Annotating Discourse Connectives And Their Arguments

Eleni Miltsakaki
University of Pennsylvania
Philadelphia, PA 19104 USA
elenimi@linc.cis.upenn.edu

Rashmi Prasad
University of Pennsylvania
Philadelphia, PA 19104 USA
rjprasad@linc.cis.upenn.edu

Aravind Joshi
University of Pennsylvania
Philadelphia, PA 19104 USA
joshi@linc.cis.upenn.edu

Bonnie Webber
University of Edinburgh
Edinburgh, EH8 9LW Scotland
bonnie@inf.ed.ac.uk

Abstract

This paper describes a new, large scale discourse-level annotation project – the Penn Discourse TreeBank (PDTB). We present an approach to annotating a level of discourse structure that is based on identifying discourse connectives and their arguments. The PDTB is being built directly on top of the Penn TreeBank and Propbank, thus supporting the extraction of useful syntactic and semantic features and providing a richer substrate for the development and evaluation of practical algorithms. We provide a detailed preliminary analysis of inter-annotator agreement – both the level of agreement and the types of inter-annotator variation.

1 Introduction

Large scale annotated corpora have played a critical role in speech and natural language research. The Penn TreeBank (PTB) is an example of such a resource with worldwide impact on natural language processing (Marcus et al., 1993). However, the PTB deals with text only at the sentence level: with the demand for more powerful NLP applications comes a need for greater richness in annotation. At the sentence level, Penn Propbank is adding predicate-argument annotation to sentences in PTB (Kingsbury and Palmer, 2002). At the discourse-level are efforts to produce corpora annotated with rhetorical relations (Carlson et al., 2003). This paper describes a more basic discourse-level annotation project – the Penn Discourse TreeBank (PDTB) – that aims to produce a large-scale corpus in which *discourse connectives* are annotated, along with their *arguments*.

There have been several approaches to describing discourse in terms of discourse relations (Mann and Thompson, 1988; Asher and Lascarides, 1998; Polanyi and van den Berg, 1996). In these approaches, the additional meaning the discourse contributes beyond the sentence derives from discourse relations. Specification of the discourse relations for a discourse thus constitutes a description of a certain level of discourse structure.

Rather than starting from (abstract) discourse relations, we describe an approach to annotating a large-scale corpus in terms of a more basic characterisation of discourse structure in terms of discourse connectives and their arguments. The motivation for such an approach stems from work by Webber and Joshi (1998), Webber et al. (1999a), Webber et al. (2000) which integrates sentence level structures with discourse level structure (using tree-adjoining grammars for both cases, LTAG and DLTAG, respectively).[1] This allows structural composition and its associated semantic composition at the sentence level to be smoothly carried over to the discourse level, a goal also shared by Gardent (1997), Schilder (1997) and Polanyi and van den Berg (1996), among others.[2]

Discourse connectives and their arguments can be successfully annotated with high reliability (cf. Section 4). This is not surprising, given that the task resembles that of annotating verbs and their arguments at the sentence level (Kingsbury and Palmer, 2002). In fact, we use a fine-grained, lexically grounded annotation in which argument labels are specific to the dis-

[1] In the PDTB annotations, we have deliberately adopted a policy to make the annotations independent of the DLTAG framework for two reasons: (1) to make the annotated corpus widely useful to researchers working in different frameworks and (2) to make the annotators' task easier, thereby increasing interannotator reliability.

[2] However, the approaches in Gardent (1997), Schilder (1997), and Polanyi and van den Berg (1996) are different in two ways: a) the process by which discourse derives compositional aspects of meaning is considered entirely separate from how clauses do so, and b) only two mechanisms are used for deriving discourse semantics – compositional semantics and inference.

course connectives involved, in much the same way as in Kingsbury and Palmer (2002). In contrast, a recent attempt (Carlson et al., 2003) at using RST-type relations for annotating a much smaller corpus has already revealed difficulties involved in reliably annotating more abstract discourse relations. Moreover, this type of annotation does not contain any record of the basis on which a relation was assigned.

The paper is organized as follows. Section 2 provides a brief overview of the fundamental ideas that provide the basis for the design of the PDTB annotation. Section 3 gives a detailed description of the annotation project, including information about the size of the corpus, completed annotations as well as annotation instructions as formulated in the guidelines. Section 4 presents data analysis based on current annotations as well as results from inter-annotator agreement. Section 5 wraps up with a summary of the work.

2 Theoretical background

The annotation project presented in this paper builds on basic ideas presented in Webber and Joshi (1998), Webber et al. (1999b) and Webber et al. (2003) – that connectives are discourse-level predicates which project predicate-argument structure on a par with verbs at the sentence level. Webber and Joshi (1998) propose a tree-adjoining grammar for discourse (DLTAG) in which compositional aspects of discourse meaning are formally defined, thus teasing apart compositional from non-compositional layers of meaning. In this framework, connectives are grouped into natural classes depending on the structure that they project at the discourse level. Subordinate and coordinating conjunctions, for example, require two arguments that can be identified structurally from adjacent units of discourse. What Webber et al. (2003) call *anaphoric discourse connectives* (some, but not all, discourse adverbials, such as “otherwise”, “instead”, “furthermore”, etc.) also require two arguments, but only one of them derives structurally. For the complete interpretation of these connectives, their other argument needs to be recovered. The crucial contribution of this framework to the design of the current project is what can be seen as a *bottom-up approach* to discourse structure. Specifically, instead of appealing to an abstract (and arbitrary) set of discourse relations whose identification involves confounding multiple sources of discourse meaning, we start with the annotation of discourse connectives and their arguments, thus exposing a clearly defined level of discourse representation.

3 Project description

The PTDB project began in November 2002. The first phase, including pilot annotations and preliminary development of guidelines, was completed in May 2003. The PDTB is expected to be released by November 2005. Intermediate versions of the annotated corpus will be made available for receiving feedback.

The PDTB corpus will include annotations of four types of connectives: subordinating conjunctions, coordinating conjunctions, adverbial connectives and implicit connectives. We specify each of these types in more detail in Section 3.1. The final number of annotations in the corpus will amount to approximately 30,000; 10,000 implicit connectives and 20,000 annotations of the 250 explicit connectives identified in the corpus. The final version of the corpus will also contain characterizations of the semantic roles associated with the arguments of each type of connective.

In this paper we present the results of annotating 10 explicit connectives, amounting to a total of 2717 annotations, as well as 386 tokens of implicit connectives. The set of 10 connectives comprises the adverbial connectives ‘therefore’, ‘as a result’, ‘instead’, ‘otherwise’, ‘nevertheless’, and the subordinate conjunctions ‘because’, ‘although’, ‘even though’, ‘when’, and ‘so that’. In all cases, annotations have been performed by four annotators. While this slows down the annotation process considerably, the nature, significance and magnitude of the project as well as the well-known complexity of discourse annotation tasks impels us to strive for maximum reliability, achieved by having the task performed by multiple annotators.[3]

Individual annotation proceeds one connective at a time. The annotation tool *WordFreak*[4] is used to identify all instances of the given connective in the corpus, and these are then annotated independently and manually by four annotators. This way, the annotators quickly gain experience with that connective and develop a better understanding of its predicate-argument characteristics. Similarly, for the annotation of implicit connectives, all instances (as specified in the guidelines, see Section 3.2) are identified one file at a time. For this task, the annotators are required to read the entire file so that they can make well-informed and reliable decisions about the implicit connectives and their arguments. In addition, after the arguments of each implicit connective have been identified, the annotators provide, if possible, an explicit connective (or other suitable expression) that best expresses the inferred relation. As with explicit connectives, annotations of implicit connectives are done by four annota-

[3] When inter-annotator consistency has stabilized, we intend to reduce the number of annotators to three, or maybe two at the minimum.

[4] *WordFreak* was developed by Tom Morton at the University of Pennsylvania. It has been substantially modified by Jeremy Lacivita to fit the needs of the PDTB project. A snapshot of the tool can be seen at http://www.cis.upenn.edu/∼pdtb.

tors.

Compared with Propbank's annotation of verb predicate-argument structures, annotation of arguments of discourse predicates is different in interesting ways. Propbank annotators have to determine the number of arguments required by each verb. In contrast, discourse connectives exhibit a clear predicate-argument structure requiring only two arguments. The main challenge we have discovered for annotating discourse connectives is determining the *extent* of their arguments. Even subordinate conjunctions whose arguments never cross a sentence boundary may sometimes be the source of disagreement between annotators.

In what follows, we present a brief overview of the classes of connectives that we annotate, followed by highlights of the annotation manual and relevant corpus examples.

3.1 Discourse connectives

We classify discourse connectives into four classes: subordinate and coordinating conjunctions, adverbials and implicit connectives. Examples of each type are given below, with their arguments shown in square brackets and the connectives, in italics.

3.1.1 Subordinate conjunctions

Subordinate conjunctions introduce clauses that are syntactically dependent on a main clause. The most common types of relations that they express are temporal (e.g., 'when', 'as soon as'), causal e.g., 'because'), concessive (e.g., 'although', 'even though'), purpose (e.g., 'so that', 'in order that') and conditional (e.g., 'if', 'unless). Clauses introduced with a subordinate conjunction may be preposed (or, more rarely, interposed) with respect to the main clause, as shown in (1).

(1) *Because* [the drought reduced U.S. stockpiles], [they have more than enough storage space for their new crop], and that permits them to wait for prices to rise.

3.1.2 Coordinating conjunctions

Coordinating conjunctions are ones such as 'and', 'but', and 'or'. Example (2) shows the annotation of an instance of the conjunction 'and'.

(2) [William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs], *and* [Gates became an industry billionaire six years after IBM adapted one of these versions in 1981].

Instances of coordinating conjunctions which coordinate nominal or other non-clausal constituents are excluded from annotation. We also exclude cases of VP-coordination because in such cases the arguments of the connective can be retrieved automatically from the syntactic layer.

3.1.3 Adverbial connectives

Adverbial connectives are sentence-modifying adverbs which express a discourse relation (Forbes, 2003). The class of adverbial connectives includes 'however', 'therefore', 'then', 'otherwise', etc. In this class, we have also included prepositional phrases with a similar sentence modifying function such as 'as a result', 'in addition', 'in fact', etc. Example (3) shows the annotation of an instance of the adverbial connective 'as a result'.

(3) ...[many analysts expected energy prices to rise at the consumer level too]. *As a result*, [many economists were expecting the consumer price index to increase significantly more than it did].

The arguments of adverbial connectives may or may not be adjacent to the sentence containing the connective. In a few cases, an argument may be found one or two paragraphs away from the connective.

3.1.4 Implicit connectives

Implicit connectives are identified between adjacent sentences with no explicit connectives.[5] The annotation of implicit connectives is intended to capture the connection between two sentences appearing in adjacent positions. For example, in (4), the two adjacent sentences are connected in a way similar to having the explicit connective "but" contrasting them. Indeed, for implicit connectives, annotators are asked to provide, when possible, an explicit connective that best describes the inferred relation. The explicit connective provided in (4) was 'in contrast'.

(4) ...[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]. *IMPLICIT*-(In contrast) [In fiscal 1984 before Mr. Gandhi came to power, only $810 million was raised].

3.2 Annotation guidelines

The annotation guidelines for PDTB have been revised considerably since the pilot phase of the project in May 2003. The current version of the guidelines is available at `http://www.cis.upenn.edu/∼pdtb`. Below we outline the basic points.

3.2.1 What counts as a discourse connective?

We count as discourse connectives (1) all subordinating and coordinating conjunctions, (2) certain adverbials, and (3) implicit connectives. The adverbials include only those which convey a relation between events or states. For example, in (5) 'as a result' conveys a cause-effect relation between the event of limiting the size of new steel

[5] There are, of course, other implicit connectives that we are not taking into account.

mills and that of the industry operating out of small, expensive and highly inefficient units. In contrast, the semantic interpretation of ‘strangely’ in (6) only requires a single event/state which it classifies in the set of *strange* events/states.[6]

(5) [In the past, the socialist policies of the government strictly limited the size of new steel mills, petrochemical plants, car factories and other industrial concerns to conserve resources and restrict the profits businessmen could make]. *As a result,* industry operated out of small, expensive, highly inefficient industrial units.

(6) Strangely, conventional wisdom inside the Beltway regards these transfer payments as “uncontrollable” or “nondiscretionary.”

The guidelines also highlight instances of lexical items with multiple functions, only one of which is as a discourse connective. For example, ‘when’ can either serve as a subordinate conjunction or introduce a relative clause modifying a nominal phrase, as in (7), where the *when*-clause modifies the nominal ‘1985’.[7]Here we again benefit from building discourse annotation on top of Penn TreeBank because the syntactic annotation of *when*-clauses distinguishes the two functions: *When*-relatives are marked as NP-modifiers adjoining to an NP, whereas adverbial *when*-clauses adjoin to a sentential node.

(7) Attorneys have argued since 1985, when the law took effect.

Similarly, some *since*-clauses function as NP modifiers as shown in (8). In such cases, ‘since’ is not annotated as a connective. As in the case of *when*-clauses, instances of NP modifying *since*-clauses can be identified in the Penn TreeBank by virtue of their syntactic annotation.

(8) In the decade since the communist nation emerged from isolation, its burgeoning trade with the West has lifted Hong Kong’s status as a regional business partner.

Finally, implicit connectives count as connectives. They are identified between adjacent sentences which do not contain any other explicit connectives. Currently, we are not annotating implicit connectives intra-sententially, such as between the matrix clause and free adjunct in Example (9). We plan to incorporate annotations of implicit intra-sentential connectives at a later stage of the project.

(9) Second, they channel monthly mortgage payments into semiannual payments, reducing the administrative burden on investors.

3.2.2 What counts as a legal argument?

Because we take discourse relations to hold between abstract objects, we require that an argument contains at least one predicate along with its arguments. Of course, a sequence of clauses or sentences may also form a legal argument, containing multiple predicates.

Because our annotations are done directly on top of the Penn TreeBank, annotators may select as an argument certain textual spans that appear to exclude one or more arguments of the predicate. These are cases in which these arguments are directly retrievable from the syntactic annotation. Thus, we are able to select only the predicates that are required for the interpretation of the discourse connective and simultaneously access their arguments for the complete interpretation of the clause while keeping the annotations of single arguments simple and maximally contiguous. In (10), for example, the relative clause is marked as one of the two arguments of the connective ‘even though’. The subject of the verb in the relative clause is directly retrievable from the Penn TreeBank annotation. Similarly, in (11) the subject of the infinitival clause is also available from the syntactic representation.

(10) Workers described “clouds of dust” [that hung over parts of the factory] *even though* [exhaust fans ventilated the air].

(11) The average maturity for funds open only to institutions, considered by some [to be a stronger indicator] *because* [those managers watch the market closely], reached a high point for the year – 33 days.

There are two exceptions to the requirement that an argument include a verb – these are nominal phrases that express an event or a state, and discourse deictics that denote an event or state. In (12), for example, the nominal phrase ‘fainting spells’ can be marked as a legal argument of the connective ‘when’ because the phrase expresses an event of fainting.

(12) Its symptoms include a cold sweat at the sound of debate, clammy hands in the face of congressional criticism, and [fainting spells] *when* [someone writes the word ”controversy.”]

Discourse deictic expressions are forms such as ‘this’ and ‘that’ that can be used to denote the interpretation of clausal textual spans from the preceding discourse. In (13), for example, ‘that’ denotes the interpretation of the sentence immediately preceding it. Our annotators are guided to make argument selections that assume that anaphoric and deictic expressions have been resolved. Thus, in (13), they are able to select ‘That’s’ as one argument of the connective ‘because’.

(13) Airline stocks typically sell at a discount of about one-third to the stock market’s price-earnings ratio – which is currently about 13 times earnings. [That’s] *because* [airline earnings, like those of auto makers, have been subject to the cyclical ups-and-downs of the economy].

[6] For a more detailed discussion of the basis for distinguishing discourse adverbials from clausal adverbials, see Forbes (2003).

[7] In cases of *when*-relatives, a *when*-clause can be annotated as **SUP** (see Section 3.2.3).

The annotators are also informed that in some cases, an argument of a connective must be *derived* from the selected textual span (Webber et al., 1999a; Webber et al., 2003). This is the case for the first argument of 'instead' in (14), which does not include the negation, although it is contained in the selected text.[8]

(14) [No price for the new shares has been set]. *Instead*, [the companies will leave it up to the marketplace to decide].

In sum, legal arguments can be groups of sentences, single sentences (a main clause and its subordinate clauses), single clauses (tensed or non-tensed), NPs that specify events or situations, and discourse deictic expressions.

3.2.3 How far does an argument extend?

One particularly significant addition to the guidelines came as a result of differences among annotators as to how large a span constituted the argument of a connective. During pilot annotations, annotators used three annotation tags: **CONN** for the connective and **ARG1** and **ARG2** for the two arguments. To this set, we have added the optional tags **SUP1**, **SUP2** (*supplementary*) for cases when the annotator wants to mark textual spans s/he considers to be useful, *supplementary* information for the interpretation an argument. Example (15) demonstrates the use of **SUP1**. Arguments are shown in square brackets, while spans providing supplementary information are shown in parentheses.

(15) *Although* [started in 1965], [Wedtech didn't really get rolling until 1975] (when Mr. Neuberger discovered the Federal Government's Section 8 minority business program).

4 Data analysis

To test the reliability of the annotation, we first considered the kappa statistic (Siegel and Castellan, 1988) which is used extensively in empirical studies of discourse (Carletta, 1996). The kappa coefficient provides an inter-annotator agreement figure for any number of annotators by measuring pairwise agreement between them and by correcting for chance expected agreement. However, the statistic requires the data tokens to be classified into discrete categories, and as a result, we could not apply it to our data since the PDTB annotation tokens cannot be classified as such. Rather, annotation in the PDTB constitutes either selection of a span of text for the arguments of connectives which can be of indeterminate length or providing explicit expressions for implicit connectives from an open-ended class of expressions.

Instead, we have assessed inter-annotator agreement in terms of agreement/disagreement on span or named expression identity for each token as a percentage of the pairs of spans or expressions that actually matched versus those that should have. For the argument annotations, we use a most conservative measure - the *exact match* criterion. In addition, we also used different diagnostics for the argument annotations for the explicit connectives, reporting percentage agreement on different classes of tokens, such as those in which the first argument (ARG1) annotations and second argument (ARG2) annotations were counted independently, as well as those in which the ARG1 and ARG2 annotations (for each connective) were counted together as a single token. For all the argument annotations, the computation of agreement excluded the *supplementary* annotations (cf. Section 3.2.3).

We present here agreement results on ARG1 and ARG2 annotations by two annotators for the annotation of ten explicit connectives, amounting to a total of 2717 annotations, and 368 annotations of implicit connectives, including agreement results on the explicit expression the annotators used in in place of the implicit connectives as well as the ARG1 and ARG2 annotations of the implicit connectives.[9] The ten explicit connectives include 5 subordinating conjunctions (*when*, *because*, *even though*, *although*, and *so that*) and 5 adverbials (*nevertheless*, *otherwise*, *instead*, *therefore*, and *as a result*).

4.1 Inter-annotator Agreement

4.1.1 Explicit connectives

For the explicit connective annotations, we used two diagnostics for measuring inter-annotator agreement. In the first diagnostic , we took the class of tokens as the total number of argument annotations, treating ARG1 and ARG2 annotations as independent tokens. The total number of tokens in this class is therefore twice the number of connective tokens, i.e, 5434. We recorded agreement using the *exact match* criterion. That is, for any ARG1 or ARG2 token, agreement was recorded as 1 when both annotators made identical textual selections for the annotation and 0 when the annotators made non-identical selections.

We achieved 90.2% agreement (4900/5434 tokens) on the annotations for this class. Agreement on only ARG1 tokens was 86.3%, and agreement on only ARG2 tokens was 94.1%. Further distribution of the agreements by connective is given in Table 1. Connectives are grouped in the table by type (subordinating conjunction (SUBCONJ) and adverbial (ADV)). The second col-

[8] For a preliminary corpus-based analysis of the arguments of 'instead', see Miltsakaki et al. (2003).

[9] Right now SUP1 and SUP2 annotations are for our use only and are not included in the current evaluations. Additional annotations by another 2 annotators are currently underway. The 2 annotators of the explicit connectives are different from the 2 annotators of the implicit connectives.

umn gives the number of agreeing tokens for each connective and the third column gives the total number of (ARG1+ARG2) tokens available for that connective. The last column gives the percent agreement for the connective in that row, i.e., as a percentage of tokens for which agreement was 1 (column 2) versus the total number of tokens for that connective (column 3).

CONNECTIVES	AGR No.	Conn. Total	%AGR
when	1877	2032	92.4%
because	1703	1824	93.4%
even though	194	206	94.1%
although	635	704	90.1%
so that	66	74	89.2%
TOTAL SUBCONJ	**4469**	**4834**	**92.4%**
nevertheless	56	94	59.6%
otherwise	44	46	95.7%
instead	172	236	72.9%
as a result	110	168	65.5%
therefore	49	56	87.5%
TOTAL ADV.	**431**	**600**	**71.8%**
OVERALL TOTAL	**4900**	**5434**	**90.2%**

Table 1: Distribution of Agreement by Connective, with ARG1 and ARG2 Annotations Counted Independently

The table shows that we achieved high agreement on argument annotations of subordinating conjunctions (92.4%). Average agreement on the adverbials was lower (71.8%). This difference between the two types is not surprising, since locating the anaphoric (ARG1) argument of adverbial connectives is believed to be a harder task than that of locating the arguments of subordinating conjunctions. For example, the anaphoric argument of the adverbial connectives may be located in some non-adjacent span of text, even several paragraphs away. Arguments of subordinating conjunctions, on the other hand, can most often be found in spans of text adjacent to the connective. The table also shows that there was uniform agreement across the different subordinating conjunctions (roughly 90%), whereas the adverbials showed more variation. In particular, agreement on *otherwise* and *therefore* was high (95.7% and 87.5% respectively), while lower for the other three adverbials, *instead* (72.9%), *as a result* (65.5%), and *nevertheless* (59.6%). This suggests either *greater variability* in how these adverbials are interpreted or *greater complexity* in their interpretation, which results in more variability when people are forced to associate an interpretation with a particular text span.

We also computed agreement using a second more conservative diagnostic in which we took the class of tokens as the total number of connective tokens (2717) so that the ARG1 and ARG2 annotations for each connective were treated together as part of the same token. Here again, we recorded agreement using the *exact match* measure. That is, for any connective token, agreement was recorded as 1 when both annotators made identical textual selections for the annotation of *both* arguments and 0 when the annotators made non-identical selections for any *one or both* arguments.

We achieved 82.8% agreement (2249/2717 tokens) on the annotations for this class. Table 2 gives the distribution of the agreements by connective. The table shows relatively lower agreements when compared with the first diagnostic, for both subordinating conjunctions (86%) as well as adverbials (57%). However, this difference is understandable since the token class as defined for this diagnostic yields a stricter measure of agreement.

CONNECTIVES	AGR No.	Conn. Total	%AGR
when	868	1016	86.4%
because	804	912	88.2%
even though	91	103	88.3%
although	288	352	81.8%
so that	27	34	79.4%
TOTAL SUBCONJ	**2078**	**2417**	**86.0%**
nevertheless	18	47	38.3%
otherwise	21	23	91.3%
instead	72	118	61.0%
as a result	38	84	45.2%
therefore	22	28	78.6%
TOTAL ADV.	**171**	**300**	**57.0%**
OVERALL TOTAL	**2249**	**2717**	**82.8%**

Table 2: Distribution of Agreement by Connective, with ARG1 and ARG2 Annotations Counted Together

We classified disagreements into 4 major types. The result of classifying the 534 disagreements from Diagnostic 1 (Table 1) is given in Table 3. The third column gives the percent of the total disagreements for each type.

DISAGREEMENT TYPE	No.	%
Missing Annotations	72	13.5%
No Overlap	30	5.6%
Partial Overlap		
Parentheticals	53	9.9%
higher verb	181	33.9%
dependent clause	182	34.1%
Other	6	1.1%
Unresolved	10	1.9%
TOTAL	534	100%

Table 3: Disagreement Classification

The majority of disagreements (79%) were due to *Partial Overlap*, which subsumes the categories *Higher Verb*, *Dependent Clause*, *Parenthetical* and *Other*. *Partial Overlap* means that there was partial overlap in the annotations selected by the two annotators. *Higher verb* includes tokens where one of the annotators included the governing predicate for the clause marked by both annotators. The higher clause occurred on the left or right periphery of the lower clause. *Dependent Clause* includes tokens where one of the annotators included extra clausal material that is syntactically dependent on the clause that was selected by both, and that occurs on the left or right periphery of the common text. *Parenthetical* means that one of the annotators included a medial parenthetical, while the other did not. The intervening text could be the main as well as the dependent clause. An example is provided below:

(16) 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].

(17) 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].

Other included tokens with partial overlap between annotations, but in addition included a combination of more than type, such as *higher verb+dependent clause*.

Note that disagreements that contain a partial overlap could be counted as agreeing tokens if we relaxed the more conservative *exact match* measure to a *partial match* measure. Our subjective view was that in several cases, the "extra" textual material, especially those fitting the *dependent clause* and *parenthetical* category did not make any significant semantic contribution in terms of their inclusion or exclusion in the argument. With the *partial match* measure, excluding these cases reduces the disagreements to half the given number, giving us 94.5% agreement overall.

The *No Overlap* tokens were cases of true disagreement in that there was no overlap in the annotations selected by the annotators. These tokens constituted 5.6% of the disagreements. Examples (18) and (19) shows the two annotations for a token in which there was no overlap in the ARG1 annotation. *Missing Annotations* also constituted a substantial proportion of the disagreements (13.5%) and was used for tokens where the annotation was missing for one annotator. Note that these don't really count as disagreement, since all connectives are pretheoretically assumed to *require* two arguments. *Unresolved* includes tokens which have introduced new issues for the annotation guidelines and cannot be resolved at this time. These include issues such as how to treat comparatives, certain types of adjunct clauses, certain types of nominalizations etc.

(18) [The word ”death” cannot be escaped entirely by the industry], but salesmen dodge it wherever possible or cloak it in euphemisms, [preferring to talk about ”savings” and ”investment”] *instead*.

(19) The word ”death” cannot be escaped entirely by the industry, but salesmen dodge it wherever possible or [cloak it in euphemisms], preferring [to talk about ”savings” and ”investment”] *instead*.

4.1.2 Implicit connectives

For the 386 tokens of implicit connectives, we analyzed inter-annotator agreement between two annotators for (a) the explicit connectives they provided in place of an implicit connective, and (b) the argument annotations of the implicit connectives.

As a preliminary step in analyzing agreement on the type of explicit connective provided by the annotators in place of an implicit connective, we considered 5 groups of connectives conveying : a) additional information (e.g., 'furthermore', 'in addition') b) cause-effect relations (e.g., 'because', 'as a result'), c) temporal relations (e.g., 'then', 'simultaneously'), d) contrastive relations (e.g., 'however', 'although'), and e) restatement or summarization (e.g., 'in other words', 'in sum').[10] Agreement was then computed on these basic groups of connectives.[11] From the total of 386 tokens of implicit connectives, 9 were excluded from the analysis due to technical error (missing annotation). For the remaining 307 tokens, we achieved 72% agreement on the type of explicit connective that best conveyed the interpretation of the implicit connective.

For the argument annotations of the implicit connectives, we present agreement results from using the first diagnostic used for the explicit connectives. That is, we counted ARG1 and ARG2 annotations as independent tokens and computed percent agreement using the *exact match* criterion. On the 772 ARG1 and ARG2 tokens, we achieved 85.1% (657/772) agreement between 2 annotators. The analysis of the 115 disagreements is given in Table 4. Note that here again, the number of disagreements reduces to half using the *partial match* measure for the *parenthetical* and *dependent clause* classes, giving us 92.6% agreement overall.

DISAGREEMENT TYPE	No.	%
Missing Annotations	6	5.2%
No Overlap	2	1.7%
Partial Overlap		
parenthetical	13	11.3%
higher verb	24	20.9%
dependent clause	44	38.3%
sentence	19	16.5%
other	3	2.6%
Unresolved	4	3.5%
TOTAL	115	100%

Table 4: Disagreement Classification for Implicit Connective ARG Annotations

[10] These groups are based on types of coherence relations derived from corpus-based distributions of connectives presented in (Knott, 1996). Initially, we also considered a group of connectives expressing hypothetical relations but no such connectives were identified in the annotations.

[11] Some polysemous connectives such as 'while' and 'in fact' appeared in more than one group.

5 Summary

In this paper we presented a new and innovative discourse-level annotation project, the Penn Discourse TreeBank (PDTB), in which discourse connectives and their arguments are annotated, thereby defining a clear level of discourse structure that can be reliably annotated for a large corpus. Our inter-annotator results confirm our expectations of high agreement and annotation reliability. At a later stage of the project, we plan to provide semantic characterizations of the arguments of connectives and resolve any cases of polysemy that might arise.

Acknowledgments

We are very grateful to Tom Morton and Jeremy Lacivita for the development and special modification of the *WordFreak* annotation tool. Special thanks to Jeremy for providing continuous technical support. Thanks are also due to our annotators, Cassie Creswell, Driya Amandita, John Laury, Emily Pawley, Alan Lee, Alex Derenzy and Steve Pettington. Also, many thanks to Katherine Forbes Riley and Jean Carletta for their comments and suggestions. Finally, we would like to thank the reviewers for their very useful comments. This work was partially supported by NSF Grant EIA 02-24417.

References

Nicholas Asher and Alex Lascarides. 1998. The semantics and pragmatics of presupposition. *Journal of Semantics*, 15(3):239–300.

Jean Carletta. 1996. Assessing agreement on classification tasks. *Computational Linguistics*, 22:249–254.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski, 2003. *Current Directions in Discourse and Dialogue*, chapter Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Kluwer Academic Publishers.

Kate Forbes. 2003. *Discourse Semantics of S-Modifying Adverbials*. Ph.D. thesis, Department of Linguistics, University of Pennsylvania.

Claire Gardent. 1997. Discourse tree adjoining grammars. Claus 89, University of the Saarlandes, Saarbrucken.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Third International Conference on Language Resources and Evaluation, LREC-02, Las Palmas, Canary Islands, Spain*.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

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

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19:313–330.

Eleni Miltsakaki, Cassandre Creswell, Kate Forbes, Aravind Joshi, and Bonnie Webber. 2003. Anaphoric arguments of discourse connectives: Semantic properties of antecedents versus non-antecedents. In *Proceedings of the Computational Treatment of Anaphora Workshop, EACL 2003, Budapest*.

Livia Polanyi and Martin van den Berg. 1996. Discourse structure and discourse interpretation. In *Proceedings of the Tenth Amsterdam Colloquium, University of Amsterdam*, pages 113–131.

Frank Schilder. 1997. Discourse tree grammar or how to get attached to a discourse? In *Proceedings of the the second International Workshop on Computational Semantics (IWCS-II), Tilburg, The Netherlands*, pages 261–273.

Sidney Siegel and N. J. Castellan. 1988. *Nonparamateric Statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition.

Bonnie Webber and Aravind Joshi. 1998. Anchoring a lexicalized tree adjoining grammar for discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal*, pages 8–92. Montreal, Canada.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999a. Discourse relations: A structural and presuppositional account using lexicalized TAG. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Maryland*, pages 41–48. College Park MD.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999b. What are little texts made of? A structural and presuppositional account using lexicalized TAG. In *Proceedings of the International Workshop on Levels of Representation in Discourse (LORID '99), Edinburgh*, pages 145–149.

Bonnie Webber, Alistair Knott, and Aravind Joshi. 2000. Multiple discourse connectives in a lexicalized grammar for discourse. In *Proceedings of the Third International Workshop on Computational Semantics*, Tilburg, The Netherlands.

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