A Short Introduction to the Penn Discourse TreeBank

Bonnie Webber, Aravind Joshi, Eleni Miltsakaki, Rashmi Prasad,
Nikhil Dinesh, Alan Lee, and Katherine Forbes[*]
bonnie@inf.ed.ac.uk

Abstract

Taking discourse connectives to be the predicates of binary discourse relations, the goal of Penn Discourse Treebank (PDTB) is to annotate the million word WSJ corpus in the Penn TreeBank with each of its discourse connectives and their arguments. The paper describes the linguistic observations and ideas that led to the PDTB, the decisions that shaped its content and the tools used in its development, its current manifestation, some issues surrounding the concept of a *discourse adverbial*, and finally, some thoughts about the future of the PDTB.

[*] Webber is at the School of Informatics, University of Edinburgh. Joshi, Miltsakaki, Prasad, Dinesh and Lee are at the Institute for Research in Cognitive Science, University of Pennsylvania. Forbes is at Dartmouth University. The Penn Discourse Treebank (PDTB) project has been supported by NSF Grant: Research Resources, EIA 02-24417 to the University of Pennsylvania (PI: Joshi).

1 Introduction

The overall goal of the Penn Discourse Treebank (PDTB) is to annotate the million word WSJ corpus in the Penn TreeBank (Marcus et al., 1993) with a layer of discourse annotation. Previous reports on this project have been presented in (Miltsakaki et al., 2004a), (Miltsakaki et al., 2004b) and (Prasad et al., 2004), where we described our annotation of discourse connectives (both explicit and implicit) and their (clausal) arguments and some early experiments on the data.

Although the idea of annotating connectives and their arguments comes from our theoretical work on discourse connectives in the framework of *lexicalised grammar* (Webber et al., 2003), the corpus itself is not tied to any particular theory. Rather, taking discourse connectives to be the predicates of binary discourse relations, the goal of the PDTB is to annotate the arguments for each token of each discourse connective[1] – for example,

(1) Even though critical, *it was just the kind of attention they were seeking*. So **they fired back at the Goldman Sachs objections in their own economics letter, “The BMC Report.”**

Here, the two arguments to this token of the conjunction *so* have been marked: the first argument being the italicised clause, and the second being the one in bold.

There are four basic benefits to be gained from a resource like the PDTB:

1. It articulates a clearly defined and relatively easily identifiable level of discourse structure, which is independent of any particular discourse theory.

2. By being able to compare this discourse annotation with syntactic annotation, we can get a better understanding of the relationship between syntactic structure and discourse structure, and hence of the relationship between clausal and discourse semantics.

[1] This is not unlike the annotation of sentence-level predicate argument structure in **PropBank** (Palmer et al., 2005), which annotates the explicit arguments for each token of each verb in a corpus.

3. It can serve as a basis for inference, contributing to more complex NLP tasks such as Question Answering, Natural Language Generation and Machine Translation.

4. It should serve as a resource for the development of robust automatic procedures for identifying connectives and their arguments.

In the rest of this paper, we will briefly describe the linguistic observations and ideas that led to the PDTB (Section 2), the decisions that shaped the content of the PDTB and the tools used in its development (Section 3) and the PDTB itself (Section 4). Section 5 discusses some issues surrounding the concept of a *discourse adverbial*. We conclude with some thoughts about the future of the PDTB.

2 Linguistic Observations and Ideas that Led to the PDTB

Informally, a range of different words and phrases have been taken to be *discourse connectives* that link together the content or purpose of adjacent textual spans – for example, the subordinate conjunction *while* in Example 2, the adverbial *otherwise* in Example 3, and the prepositional phrase (PP) *as a result* in Example 4.

(2) John eats porridge for breakfast, while Mary eats muesli.

(3) Eat your porridge. Otherwise you're not going to football practice.

(4) You've eaten your porridge every day this week. As a result, I'm going to give you the iPOD I promised you.

But, as the next set of examples show, even discourse connectives in the same matrix clause don't necessarily link the same discourse elements. In Example (5), there are two adjacent discourse connectives – the subordinate conjunction *because* and the adverbial *then*.

(5) a. John loves Barolo. So he ordered three cases of the '97. But *he had to cancel the order* because **then he discovered he was broke**.

b. John loves Barolo. So *he ordered three cases of the '97*. But he had to cancel the order because then **he discovered he was broke**.

Here (5a) shows the two arguments to *because* (the cancelling and the discovery, the former taken to be the result of the latter), and (5b) shows the two arguments of *then* (the ordering and the discovery, the latter following on from the former). As can be seen, their first arguments are not the same.

Similarly, there are two discourse connectives in the text shown in Example (6) – the conjunction *but* and the adverbial *instead*.

(6) a. *Buyers can look forward to double-digit annual returns if they are right*. <u>But</u> **they will have disappointing returns or even losses if interest rates rise instead**.

b. Buyers can look forward to double-digit annual returns if *they are right*. But they will have disappointing returns or even losses if **interest rates rise** <u>instead</u>.

Here (6a) shows the two arguments to *but* recorded by the PDTB annotators (having disappointing results and looking forward), while (6b) shows the two quite different arguments to *instead* (interest rates rising instead of the buyers being right).

Since any higher-level or interpretative theory of discourse structure should reflect what it is that discourse connectives actually connect, what can be learned from the PDTB will provide a basis for identifying these discourse-level predicate-arguments patterns in other texts as well.

3 Decisions that have shaped the PDTB and its Annotation Tool

Having decided to annotate the discourse connectives in a corpus, we were faced with several further decisions, including (1) what corpus to annotate; (2) how to organise the process of annotation; (3) what to treat as connectives; and (4) what form of annotation to use. We address each of these decisions briefly in this paper. More extensive discussion can be found in the guide to PDTB Annotation at http://www.cis.upenn.edu/˜pdtb/manual/pdtb-tutorial.pdf.

3.1 What corpus to annotate?

From the start, we decided to annotate the same Penn WSJ corpus as the Penn TreeBank (Marcus et al., 1993) and PropBank (Palmer et al., 2005), as this would then permit alignment of three types of annotation (syntactic, semantic and discourse) when annotation was complete.

The Penn WSJ corpus exists in several forms: "raw" tokenised text, text tagged with part-of-speech, and TreeBank parsed text. We decided to annotate over the raw text, rather than syntactic trees. Although this then required the annotators to recognise and reject tokens that were not functioning as connectives – for example, tokens of *when* as a relative pronoun, as in

(7) Georgia-Pacific's sale climed to $9.5 billion last year, compared with $6 billion in 1983, WHEN Mr. Hahn took the reins.

or as the head of an argument to a verb, as in

(8) The maker of chemical and industrial material didn't say how much it would pay or WHEN it would make the transactions.

still, by annotating over raw text, the PDTB could avoid errors and/or inconsistencies in the PTB, and allow for cases where discourse arguments would not align with syntactic structures (Section 4.2; also Dinesh et al., 2005).

3.2 How to organise the process of annotation?

We decided to have the annotators proceed through the corpus, one connective at a time, annotating every instance of the word or phrase that functioned as a connective in the corpus before going on to the next connective in the list. Although this meant that no part of the corpus would be fully annotated until all connectives were covered, it had the benefit of speeding annotation by allowing annotators to immediately exploit the experience they were gaining in annotating a connective.

Annotation was divided into ten phases, in each of which the annotators were given a new list of words and phrases that can function as discourse connectives. They would proceed sequentially through the list. For the current word or phrase, they would be led through the corpus by the annotation tool **WordFreak** (Figure 2) to the next instance

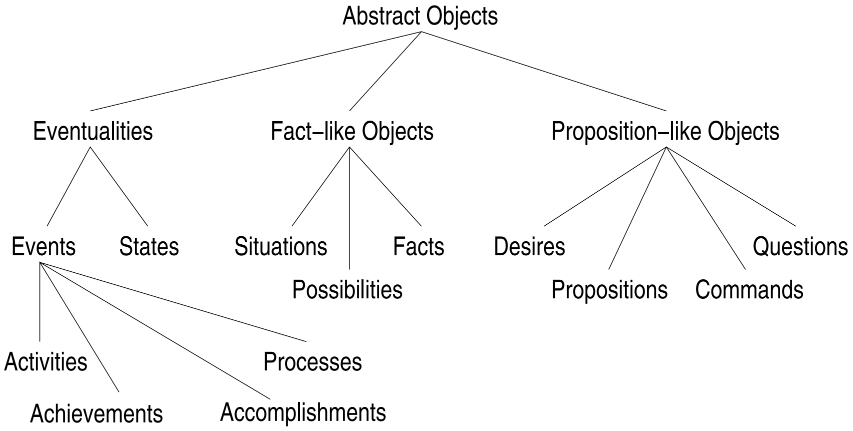


Figure 1: Hierarchy of Abstract Objects (from Asher, 1993)

of that word or phrase. At this point, they could either annotate its arguments, or reject it as not functioning as a discourse connective. When all instances of that word or phrase in the corpus were exhausted, they would turn to the next word or phrase in the list. When annotation of all connectives on the list was complete, they could move on to the list selected for the next phase.

In none of the work we carried out prior to starting annotation of the PDTB did we find an English discourse connective that had anything other than **two** arguments. (In this, discourse connectives are unlike verbs, which can and do take more than two arguments.) Thus, the process of annotation consistently was one of identifying the two arguments to each token of the connective currently being annotated.

Because there are, as yet, no generally accepted *abstract semantic categories* for classifying the arguments to discourse connectives as there are for verbs (eg., *agent*, *patient*, *theme*, etc.), the two arguments to a discourse connective were simply labelled *Arg2*, for the argument that appears in the clause that is syntactically bound to the connective, and *Arg1*, for the other argument. In examples used in this paper, the text whose interpretation is the basis for *Arg1* appears in italics, while that of *Arg2* appears in bold.

3.3 What to consider connectives?

Formally, we take a discourse connective to be a word, phrase or pair of phrases whose interpretation conveys a semantic relationship between two *Abstract Objects* (Asher, 1993) of contextually appropriate types (Figure 1). While every clause has an abstract object (AO) interpretation (and often, more than one), other syntactic entities can also have AO interpretations, including nominalisations, discourse deictics (*this, that*), sentences, and even sentence sequences. The classes of words or phrases conveying relationships between such syntactic entities include:

- Subordinating conjunctions (e.g. *when, because, as soon as, now that,* etc.), both bare and with a modifier (e.g. *only if, just because, even though, mainly when*)
- Coordinating conjunctions (*and, but, or, nor*)
- Subordinators (e.g. *provided (that), in order that, except (that), notwithstanding that*[2]
- Discourse adverbials, including both adverbs (e.g. *instead, therefore*), and prepositional phrases (e.g. *on the other hand, as a result*).

Over 90 different connectives have already been annotated. Section 5 contains a brief discussion of how discourse adverbials were identified.

For practical reasons, we decided not to annotate those PP discourse adverbials containing a noun phrase (NP) with a demonstrative determiner (e.g., *for this reason, in this respect, in that case*) or any comparative adverbial other than *earlier* or *later* (e.g., *more importantly, less controversially*). This is because we assumed that the former would be caught if and when all demonstrative pronouns and NPs in the Wall Street Journal corpus were annotated for coreference, and the latter, if and when all comparatives were annotated for the target they were being compared to.

[2] (Huddleston and Pullum, 2002) takes all conjunctions and subordinators to be *prepositions* that can take clauses, as well as various types of phrases, as arguments.

Also for practical reasons, we decided to postpone annotation of most subordinators. However, we are exploring the possibility of extracting their arguments (semi-)automatically from the Penn TreeBank (PTB), similar to the coordinating conjunctions appearing in VP coordinations (which have also not yet been annotated manually).

Finally, it is important to note that PDTB does not annotate *cue phrases* used for discourse management such as *now, so, anyway,* and *well,* since in managing the discourse, they do not convey a particular semantic relationship between clausal (AO) interpretations. In the case of *so,* this means that the annotators have to exclude tokens functioning as an intensifier (“so large”), or as part of a VP “do so” construction, or as a cue phrase (“So what should we do now?”), while annotating those tokens that function as connectives.

3.4 What form of annotation to use?

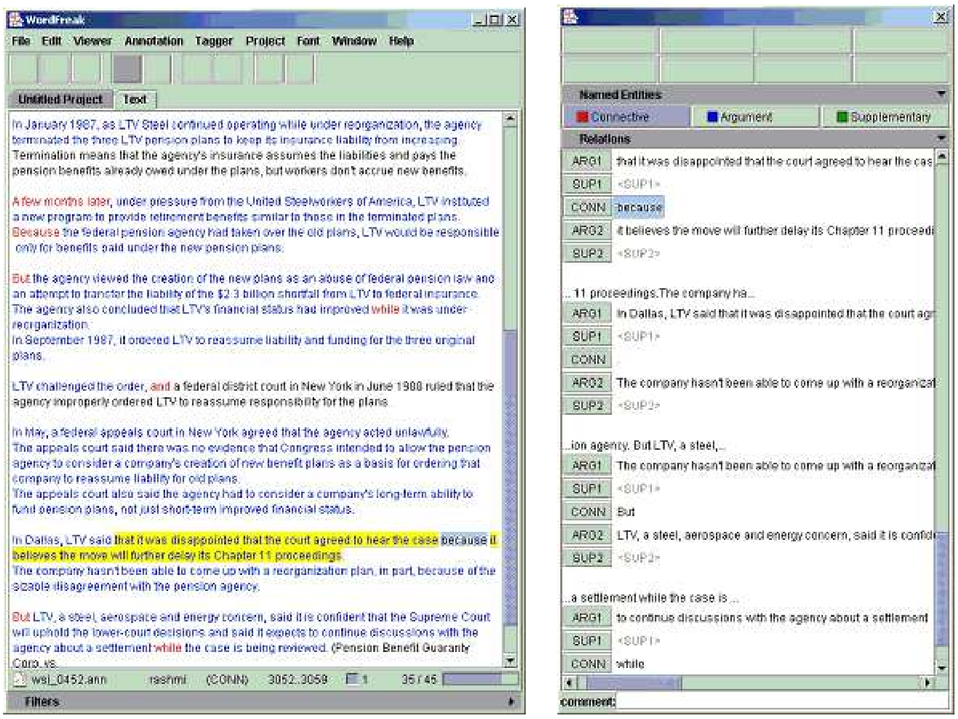
We decided to use standoff, rather than in-line (XML) annotation for the PDTB, not only because it would produce clearer annotation, but also because it was clear that the arguments of different connectives could overlap one another in ways that would violate the syntax of XML. Standoff annotation avoided this problem, as well as that of discontinuous arguments that annotators wanted to record (as in *Arg1* in Example 9) and of connectives occuring in disjoint pairs (as in Example 10).

(9) *But*, says Mr. Dinkins, *he did get an office*. So **he shouldn’t complain**.

(10) On the one hand, *Mr. Giuliani wants to cut into Mr. Dinkins’s credibility*. On the other, **he seeks to convince voters he’s the new Fiorello LaGuardia – affable, good-natured and ready to lead New York out of the mess it’s in**.

3.5 Annotation Tool

PDTB annotation was done using a customised version of the WordFreak tool developed by Tom Morton and Jeremy Lacivita, and available at http://sourceforge.net/projects/wordfreak. Additional tools for adjudication and technical corrections were developed by Alan Lee.



In January 1987, as LTV Steel continued operating while under reorganization, the agency terminated the three LTV pension plans to keep its insurance liability from increasing. Termination means that the agency's insurance assumes the liabilities and pays the pension benefits already owed under the plans, but workers don't accrue new benefits.

A few months later, under pressure from the United Steelworkers of America, LTV instituted a new program to provide retirement benefits similar to those in the terminated plans. Because the federal pension agency had taken over the old plans, LTV would be responsible only for benefits paid under the new pension plans.

But the agency viewed the creation of the new plans as an abuse of federal pension law and an attempt to transfer the liability of the $2.3 billion shortfall from LTV to federal insurance. The agency also concluded that LTV's financial status had improved while it was under reorganization.
In September 1987, it ordered LTV to reassume liability and funding for the three original plans.

LTV challenged the order, and a federal district court in New York in June 1988 ruled that the agency improperly ordered LTV to reassume responsibility for the plans.

In May, a federal appeals court in New York agreed that the agency acted unlawfully.
The appeals court said there was no evidence that Congress intended to allow the pension agency to consider a company's creation of new benefit plans as a basis for ordering that company to reassume liability for old plans.
The appeals court also said the agency had to consider a company's long-term ability to fund pension plans, not just short-term improved financial status.

In Dallas, LTV said that it was disappointed that the court agreed to hear the case because it believes the move will further delay its Chapter 11 proceedings.
The company hasn't been able to come up with a reorganization plan, in part, because of the sizable disagreement with the pension agency.

But LTV, a steel, aerospace and energy concern, said it is confident that the Supreme Court will uphold the lower-court decisions and said it expects to continue discussions with the agency about a settlement while the case is being reviewed. (Pension Benefit Guaranty Corp. vs.

wsj_0452.ann rashmi (CONN) 3052..3059 1 35 / 45

ARG1 that it was disappointed that the court agreed to hear the cas
SUP1 <SUP1>
CONN because
ARG2 it believes the move will further delay its Chapter 11 proceedi
SUP2 <SUP2>

... 11 proceedings.The company ha...
ARG1 In Dallas, LTV said that it was disappointed that the court agr
SUP1 <SUP1>
CONN .
ARG2 The company hasn't been able to come up with a reorganizat
SUP2 <SUP2>

...ion agency. But LTV, a steel,...
ARG1 The company hasn't been able to come up with a reorganizat
SUP1 <SUP1>
CONN But
ARG2 LTV, a steel, aerospace and energy concern, said it is confid
SUP2 <SUP2>

...a settlement while the case is ...
ARG1 to continue discussions with the agency about a settlement
SUP1 <SUP1>
CONN while

comment:

Figure 2: The WordFreak Tool

4 The PDTB itself

Annotation was initially done by four (4) separate annotators, each using a copy of **WordFreak**. After verifying the level of inter-annotator agreement (discussed briefly below), we decreased the quantity of parallel annotation to two annotators per example.

A study of the inter-annotator reliability achieved in annotating the arguments to connectives is presented in detail in (Miltsakaki et al., 2004a). Briefly, the study was carried out on the annotation of ten explicit connectives (five subordinate conjunctions and five adverbials), comprising 2717 tokens. An independent assessment of agreement was carried out on *Arg1* and for *Arg2* – that is, 2717 instances of *Arg1* and 2717 instances of *Arg2*. Overall, exact match agreement stood at 90.2% – 92.4% exact match agreement on subordinating conjunctions and 71.8% exact match agreement on adverbials. Assessment of partial overlap showed the annotator agreement at 94.5% overall.

When argument annotations overlapped but didn't match exactly, it was due to differential inclusion of:

- a clause's governing verb
- a dependent clause at the periphery of an argument
- a parenthetical in the middle of an argument – e.g.

(11) Annotator 1: Bankers said *warrants for Hong Kong stocks are attractive* because **they give foreign investors**, wary of volatility in the colony's stock market, **an opportunity to buy shares without taking too great a risk**.

(12) Annotator 2: Bankers said *warrants for Hong Kong stocks are attractive* because **they give foreign investors, wary of volatility in the colony's stock market, an opportunity to buy shares without taking too great a risk**.

Such differences among how much of a span an annotator took to constitute the argument to a connective led us to a *Minimality Principle*:

> Select as an argument only what is "minimally" necessary to interpret the relation established by the connective.
>
> Anything else that is felt to be useful information for that interpretation, mark as SUP1 (supplementary to ARG1) or SUP2 (supplementary to ARG2).

For the example above, this led to the inner parenthetical in *Arg2* being marked SUP2 – viz.

(13) Bankers said *warrants for Hong Kong stocks are attractive* because **they give foreign investors** ($_{\text{SUP2}}$wary of volatility in the colony's stock market), **an opportunity to buy shares without taking too great a risk**.

Example 14 illustrates an example of SUP1:

(14) Although **started in 1965**, *Wedtech didn't really get rolling until 1975* ($_{\text{SUP1}}$ when Mr. Neuberger discovered the Federal Government's Section 8 minority business program).

Even though we decided on practical grounds (Section 3.3) not to carry out certain annotation that would clearly fall within the remit of the PDTB, we also decided to expand its remit somewhat and annotate some features closely related to discourse connectives and their arguments. We felt that these additional features would both provide useful information and facilitate possible useful future directions of the PDTB.

4.1 Implicit Connectives

As many people have remarked, more often than not, there is **no** discourse connective explicitly connecting a clause to something in the previous discourse. So as a step towards establishing the conditions under which a particular semantic relation between *abstract object* (AO) interpretations is realised explicitly with a discourse connective and when not, we decided to look at sentence boundaries within a paragraph that are unmarked by any explicit connective. At those points, we would ask the annotators to record the connective that conveyed the implicit relationship that they saw as holding between arguments expressed in the adjacent sentences. (Annotators were also allowed to provide more than one connective if they perceived multiple "simultaneous" interpretations.) Thus, here annotators were recording both what they took to be *Arg1* and *Arg2* and what they took to be the one (or more) implicit connectives holding between them. Here are several examples:

(15) *The $6 billion that some 40 companies are looking to raise in the year ending March 31 compares with only $2.7 billion raised on the capital market in the previous fiscal year*. **In fiscal 1984 before Mr. Gandhi came to power, only $810 million was raised**.

(16) *The small, wiry Mr. Morishita comes across as an outspoken man of the world.* ($_{SUP2}$Stretching his arms in his silky white shirt and squeaking his black shoes) **he lectures a visitor about the way to sell American real estate and boasts about his friendship with Margaret Thatcher's son**.

(17) *"We like to make our own judgments" about Mr. Morishita,* says

Christopher Davidge, Christies' group managing director." **People have a different reputation country by country**."

(18) *The gruff financier recently started socializing in upper-class circles*. Although he says he wasn't keen on going, **last year he attended a New York gala where his daughter made her debut**.

In Example 15, annotators recorded the connective *in contrast* as expressing the relationship between the adjacent sentences, while in Example 16, one of the annotators recorded both *when* and *for example* as conveying the relationship taken to hold between the adjacent sentences. Both Examples 17 and 18 illustrate the fact that annotators do not need to mark the entire sentence on either side of the implicit connective(s) as serving arguments. Example 17 has been annotated with the implicit connective *because*, while Example 18 has been annotated with the implicit connective *for example*.

While annotating the implicit connective between adjacent sentences, for practical reasons, we decided to delay annotation of implicit connectives between adjacent clauses within the same sentence. This can happen in sentences containing both a main clause and one or more *free adjuncts*. As between adjacent sentences, different relationships can hold between these clauses (Webber and Di Eugenio, 1990).

(19) The market for export financing was liberalized in the mid-1980s, forcing the bank to face competition.

(20) Mr. Cathcart says he has had "a lot of fun" at Kidder, adding the crack about his being a "tool-and-die man" never bothered him.

So in Example 19, the event expressed in free adjunct is a consequence of that expressed in the main clause (which might be annotated with an implicit *so*), while in Example 20 the event expressed in the free adjunct merely follows that expressed in the main clause (which might be annotated with an implicit *then*). While these relations are clearly of interest, it was nevertheless felt that such annotation could wait until initial annotation of the PDTB was complete.

Also for practical reasons, we decided to only annotate *implicit connectives* between adjacent sentences in the same paragraph that had no *explicit connective* between them. While theoretically, the presence of a discourse adverbial in a sentence does not preclude the presense of either another explicit connective with the previous text (Example 21) or an implicit connective (Example 22), again our annotation resources (time, money and annotators) were not sufficient to do this as well.

(21) If the light is red, *stop* because <u>otherwise</u> **you'll get a ticket**.

(22) If the light is red, *stop*. <u>Otherwise</u> **you'll get a ticket**.

Additionally, the PDTB does not annotate *implicit connectives* between non-adjacent sentences, even if such a relationship clearly holds. For example, even if the discourse adverbial *then* were removed from Example 5

(23) a. John loves Barolo.
b. So he ordered three cases of the '97.
c. But he had to cancel the order
d. because he discovered he was broke.

the event expressed by clause (23d) would still be understood as holding after that expressed by clause (23b). Nevertheless, we neither require nor allow the annotators to annotate the one or more implicit connectives that express the connection holding between clauses (23b) and (23d). Again, this is a practical decision rather than one that has any deep significance.

4.2 Attribution

Attribution has to do with ascribing beliefs and assertions expressed in text to the agent(s) holding or making them (Wiebe, 2002, Wiebe et al., 2005). If we consider the issue of attribution with respect to discourse connectives and their arguments, there are broadly two possibilities:

Case 1 A discourse connective and both its arguments are attributed to the same source.

Case 2 One or both arguments have a different attribution value from the discourse connective.

Initially, we have only distinguished between two different sources of attribution: the writer of the article (**WA**, for “Writer Attribution”) and some agent that s/he is writing about (**SA** or “Speaker Attribution”). Speaker attribution of a connective and both its arguments can involve either quoted and indirect speech, as in Examples 24 and 25, respectively.

(24) “Now, Philip Morris Kraft General Foods’ parent company is committed to the coffee business and to increased advertising for Maxwell House,” says Dick Mayer, president of the General Foods USA division. “<u>Even though</u> **brand loyalty is rather strong for coffee**, *we need advertising to maintain and strengthen it*.”

(25) Like other large Valley companies, Intel also noted that *it has factories in several parts of the nation,* <u>so that</u> **a breakdown at one location shouldn’t leave customers in a total pinch**.

And attribution is annotated for implicit connectives (Example 26), as well as for the explicit connectives illustrated above.

(26) “People say they swim, and that may mean they’ve been to the beach this year,” said Fitness and Sports. “*It’s hard to know if people are responding truthfully*. <u>IMPLICIT-because</u> **People are too embarrassed to say they haven’t done anything**.”

Writer attribution (**WA**) of an explicit connective and both its arguments is illustrated in Example 10, and that of an implicit connective is illustrated in Examples 15 and 16.

When a connective and both its arguments have the same attribution, only the connective is explicitly annotated as **WA** or **SA**, as appropriate, and the arguments, as **IN** (for “inherited attribution”).

There are other instances (**Case 2**), where one or both arguments have a different attribution value from their associated discourse connective, as in Example 27, where the annotators have attributed *Arg1* to the writer (**WA**) and *Arg2* to another speaker (here, the spokeswoman, although that is not recorded as part of the annotation).

(27) *The current distribution arrangement ends in March 1990,* although Delmed said **it will continue to provide some supplies of the peritoneal dialysis products to National Medical**, the spokeswoman said.

Finally, when attribution of a connective or its arguments is uncertain, annotators have been told to attribute the uncertain element to the writer. For example, in Example 27, one cannot tell whether the relation headed by *although* should be attributed to the spokeswoman or the author of the text. As a default, it is attributed to the writer.

The attribution tags in the PDTB are currently being further refined (while stiill maintaining the basic distinction between speaker and writer attribution) to include further distinctions between, for example, verbs of saying and verbs of propositional attitude, and to represent the interaction of verbs of attribution with negation and factuality. For further discussion of attribution, see (Dinesh et al., 2005).

4.3 Sense

Given that many discourse connectives are known to have more than one sense (e.g., *since, while, if, when, because*) and given our professed interest in the inferences that can be drawn from the occurence of a discourse connective and its arguments, it seemed natural to consider trying to annotate connectives in the PDTB for sense as well as for arguments. Fortunately, with all connectives being binary, one can annotate their arguments *independently* of annotating their sense: Different senses of an ambiguous connective will not have different arguments.

The second release of the PDTB aims to provide a gold standard for a sense annotation of those connectives whose senses can be broadly distinguished. For example, this seems possible for the subordinate conjunction *since*, which can be seen to have a purely temporal sense, as in

(28) *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**.

or a purely causal sense, as in

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

or both a temporal and a causal sense simultaneously (T/C), as in

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

Looking at all instances of *since* in the PDTB on which there was annotator agreement on the arguments, the different senses were taken to occur with the following frequency:

	Annot. 1	Annot.2
Temporal	74 (39.8%)	76 (40.9%)
Causal	90 (48.4%)	93 (50%)
T/C	21 (11.3%)	16 (8.6%)
Uncertain	1 (0.5%)	1 (0.5%)
Total	186	186

In this sense annotation of *since*, inter-annotator agreement was high, with 169 instances of exact agreement (90.9%); 14 instances of partial agreement (7.5%), with one annotator assigning a T/C label and the other, either temporal or causal, but not both. There were only 3 instances of pure disagreement (1.6%).

Much more work will be done on sense annotation before the second release of the PDTB in 2006.

5 Some issues surrounding discourse adverbials

The concept of a *discourse adverbial* is one that we have introduced to cover those adverbials that function as discourse connectives. A relevant question is whether one can characterise this set in some way other than by simply listing them.

Syntactically, adverbials can be modifiers of adjectives or adverbs (e.g., *blindingly obvious*), verbs or verb phrases (e.g., *run quickly*, *wash one's hands frequently*) or clauses. While all discourse adverbials fall into the latter class, they don't exhaust it. In order to distinguish discourse adverbials from other clause-modifying adverbials, we consider how their interpretation relates to the AO interpretation of their matrix clause.

Clausal adverbials have only one AO involved in their interpretation – ie., the interpretation of their matrix clause.

(31) a. *Frequently*, clients express interest in paintings but don't end up bidding, so we don't know who the potential buyer will be. (**AO**: EVENTUALITY)

b. *In truth*, lacking the capital to write off their mistakes or to build a navy, the banks have no alternative but to go along. (**AO**: PROPOSITION)

c. *Personally*, I'm irked by its combination of ponderousness and timidity, which adds up to an utter lack of drama. (**AO**: BELIEF)

In contrast, the interpretation of **discourse adverbials** involves two AO arguments, the second derived from a (usually clausal) constituent in the previous discourse.[3] Empirical support for this claim comes from analysing the the 13823 S-initial S-adjoined ADVP and PP adverbials in the WSJ and Brown corpora (Forbes, 2003). Forbes identified seven different groups of adverbials whose interpretations involved two AO arguments.

1. PP adverbials with a demonstrative NP internal argument, such as *in that/this case* (25), *at that/this point* (21), *by that/this time* (13), and *in that way* (12). (The number in parenthesis is the total number of S-initial and S-adjoined tokens found in the WSJ and Brown corpora.) The referent of that internal arg is the second AO argument to the adverbial. (Since the PDTB uses the label *Arg2* for the clause containing the connective and *Arg1* for the other argument, in the following examples, the text from which the second AO argument of the connective derives is indicated in *Italics*.)

[3] We have argued elsewhere that this derivation is similar to other forms of anaphor resolution (Webber et al., 2001, Webber et al., 2003).

(32) *GM is likely to reach the cooperative operating pact it has been seeking in about two weeks*, knowledgeable individuals say. At that point, **investors may face a long, bumpy ride**.

2. PP adverbials with definite NP internal argument, such as *at the same time* (71), *at the time* (17), *in the end* (20), and *in the meantime* (14). Here, the referent of that internal arg is the second AO argument to the adverbial.

(33) *The debt-laden parent has been under pressure from large shareholders to boost the company's share price*. At the same time **it has been caught in an earnings squeeze**.

3. PP adverbials with indefinite/generic relational NP as internal argument, such as *in addition* (204), paraphrasable as *"in addition to that"*; *for example* (167), paraphrasable as *"as an example of that"*; *as a result* (84), paraphrasable as *"as a result of that"*; and *for instance* (70), paraphrasable as *"as an instance of that"*. The missing argument to that relational NP (i.e., the referent of "that") is the second AO argument to the adverbial.

(34) Despite the economic slowdown, *there are few clear signs that growth is coming to a halt*. As a result, **Fed officials may be divided over whether to ease credit**.

4. Deictic ADV adverbials, such as *then* (292), paraphrasable as *"at that point"*; *now* (189), paraphrasable as *"at this point"*; *thus* (114), paraphrasable as *"as a result of this"*; *yet* (80), paraphrasable as *"despite this"* and *therefore* (48), paraphrasable as *"as a result of this"*. In this case, the referent of the deictic NP in the paraphrase is the second AO argument to the adverbial.

(35) Prosecutors have told Mr. Antar's attorneys that *they believe Mr. Antar's allegedly ill-gotten gains are so great that any money he has used to pay attorneys derives from illegal activities*. Therefore, they said, **the money can be taken from the lawyers even after they are paid**.

5. Comparative ADV adverbials, such as *moreover* (53), paraphrasable as *"more than this/that"*; *furthermore* (31), paraphrasable as *"more than this/that"*; *later* (30), paraphrasable as *"later than this/that"*, and *otherwise* (19), paraphrasable as *"other than this/that"*. The target of the comparison is the second AO argument to the adverbial.

   (36) *"Just say the offices are tastefully appointed,"* he says. "Otherwise, **the regulators will take it for decadence**, and nowadays everything's got to be pristine."

6. Idiosyncratic relational ADV adverbials, such as *similarly* (12), paraphrasable as *"similar to this/that"*; *accordingly* (12), paraphrasable as *"in accordance with this/that"*; *simultaneously* (8), paraphrasable as *"at the same time as this/that"* and *consequently* (8), paraphrasable as *"as a consequence of this/that"*. The missing argument to that relation is the second AO argument to the adverbial.

   (37) *UCLA OAIC sponsored research projects share a common theme, "linking interventional research to basic science."* Accordingly, **each research project relates a current or potential clinical intervention to a basic science**.

7. Set-evoking ADV adverbials, such as *finally* (49), *first* (34), *usually* (14), *occasionally* (8), and *secondly* (5). Here, the second AO argument is the set to which matrix interpretation belongs.

   (38) *A number of issues still need to be resolved before Canadian regulators give any project the final go-ahead.* First, **the price of natural gas will have to almost double**.

Examples of all but the first group have been annotated in the PDTB. (As noted in Section 3.3, we have decided, for practical reasons, not to annotate PP discourse adverbials containing a demonstrative NP, as the referents of all demonstrative NPs should be annotated in future work.) But what about other adverbials that occur S-initially and S-adjoined, whose interpretations involve only a single AO (making the *clausal* *adverbials*), but that nevertheless seem to relate sentence/clauses to the previous discourse? What does this impression arise from, and should they not be considered discourse adverbials as well?

5.1 Pragmatic Implicature

Aijmer and Simon-Vandenbergen (2004) consider that the use of clausal adverbials such as *actually*, *in fact* and *indeed* may indirectly convey through *pragmatic implicature*, that a discourse relation holds between adjacent discourse units. Why should this be so? That is, why assert that some proposition is true when all a speaker's claims are supposed to be true? And why should doing so convey or reinforce some discourse connection between the matrix clause and the preceding discourse?

Consider example 39.

(39) Keeping the listed price at a dollar is primarily a convenience. Actually, the funds do fluctuate, but beyond the third decimal place. Rounding-off keeps them at $1. [wsj_1507]

(Aijmer and Simon-Vandenbergen, 2004) say that *actually* implicates here that the matrix clause is the basis for the previous claim, just as the subordinate conjunction *because* would if it were present. But notice that many readers would infer the same relation as holding between the matrix clause and the previous discourse **without** *actually*, and that *because* often co-occurs explicitly with *actually* and *in fact*. Thus, it might be more accurate to take the clausal adverb as simply calling attention to the truth of matrix clause in connection with the role it plays in an otherwise signalled explicit or implicit relation to the previous clause.

Aijmer and Simon-Vanderbergen note two other relations to the previous discourse that can be signalled by the adverbials *actually* and *in fact*. In examples such as

(40) Indeed, as I understand it, the paper considered by the Bureau referred to the Inter-Group on Ageing as having been "recently established". In fact, Mr. President, the Inter-Group on Ageing was established in 1984.

they say that asserting factuality implicates that the matrix clause contrasts with the previous claim, just as it would if the conjunction *but* were present. But again, many readers would infer the same relation holds here **without** *actually*, and the conjunction *but* often co-occurs explicitly with *actually* and *in fact*. So again, it might be more accurate to take the clausal adverb as simply calling attention to the truth of matrix clause in connection with the role it plays in an otherwise signalled explicit or implicit relation to the previous clause.

Finally, in examples such as

(41) Virtually word for word, the notes matched questions and answers on the social-studies section of the test the student was taking. In fact, the student had the answers to almost all of the 40 questions in that section.

Aijmer and Simon-Vanderbergen say that asserting factuality implicates that the matrix clause strengthens the previous claim, sitting in the same semantic field as the adverbials *what's more* and *indeed*. Now, without the adverbial, most readers would still see the second sentence in Example 41 as elaborating the claim in the first (i.e., rather than making some distinct claim). While strengthening is more specific than simply elaborating, one might still say that the basic relation between the two clauses is conveyed by other cues, and that this basic relation is merely further specified by the assertion of factuality through *in fact*.

It is interesting to note that while *in fact*, *actually* and *indeed* co-occur with other connectives such as *so* and *because*, it does not appear to be the case that alone, any of these adverbials ever conveys the sense of these other connectives (either through pragmatic implicature, or by reinforcing a sense conveyed by other means). There is clearly a story here that needs better telling.

5.2 Information Structure

A second reason that a clausal adverbial in S-initial position may seem to relate its matrix clause to the preceding discourse comes from *Information Structure (IS)*. Following Steedman (Steedman, 2000), the intonation pattern of any utterance establishes the following aspects of IS:

- *theme/rheme*, where *theme* conveys presupposed information that can be recovered either from the prior discourse or through accommodation, and *rheme* conveys new information;
- *background/focus*, where *background* within the theme or rheme indicates information that is already given, while *focus* within the theme or rheme conveys information that is to be distinguished from other alternatives in the context.

IS thus provides another mechanism for linking an utterance to the previous discourse.

(Forbes, 2003) shows that if a clausal adverbial is assigned a *theme-related* role in *IS*, it may appear that its content is what links the interpretation of its matrix clause with the previous discourse, when it is really *IS* that is doing so – cf.

(42) John Cooper Powys used to be a popular writer. What is the current view?
A. Nowadays, his books are rarely read. (*Theme Focus*)

(43) Robert Ashton Lister, the founder of modern teak furniture, sourced much of the wood he used from old British warships but nowadays all the timber comes from managed forests.(*Contrastive Theme*)

In (Steedman, 2000), a contrastive theme can require the hearer to accommodate the theme it is being contrasted with, if it is not already in the discourse context. This theme is the other AO to the contrast discourse relation.

Other S-initial S-attached clausal adverbials found in the Penn WSJ Corpus include *clearly* (69 tokens), *surely* (19 tokens), *for now* (30 tokens), *in essense* (3 tokens), *unfortunately*, *undoubtedly* (10 tokens), *admittedly* (3 tokens), *of course* (81 tokens), *notably* (5 tokens), *naturally* (13 tokens) and *presently* (3 tokens). When the work of annotating implicit connectives (Section 4.1) in the PDTB is complete, it will be instructive to analyse the implicit connectives taken to hold at those boundaries marked by clausal adverbials, and try to better understand what, if any, role they play in the relation(s) that are taken to hold between clauses and how they play it.

6 Conclusion

As we have tried to show in this brief paper, the Penn Discourse TreeBank (PDTB) focusses on a clearly defined and relatively easily identifiable level of discourse structure, which is independent of any particular discourse theory. This level may also be language independent, with Discourse TreeBank annotation useful for languages other than English. We have learned of similar efforts now being undertaken for the annotation of German text (Manfred Stede, University of Potsdam) and Danish text (Dan Hardt, Copenhagen Business School).

The first release of the PDTB will be in November 2005, containing $\sim$16k tokens of explicit connectives and their arguments, and $\sim$4k tokens of implicit connectives and their arguments (three sections of the WSJ corpus). A second release is planned for 2006, containing sense annotation.

We are keen to share our technology and insights with researchers interested in developing Discourse TreeBanks for other languages, as well as researchers interested in using the PDTB. Further information on the PDTB can be found at http://www.ircs.upenn.edu/˜pdtb/.

References

Aijmer and Simon-Vandenbergen, 2004 Aijmer, K. and Simon-Vandenbergen, A.-M. (2004). A model and a methodology for the study of pragmatic markers. *Journal of Pragmatics*, 36:1781–1805.

Asher, 1993 Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Boston MA.

Dinesh et al., 2005 Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2005). Attribution and the (non-)alignment of syntactic and discourse arguments of connectives. In *ACL Workshop on Frontiers in Corpus Annotation*, Ann Arbor MI.

Forbes, 2003 Forbes, K. (2003). *Discourse Semantics of S-Modifying Adverbials*. PhD thesis, Department of Linguistics, University of Pennsylvania.

Huddleston and Pullum, 2002 Huddleston, R. and Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge UK.

Marcus et al., 1993 Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large scale annotated corpus of English: The Penn TreeBank. *Computational Linguistics*, 19:313–330.

Miltsakaki et al., 2004a Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004a). Annotating discourse connectives and their arguments. In *NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston MA.

Miltsakaki et al., 2004b Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004b). The penn discourse treebank. In *LREC*, Lisbon, Portugal.

Palmer et al., 2005 Palmer, M., Gildea, D., and Kingbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Prasad et al., 2004 Prasad, R., Miltsakaki, E., Joshi, A., and Webber, B. (2004). Annotation and data mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Steedman, 2000 Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 34:649–689.

Webber and Di Eugenio, 1990 Webber, B. and Di Eugenio, B. (1990). Free Adjuncts in Natural Language Instructions. In *COLING90, Proceedings of the 13th International Conference on Computational Linguistics*, pages 395–400, Helsinki, Finland.

Webber et al., 2001 Webber, B., Knott, A., and Joshi, A. (2001). Multiple discourse connectives in a lexicalized grammar for discourse. In Bunt, H., Muskens, R., and Thijsse, E., editors, *Computing Meaning (Volume 2)*, pages 229–249. Kluwer.

Webber et al., 2003 Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29:545–587.

Wiebe, 2002 Wiebe, J. (2002). Instructions for annotating opinions in newspaper articles. Technical Report TR-02-101, Department of Computer Science, University of Pittsburgh.

Wiebe et al., 2005 Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).