The Penn Discourse TreeBank 2.0.

Rashmi Prasad*, Nikhil Dinesh*, Alan Lee*, Eleni Miltsakaki*, Livio Robaldo†
Aravind Joshi*, Bonnie Webber**

*University of Pennsylvania
Philadelphia, PA, USA
rjprasad, nikhild, aleewk, elenimi,joshi@seas.upenn.edu

†University of Torino
Torino, Italy
livio.robaldo@di.unito.it

**University of Edinburgh
Edinburgh, Scotland
bonnie@inf.ed.ac.uk

Abstract

We present the second version of the Penn Discourse Treebank, PDTB-2.0, describing its lexically-grounded annotations of discourse relations and their two abstract object arguments over the 1 million word Wall Street Journal corpus. We describe all aspects of the annotation, including (a) the argument structure of discourse relations, (b) the sense annotation of the relations, and (c) the attribution of discourse relations and each of their arguments. We list the differences between PDTB-1.0 and PDTB-2.0. We present representative statistics for several aspects of the annotation in the corpus.

1. Introduction

An increasing interest in moving human language technology beyond the level of the sentence in text summarization, question answering, and natural language generation (NLG) *inter alia* has recently led to the development of several resources that are richly annotated at the discourse level. Among these is the Penn Discourse TreeBank (PDTB)[1], a large-scale resource of annotated discourse relations and their arguments over the 1 million word Wall Street Journal (WSJ) Corpus. Since the sentence-level syntactic annotations of the Penn Treebank (Marcus et al., 1993) and the predicate-argument annotations of the Propbank (Palmer et al., 2005) have been done over the same target corpus, the PDTB thus provides a richer substrate for the development and evaluation of practical algorithms while supporting the extraction of useful features pertaining to syntax, semantics and discourse all at once.[2] The PDTB is the first to follow a lexically-grounded approach to the annotation of discourse relations (Webber et al., 2003). Discourse relations, when realized explicitly in the text, are annotated by marking the necessary lexical items - called *discourse connectives* - expressing them, thus supporting their automatic identification. For example, the causal relation in (1) is annotated by marking the discourse connective *as a result* as the expression of the relation.

(1) *U.S. Trust, a 136-year-old institution that is one of the earliest high-net worth banks in the U.S., has faced intensifying competition from other firms that have established, and heavily promoted, private-banking businesses of their own.* As a result, **U.S. Trust's earnings have been hurt**.

The PDTB is also unique in adopting a theory-neutral approach to the annotation, making no commitments to what kinds of high-level structures may be created from the low-level annotations of relations and their arguments. This approach has the appeal of allowing the corpus to be useful for researchers working within different frameworks while at the same time providing a resource to validate the various existing theories of discourse structure (Mann and Thompson, 1988; Webber and Joshi, 1998; Wolf and Gibson, 2005). This theory neutrality also permits investigation of the general question of how structure at the sentence level relates to structure at the discourse level, at least that part of the discourse structure that is "parallel" to the sentence structure (Lee et al., 2006).

In addition to the argument structure of discourse relations, the PDTB provides sense labels for each relation following a hierarchical classification scheme. Annotation of senses highlights the polysemy of connectives, making the PDTB useful for sense disambiguation tasks (Miltsakaki et al., 2005). Finally, the PDTB separately annotates the attribution of each discourse relation and of each of its two arguments. While attribution is a relation between agents and abstract objects and thus not a discourse relation, it has been annotated in the PDTB because (a) it is useful for applications such as subjectivity analysis and multi-perspective QA (Prasad et al., 2006), and (b) it exhibits an interesting and complex interaction between sentence-level structure and discourse structure (Dinesh et al., 2005).

The first preliminary release of the PDTB was in April 2006. A significantly extended version was released as PDTB-2.0 in February 2008, through the Linguistic Data Consortium (LDC).[3] This paper describes the annotations

[1] `http://www.seas.upenn.edu/~pdtb`.

[2] The PDTB corpus provides alignments of its annotations with the Penn Treebank but not the Propbank.

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T05. Along with the corpus, the PDTB LDC dis-

found in the PDTB-2.0. Each aspect of the annotation described is accompanied by representative statistics from the corpus and the reliability standards. Section 2. describes the annotation of discourse relations and their arguments. Section 3. presents the sense annotation, and Section 4. describes the annotation of attribution. Section 5. highlights differences between PDTB-1.0 and PDTB-2.0, and future work is mentioned in Section 6.

2. Annotation of Discourse Relations and their Arguments

Discourse relations in the PDTB can be broadly characterized into two types depending on how the relations are realized in text. The first type involves relations realized explicitly by "Explicit" connectives that are drawn from syntactically well-defined classes – e.g., the adverbial *as a result* in Example (1). Arguments of Explicit connectives are unconstrained in terms of their location, and can be found anywhere in the text. The second type involves relations between two adjacent sentences in the absense of an Explicit connective. In all cases, discourse relations are assumed to hold between two and only two arguments. Because there are no generally accepted abstract semantic categories for classifying the arguments to discourse relations as have been suggested for verbs (e.g., agent, patient, theme, etc.), the two arguments to a connective are simply labelled Arg1 and Arg2. In the case of Explicit connectives, Arg2 is the argument to which the connective is syntactically bound, and Arg1 is the other argument. In the case of relations between adjacent sentences, Arg1 and Arg2 reflect the linear order of the arguments, with Arg1 before Arg2.

Explicit connectives are drawn from three grammatical classes: subordinating conjunctions (e.g., *because*, *when*, etc.), coordinating conjunctions (e.g., *and*, *or*, etc.), and discourse adverbials (e.g., *for example*, *instead*, etc.). Modified and conjoined forms of connectives are also annotated, such as *only because*, *if and when*, as well as a small set of parallel connectives, such as *either..or*, *on the one hand..on the other hand* etc. For subordinating conjunctions, since the subordinate clause is bound to the connective, it is always Arg2. Hence the linear order of the arguments can be Arg1-Arg2 (Ex. 2) or Arg2-Arg1 (Ex. 3), or Arg2 may appear embedded in Arg1 (Ex. 4). (In the examples in this paper, the connective is underlined, Arg1 is in italics, and Arg2 is in bold.)

(2) *Third-quarter sales in Europe were exceptionally strong*, boosted by promotional programs and new products – although **weaker foreign currencies reduced the company's earnings**.

(3) Michelle lives in a hotel room, and although **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it*.

(4) *Most oil companies*, when **they set exploration and production budgets for this year**, *forecast revenue of $15 for each barrel of crude produced*.

The order of the arguments for adverbials and coordinating conjunctions is typically Arg1-Arg2 since Arg1 usually appears in the prior discourse. But as Example (5) shows, Arg1 of an adverbial can appear embedded within Arg2.

(5) As an indicator of the tight grain supply situation in the U.S., market analysts said that **late Tuesday the Chinese government**, *which often buys U.S. grains in quantity*, **turned** instead **to Britain to buy 500,000 metric tons of wheat**.

Arguments of Explicit connectives are not constrained to be single clauses or single sentences: They can be associated with multiple clauses or sentences. However, a *minimality principle* requires an argument to contain the minimal amount of information needed to complete the interpretation of the relation. Any other span of text that is perceived to be relevant (but not necessary) to the interpretation of arguments is annotated as *supplementary information*, labelled Sup1, for material supplementary to Arg1, and Sup2, for material supplementary to Arg2. Also as a consequence of the abstract object characterization of arguments, arguments may be denoted by non-clausal units such as *nominalizations* that have an event interpretation, and *discourse deictics* (*this*, *that*) that refer to abstract objects.

Of interest to both the theoretical understanding of discourse and practical applications in extractive summarization and NLG are the *location* and *extent* of the arguments of Explicit connectives. Since Arg2 is the argument with which an Explicit connective is syntactically associated, identifying the Arg2 of an Explicit connective is less problematic than identifying Arg1.[4] The variability in the *location* and *extent* of the Arg1 of Explicit connectives is shown in Table 1.

Looking first at the distance of Arg1 from its connective (the rows of the Table), one sees that 9% (1666/18459) are located in one or more previous non-adjacent sentences (NAPS). Further analysis of these NAPS cases shows that (1) two involve subordinating conjunctions - *Because* and *Ever since*, both appearing sentence-initially; (2) 31.6% (527/1666) involve S-initial and 12.5% (209/1666), S-medial discourse adverbials, which Webber et al. (2003) and Forbes-Riley et al. (2006) argue should be taken to be anaphoric and hence should have no problem with their anaphoric argument being located a distance away; (3) 24% (400/1666) involve S-initial coordinating conjunctions, which are discussed in Webber and Prasad (To appear); and (4) 31.7% (528/1666) involve S-medial tokens of *also*, which has been taken by formal semanticists to be a presupposition carrier and is worth analysing further in this light.

Table 1 also reports on the extent of Arg1, by considering (a) whether Arg1 is contained in a single sentence (Single) or spans across multiple sentences (Mult), and (b) whether it comprises the complete sentence(s) (Full) or only part of the sentence(s) (Partial). The first thing to note is that

tribution also provides the PDTB-2.0 annotation manual (PDTB-Group, 2008), links to some tools for browsing and querying the corpus, and papers and slides related to the project.

[4] While Arg2 of an Explicit connective always appears in the same sentence as the connective, there are a few cases (114/18459) where it extends beyond the connective's sentence to include additional sentences in the subsequent discourse.

	SingleFull	SinglePartial	MultFull	MultPartial	Total
SS	0	11224	0	12	11236
IPS	3192	1880	370	107	5549
NAPS	993	551	71	51	1666
FS	2	0	1	5	8
Total	4187	13655	442	175	18459

Table 1: Distribution of the location (rows) and extent (columns) of Arg1 of Explicit connectives. SS = same sentence as the connective; IPS = immediately previous sentence; NAPS = non-adjacent previous sentence; FS = some sentence following the sentence containing the connective; SingleFull = Single Full sentence; SinglePartial = Part of single sentence; MultFull = Multiple full sentences; MultPartial = Parts of multiple sentences.

even when Arg1 is in the same sentence (SS) as its connective, it may still extend across multiple sentences (cf. SS/MultPartial). Arg1 is also seen to extend across multiple sentences when it is located in the immediately previous sentence (cf. IPS/MultFull and IPS/MultPartial), a non-adjacent previous sentence (cf. NAPS/MultFull and NAPS/MultPartial) and in a subsequent sentence (cf. FS/MultFull and FS/MultPartial). Secondly, Arg1 is found to comprise partial sentences in all cases, to a significant extent: 100% for SS, 37% for IPS, 36% for NAPS, and 62.5% for FS. Further studies on partial sentence Arg1s require further study, but we note here that although the proportion of partial SS Arg1s is expectedly high since the connective and Arg2 are also in the same sentence, Dinesh et al. (2005) provide initial evidence that SS Arg1s can be even less than the material that remains when the connective and Arg2 are removed.

Discourse relations are not always realized as Explicit connectives, and are left to be *inferred* by the reader in many cases. In Example (6), a causal relation is inferred between *raising cash positions to record levels* and *high cash positions helping to buffer a fund*, even though no Explicit connective appears in the text to express this relation.

(6) But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls.**

In the PDTB, such inferred relations are annotated by *inserting* a connective expression — called an "Implicit" connective — that best expresses the inferred relation. So in Example (6), the Implicit connective *because* was inserted to capture the inferred causal relation.

The decision to *lexically encode* inferred relations in this way was made with the aim of achieving high reliability among annotators while avoiding the difficult task of training them to reason about pre-defined abstract relations. The annotation of inferred relations was thus done intuitively, and involved reading adjacent sentences (but in the context of the entire prior text), making a decision about whether or not a relation could be inferred between them, and providing an appropriate Implicit connective to express the inferred relation, if any. Three distinct labels, "AltLex", "EntRel" and "NoRel", were used for cases where an Implicit connective could *not* be provided: AltLex for cases where the insertion of an Implicit connective to express an inferred relation led to a *redundancy* due to the relation being alternatively lexicalized by some non-connective expression (Example 7); EntRel for cases where only an *entity-based coherence* relation could be perceived between the sentences (Example 8); and NoRel for cases where no discourse relation or entity-based relation could be perceived between the sentences (Example 9).

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

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

(9) *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million, according to Intel**.

Implicit connectives are annotated between all successive pairs of sentences within paragraphs, but they are also annotated intra-sententially between complete clauses delimited by semi-colon (";") or colon (":").

PDTB Relations	No. of tokens
Explicit	18459
Implicit	16224
AltLex	624
EntRel	5210
NoRel	254
Total	40600

Table 2: Distribution of Relations in PDTB-2.0

There are a total of 40600 tokens of relations annotated in PDTB-2.0. Table 2 gives the distribution of the relations annotated variously as Explicit, Implicit, AltLex, EntRel, and NoRel. There are 100 types of Explicit connectives annotated, with their modified forms treated as belonging to the same type as the unmodified forms. Types for the Implicit connectives total 102.[5]

[5]The token totals for the Implicit connectives differ depending on whether multiple implicit connectives (Webber et al., 2003) annotated between adjacent sentences are counted together or sepa-

Discourse relations and their arguments in the entire corpus were annotated first by two annotators per token and then adjudicated by a team of four experts. Inter-annotator agreement measured agreement between the extent of the argument spans selected, and was assessed on a subset of the corpus. With an *exact match* metric, agreement achieved was 90.2% for the arguments of Explicit connctives, and 85.1% for the arguments of the Implicit connectives. With a *partial match* metric, agreement achieved was 94.5% for the former, and 92.6% for the latter.[6] A more detailed study of the inter-annotator disagreements and reliability metrics is given in Miltsakaki et al. (2004).

3. Sense Annotation

PDTB provides sense annotations for the Explicit connectives, Implicit connectives, and AltLex relations. Just like verbs, discourse connectives can have more than one sense, depending at least on the context and the content of the arguments. For example, *since* seems to have three senses, one purely 'Temporal' (10), another purely 'Causal' (11) and a third both 'Causal' and 'Temporal' (12).

(10) *The Mountain View, Calif., company has been receiving 1,000 calls a day about the product* since **it was demonstrated at a computer publishing conference several weeks ago**.

(11) *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand**.

(12) *Domestic car sales have plunged 19%* since **the Big Three ended many of their programs Sept. 30**.

In such cases, the purpose of the sense annotation is to indicate which sense holds. In all cases, sense tags provide a semantic description of the relation between the arguments of connectives. Multiple sense tags are provided when the annotators have identified more than one simultaneous interpretation. Since arguments may also be related in ways that do not have corresponding sense tags, sense annotation specifies at least one, but not necessarily all, the semantic relations holding between the arguments.

The tagset of senses is organized hierarchically (cf. Figure 1), with three levels: *class*, *type* and *subtype*. The top level (*class*) has four types representing four major semantic classes: "TEMPORAL", "CONTINGENCY", "COMPARISON" and "EXPANSION". For each class, a set of *types* is defined to further refine the sense. For example, "CONTINGENCY" has two types "Cause" (relating two situations via a direct cause-effect relation) and "Condition" (relating a hypothetical scenario with its possible consequences). A third level of *subtype* specifies the semantic contribution of each argument. For "CONTINGENCY", its "Cause" type has two subtypes -"reason" (which applies when the connective indicates that the situation specified in Arg2 is interpreted as the cause of the situation specified in Arg1, as with the connective *because*) and "result" (which is used when the connective indicates that the situation described in Arg2 is interpreted as the result of the situation presented in Arg1, as with *as a result*).

Connectives can also be used to relate the *use* of the arguments of a connective to one another or the use of one argument with the sense of the other. For these *rhetorical* or *pragmatic* uses of connectives, we have defined *pragmatic* sense tags – specifically, "Pragmatic Cause", "Pragmatic Condition", "Pragmatic Contrast" and "Pragmatic Concession" (shown in italics in Figure 1).

For most types and subtypes, we also provide some hints about their possible formal semantics in the PDTB-2.0 manual citepdtb-2-tech-rept08. Although these formal descriptions are not provided in the annotation, they can be viewed as a starting point for the definition of an integrated logical framework able to deal with the semantics of discourse connectives. The descriptions were also meant to help annotators choose proper sense tags.

The hierarchical organization of the sense tags serves two purposes. First, it addresses well-known issues of inter-annotator reliability, by allowing the annotators to select a tag they are comfortable with: They are not forced to make fine distinctions between sense tags if they are not confident that there is sufficient information in the discourse to support them. Secondly, useful inferences can be derived at all levels. For example, when the CLASS tag "TEMPORAL" is used, as in (13), it is inferred that it cannot be determined whether the events described in Arg1 and Arg2 are ordered or overlapping. More generally, a higher level tag allows the inference that a more specified interpretation cannot be determined, and a lower lever tag allows the inference that interpretations described by its sister tags cannot be entertained simultaneously.

(13) *Fujitsu Ltd.'s top executive took the unusual step of publicly apologizing for his company's making bids of just one yen for several local government projects,* while **computer rival NEC Corp. made a written apology for indulging in the same practice**.

The hierarchical organization of sense labels in PDTB reflects our understanding that there is a small, core set of semantic relations that can hold between the situations described in the arguments of connectives (Kehler, 2002). This core set of relations, represented at the CLASS level in PDTB, can be refined in many ways, yielding an open set of possible relations. An important difference between positing a list of relations and a hierarchically related set is that in a hierarchical structure added relations inherit the core meaning of their parent. This is the basic difference between PDTB and the RST-style (Mann and Thompson, 1988) of labeling used for the RST-corpus (Carlson et al., 2001). The decisions made for the specific number of levels and tags in PDTB were guided by the data and the nature of the task.

The PDTB corpus was sense annotated by two annotators, with inter-annotator agreement computed for the three tagging levels. At class level, we noted disagreement when the two annotators picked a subtype, type or class tag of different classes. At type level, we noted disagreement when the annotators picked different types of the same class,

rately. Table 2 reports the total for when they are counted separately. When counted together as a single token, the total is 16053, which accounts for 171 instances of multiple connectives.

[6] The identification of AltLex, EntRel and NoRel as essential categories was developed during the adjudication stage, to account for cases where an Implicit connective could not be provided.

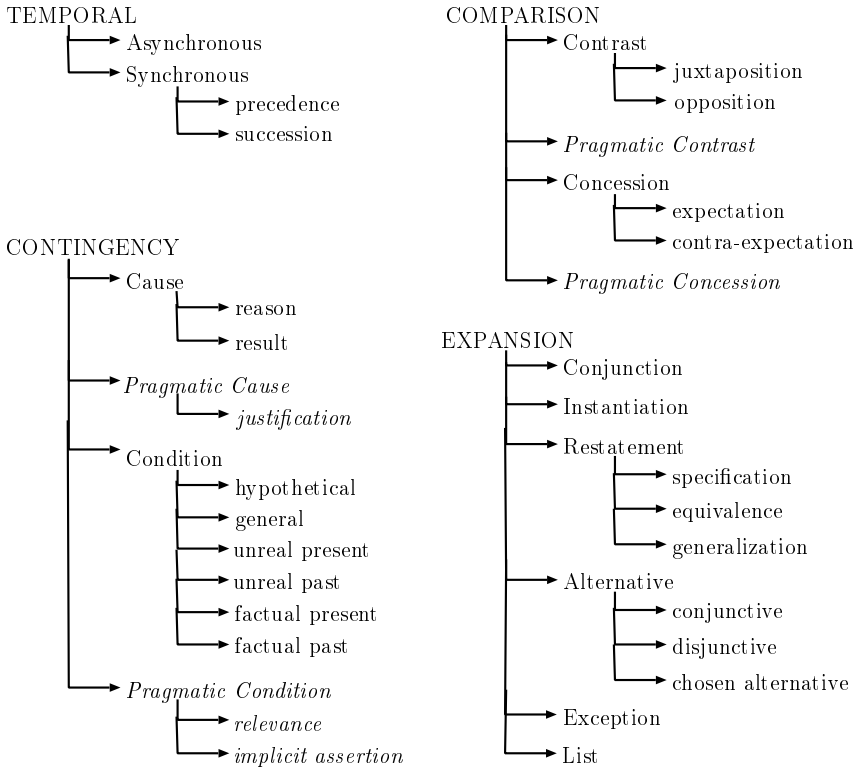

Figure 1: Hierarchy of sense tags

e.g., "Contrast" vs "Concession". Cases when one annotator picked a class level tag, e.g., "COMPARISON", and the other picked a type level tag of the same class, e.g., "Contrast", did not count as disagreement. At the subtype level, disagreement was noted when the two annotators picked different subtypes, e.g., "expectation" vs. "contra-expectation". Higher level disagreement was counted as disagreement at all the levels below. Inter-annotator agreement is shown in Table 3. Percent agreement, computed for five sections (5092 tokens), is shown for each level. Agreement is high for all levels, ranging from 94% at the class level to 80% at the subtype level.

Class level disagreement was adjudicated by a team of three experts. Disagreement at lower levels was resolved by providing a sense tag from the immediately higher level. For example, if one annotator tagged a token with the type "Concession" and the other, with the type "Contrast", the disagreement was resolved by applying the higher level tag "COMPARISON". This decision was based on the assumption that both interpretations were possible, making it hard to determine with confidence which one was intended.

LEVEL	% AGREEMENT
CLASS	94%
TYPE	84%
SUBTYPE	80%

Table 3: Inter-annotator agreement

Table 4 shows the distribution of "CLASS" level tags in the corpus. Each "CLASS" count includes all the annotations of the specified "CLASS" tag and all its types and subtypes. The total of Explicit, Implicit and AltLex tokens is shown in parentheses at the top row. The total of sense tags applied to these categories is shown at the bottom of the table. The numbers differ because some tokens may have been annotated with two senses.

Table 5 shows the top ten most polysemous connectives and the distribution of their sense tags. The total number of tokens whose sense tags occurred less than ten times are shown as *other*. The connectives *after*, *since* and *when*, which typically relate non-simultaneous situations, are ambiguous between "TEMPORAL" and "CONTINGENCY" senses. The connectives *while* and *meanwhile*, which typically relate simultaneous situations, are ambiguous between the "TEMPORAL" and "COMPARISON" senses. The connectives *but*, *however* and *although* are ambiguous between the "Contrast" and "Concession" types and subtypes of "COMPARISON" but rarely between different classes of senses. The connective *if* is ambiguous between subtypes of "Condition" and some pragmatic uses.

4. Attribution Annotation

Recent work (Wiebe et al., 2005; Prasad et al., 2005) has shown the importance of attributing beliefs and assertions expressed in text to the agent(s) holding or making them. Such attributions are a common feature in the PDTB corpus which belongs to the news domain. Since the discourse relations in the PDTB are annotated between abstract objects, with the relations themselves denoting a class of abstract objects (called "relational propositions" (Mann and Thompson, 1988)), one can distinguish a variety of cases depending on the *attribution* of the discourse relation or its arguments: that is, whether the relation and its arguments are attributed to the writer (e.g., attribution to the writer in

"CLASS"	**Explicit** (18459)	**Implicit** (16224)	**AltLex** (624)	**Total**
"TEMPORAL"	3612	950	88	4650
"CONTINGENCY"	3581	4185	276	8042
"COMPARISON"	5516	2832	46	8394
"EXPANSION"	6424	8861	221	15506
Total	19133	16828	634	36592

Table 4: Distribution of "CLASS" sense tags

Connective	**Senses**
after	succession (523), succession-reason (50), other (4)
since	reason (94), succession (78), succession-reason (10), other (2)
when	Synchrony (477), succession (157), general (100), succession-reason (65), Synchrony-general (50), Synchrony-reason (39), hypothetical (11), implicit assertion (11), Synchrony-hypothetical (10), other (69)
while	juxtaposition (182), Synchrony (154), Contrast (120), expectation (79), opposition (78), Conjunction (39), Synchrony-juxtaposition (26), Synchrony-Conjunction (21), Synchrony-Contrast(22), COMPARISON (18), Synchrony-opposition (11), other (31)
meanwhile	Synchrony-Conjunction (92), Synchrony (26), Conjunction (25), Synchrony-juxtaposition (15), other(35)
but	Contrast (1609), juxtaposition (636), contra-expectation (494), COMPARISTON (260), opposition (174), Conjunction (63), Conjunction-Pragmatic contrast (14), Pragmatic-contrast (14), other (32)
however	Contrast (254), juxtaposition (89), contra-expectation (70), COMPARISON (49), opposition (31), other (12)
although	expectation (132), Contrast (114) juxtaposition (34), contra-expectation (21), COMPARISON (16), opposition (9), other (2)
and	Conjunction (2543), List (210), result-Conjunction (138), result (38), precedence-Conjunction (30), juxtaposition (11), other(30)
if	hypothetical (682), general (175), unreal present (122), factual present (73), unreal past (53), expectation (34), implicit assertion (29), relevance (20), other (31)

Table 5: Top ten polysemous connectives (Explicit)

Example (14)) of the text or someone other than the writer (e.g., attribution to Bill Biedermann in Example (15)), as well as whether the relation and its arguments are attributed differently to different sources (e.g., attribution of relation and Arg1 to writer, and Arg2 to the purchasing agents in Example (16)). (In the examples, text spans corresponding to the attribution phrase are shown boxed.)

(14) Since **the British auto maker became a takeover target last month**, its ADRs have jumped about 78%.

(15) "*The public is buying the market* when **in reality there is plenty of grain to be shipped**," said Bill Biedermann, Allendale Inc. director .

(16) *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**.

Furthermore, as shown in Dinesh et al. (2005), attribution is a major source of the mismatch between the syntactic and discourse arguments of connectives. That is, simply taking the syntactic arguments of *while* in Example (16) to be its discourse arguments yields an incorrect semantic interpretation unless the attribution is factored out of the description. Thus recognizing attributions within discourse relations is an important task for deriving the correct interpretation of the relations. In the PDTB, each discourse relation and its two arguments are annotated for attribution.[7]

The annotation scheme marks the text span corresponding to the attribution phrase, and isolates four key properties of attribution, to be annotated as features:

Source. The "source" feature distinguishes between different types of agents: (a) the writer of the text ("Wr"), (b) some specific agent introduced in the text ("Ot" for other), and (c) some arbitrary ("Arb") individual(s) indicated via a non-specific reference in the text.
Results from preliminary annotations show that a significant proportion (34%) of the annotated discourse relations – for both explicit and implicit connectives – have some non-writer agent as the source, for either the relation or one or both arguments. Thus one cannot simply attribute discourse relations and their arguments to the writer of news text, without being wrong two-thirds of the time. The distributions also show that there are a large number of cases in which the components of the relations are attributed to different sources, suggesting that recognition of attributions may be a complex task for which an annotated corpus will be useful.

Type. The "type" feature encodes the nature of the relationship between agents and abstract objects, by distinguishing abstract objects into four sub-types: *assertion propositions*, *belief propositions*, *facts* and *eventualities*. Both assertion propositions and belief propositions involve

[7] Attribution is annotated for Explicit connectives, Implicit connectives, and AltLex relations but not for EntRel and NoRel, since the former but not the latter indicate the presense of discourse relations.

attribution to an agent of his/her commitment towards the truth of a proposition, but they differ in the degree of the commitment. Facts involve attribution to an agent of an evaluation towards or knowledge of a proposition whose truth is taken for granted (i.e., a presupposed proposition). And eventualities, when they occur with attribution, convey an agent's intention/attitude towards a considered event, state, or action. In the annotation scheme, each of these are annotated in terms of the type of attributive expressions used to convey the type of the abstract object: (a) verbs of communication (annotated as "Comm") (Levin, 1993), such as *say*, *mention*, *claim*, *argue*, *explain* etc., for assertions; (b) propositional attitude verbs (annotated as "PAtt") (Hintikka, 1971), such as *believe*, *think*, *expect*, *suppose*, *imagine* etc., for beliefs; (c) factive and semi-factive verbs (annotated as "Ftv") (Kiparsky and Kiparsky, 1971; Karttunen, 1971), such as *regret*, *forget*, *remember*, *know*, *see*, *hear* etc., for facts, and (d) control verbs (annotated as "Ctrl") (Sag and Pollard, 1991), such as *persuade*, *permit*, *promise*, *intend*, *want*, *expect* etc., for eventualities. Looking at the "type" of the relations in PDTB-2.0, with the Explicit, Implicit and AltLex tokens counted together (35136 tokens),[8] assertions account for a majority of the relations (34544/35136), with the other types distributed as follows: 254 for beliefs, 86 for factives, and 252 for eventualities.

Scopal polarity. The "scopal polarity" feature is annotated on relations and their arguments to identify cases when verbs of attribution are negated on the surface - syntactically (e.g., *didn't say*, *don't think*) or lexically (e.g., *denied*), but when the negation in fact **reverses** the polarity of the attributed relation or argument content (Horn, 1978). Example (17) illustrates such a case. The *but* clause entails an interpretation such as "I think it's not a main consideration", for which the negation must take narrow scope over the embedded clause rather than the higher clause.

(17) "*Having the dividend increases is a supportive element in the market outlook,* but I don't think **it's a main consideration**," he says.

To capture such entailments, "Neg" is marked on a relation or argument when surface negation on a higher attribution takes narrow scope over it. Thus, in Example (17), *scopal polarity* is marked as "Neg" for Arg2. "Null" is used as the default, in the absence of neg-lowered interpretations. For Explicit connectives, "Neg" is annotated in PDTB-2.0 a total of 35 times: on the relation in 3 instances, on Arg1 in 20 instances, and on Arg2 in 12 instances. For Implicit connectives, "Neg" is annotated in 44 cases, none of which involve the relation itself: 23 of these are marked on Arg1 and 24 on Arg2. "Neg" does not appear at all for AltLex relations.

Determinacy. The "determinacy" feature captures the fact that the attribution over a relation or argument can itself be cancelled in particular contexts, such as within the scope of negations, conditionals, or infinitivals. Such indeterminacy is indicated by the value "Indet", while determinate attributions are simply marked by the default "Null". The annotation in Example (18) illustrates a case of indeterminacy of the (belief) attribution on the relation. Here, it is not that a belief or opinion about "our teachers educating our children better if only they got a few thousand dollars a year more" is being attributed to anyone. Rather, the attribution is only being conjectured as a possibility. For Explicit connectives, "Indet" is annotated in PDTB-2.0 a total of 109 times: on the relation in 83 instances, on Arg1 in 22 instances, and on Arg2 in 4 instances. For Implicit connectives, "Indet" is annotated in only 5 cases, none of which involve the relation itself: 2 of these are marked on Arg1 and 3 on Arg2. "Indet" does not appear at all for AltLex relations.

(18) It is silly libel on our teachers to think *they would educate our children better* if only **they got a few thousand dollars a year more**.

The PDTB was annotated for attribution after the annotation of the argument structure and before the annotation of senses. It was done by a single expert. We noted in Section 1. that attribution is not treated as a discourse relation because the objects related by attribution are not the same as those related by discourse relations: the former relate agents and abstract objects, whereas the latter relate two abstract objects. One confounding consequence of treating them the same way in the RST-Bank (Carlson et al., 2001) and GraphBank (Wolf and Gibson, 2005) is that the resulting structures are unecessarily complex. Attribution arcs in (Wolf and Gibson, 2005) account for a significant proportion of structural crossings in the discourse structure. However, the approach taken in the PDTB shows these crossings to be spurious, once attribution is appropriately represented.

5. Differences between PDTB-1.0. and PDTB-2.0

Differences between PDTB-1.0 and PDTB-2.0 include:

1. Annotations of implicit relations across the entire corpus. In PDTB-1.0, implicit relations were annotated for only 3 WSJ sections.
2. Revision of the annotation scheme for senses. PDTB-1.0 used a simpler 7-way broad classification.
3. Sense annotations of all Explicit connectives, Implicit connectives, and AltLex relations. In PDTB-1.0, sense annotations were provided for only implicit relations.
4. Revision of the annotation scheme for attribution. PDTB-1.0 used a simpler scheme, making no distinctions between abstract object types, as well as between the different scopal behaviors distinguished in PDTB-2.0 as scopal polarity and determinacy.

Complete descriptions of schemes used across the two versions of the corpus can be found in the annotation manuals released along with the two versions (PDTB-Group, 2006; PDTB-Group, 2008). While PDTB-1.0 is no longer available, the differences listed above may be useful for people who have initiated research on the earlier version. More detailed distributions of annotations in PDTB-2.0 can be also found in the Appendix sections of the PDTB-2.0 manual.

[8] Here, multiple Implicit connectives are counted together - see Footnote 5.

6. Conclusion

We described the second version of the Penn Discourse Treebank, PDTB-2.0, an annotated corpus of discourse relations and their arguments over the 1 million word Wall Street Journal corpus, and containing alignments with the Penn Treebank annotated over the same corpus. PDTB-2.0 was released by the Linguistic Data Consortium in February 2008. Future work will focus *inter alia* on using it to address questions of the complexity of discourse structure in relation to sentence structure, and for natural language applications, such as summarization and generation.

7. Acknowledgements

This work was partially supported by NSF grants EIA-02-24417, EIA-05-63063, and IIS-07-05671.

8. References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*.

Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Katherine Forbes-Riley, Bonnie Webber, and Aravind Joshi. 2006. Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, 23:55–106.

Jaakko Hintikka. 1971. Semantics for propositional attitudes. In L. Linsky, editor, *Reference and Modality*, pages 145–167. Oxford.

Laurence Horn. 1978. Remarks on neg-raising. In Peter Cole, editor, *Syntax and Semantics 9: Pragmatics*. Academic Press, New York.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications, Stanford.

Carol Kiparsky and Paul Kiparsky. 1971. Fact. In D. D. Steinberg and L. A. Jakobovits, editors, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, pages 345–369. Cambridge University Press, Cambridge.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories*.

Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotation and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*.

Martha Palmer, Daniel Guildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for NLG*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. Annotating attribution in the Penn Discourse TreeBank. In *Proceedings of the COLING/ACL Workshop on Sentiment and Subjectivity in Text*, pages 31–38.

Ivan A. Sag and Carl Pollard. 1991. An integrated theory of complement control. *Language*, 67(1):63–113.

PDTB-Group. 2006. The Penn Discourse TreeBank 1.0 Annotation Manual. Technical Report IRCS-06-01, Institute for Research in Cognitive Science, University of Pennsylvania.

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.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree-adjoining grammar for discourse. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 86–92. Association for Computational Linguistics.

Bonnie Webber and Rashmi Prasad. To appear. Discourse structure: Swings and roundabouts. In Cathrine Fabricius-Hansen and Bergljot Behrens, editors, *Linguistics, Special issue on Structuring Information in Discourse: the Explicit/Implicit Dimension*.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

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