Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives

Nikhil Dinesh and **Alan Lee** and **Eleni Miltsakaki** and **Rashmi Prasad** and **Aravind Joshi**
University of Pennsylvania
Philadelphia, PA 19104 USA
{nikhild,aleewk,elenimi,rjprasad,joshi}@linc.cis.upenn.edu

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

Abstract

The annotations of the Penn Discourse Treebank (PDTB) include (1) discourse connectives and their arguments, and (2) *attribution* of each argument of each connective and of the relation it denotes. Because the PDTB covers the same text as the Penn TreeBank WSJ corpus, syntactic and discourse annotation can be compared. This has revealed significant differences between syntactic structure and discourse structure, in terms of the arguments of connectives, due in large part to attribution. We describe these differences, an algorithm for detecting them, and finally some experimental results. These results have implications for automating discourse annotation based on syntactic annotation.

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 annotations. A preliminary report on this project was presented at the 2004 workshop on *Frontiers in Corpus Annotation* (Miltsakaki et al., 2004a), where we described our annotation of discourse connectives (both explicit and implicit) along with their (clausal) arguments.

Further work done since then includes the annotation of *attribution*: that is, who has expressed each argument to a discourse connective (the writer or some other speaker or author) and who has expressed the discourse relation itself. These ascriptions need not be the same. Of particular interest is the fact that attribution may or may not play a role in the relation established by a connective. This may lead to *a lack of congruence between arguments at the syntactic and the discourse levels*. The issue of congruence is of interest both from the perspective of annotation (where it means that, even within a single sentence, one cannot merely transfer the annotation of syntactic arguments of a subordinate or coordinate conjunction to its discourse arguments), and from the perspective of inferences that these annotations will support in future applications of the PDTB.

The paper is organized as follows. We give a brief overview of the annotation of connectives and their arguments in the PDTB in Section 2. In Section 3, we describe the annotation of the attribution of the arguments of a connective and the relation it conveys. In Sections 4 and 5, we describe mismatches that arise between the discourse arguments of a connective and the syntactic annotation as provided by the Penn TreeBank (PTB), in the cases where all the arguments of the connective are in the same sentence. In Section 6, we will discuss some implications of these issues for the theory and practice of discourse annotation and their relevance even at the level of sentence-bound annotation.

2 Overview of the PDTB

The PDTB builds on the DLTAG approach to discourse structure (Webber and Joshi, 1998; Webber et al., 1999; Webber et al., 2003) in which connectives are discourse-level predicates which project predicate-argument structure on a par with verbs at

the sentence level. Initial work on the PDTB has been described in Miltsakaki et al. (2004a), Miltsakaki et al. (2004b), Prasad et al. (2004).

The key contribution of the PDTB design framework is its *bottom-up approach* to discourse structure: Instead of appealing to an abstract (and arbitrary) set of discourse relations whose identification may confound 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.

The PDTB annotates as *explicit discourse connectives* all subordinating conjunctions, coordinating conjunctions and discourse adverbials. These predicates establish relations between two *abstract objects* such as events, states and propositions (Asher, 1993).[1]

We use Conn to denote the connective, and Arg1 and Arg2 to denote the textual spans from which the abstract object arguments are computed.[2] In (1), the subordinating conjunction *since* establishes a temporal relation between the event of the earthquake hitting and a state where no music is played by a certain woman. In all the examples in this paper, as in (1), Arg1 is italicized, Arg2 is in boldface, and Conn is underlined.

(1) *She hasn't played any music* since **the earthquake hit**.

What counts as a legal argument? Since we take discourse relations to hold between *abstract objects*, we require that an argument contains at least one clause-level predication (usually a verb – tensed or untensed), though it may span as much as a sequence of clauses or sentences. The two exceptions are nominal phrases that express an event or a state, and discourse deictics that denote an abstract object.

[1]For example, discourse adverbials like *as a result* are distinguished from clausal adverbials like *strangely* which require only a single abstract object (Forbes, 2003).

[2]Each connective has exactly two arguments. The argument that appears in the clause syntactically associated with the connective, we call Arg2. The other argument is called Arg1. Both Arg1 and Arg2 can be in the same sentence, as is the case for subordinating conjunctions (e.g., *because*). The linear order of the arguments will be Arg2 Arg1 if the subordinate clause appears sentence initially; Arg1 Arg2 if the subordinate clause appears sentence finally; and undefined if it appears sentence medially. For an adverbial connective like *however*, Arg1 is in the prior discourse. Hence, the linear order of its arguments will be Arg1 Arg2.

Because our annotation is on the same corpus as the PTB, annotators may select as arguments textual spans that omit content that can be recovered from syntax. In (2), for example, the relative clause is selected as Arg1 of *even though*, and its subject can be recovered from its syntactic analysis in the PTB. In (3), the subject of the infinitival clause in Arg1 is similarly available.

(2) Workers described "clouds of blue dust" *that hung over parts of the factory* even though **exhaust fans ventilated the air**.

(3) 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.

The PDTB also annotates *implicit connectives* between adjacent sentences where no explicit connective occurs. For example, in (4), the two sentences are contrasted in a way similar to having an explicit connective like *but* occurring between them. Annotators are asked to provide, when possible, an explicit connective that best describes the relation, and in this case *in contrast* was chosen.

(4) *The $6 billion that some 40 companies are looking to raise in the year ending March 21 compares with only $2.7 billion raise on the capital market in the previous year*. IMPLICIT - in contrast **In fiscal 1984, before Mr. Gandhi came into power, only $810 million was raised**.

When complete, the PDTB will contain approximately 35K annotations: 15K annotations of the 100 explicit connectives identified in the corpus and 20K annotations of implicit connectives.[3]

3 Annotation of attribution

Wiebe and her colleagues have pointed out the importance of ascribing beliefs and assertions expressed in text to the agent(s) holding or making them (Riloff and Wiebe, 2003; Wiebe et al., 2004; Wiebe et al., 2005). They have also gone a considerable way towards specifying how such subjective material should be annotated (Wiebe, 2002). Since we take discourse connectives to convey semantic predicate-argument relations between abstract objects, one can distinguish a variety of cases depending on the *attribution* of the discourse relation or its

[3]The annotation guidelines for the PDTB are available at `http://www.cis.upenn.edu/~pdtb`.

arguments; that is, whether the relation or arguments are ascribed to the author of the text or someone other than the author.

Case 1: The relation and both arguments are attributed to the same source. In (5), the concessive relation between Arg1 and Arg2, anchored on the connective *even though* is attributed to the speaker *Dick Mayer*, because he is quoted as having said it. Even where a connective and its arguments are not included in a single quotation, the attribution can still be marked explicitly as shown in (6), where only Arg2 is quoted directly but both Arg1 and Arg2 can be attibuted to *Mr. Prideaux*. Attribution to some speaker can also be marked in reported speech as shown in the annotation of *so that* in (7).

(5) "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. "Even though **brand loyalty is rather strong for coffee**, *we need advertising to maintain and strengthen it*."

(6) *B.A.T isn't predicting a postponement* because **the units "are quality businesses and we are encouraged by the breadth of inquiries**," said Mr. Prideaux.

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

Wherever there is a clear indication that a relation is attributed to someone other than the author of the text, we annotate the relation with the feature value **SA** for "speaker attribution" which is the case for (5), (6), and (7). The arguments in these examples are given the feature value **IN** to indicate that they "inherit" the attribution of the relation. If the relation and its arguments are attributed to the writer, they are given the feature values **WA** and **IN** respectively.

Relations are attributed to the writer of the text by default. Such cases include many instances of relations whose attribution is ambiguous between the writer or some other speaker. In (8), for example, we cannot tell if the relation anchored on *although* is attributed to the *spokeswoman* or the author of the text. As a default, we always take it to be attributed to the writer.

Case 2: One or both arguments have a different attribution value from the relation. While the default value for the attribution of an argument is the attribution of its relation, it can differ as in (8). Here, as indicated above, the relation is attributed to the writer (annotated **WA**) by default, but Arg2 is attributed to Delmed (annotated **SA**, for some speaker other than the writer, and other than the one establishing the relation).

(8) *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.

Annotating the corpus with attribution is necessary because in many cases the text containing the source of attribution is located in a different sentence. Such is the case for (5) where the relation conveyed by *even though*, and its arguments are attributed to *Dick Mayer*.

We are also adding attribution values to the annotation of the implicit connectives. Implicit connectives express relations that are *inferred* by the reader. In such cases, the author *intends* for the reader to *infer* a discourse relation. As with explicit connectives, we have found it useful to distinguish implicit relations intended by the writer of the article from those intended by some other author or speaker. To give an example, the implicit relation in (9) is attributed to the writer. However, in (10) both Arg1 and Arg2 have been expressed by the speaker whose speech is being quoted. In this case, the implicit relation is attributed to the speaker.

(9) *Investors in stock funds didn't panic the weekend after mid-October's 190-point market plunge*. IMPLICIT-instead **Most of those who left stock funds simply switched into money market funds**.

(10) "People say they swim, and that may mean they've been to the beach this year," Fitness and Sports. "*It's hard to know if people are responding truthfully*. IMPLICIT-because **People are too embarrassed to say they haven't done anything**."

The annotation of attribution is currently underway. The final version of the PDTB will include annotations of attribution for all the annotated connectives and their arguments.

Note that in the Rhetorical Structure Theory (RST) annotation scheme (Carlson et al., 2003), attribution is treated as a discourse relation. We, on the other hand, do not treat attribution as a discourse

relation. In PDTB, discourse relations (associated with an explicit or implicit connective) hold between two abstracts objects, such as events, states, etc. Attribution relates a proposition to an entity, not to another proposition, event, etc. This is an important difference between the two frameworks. One consequence of this difference is briefly discussed in Footnote 4 in the next section.

4 Arguments of Subordinating Conjunctions in the PTB

A natural question that arises with the annotation of arguments of subordinating conjunctions (SUBCONJS) in the PDTB is *to what extent they can be detected directly from the syntactic annotation in the PTB*. In the simplest case, Arg2 of a SUBCONJ is its complement in the syntactic representation. This is indeed the case for (11), where *since* is analyzed as a preposition in the PTB taking an S complement which is Arg2 in the PDTB, as shown in Figure 1.

(11) Since **the budget measures cash flow**, *a new $1 direct loan is treated as a $1 expenditure*.

Furthermore, in (11), *since* together with its complement (Arg2) is analyzed as an SBAR which modifies the clause *a new $1 direct loan is treated as a $1 expenditure*, and this clause is Arg1 in the PDTB.

Can the arguments always be detected in this way? In this section, we present statistics showing that this is not the case and an analysis that shows that this lack of congruence between the PDTB and the PTB is not just a matter of annotator disagreement.

Consider example (12), where the PTB requires annotators to include the verb of attribution *said* and its subject *Delmed* in the complement of *although*. But *although* as a discourse connective denies the expectation that the supply of dialysis products will be discontinued when the distribution arrangement ends. It does **not** convey the expectation that Delmed will not say such things. On the other hand, in (13), the contrast established by *while* is between the opinions of two entities i.e., *advocates* and *their opponents*.[4]

(12) *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.

(13) *Advocates said the 90-cent-an-hour rise, to $4.25 an hour by April 1991, is too small for the working poor*, while **opponents argued that the increase will still hurt small business and cost many thousands of jobs**.

In Section 5, we will identify additional cases. What we will then argue is that it will be insufficient to train an algorithm for identifying discourse arguments simply on the basis of syntactically analysed text.

We now present preliminary measurements of these and other *mismatches* between the two corpora for SUBCONJS. To do this we describe a procedural algorithm which builds on the idea presented at the start of this section. The statistics are preliminary in that only the annotations of a single annotator were considered, and we have not attempted to exclude cases in which annotators disagree.

We consider only those SUBCONJS for which both arguments are located in the same sentence as the connective (which is the case for approximately 99% of the annotated instances). The syntactic configuration of such relations pattern in a way shown in Figure 1. Note that it is not necessary for any of $Conn$, $Arg1$, or $Arg2$ to have a single node in the parse tree that dominates it exactly. In Figure 1 we do obtain a single node for $Conn$, and $Arg2$ but for $Arg1$, it is the set of nodes $\{NP, VP\}$ that dominate it exactly. Connectives like *so that*, and *even if* are not dominated by a single node, and cases where the annotator has decided that a (parenthetical) clausal element is not minimally necessary to the interpretation of $Arg2$ will necessitate choosing multiple nodes that dominate $Arg2$ exactly.

Given the node(s) in the parse tree that dominate $Conn$ ($\{IN\}$ in Figure 1), the algorithm we present tries to find node(s) in the parse tree that dominate $Arg1$ and $Arg2$ exactly using the operation of **tree subtraction** (Sections 4.1, and 4.2). We then discuss its execution on (11) in Section 4.3.

[4]This distinction is hard to capture in an RST-based parsing framework (Marcu, 2000). According to the RST-based annotation scheme (Carlson et al., 2003) 'although Delmed said' and 'while opponents argued' are elementary discourse units annotated in the same way: as satellites of the relation *Attribution*. RST does not recognize that satellite segments, such as the ones given above, sometimes participate in a higher RST relation along with their nuclei and sometimes not.

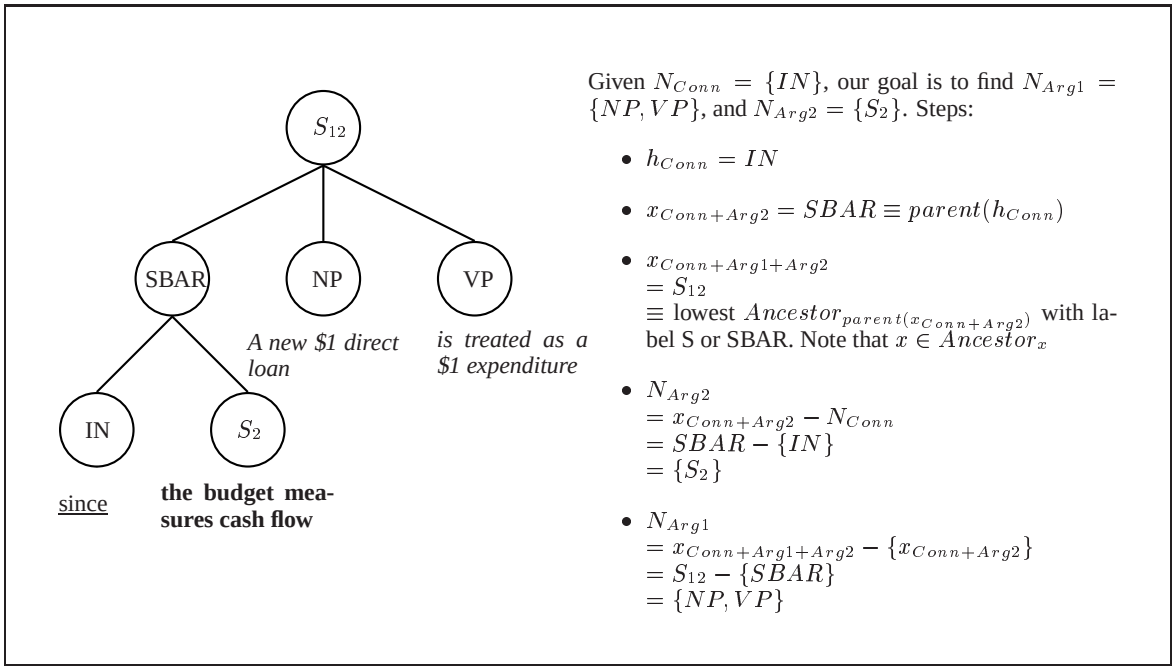

Figure 1: The syntactic configuration for (11), and the execution of the tree subtraction algorithm on this configuration.

4.1 Tree subtraction

We will now define the operation of tree subtraction the graphical intuition for which is given in Figure 2. Let T be the set of nodes in the tree.

Definition 4.1. *The ancestors of any node $t \in T$, denoted by $Ancestor_t \subseteq T$ is a set of nodes such that $t \in Ancestor_t$ and $parent(u, t) \Rightarrow ([u \in Ancestor_t] \wedge [Ancestor_u \subset Ancestor_t])$*

Definition 4.2. *Consider a node $x \in T$, and a set of nodes $Y \subset T - \{x\}$, we define the set $Z' = \{n | n \in T - \{x\} \wedge x \in Ancestor_n \wedge (\forall y \in Y, y \notin Ancestor_n \wedge n \notin Ancestor_y)\}$. Given such an x and Y, the operation of tree subtraction gives a set of nodes Z such that, $Z = \{z_1 | z_1 \in Z' \wedge (\forall z_2 \in Z', z_2 \notin (Ancestor_{z_1} - \{z_1\}))\}$*

We denote this by $x - Y = Z$.

The nodes $z \in Z$ are the highest descendants of x, which do not dominate any node $y \in Y$ and are not dominated by any node in Y.

4.2 Algorithm to detect the arguments

For any $t \in T$, let L_t denote the set of leaves(or terminals) dominated by t and for $A \subseteq T$ we denote the set of leaves dominated by A as $L_A = \bigcup_{\forall a \in A} L_a$.

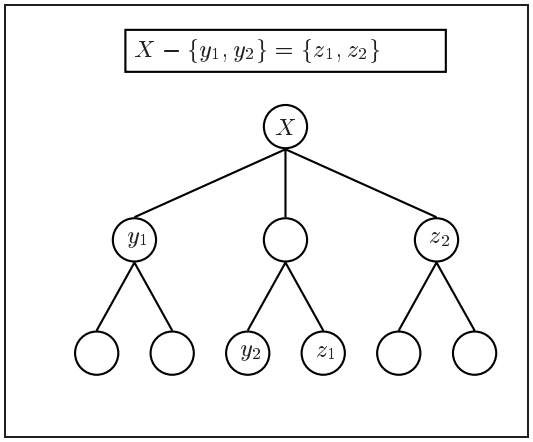

Figure 2: Tree subtraction $x - Y = Z$

For any set of leaves L we define N'_L to be a set of nodes of maximum cardinality such that $L_{N'_L} = \bigcup_{\forall n \in N'_L} L_n = L$

The set $N_L = \{n_1 | n_1 \in N'_L \wedge (\forall n_2 \in N'_L, n_2 \notin (Ancestor_{n_1} - \{n_1\}))\}$. We can think of Conn, Arg1 and Arg2 each as a set of leaves and we use N_{Conn}, N_{Arg1} and N_{Arg2} to denote the set of highest nodes which dominate them respectively.

Given N_{Conn}, our task is then to find N_{Arg1} and

N_{Arg2}. The algorithm does the following:

1. Let h_{Conn}(the head) be the last node in N_{Conn} in an in-order traversal of the tree.
2. $x_{Conn+Arg2} \equiv parent(h_{Conn})$
3. Repeat while $parent(x_{Conn+Arg2})$ has label S or SBAR, and has only two children:

   $x_{Conn+Arg2} = parent(x_{Conn+Arg2})$

   This ensures the inclusion of complementizers and subordinating conjuctions associated with the clause in Arg1. The convention adopted by the PDTB was to include such elements in the clause with which they were associated.
4. $x_{Conn+Arg1+Arg2}$ is the lowest node with label S or SBAR such that:

   $x_{Conn+Arg1+Arg2} \in Ancestor_{parent(x_{Conn+Arg2})}$
5. Repeat while $parent(x_{Conn+Arg1+Arg2})$ has label S or SBAR, and has only two children:

   $x_{Conn+Arg1+Arg2} = parent(x_{Conn+Arg1+Arg2})$
6. $N_{Arg2} = x_{Conn+Arg2} - N_{Conn}$ (tree subtraction)
7. $N_{Arg1} = x_{Conn+Arg1+Arg2} - \{x_{Conn+Arg2}\}$ (tree subtraction)

4.3 Executing the algorithm on (11)

The idea behind the algorithm is as follows. Since we may not be able to find a single node that dominates $Conn$, $Arg1$, and/or $Arg2$ exactly, we attempt to find a node that dominates $Conn$ and $Arg2$ together denoted by $x_{Conn+Arg2}$ ($SBAR$ in Figure 1), and a node that dominates $Conn$, $Arg1$ and $Arg2$ together denoted by $x_{Conn+Arg1+Arg2}$ (S_{12} in Figure 1). Note that this is an approximation, and there may be no single node that dominates $Conn$, and $Arg2$ exactly.

Given $x_{Conn+Arg2}$ the idea is to remove all the material corresponding to $Conn$ (N_{Conn}) under that node and call the rest of the material $Arg2$. This is what the operation of tree subtraction gives us, i.e., $x_{Conn+Arg2} - N_{Conn}$ which is $\{S_2\}$ in Figure 1.

Similarly, given $x_{Conn+Arg1+Arg2}$ we would like to remove the material corresponding to $Conn$ and $Arg2$ and $\{x_{Conn+Arg2}\}$ is that material. $x_{Conn+Arg1+Arg2} - \{x_{Conn+Arg2}\}$ gives us the nodes $\{NP, VP\}$ which is the desired $Arg1$.

5 Evaluation of the tree subtraction algorithm

Describing the mismatches between the syntactic and discourse levels of annotation requires a detailed analysis of the cases where the tree subtraction algorithm does not detect the same arguments as annotated by the PDTB. Hence this first set of experiments was carried out *only* on Sections 00-01 of the WSJ corpus (about 3500 sentences), which is accepted by the community to be development data.

First, the tree subtraction algorithm was run on the PTB annotations in these two sections. The arguments detected by the algorithm were classified as: (a) **Exact**, if the argument detected by the algorithm exactly matches the annotation; (b) **Extra Material**, if the argument detected contains some additional material in comparison with the annotation; and (c) **Omitted Material**, if some annotated material was not included in the argument detected. The results are summarized in Table 1.

Argument	Exact	Extra Material	Omitted Material
Arg1	82.5% (353)	12.6% (54)	4.9% (21)
Arg2	93.7% (401)	2.6% (11)	3.7% (16)

Table 1: Tree subtraction on the PTB annotations for SUBCONJS. Section 00-01(428 instances)

5.1 Analysis of the results in Table 1

5.1.1 Extra Material

There were 54 (11) cases where Arg1 (Arg2) in the PTB (obtained via tree subtraction) contained more material than the corresponding annotation in the PDTB. We describe only the cases for Arg1, since they were a superset of the cases for Arg2.

Second VP-coordinate - In these cases, Arg1 of the SUBCONJ was associated with the second of two coordinated VPs. Example (14) is the relation annotated by the PDTB, while (15) is the relation produced by tree subtraction.

(14) She became an abortionist accidentally, *and continued* because **it enabled her to buy jam, cocoa and other war-rationed goodies**.

(15) *She became an abortionist accidentally, and continued* because **it enabled her to buy jam, cocoa and other war-rationed goodies**.

Such mismatches can be either due to the fact that the algorithm looks only for nodes of type S or SBAR, or due to disagreement between the PTB and PDTB. Further investigation is needed to under-

stand this issue more precisely.[5] The percentage of such mismatches (with respect to the total number of cases of extra material) is recorded in the first column of Table 2, along with the number of instances in parentheses.

Lower Verb - These are cases of a true mismatch between the PDTB and the PTB, where the PDTB has associated Arg1 with a lower clause than the PTB. 9 of the 13 "lower verb" cases for Arg1 were due to *verbs of attribution*, as in (12). (The percentage of "lower verb" mismatches is given in the second column of Table 2, along with the number of instances in parentheses.)

Clausal Adjuncts - Finally, we considered cases where clause(s) judged not to be minimally necessary to the interpretation of Arg1 were included. (16) shows the relation annotated by the PDTB, where the subordinate clause headed by *partly because* is not part of Arg1, but the tree subtraction algorithm includes it as shown in (17).

(16) When **Ms. Evans took her job**, *several important divisions that had reported to her predecessor weren't included* partly because she didn't wish to be a full administrator.

(17) When **Ms. Evans took her job**, *several important divisions that had reported to her predecessor weren't included partly because she didn't wish to be a full administrator*.

To get an idea of the number of cases where a single irrelevant clause was included, we determined the number of instances for which pruning out one node from Arg1 resulted in an exact match. This is given in the third column of Table 2. The second row of Table 2 illustrates the same information for Arg2. Most of these are instances where irrelevant clauses were included in the argument detected from the PTB.

Argument	Second VP Coordinate	Lower Verb	One Node Pruned	Other
Arg1	16.7% (9)	24.1% (13)	31.5% (17)	27.7% (15)
Arg2	0% (0)	9.1% (1)	72.7% (8)	18.2% (2)

Table 2: Cases which result in extra material being included in the arguments.

[5]It is also possible for the PDTB to associate an argument with only the first of two coordinated VPs, but the number of such cases were insignificant.

5.1.2 Omitted Material

The main source of these errors in Arg1 are the **higher verb** cases. Here the PDTB has associated Arg1 with a higher clause than the PTB. Examples (18) and (19) show the annotated and algorithmically produced relations respectively. This is the inverse of the aforementioned *lower verb* cases, and the majority of these cases are due to the verb of attribution being a part of the relation.

(18) *Longer maturities are thought to indicate declining interest rates* because **they permit portfolio managers to retain relatively higher rates for a longer period**.

(19) Longer maturities are thought *to indicate declining* interest rates because **they permit portfolio managers to retain relatively higher rates for a longer period**.

To get an approximate idea of these errors, we checked if selecting a higher S or SBAR made the Arg1 exact or include extra material. These are the columns **Two up exact** and **Two up extra included** in Table 3. At this time, we lack a precise understanding of the remaining mismatches in Arg1, and the ones resulting in material being omitted from Arg2.

Argument	Two up exact	Two up extra included	Other
Arg1	47.6% (10)	14.3% (3)	28.1% (8)

Table 3: Cases which result in material being omitted from Arg1 as a result of excluding a higher verb

5.2 Additional experiments

We also evaluated the performance of the tree subtraction procedure on the PTB annotations on Sections 02-24 of the WSJ corpus, and the results are summarized in Table 4.

Argument	Exact	Extra Material	Omitted Material
Arg1	76.1%	17.6%	6.3%
Arg2	92.5%	3.6%	3.9%

Table 4: Tree subtraction on PTB annotations for the SUBCONJS(approx. 5K instances). Sections 02-24

Finally we evaluated the algorithm on the output of a statistical parser. The parser implementation in (Bikel, 2002) was used in this experiment and it was run in a mode which emulated the Collins (1997) parser. The parser was trained on Sections 02-21 and Sections 22-24 were used as test data, where

the parser was run and the tree subtraction algorithm was run on its output. The results are summarized in Table 5.

Argument	Exact	Extra Material	Omitted Material
Arg1	65.5%	25.2%	9.3%
Arg2	84.7%	0%	15.3%

Table 5: Tree subtraction on the output of a statistical parser (approx. 600 instances). Sections 22-24.

6 Conclusions

While it is clear that discourse annotation goes beyond syntactic annotation, one might have thought that at least for the annotation of arguments of subordinating conjunctions, these two levels of annotation would converge. However, we have shown that this is not always the case. We have also described an algorithm for discovering such divergences, which can serve as a useful baseline for future efforts to detect the arguments with greater accuracy. The statistics presented suggest that the annotation of the discourse arguments of the subordinating conjunctions needs to proceed separately from syntactic annotation – certainly when annotating other English corpora and very possibly for other languages as well.

A major source of the mismatches between syntax and discourse is the effect of attribution, either that of the arguments or of the relation denoted by the connective. We believe that the annotation of attribution in the PDTB will prove to be a useful aid to applications that need to detect the relations conveyed by discourse connectives with a high degree of reliability, as well as in constraining the inferences that may be drawn with respect to the writer's commitment to the relation or the arguments. The results in this paper also raise the more general question of whether there may be other mismatches between the syntactic and discourse annotations at the sentence level.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Press.

Daniel Bikel. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *HLT*.

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, pages 85–112. Kluwer Academic Publishers.

Michael Collins. 1997. Three Generative, Lexicalized Models for Statistical Parsing. In *35th Annual Meeting of the ACL*.

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

Daniel Marcu. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach. *Computational Linguistics*, 26(3):395–448.

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

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004a. Annotating Discourse Connectives and their Arguments. In *the HLT/NAACL workshop on Frontiers in Corpus Annotation*, Boston, MA.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004b. The Penn Discourse Treebank. In *the Language Resources and Evaluation Conference*, Lisbon, Portugal.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and Data Mining of the Penn Discourse TreeBank. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain.

Ellen Riloff and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP '03)*, pages 105–112, Sapporo, Japan.

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, Canada, August.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse Relations: A Structural and Presuppositional Account using Lexicalized TAG. In *ACL*, College Park, MD, June.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and Discourse Structure. *Computational Linguistics*, 29(4):545–87.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

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

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