The Penn Discourse TreeBank as a Resource for Natural Language Generation

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki
Institute for Research in Cognitive Science, University of Pennsylvania
rjprasad,joshi,nikhild,aleewk,elenimi@linc.cis.upenn.edu
Bonnie Webber
Division of Informatics, University of Edinburgh
bonnie@inf.ed.ac.uk

Abstract

While many advances have been made in Natural Language Generation (NLG), the scope of the field has been somewhat restricted because of the lack of annotated corpora from which properties of texts can be automatically acquired and applied towards the development of generation systems. In this paper, we describe how the Penn Discourse TreeBank (PDTB) can serve as a valuable large scale annotated corpus resource for furthering research in NLG and for inducing models for the development of NLG systems. The PDTB is annotated for discourse relations, and encodes explicitly the elements of these relations: explicit and implicit discourse connectives, denoting the predicates of the relations, and text spans, denoting the arguments of the relations. Connectives and arguments are also annotated with features and spans related to attribution, and each connective will be annotated with labels standing for the projected discourse relation, including sense distinctions for polysemous connectives. We exemplify the use of the corpus for two tasks in NLG: the realization of discourse relations during sentence planning, and the representation and realization of attribution.

1 Introduction

Many NLG systems, especially those that follow the pipelined architecture [Reiter, 1994] comprising modules for *content determination* (aka *text planning*), *microplanning* (aka *sentence planning*) and *surface realization*, do not specify discourse relations between elementary content units, ECUs (i.e., propositions denoting the simplest predication over entities), in the output of the text planner (e.g., [Rambow and Korelsky, 1992; Walker *et al.*, 2001]). Some that do represent discourse relations, do so by using (pre-defined) schemas - broadly following [McKeown, 1985] - and assume a one-to-one mapping between discourse relations and discourse connectives (e.g., [Davey, 1979; Hovy, 1987]).[1] The shortcoming of such systems is that their flexibility is severely restricted by the schemas because of which they are also not easily portable to other domains.

In contrast, an increasing number of systems have extended their text planning components to represent discourse relations in the text plan – e.g., [Hovy, 1993; Mellish *et al.*, 1998; Walker *et al.*, 2003]. However, with the lack of a complete understanding of discourse relations and of the ways they can be realized in text, what such systems now require are corpus resources from which to derive knowledge needed by the next module in the pipeline (*sentence planning*) to model the interaction of aggregation and discourse relations.[2] In this paper, we discuss what the PDTB [Miltsakaki *et al.*, 2004; Prasad *et al.*, 2004; Webber *et al.*, 2005] can contribute to natural language generation, focusing on the sentence planning task of discourse relation lexicalization (DR-lexicalization), including *occurrence*, *selection*, and *placement* [Moser and Moore, 1995], and on the representation of attribution in the text plan as well as its interaction with aggregation.

In Section 2, we give a brief overview of the Penn Discourse TreeBank annotations. In Section 3 we discuss the relevance of the PDTB for sentence planning tasks, addressing the DR-lexicalization problems of *occurrence*, *selection*, and *placement* in detail. In Section 4, we discuss the attribution annotations in the PDTB and show how they can be useful for the representation of attribution for content determination, as well as for their realization during sentence planning. We summarize in Section 5.

2 The Penn Discourse TreeBank

The PDTB contains annotations of *explicit* and *implicit* discourse connectives and their arguments on the 1 million word Wall Street Journal corpus. Following the views toward discourse structure in [Webber *et al.*, 2003], the PDTB treats discourse connectives as discourse-level predicates that take **two** *abstract objects* such as events, states, and propositions

[1] [Rambow and Korelsky, 1992] also use schemas but do not represent discourse relations between the units.

[2] Here, as for the rest of this paper, we define aggregation in its broadest possible sense, to stand for any syntactic transformation that can be applied to two *lexicalized* content units (CUs), including the PERIOD operation, which is applied to generate two input CUs as two sentences, as well as the (non-)realization of discourse relations. For the purpose of this paper, we assume that content units are lexicalized before aggregation, but that discourse relations are lexicalized during or after aggregation.

[Asher, 1993] as their arguments. For example, in (1), the subordinating conjunction *since* is a discourse connective that establishes a TEMPORAL relation between the event of the earthquake hitting and a state where no music is played by a certain woman.[3]

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

The following four classes of explicit connectives are annotated in the PDTB (Examples provided for each class are only a few of those annotated in the PDTB - see Section 2.4.):

- *subordinating conjunctions*, both bare (e.g., *because*, *when*, *since*, *although*) and with a modifier (e.g., *only because*, *particularly since*, *even after*),
- *subordinators* (e.g., *in order that*, *except that*),[4]
- *coordinating conjunctions* (e.g., *and*, *or*, *nor*), and
- *discourse adverbials* (e.g., *however*, *otherwise*, *then*).[5,6]

Because there are, as yet, no generally accepted abstract semantic categories for classifying the arguments to discourse connectives as have been suggested for verbs (e.g., *agent*, *patient*, *theme*, etc.), the two arguments to a discourse connective are 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. For the subordinating conjunctions, since the subordinate clause is bound to the connective, *Arg2* corresponds to the subordinate clause, and hence the linear order of the arguments can be *Arg1-Arg2* (Ex. 2), *Arg2-Arg1* (Ex. 3), or *Arg2* may appear embedded in *Arg1* (Ex. 4), depending on the relative position of the subordinate clause with respect to its governing matrix clause.

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

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

(4) *Most oil companies*, when **they set exploration and production budgets for this year**, *forecast revenue of $15 for each barrel of crude produced*.[7]

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

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

Abstract objects can be arbitrarily complex in the PDTB so that arguments of connectives can be associated with single clauses, multiple clauses, single sentences, or multiple sentences. However, a *minimality principle* requires an argument to contain the minimal amount of information needed to complete the interpretation of the relation. Any other span of text that is perceived to be relevant (but not necessary) in some way to the interpretation of arguments is annotated as *supplementary information*, labelled *Sup1*, for material supplementary to *Arg1*, and *Sup2*, for material supplementary to *Arg2*.

Also as a consequence of the abstract object characterization of arguments, arguments may be denoted by non-clausal units such as *nominalizations* that have an event interpretation, and *discourse deictics* (*this*, *that*) that refer to abstract objects.

2.1 Implicit Connectives

Implicit connectives are annotated in the PDTB between adjacent sentences when no connective appears explicitly to relate the second sentence to the first. For example, in (6), the second sentence is related to the first via an EXPLANATION relation (i.e, Mr. Breeden's wise perception of the ways of Washington is being used as an explanation for the assertion that he may be able to succeed), but this relation is not expressed explicitly.

(6) *Also unlike Mr. Ruder, Mr. Breeden appears to be in a position to get somewhere with his agenda*. IMPLICIT=BECAUSE **As a former White House aide who worked closely with Congress, he is savvy in the ways of Washington**.

Annotation at such points consists of a record of an explicit connective that "best" conveys the implicit relation perceived as holding between the adjacent sentences. For the implicit relation perceived in Example (6), *because* is recorded as the connective. In order to account for multiple "simultaneous" relations between the same two abstract objects, there may also be more than one connective between the sentences. Example (7) shows an annotation where two relations were perceived as holding simultaneously, and for which the connectives *when* and *for example* were recorded.

[3] The assumption of the arity constraint of the arguments has been upheld in all the annotation done thus far. Discourse-level predicate-argument structures are therefore unlike the predicate-argument structures of verbs at the sentence-level (PROPBANK [Kingsbury and Palmer, 2002]), where verbs can take any number of arguments.

[4] The class of subordinators was added at a later stage.

[5] *Discourse adverbials* are to be distinguished from *clausal adverbials* (see [Forbes, 2003]).

[6] Discourse markers such as *well*, *anyway*, *now*, etc., that signal the organizational or focus structure of the discourse, are not annotated.

[7] As this example shows, annotations in the PDTB can be discontinuous, a feature allowed by *WordFreak*, the discourse annotation tool (developed by Tom Morton and Jeremy Lacivita). Discontinuous annotation is possible for connectives as well, such as for *on the one hand . . . on the other hand*.

(7) *The small, wiry Mr. Morishita comes across as an outspoken man of the world.* IMPLICIT=WHEN IMPLICIT=FOR EXAMPLE ($_{sup2}$ 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**.

As examples (6) and (7) show, the annotation of implicit connectives also includes the marking of the textual span from the two adjacent sentences that are the arguments of the inferred implicit relation. That the spans selected for the two arguments need not (trivially) constitute the entire sentence can be seen in Example (7).

At the current stage of the project, implicit connectives between adjacent sentences across paragraphs, and intra-sentential connectives (such as those occurring with *free adjuncts*) are not annotated.

2.2 Sense Annotation

All explicit and implicit connectives in the PDTB will be annotated with labels for the discourse relation that they denote, including sense distinctions for polysemous connectives (e.g., *since, while, if, when, because*). For example, *since* seems to have three different senses, one purely TEMPORAL (as in Ex. 8), another purely CAUSAL (as in Ex. 9) and a third both CAUSAL and TEMPORAL (as in Ex. 10).

(8) *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**.

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

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

2.3 Attribution Annotation

Attribution, which has to do with ascribing beliefs and assertions expressed in text to the agent(s) holding or making them, is annotated in the PDTB to primarily distinguish between two different sources of attribution, the Writer of the text ("Writer attribution"), or some other Speaker (or Agent) mentioned by the Writer ("Speaker Attribution"). With respect to attribution associated with discourse connectives and their arguments, there are broadly two possibilities:[8]

Case 1 A discourse connective and both its arguments are attributed to the same source, either the Writer, as in Example (1), or the Speaker (Bill Biedermann) in Example (11):

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

Case 2 One or both arguments have a different attribution value from the discourse connective. In Example (12), the connective and *Arg1* are attributed to the Writer, whereas *Arg2* is attributed to another Speaker (here, the purchasing agents):[9]

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

Attribution tags in the PDTB are currently being further refined (while still 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. In addition, the second release of the PDTB will also record the text span associated with the source and type of attribution. For further discussion of attribution annotation in PDTB, see [Dinesh *et al.*, 2005].

2.4 Summary and Project Goals

The first release of the PDTB (November 2005) will contain approx. 16K annotations of explicit connectives (approx. 6000 subordinating conjunctions, 5000 discourse adverbials and 5000 coordinating conjunctions) and approx. 20K annotations of implicit connectives. There are over 90 different types of connectives.

3 PDTB and Sentence Planning

Following the introduction of *sentence planning* as an independent intermediate stage [Rambow and Korelsky, 1992] in the traditional two-way split of NLG systems into a content determination component and a realization component, discourse connectives have invited a great deal of research in NLG, as they are related simultaneously to the *aggregation* and *lexicalization* tasks in sentence planning. Assuming the basic three-way pipelined architecture [Reiter, 1994], the input to the sentence planning component is, thus, taken to be a hierarchically ordered text plan structure that encompasses all the elementary content units (ECUs) that the system has decided to generate, as the leaves of the structure, with the internal nodes specifying the discourse relations holding between ECUs or groups of ECUs.[10]

In the pipelined architecture, sentence planning is relieved of decisions related to content determination, specifically that of determining the discourse relation that holds between the CUs, so that it can focus on the problem of how to *express* the discourse relations. Work on connective usage [Moser and Moore, 1995] has identified three separate but related decision making processes during sentence planning, for the generation of discourse connectives: (a) *occurrence*, i.e., whether to generate a connective or not; (b) *selection*, i.e., which connective to generate; and (c) *placement*, i.e., where to place the connective.

[8] Attribution is annotated for both explicit and implicit connectives.

[9] When attribution of a connective or its arguments is uncertain, the attribution is assigned to the Writer as a Default.

[10] The structure of the text plan is the same irrespective of whether the text planning task is done in a schema-based top-down manner [McKeown, 1985] or in a bottom-up manner [Marcu, 1997].

While many insightful studies have been carried out on discourse connectives for generation purposes, they have either singled out a few connectives (e.g., [Elhadad and McKeown, 1990; Dorr and Gaasterland, 1995; Rösner and Stede, 1992], or proposed heuristics based on a small number of constructed examples (e.g., [Scott and Souza, 1990]), or proposed classification-based lexicons that are very hand-intensive to build, (e.g., [Grote and Stede, 1998; Knott and Mellish, 1996]), especially in a multilingual context. In contrast, corpora annotated with information about connectives can provide a useful knowledge source from which to automatically induce properties of connectives designing sentence planning tasks. In the rest of this section, we discuss how the PDTB annotations of discourse connectives and their arguments can be useful towards the three tasks of connective generation discussed above.

3.1 Occurrence

As corpus studies [Moser and Moore, 1995; Williams and Reiter, 2003] have shown, more often than not, there is **no** discourse connective explicitly connecting a clause to the previous discourse. When combining two CUs during aggregation, the sentence planner has to make a choice about whether to generate a connective or not. The two important questions to ask here is whether there are significant constraints on the lexicalization of discourse relations, and whether some or all of these constraints can be identified directly from annotated corpora. For example, what is the reason for lexicalizing the CONSEQUENCE relation in (13) and not in (14)?

(13) *The three men worked together on the so-called Brady Commission, headed by Mr. Brady, which was established after the 1987 crash to examine the market's collapse*. As a result **they have extensive knowledge in financial markets, and financial market crises**.

(14) From 1984 to 1987, its (Iverson's) earnings soared six-fold, to $3.8 million, on a seven-fold increase in revenue, to $44.1 million. *But in 1988, it ran into a buzz saw: a Defense Department spending freeze*. IMPLICIT=AS A RESULT **Iverson's earnings plunged 70% to $1.2 million**.

Some research has shown that the choice of whether or not to lexicalize a relation is indeed governed by constraints that can be built into a sentence planner: while some constraints may require deep reasoning over world knowledge and properties of the content units, some are more directly associated with the surface properties of the content units and the text plan. For example, on the one hand, [Amsili and Rossari, 1998] show that in French, the use of a connective to express a CAUSAL relation between eventualities can be constrained by the interaction of the (Vendlerian) *aspectual classes* of the two eventualities as well as the order in which the eventualities appear in the CAUSAL relation. On the other hand, corpus-based research [Williams and Reiter, 2003] has shown that there are statistically significant differences across *classes of connectives* with respect to *how frequently they are lexicalized*.

Like other corpus based work, the PDTB also offers the opportunity to find statistically significant patterns. However, it also offers much more, since, unlike previous studies, the size of the corpus is much larger, and since other layers of annotation on the same text are also available, namely the syntactic annotation of the Penn TreeBank and the semantic annotation of the PropBank. A generation system with an architecture such as the one we have assumed here, provides a syntactic and semantic analysis of the content units to the sentence planning component, so models induced from the PDTB will provide a much richer set of constraints to constitute the criteria for the (non-)lexical occurrence of discourse relations.

Finally, for the occurrence task, it is also useful that the PDTB specially identifies cases where there is no explicit phrase that can be inserted in place of the purported implicit relation between adjacent sentences. These cases were analyzed and distinguished as three types: (a) NOREL, where no discourse relation was inferred between the adjacent sentences (Ex. 15), (b) NOCONN-ENT, where the relation was perceived to be one established by elaboration via entity description (Ex. 16), and finally, (c) NOCONN, where some relation - other than the entity elaboration relation - was perceived, but for one of several reasons, including redundancy, the use of an explicit connective in the text sounded unacceptable (Ex. 17).[11]

(15) *The transaction has been approved by Kyle's board, but requires the approval of the company's shareholders*. IMPLICIT=NOREL **Kyle manufactures electronic components**.

(16) *C.B. Rogers Jr. was named chief executive officer of this business information concern*. IMPLICIT=NOCONN-ENT **Mr. Rogers, 60 years old, succeeds J.V. White, 64, who will remain chairman and chairman of the executive committee**.

(17) *In the 1920s, a young schoolteacher, John T. Scopes, volunteered to be a guinea pig in a test case sponsored by the American Civil Liberties Union to challenge a ban on the teaching of evolution imposed by the Tennessee Legislature*. IMPLICIT-NOCONN **The result was a world-famous trial exposing profound cultural conflicts in American life between the "smart set,"** . . . **and the religious fundamentalists,** . . .

In sum, we hope that the implicit connective annotations in the PDTB will encourage research and experiments that will support decisions related to connective occurrence in NLG.

3.2 Selection

The problem of lexical choice for connectives is well recognized: a given discourse relation can be expressed with a variety of connectives, but there are subtle syntactic, semantic, pragmatic, and stylistic factors that preclude the use of any of a class of connectives in a given context.

[11] One reason that an explicit connective might sound redundant is if the relation is already lexicalised elsewhere in the clause - for example, in the subject (as in Ex. 17) or the verb, as in "This [*resulted in, led to*] a world-famous trial . . . ."

The design of the PDTB annotations provides direct access to the discourse connectives and their arguments (since the annotations are anchored on the connectives), and together with the other layers of annotation (PTB and PropBank), can allow inferences to be drawn easily from the observed patterns. Claims made in the literature about particular connectives can also be empirically tested. In some studies that we have conducted, there are indications that some well-known accounts of connectives are not supported by the PDTB annotations. For example, [Elhadad and McKeown, 1990] argue that while the CAUSE relation can be expressed by both *because* and *since*, the two connectives are not freely interchangeable, and that they differ in whether the information is known to the receiver or not, in that *because* introduces *new* information whereas *since* presents *given* information. Crucially, their account is based on an elsewhere claimed tendency [Quirk *et al.*, 1972] for *because* and *since* to be in complementary distribution, with *because* appearing postposed and *since* appearing preposed, and the notion that *new information* tends to be placed towards the end of a clause [Halliday, 1985]. The PDTB annotations show that while *because* does tend to appear postposed, (see [Prasad *et al.*, 2004]), the 90 confirmed instances of CAUSAL *since* are distributed equally in pre- and postposed position, suggesting a clarification of the above correlation between information status and clause order.

In addition, in earlier work [Prasad *et al.*, 2004], we had integrated a subset of the PDTB annotations with the PTB syntactic annotations, and found that *although* and *even though*, which denote a CONCESSION relation and are thought to be undifferentiable except for *even though* carrying "emphasis" [Huddleston and Pullum, 2002], behaved quite differently with respect to the relative position of their arguments. *Although* clauses were more frequently preposed, whereas *even though* clauses were more frequently postposed. To this, we have now added results obtained for *though*. Table 1 shows the argument-order distribution for *although*, *even though*, and *though*.[12]

CONN	*Arg2* Postposed	*Arg2* Preposed	Total
although	129 (37%)	218 (63%)	347
even though	77 (75%)	26 (25%)	103
though	97 (70%)	42 (30%)	139
Total	303 (51%)	286 (49%)	589

Table 1: Argument Order for *although*, *even though*, and *though*.

Table 1 shows that *even though* and *though* pattern alike, and that their variation with *although* is highly significant. The former occur postposed about 72% of the time and preposed about 28% of the time, the opposite of *although* (see Table 2.) Further analysis of these connectives is needed to determine what the variation might correlate with.[13]

Some recent studies (e.g., [Hutchinson, 2005]) have tried to model the substitutability of discourse connectives based on corpus data. However, the model uses only lexical cooccurrences. We believe that better models could be obtained with corpora such as the PDTB that are aligned with other levels of (syntactic and semantic) analysis. This would be especially beneficial for generation approaches (such as is assumed here) that carry out the task of lexical choice for connectives *after* the abstract syntactic specification for the connective's arguments have already been constructed. This means that syntactic and semantic features of the CUs can play a role in modeling the use of connectives.

CONN	*Arg2* Postposed	*Arg2* Preposed	Total
although	129 (37%)	218 (63%)	347
(even) though	174 (72%)	68 (28%)	242
Total	303 (51%)	286 (49%)	589

Table 2: Argument Order for *although* and *(even) though*, with *though* and *even though* combined.

3.3 Placement

When relating two CUs, discourse connectives are syntactically bound to one of the CUs (called *Arg2* in the PDTB), so the sentence planner needs to make a decision about (a) which CU to associate the connective with, and (b) where to place the connective in the CU.

The first decision can mostly be made very simply by reference to the relative order of the CUs and the syntactic class of the connective, if it is assumed that linear ordering of the CUs during aggregation is done prior to DR-lexicalization (see Footnote 2). For instance, in both Examples (18) and (19),[14] a CONCESSION relation holds between the two CUs, in that the assertion of John being smart denies the expectation raised by the other assertion, that John is not smart. However, in each case, the linear ordering of the CUs is taken as given for the placement task, and the decision of where to place the connective depends on the relative position of the CU that raises the expectation to be denied, and the syntactic class of the connective selected. If *although*, a subordinating conjunction, is selected, it must be associated with the CU that raises the expectation, whereas if *but*, a coordinating conjunction, is selected, it must be associated with the CU that denies the expectation.

(18) Although John failed the exam, he is smart.

(19) John failed the exam but he is smart.

The second decision, however, is more difficult to make, and relates to discourse adverbials. Unlike subordinating conjunctions and coordinating conjunctions, which can modify their (*Arg2*) CU clause only in initial position, discourse adverbials can occur in several positions in the clause. The examples below show the connective *as a result* appearing in initial position (Ex. 20), in medial position (Ex. 21), and in final position (Ex. 22) in the clause.

[12] The tokens for *though* exlcude its adverbial occurrences.

[13] The similar behavior of *even though* and *though* also suggests that the claim about "emphasis" being the sole distinguishing feature might still stand, but only for these two connectives.

[14] These examples are adapted from [Elhadad and McKeown, 1990].

(20) *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**.

(21) *The chief culprits*, he says, *are big companies and business groups that buy huge amounts of land "not for their corporate use, but for resale at huge profit."* . . . **The Ministry of Finance**, as a result, **has proposed a series of measures that would restrict business investment in real estate** . . .

(22) *Polyvinyl chloride capacity "has overtaken demand* and **we are experiencing reduced profit margins** as a result.", . . .

In previous work [Prasad *et al.*, 2004], we conducted experiments on 5 adverbials (*as a result*, *instead*, *nevertheless*, *otherwise*, and *therefore*), looking at the position in which the connective was realized in the *Arg2* CU clause, and we found that the connectives in this set occurred predominantly in initial position in their clause. However, most of the examples collected for these experiments had the connective in initial position, so we want to re-run the experiments when we have further data.

4 PDTB and the Representation and Realization of Attribution

Taking the theoretical view of language as goal-driven communication, most working NLG systems are built within restricted domains with clearly defined communicative goals. This means that while some tasks, such as some aspects of sentence planning and realization, are modeled in a general way and can be extended across applications, other tasks such as content determination and text planning are driven by the needs of the domain, in particular the information content of the domain. One of the first tasks of NLG systems is thus *domain modeling*, i.e., analyzing target texts and declaring the different types of information that need to be conveyed within the domain, and at what level of granularity. For example, in the restaurant review domain [Walker *et al.*, 2003], the primary kind of entity is *restaurant* with properties like *food*, *service*, and *atmosphere* defined over these entities. In contrast, the weather domain [Reiter and Dale, 2000] includes *time-span* entities, with properties like *rainfall* defined over these entities.

In the News domain, applications that deal with the generation of News reports have to model more complex types of entities and relations, such as the relation of attribution, which is a relation of "ownersip" between abstract objects and individuals or agents. Because they are News reports, writers of these texts are concerned with ascribing beliefs held and statements made to the correct sources (including themselves), and relatedly, with preventing the false inference (on the part of the reader) that a certain piece of information that is being conveyed is a commonly known fact or commonly held view). For example, there is a big difference between the way the basic information in the following two sentences is presented, because it leads the reader of the report to make different inferences about the facts.

(23) The chief culprits, he (Mr. Lee) says, are big companies and business groups that buy huge amounts of land "not for their corporate use, but for resale at huge profit."

(24) The chief culprits are big companies and business groups that buy huge amounts of land "not for their corporate use, but for resale at huge profit."

In Example (23), the assertion that the chief culprits are big companies and business groups is understood as "fact", but from Mr. Lee's point of view.[15] On the other hand, Example (24) strongly suggests an interpretation where the same assertion is considered to be a well-known fact (and hence to be "true").

Even though the attribution relation in the domain model is expressed between all abstract objects and agent entities that are related in this way, the relation information need not necessarily continue on to successive stages of content determination and text planning, at least not in the same manner. For one thing, the content planner may decide to present some information as well-known fact even though it was attributed to some agent. This is somewhat evident from the texts themselves, where many sentences suggest that the writer holds a strong belief that the information from some external source is true, or is confident about its truth, and has decided to drop the attribution, i.e, not realize it. This is a common idea in the definition of ECUs in generation systems, where going from the information contained in the domain model to the definition of the ECUs involves making decisions about what to convey and what not to convey (depending on the overall and immediate communicative goals), especially since the idea is that all pieces of the defined ECUs have to be realized in one way or another.

Secondly, even if attributions are included in the ECUs for realization, the content planner can, and must, have a way of representing them in at least two different ways. Evidence for this comes from the PDTB annotation of attribution on discourse connectives and their arguments. Assuming that the text plan encodes the ECUs and the discourse relations holding between them, a discourse relation may hold either between the attributions themselves or just between the abstract object arguments of the attribution. These two possiblities are shown in Examples (25) and (26):

(25) When **Mr. Green won a $240,000 verdict in a land condemnation case against the state in June 1983**, he says *Judge O'Kicki unexpectedly awarded him an additional $100,000*.

(26) *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**.

[15] In the examples in this section, the text span associated with attribution is shown boxed for illustrative purposes only: the guidelines for annotating attribution spans have not been decided yet.

In Example (25), the TEMPORAL discourse relation is expressed between the eventuality of Mr. Green winning the verdict and the Judge giving him an additional award. The discourse relation does not entail the interpretation of the attribution relation. On the other hand, Example (26) shows that the CONTRAST relation holds between the agent arguments of the attribution relation, which means that the attribution relation must be part of the contrast as well. These examples therefore show that attribution has both an "owner" and a scope, and that both must be correctly represented in the text plan, for appropriate realization.

The attribution relation can be defined over discourse relations as well, as seen in Example (27), where the TEMPORAL relation between the two arguments is presumably also being quoted and thus attributed to some Speaker (some managing director). If discourse relations can indeed be objects of attribution, then it suggests a further extension of the abstract object ontology (in the domain model) and its representation by the content planner.

(27) "When **the airline information came through**, *it cracked every model we had for the marketplace*," said a managing director at one of the largest program-trading firms.

With the above examples, we have illustrated one aspect of the annotation in the PDTB that has a bearing on how NLG systems working within the News report domain must model and represent attribution. The completed attribution annotations in the PDTB corpus will provide a useful resource as a target corpus to confirm the above hypotheses, and to discover other aspects relevant to the representation of attribution.

Carrying on from the stages of domain modeling and content determination, the PDTB annotation shows that the attribution relation continues to affect sentence planning decisions as well. The attribution relation can be realized in a variety of ways: Speaker attribution of a connective and both its arguments can involve either quoted or indirect speech, as in Examples (28) and (29), respectively.

(28) "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*."

(29) 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**.

Example (28) also shows that the attribution may be unrealized when the abstract object argument of the attribution is expressed in direct quotes, leaving it to the reader to recover the attribution anaphorically from the preceding sentence.

Finally, Example (30) shows that both arguments of the CONTRAST relation signaled by *nevertheless* are attributed to a Speaker, Mr. Robinson, whereas the relation itself is attributed to the Writer.

(30) Mr. Robinson . . .said *Plant Genetic's success in creating genetically engineered male steriles doesn't automatically mean it would be simple to create hybrids in all crops*. . . . Nevertheless, he said, **he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton**.

Space does not permit us to present the many more ways in which the attribution relation is realized in the PDTB. We hope that the examples provided here will encourage researchers to exploit the PDTB to automatically discover the full range of variation that seems to be present, and model the conditions under which the different realizations are generated. Apart from the interaction of attribution with aggregation, since the PDTB annotation will contain the span of text associated with attribution, it will also provide a valuable resource for determining the different ways in which different types of attribution can be lexicalized, for example, for the choice between different verbs of saying, such as *say* and *note*, and to model the conditions under which one may be used as against the other.

5 Summary

In this paper, we have described how the Penn Discourse TreeBank (PDTB), a large-scale multi-layered annotated corpus of discourse relations, can contribute as a resource towards research and development in NLG. We gave an overview of the types of annotation in the PDTB: explicit and implicit discourse connectives and their arguments, sense distinctions for connectives, and attribution associated with connectives and their arguments. We showed the relevance and use of the annotations for the sentence planning tasks of occurrence, selection, and placement, giving results from experiments conducted on the PDTB, and showing how the results could be integrated into an NLG sentence planner. We also showed that empirical research on the PDTB can be used to test theoretical claims made in the literature about discourse connectives. We then described the role of attribution for domain modeling of applications that deal with generation of WSJ style texts and demonstrated the utility of the PDTB annotations for the representation and realization of attribution.

Acknowledgements

The Penn Discourse TreeBank project is partially supported by NSF Grant: Research Resources, EIA 02-24417 to the University of Pennsylvania (PI: A. Joshi).

References

[Amsili and Rossari, 1998] Pascal Amsili and Corinne Rossari. Tense and connective constraints on the expression of causality. In *Proc. COLING-ACL*, pages 48–54, 1998.

[Asher, 1993] Nicholas Asher. *Reference to Abstract Objects*. Kluwer, Dordrecht, 1993.

[Davey, 1979] Anthony Davey. *Discourse Production*. Edinburgh Univ. Press, 1979.

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

[Dorr and Gaasterland, 1995] Bonnie J. Dorr and Terry Gaasterland. Selecting tense, aspect and connecting words in language generation. In *Proc. IJCAI*, pages 1299–1305, 1995.

[Elhadad and McKeown, 1990] Michael Elhadad and Kathleen R. McKeown. Generating connectives. In *Proc. COLING*, volume 3, pages 97–101, 1990.

[Forbes, 2003] Katherine Forbes. *Discourse Semantics of S-modifying Adverbials*. PhD thesis, Univ. of Penn., 2003.

[Grote and Stede, 1998] Brigitte Grote and Manfred Stede. Discourse marker choice in sentence planning. In *Proc. INLG*, pages 128–137, 1998.

[Halliday, 1985] Michael A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.

[Hovy, 1987] Eduard Hovy. *Generating Natural Language under Pragmatic Constraints*. PhD thesis, Yale Univ., 1987.

[Hovy, 1993] Eduard H. Hovy. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63:341–385, 1993.

[Huddleston and Pullum, 2002] Ronald Huddleston and Geoffrey Pullum. *The Cambridge Grammar of the English Language*. Cambridge Univ. Press, Cambridge, UK, 2002.

[Hutchinson, 2005] Ben Hutchinson. Modeling the substitutability of discourse connectives. In *Proc. ACL*, 2005.

[Kingsbury and Palmer, 2002] Paul Kingsbury and Martha Palmer. From TreeBank to PropBank. In *Proc. LREC*, 2002.

[Knott and Mellish, 1996] Alistair Knott and Chris Mellish. A feature-based account of the relations signalled by sentence and clause connectives. *Language and Speech*, 39(2-3):143–183, 1996.

[Marcu, 1997] Daniel Marcu. From local to global coherence: a bottom-up approach to text planning. In *Proc. AAAI*, pages 629–635, 1997.

[McKeown, 1985] Kathleen R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge Univ. Press, Cambridge, U.K., 1985.

[Mellish *et al.*, 1998] Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. An architecture for opportunistic text generation. In *Proc. INLG*, pages 28–37, 1998.

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

[Moser and Moore, 1995] Megan G. Moser and Johanna D. Moore. Using discourse analysis and automatic text generation to the study of cue usage. In *Proc. AAAI Symposium on Empirical Methods in Discourse Interpretation and Organization*, pages 92–98, 1995.

[Prasad *et al.*, 2004] Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. Annotation and data mining of the Penn Discourse Treebank. In *Proc. ACL Workshop on Discourse Annotation*, pages 88–95, 2004.

[Quirk *et al.*, 1972] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Grammar of Contemporary English*. Longman, London, 1972.

[Rambow and Korelsky, 1992] Owen Rambow and Tanya Korelsky. Applied text generation. In *Proc. ANLP*, pages 40–47, 1992.

[Reiter and Dale, 2000] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge Univ. Press, 2000.

[Reiter, 1994] Ehud Reiter. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proc. INLG*, pages 163–170, 1994.

[Rösner and Stede, 1992] Dietmar Rösner and Manfred Stede. Customizing RST for the automatic production of technical manuals. In R. Dale, E. Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation. Proc. INLG*, pages 199–214, Heidelberg, 1992. Springer.

[Scott and Souza, 1990] Donia R. Scott and Clarisse Sieckenius de Souza. Getting the message across in RST-based text generation. In R. Dale, C. Mellish, and M. Zock, editors, *Current Research in Natural Language Generation*, pages 47–73. Academic Press, 1990.

[Walker *et al.*, 2001] Marilyn Walker, Owen Rambow, and Monica Rogati. SPoT: A trainable sentence planner. In *Proc. NAACL*, pages 17–24, 2001.

[Walker *et al.*, 2003] Marilyn Walker, Rashmi Prasad, and Amanda Stent. A trainable generator for recommendations in multimodal dialogue. In *Proc. EUROSPEECH*, pages 1697–1701, 2003.

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

[Webber *et al.*, 2005] Bonnie Webber, Aravind Joshi, Eleni Miltsakaki, Rashmi Prasad, Nikhil Dinesh, Alan Lee, and Kate Forbes. A short introduction to the Penn Discourse TreeBank. In *Copenhagen Working Papers in Language and Speech Processing*. 2005.

[Williams and Reiter, 2003] Sandra Williams and Ehud Reiter. A corpus analysis of discourse relations for natural language generation. In *Proc. Corpus Linguistics*, pages 899–908, 2003.