Zeyrek, Amalia Mendes, and Murathan Kurfali. 2018. Multi-lingual extension of PDTB-style annotation: The case of TED multi-lingual discourse bank. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC-18*, Miyazaki, Japan.
- Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations. In *Proceedings of Language Resources and Evaluation (LREC)*, pages 397–431.