Annotation and Data Mining of the Penn Discourse TreeBank

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

Eleni Miltsakaki
University of Pennsylvania
Philadelphia, PA 19104 USA
elenimi@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

The Penn Discourse TreeBank (PDTB) is a new resource built on top of the Penn *Wall Street Journal* corpus, in which discourse connectives are annotated along with their arguments. Its use of stand-off annotation allows integration with a stand-off version of the Penn TreeBank (syntactic structure) and PropBank (verbs and their arguments), which adds value for both linguistic discovery and discourse modeling. Here we describe the PDTB and some experiments in linguistic discovery based on the PDTB alone, as well as on the linked PTB and PDTB corpora.

1 Introduction

Large scale annotated corpora such as the Penn TreeBank (Marcus et al., 1993) have played a central role in speech and natural language research. However, with the demand for more powerful NLP applications comes a need for greater richness in annotation – hence, the development of PropBank (Kingsbury and Palmer, 2002), which adds basic semantics to the PTB in the form of verb predicate-argument annotation and eventually similar annotation of nominalizations. We have been developing yet another annotation layer above these both. The Penn Discourse TreeBank (PDTB) adds low-level discourse structure and semantics through the annotation of discourse connectives and their arguments, using connective-specific semantic role labels. With this added knowledge, the PDTB (together with the PTB and PropBank) should support more in-depth NLP research and more powerful applications.

Work on the PDTB is grounded in a lexicalized approach to discourse – DLTAG (Webber and Joshi, 1998; Webber et al., 1999a; Webber et al., 2000; Webber et al., 2003). Here, low-level discourse structure and semantics are taken to result (in part) from composing elementary predicate-argument relations whose predicates come mainly from discourse connectives[1] and whose arguments come from units of discourse – clausal, sentential or multi-sentential units. The PDTB therefore differs from the RST-annotated corpus (Carlson et al., 2003) which starts with (abstract) rhetorical relations (Mann and Thompson, 1988) and annotates a subset of the Penn WSJ corpus with those relations that can be taken to hold between (primarily) pairs of discourse spans identified in the corpus.

The current paper focuses on what can be discovered through analyzing PDTB annotation, both on its own and together with the Penn TreeBank. Section 2 of the paper briefly reviews the theoretical background of the project, its current state, the guidelines given to annotators, the annotation tool they used (*WordFreak*), and the extent of inter-annotator agreement. Section 3 shows how we have used PDTB annotation, along with the PTB, to extract several features pertaining to discourse connectives and their arguments, and discusses the relevance of these features for NLP research and applications. Section 4 concludes with the summary.

2 Project overview

2.1 Theoretical background

The PDTB project 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 ar-

[1] Despite this, we have deliberately adopted a policy of having the annotations *independent* of the DLTAG structural descriptions for two reasons: (1) to make the annotated corpus useful to researchers working in different frameworks and (2) to simplify the annotators' task, thereby increasing inter-annotator reliability.

guments that can be identified structurally from adjacent units of discourse. What Webber et al. (2003) call *anaphoric connectives* (discourse adverbials, such as *otherwise*, *instead*, *furthermore*, etc.) also require two arguments – one derived structurally, and the other derived anaphorically from the preceding discourse. The crucial contribution of this framework to the design of PDTB 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 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.

2.2 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, and we expect to release the PDTB by November 2005. Intermediate versions of the annotated corpus will be made available for feedback from the community.

The PDTB corpus will include annotations of four types of connectives: subordinating and coordinating conjunctions, adverbial connectives and implicit connectives. The final number of annotations will amount to approximately 30K: 20K annotations of the 250 types explicit connectives identified in the corpus and 10K annotations of implicit connectives. The final version of the corpus will also characterize the semantic role of each argument.

To date, we have annotated 10 explicit connectives (*therefore*, *as a result*, *instead*, *otherwise*, *nevertheless*, *because*, *although*, *even though*, *when*, *so that*), amounting to a total of 2717 annotations, as well as 386 tokens of implicit connectives. Annotations have been performed by two to four annotators.

2.3 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 basic points from the guidelines.

What counts as a discourse connective? We count as discourse connectives (1) all subordinating and coordinating conjunctions, (2) all discourse adverbials, and (3) all inter-sentential implicit connectives. Discourse adverbials include only those adverbials which convey relationships between two *abstract objects* such as events, states, propositions, etc. (Asher, 1993). For instance, in Example 1, *as a result* conveys a cause-effect relation between the event of limiting the size of industries and that of industries operating out of small, expensive, and inefficient units. In contrast, the semantic interpretation of the clausal adverbial *strangely* in Example 2 only requires a single event/state which it classifies in the set of *strange* events/states.[2]

(1) [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].

(2) Strangely, conventional wisdom inside the Beltway regards these transfer payments as "uncontrollable" or "nondiscretionary."

Implicit connectives are taken to occur between adjacent sentences not related by any explicit connective. They are annotated with whatever explicit connective the annotator feels could be inserted, with the original meaning retained. Assessment of inter-annotator agreement groups these annotations into five coarse classes (Miltsakaki et al., 2004). Currently, we are not annotating implicit connectives intra-sententially (such as between a main clause and a free adjunct) or across paragraphs.

What counts as a legal argument? The simplest argument to a connective is what we take to be the minimum unit of discourse. Because we take discourse relations to hold between *abstract objects*, we require that an argument contain 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*.

What we describe to annotators as *arguments* to discourse connectives are actually the textual span from which the argument is *derived* (Webber et al., 1999a; Webber et al., 2003). This is especially clear in the case of the first argument of *instead* in (3), which does not actually include the negation, although it is part of the selected text.[3]

[2] For a more detailed discussion of how discourse adverbials can be distinguished from clausal adverbials, see Forbes (2003).

[3] For a corpus-based study of the arguments of *instead*, see (Miltsakaki et al., 2003).

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

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 two optional tags, **SUP1** and **SUP2** (*supplementary*), for cases when the annotator wants to mark textual spans s/he considers to be useful, *supplementary* information for the interpretation of an argument. Examples (4) and (5) demonstrate its use. The arguments are shown in square brackets, while spans providing supplementary information are shown in parentheses.[4]

(4) *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).

(5) *Because* [mutual fund trades don't take effect until the market close] (– in this case, at 4 p.m. –) [these shareholders effectively stayed put].

2.4 Inter-Annotation Reliability

An extensive discussion of inter-annotator reliability in the PDTB is presented in (Miltsakaki et al., 2004). The three things that are relevant to the discussion here of using the PDTB for linguistic discovery are (1) the agreement criterion, (2) the level of inter-annotator agreement, and (3) the types of inter-annotator variation.

With respect to agreement, we did not use the kappa statistic (Siegel and Castellan, 1988) because it requires the data tokens to be classified into discrete categories and PDTB annotation involves selecting a span of text whose length is not prescribed *a priori*.[5] Instead of kappa, we assessed inter-annotator agreement using an *exact match* criterion: 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.

Treating ARG1 and ARG2 annotations as independent tokens for assessment, the total number of inter-annotator judgments assessed for explicit connectives was twice the number of connective tokens, i.e, 5434. In this measure, we achieved a high-level of agreement on the arguments to subordinate conjunctions (92.4%), while lower agreement on adverbials (71.8%).[6] 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.

A detailed analysis of inter-annotator variation shows that most of the disagreements (79%) involved *Partial Overlap* – that is, text that is common to what is selected separately by each annotator. *Partial overlap* subsumes categories such as (a) *higher verb*, where one of the annotators included some extra clausal material that contained a higher governing predicate, (b) *dependent clause*, where one of the annotators included extra clausal material which was syntactically dependent on the clause selected by both, and (c) *parenthetical*, where one of the annotators included text that occurred in the middle of the other annotator's selection. Example 6 illustrates a case of *higher verb* disagreement.

(6) a. [he knew the RDF was neither rapid nor deployable nor a force] – *even though* [it cost $8 billion or $10 billion a year].

b. he knew [the RDF was neither rapid nor deployable nor a force] – *even though* [it cost $8 billion or $10 billion a year].

The *partial overlap* disagreements are important with respect to the experiments described in the next section, because most of this variation turns out to be irrelevant to the experiments. We will elaborate on this further in the next section.

3 Data Mining

PDTB annotation indicates two things: the *arguments* of each explicit discourse connective and the *lexical tokens* that actually play a role as discourse connectives. It should be clear that the former

[4] **SUP** annotations have not been used in the current experiments.

[5] Carlson et al. (2003) avoid this by using two sets of categories: one set in which there is a separate category for each span that could constitute an elementary discourse unit, and one set in which there is only a separate category for each span that at least one annotator has selected. Because the arguments of connectives tend to be longer and hence more variable than the elementary spans used in the RST-corpus, we do not see any gain from introducing the first set of categories, and the second set is equivalent to our *exact match* criterion.

[6] In Miltsakaki et al. (2004), we have reported on the annotation of implicit connectives as well. We achieved 72% agreement on the use of explicit expressions in place of the implicit connectives. More details on the implicit connective annotation can be found in this work.

cannot be derived automatically from existing resources, since determining the size and location of the arguments is not simply a matter of sentential syntax or verb predicate argument relations. But the latter is also a non-trivial feature because every lexical item that functions as a discourse connective also has a range of other functions. While some of these functions correlate with POS-tags other than those used in annotating connectives, the PTB POS-tags themselves cannot always be reliably distinguished, given inconsistencies in how the lexical items are analyzed.

We believe that the PDTB annotation can contribute to a range of linguistic discovery and language modeling tasks, such as

- providing empirical evidence for the DLTAG claim that discourse adverbials get one argument anaphorically, while structural connectives such as conjunctions establish relations between adjacent units of text (Creswell et al., 2002).
- acquiring common usage patterns of connectives and identifying their dependencies, in order to support "natural" choices in Natural Language Generation (di Eugenio et al., 1997; Moser and Moore, 1995; Williams and Reiter, 2003).
- developing decision procedures for resolving and interpreting discourse adverbials (Miltsakaki et al., 2003) which can be built on top of discourse parsing systems (Forbes et al., 2003).
- developing "word sense disambiguation" procedures for distinguishing among different senses of a connective and hence interpreting connectives correctly (e.g., distinguishing between temporal and explanatory *since*, between hypothetical and counterfactual *if*, between epistemic and semantic *because*, etc.)
- providing empirical evidence for theories of anaphoric phenomena such as *verb phrase ellipsis* that see them as sensitive to the type of discourse relation in which they are expressed (Hardt and Romero, 2002; Kehler, 2002).

The value of carrying out such studies using a single corpus with multiple layers of annotation is that relationships between phenomena are clearer. (The downside is focusing on a single genre – newspaper text – and a particular "house style" – that of the *Wall Street Journal.* However, developing the PDTB may help facilitate the production of more such corpora, through an initial pass of automatic annotation, followed by manual correction, much as was done in developing the PTB (Marcus et al., 1993).)

Here we present some preliminary experiments we have carried out on the current version of the PDTB. We automatically extracted features associated with discourse connectives and their arguments, both from the PDTB annotation alone as well as from the integrated annotation of the PDTB and PTB. The findings reveal novel patterns regarding the location and size of the arguments of discourse connectives and suggest additional experiments.

The multi-layered annotations for PDTB, PTB (and soon to be available PropBank) are rendered in XML within a "stand-off" annotation architecture in which multiple (independently conducted) annotations refer to the same primary document. *WordFreak* directly renders the PDTB annotations in the stand-off XML representation, but for the syntactic layer, the PTB phrase structure constituent annotations had to first be converted to the XML stand-off representation.[7]

For preparing the connective tokens for data mining, we started with the 2717 annotations for the 10 explicit connectives reported in Section 2.2 and extracted those tokens on which we achieved full "exact match" agreement as well as "partial overlap" agreement on both the arguments (cf. Section 2.4). We felt justified in combining both sets because "partial overlap" disagreements, which occurred mostly within sentences, did not make any overall difference to the features that were extracted. The total number of tokens we obtained from this was 2688. 51 tokens on this set had to be thrown out since the official release of the Penn TreeBank did not have the corresponding syntactic annotations for these tokens.[8] From the remaining 2637 tokens, we extracted two sets of features, one for adverbials (229 tokens) and the other for subordinating conjunctions (2408 tokens).

For the adverbials, we wanted to determine whether the results reported in earlier work (Creswell et al., 2002) held up. Among other things, this work examined whether (1) anaphoric arguments could be reliably annotated, to facilitate the development of robust anaphora resolution algorithms, and (2) there were differences be-

[7] Thanks to Jeremy Lacivita for implementing the representation of PTB in stand-off XML form. The stand-off representation of PTB will be released together with the PDTB corpus.

[8] Researchers who are currently conducting or are planning to conduct multi-layered annotations or experiments with the Penn TreeBank should be aware that the official release contains more source and PoS-tagged files than the parsed files. Future annotations of the PDTB will only be performed on texts that are parsed.

tween the type, size and location of the arguments of anaphoric (adverbial) connectives and those of structural connectives.

The high inter-annotator agreement reported in this earlier study has now been confirmed by the PDTB annotation (cf. Section 2.4). As for the other, we automatically extracted some of the same features that were hand-annotated in Creswell et al. (2002) to determine the distribution of these connectives with respect to their position (**POS**) and the size and location (**LOC**) of their anaphoric arguments. These features are further described below:

POS: pertains to the *position of the connective in its host argument*, i.e., the argument in which it occurs.[9] **POS** can take three defined values: INIT for argument-initial position (Examples 7-9), MED for argument-medial position (Examples 10-11), and FINAL for argument-final position (Examples 12 and 13). Note that the host argument of the connective is a sentence in Example 8 and 9, a VP conjunct in Example 7, a free adjunct in Example 10, the main clause of a sentence in Example 11, a subordinate clause in Example 12, and finally, the first of the two coordinated sentences in Example 13.

LOC: pertains to the *size and location of the anaphoric argument* of the connective. **LOC** can take four defined values: **SS** for when the anaphoric argument occurs in the same sentence as the connective (Examples 7, 10 and 11), **PS** for when the argument occurs in the immediately previous sentence (Examples 12 and 13), **PP** for when the argument occurs in the immediately preceding sequence of sentences (Example 8), and **NC** for when the argument occurs in some non-contiguous sentence(s) (Example 9). A *sentence* is defined as minimally a main clause and all of its attached subordinate clauses, if any. Coordinated main clauses, by this definition, are treated as separate sentences. Note that according to the definition of the **LOC** feature, the anaphoric argument may constitute the entire sentence(s), as in Examples 8, 9 and 13, or it may be part of the sentence(s), as in Examples 7 and 10-12.

An important aspect of the **LOC** feature is that it involved the multi-layering of PDTB and PTB, since the PDTB itself contains no information about syntactic constituency or even sentence boundaries. For deriving the **LOC** feature values, we needed information not only about the sentence boundaries of texts, but also about coordinated clause boundaries, which requires accessing sentence-internal constituents.

(7) **INIT-SS:** Rep. John LaFalce (D., N.Y.) said Mr. Johnson refused [to testify jointly with Mr. Mulford] and *instead* [asked to appear after the Treasury official had completed his testimony].

(8) **INIT-PP:** [But Mr. Treybig questions whether that will be enough to stop Tandem's first mainframe from taking on some of the functions that large organizations previously sought from Big Blue's machines. ”The answer isn't price reductions, but new systems,” he said]. *Nevertheless*, [Tandem faces a variety of challenges, the biggest being that customers generally view the company's computers as complementary to IBM's mainframes].

(9) **INIT-NC:** [For years, costume jewelry makers fought a losing battle]. Jewelry displays in department stores were often cluttered and uninspired. And the merchandise was, well, fake. *As a result*, [marketers of faux gems steadily lost space in department stores to more fashionable rivals – cosmetics makers].

(10) **MED-SS:** Investors usually don't want [to take physical delivery of a contract], [preferring *instead* to profit from its price swings and then end any obligation to take delivery or make delivery as it nears expiration].

(11) **MED-SS:** Although [bond prices weren't as volatile on Tuesday trading as stock prices], [traders *nevertheless* said action also was much slower yesterday in the Treasury market].

(12) **FIN-PS:** Buyers can look forward to double-digit annual returns if [they are right]. But they will have disappointing returns or even losses if [interest rates rise] *instead*.

(13) **FIN-PS:** [Tons of delectably rotting potatoes, barley and wheat will fill damp barns across the land as thousands of farmers turn the state's buyers away]. [Many a piglet won't be born] *as a result*, and many a ham will never hang in a butcher shop.

The distribution of the **POS** feature values across the different connectives, given in Table 1, shows that the connectives in this set occurred predominantly in the initial position of their host argument. The question of whether or not these different positions correlate with any aspect of the information structure of the arguments (Forbes et al., 2003; Kruijff-Korbayová and Webber, 2001) is, however, an open one and will need to be explored further with the PDTB annotations.

INIT	MED	FIN	TOTAL
201 (87.8%)	13 (5.7%)	15 (6.5%)	229

Table 1: Distribution of the Position (**POS**) of Discourse Adverbials

[9] We achieved 94.1% agreement on the host argument (ARG2) annotations.

CONN	SS		PS		PP		NC		Total
nevertheless	3	(9.7%)	17	(54.8%)	3	(9.7%)	8	(25.8%)	31
otherwise	2	(11.1%)	14	(77.8%)	1	(5.6%)	1	(5.6%)	18
as a result	3	(4.8%)	44	(69.8%)	5	(7.9%)	12	(19%)	63
therefore	11	(55%)	7	(35%)	1	(5%)	1	(5%)	20
instead	22	(22.7%)	62	(63.9%)	2	(2.1%)	11	(11.3%)	97
TOTAL	**41**	**(17.9%)**	**144**	**(62.9%)**	**12**	**(5.2%)**	**33**	**(14.4%)**	**229**

Table 2: Distribution for Location (**LOC**) of Anaphoric Argument of Adverbial Connectives

The distribution of the **LOC** values across the different connectives is shown in Table 2. We first look at all the connectives taken together (i.e., the final **TOTAL** row) and focus on differences in **LOC** and what such differences suggest.

The first thing that is evident from the **TOTAL** row in Table 2 is the significant proportion of ARG1 tokens that occur in a position non-adjacent to the discourse adverbial (**NC** = 14.4%). This accords with the results in (Creswell et al., 2002), in terms of providing evidence that discourse adverbials (unlike structural connectives) are not getting both their arguments from structurally defined positions.

The second point that is evident from the **TOTAL** row is the significant proportion of ARG1 tokens in **SS** location. This includes instances of ARG1 in complement clauses (Example 7), subordinate clauses (Example 11), relative clauses (both restrictive and non-restrictive, as in Example 14), preceding VP conjuncts (Example 15), and from main clauses, where the adverbial is attached to a free adjunct, as in Example 16.

(14) [$_i$ The British pound], [pressured by last week's resignations of key Thatcher administration officials], *nevertheless* [$_i$ rose Monday to $1.5820 from Friday's $1.5795].[10]

(15) Appealing to a young audience, [he scraps an old reference to Ozzie and Harriet] and *instead* [quotes the Grateful Dead].

(16) [The transmogrified brokers never let the C-word cross their lips], *instead* [stressing such terms as "safe," "insured" and "guaranteed"].

While one might want to argue that the latter is no different from adjacent full clauses and hence should be treated the same as a location in the previous sentence (i.e., **LOC**=**PS**), the other **SS** cases provide additional evidence for an anaphoric analysis of these discourse adverbials since there already exists a separate structural relation in each case. Furthermore, in Example 7, the arguments of the conjunction *and*, though not yet addressed by our annotators, differ from the arguments of *instead.* Any attempt to treat *instead* as a structural connective will produce a syntactic analysis with crossing branches – a source of both theoretical and practical (parsing) problems (Forbes et al., 2003).

Turning now to the individual analysis of adverbials, Table 2 shows that the 4 connectives other than *therefore* pattern rather similarly with respect to the location of the anaphoric argument (**SS**, **PS**, **PP**, **NC**). All of them except *therefore* have their antecedent predominantly in the previous sentence (between 54.8% and 77.8%). The question is whether the difference in how *therefore* patterns – i.e., drawing its antecedent 55% of the time from the same sentence – is simply a consequence of having such few data points (i.e., only 20) or a matter of "house style" (with all the examples from the *Wall Street Journal*) or a difference that is theoretically motivated. If the answer lies in house style or theory, then it is relevant to work in natural language generation. Further annotation and analysis of adverbials and their arguments in the PDTB will provide more information as to this puzzle.

At the start of this section, we indicated five different areas in which PDTB annotation could contribute to linguistic discovery and language modeling. This data mining experiment illustrates the first three, as well as providing information relevant to further development of discourse parsing systems and natural language generation systems. For future work, we intend to explore further the extraction and study of other features related to discourse adverbials. Two features that we are currently working to extract automatically pertain to (a) the cooccurrence of discourse adverbials with other connectives in the host argument, and (b) the syntactic type and depth of the anaphoric arguments, such as whether the argument was a finite or non-finite complement clause, a relative clause, or a finite or non-finite subordinate clause etc.

For the subordinating conjunctions (Table 3), we extracted features pertaining to the relative position of the two arguments of the conjunction. Subordinating conjunctions often take their arguments in the same sentence with the subordinate clause as one argument and the main clause as its other ar-

[10] The subscripts on the bracketed spans in this example indicate discontinuous parts of the host argument of *nevertheless.*

gument. However, the subordinate clause can either occur to the right of the main clause, i.e., postposed, as in Example 17, or it can occur preposed, i.e., before the main clause, as in Example 18.

(17) **ARG1-ARG2:** But Sen. McCain says [Mr. Keating broke off their friendship abruptly in 1987], *because* [the senator refused to press the thrift executive's case as vigorously as Mr. Keating wanted].

(18) **ARG2-ARG1:** *Because* [Swiss and EC insurers are widely present on each other's markets], [the accord isn't expected to substantially increase near-term competition].

The distribution of the relative position of the arguments of these connectives, given in Table 3, shows significant differences across the connectives.

CONN	ARG1-ARG2	ARG2-ARG1	Total
when	545 (54%)	465 (46%)	1010
because	822 (90%)	93 (10%)	915
even though	77 (75%)	26 (25%)	103
although	129 (37%)	218 (63%)	347
so that	33 (100%)	0 (0%)	33
Total	1606 (67%)	802 (33%)	2408

Table 3: Distribution for Argument order for Subordinating Conjunctions

There are a few interesting things to note here. First, even if one considers only the four subordinating conjunctions with $>$100 tokens, no two of them pattern in the same way.

Second, with *when*, the almost equal distribution of preposed and postposed tokens suggests either *free variation* of the two patterns or *different uses* of the two patterns, with each use favoring a different pattern. The latter would accord with a theoretical distinction that has been made between postposed *when* expressing a purely temporal relation between the two clauses, and preposed *when* expressing a contingent relation between them (Moens and Steedman, 1988). Integrated evidence from the PTB and PropBank may help distinguish the two possibilities.

Third, there is a striking contrast between the patterning of *although* and *even though*, especially if one assumes that *even though* (like *even when*, *even after*, *even if*, etc.) involves application of the topicalizer *even* to the subordinate clause, just as it can apply to other constituents. Further annotation and analysis of the PDTB will reveal whether all subordinating conjunctions that co-occur with *even* pattern like *even though*, or whether this is specific to the concessive.

Finally, when Williams and Reiter (2003) examined 342 texts from the RST annotation of the Penn TreeBank corpus (Carlson et al., 2003), they reported that 77% of the instances of *concessive relations* that they examined appeared in the order ARG2-ARG1. (The eleven instances of *although* that they examined and the three instances of *even though* appeared in *concessive relations*, along with instances of *but*, *despite*, *however*, etc.) If we were to collapse together all instances of *although* and *even though* annotated in the PDTB (totalling 450), we would find that 46% (206) patterned as ARG1-ARG2, and 54% of them (244) patterned as ARG2-ARG1. This might lead us to draw a similar conclusion to Williams and Reiter (2003). But it would also disguise the fact noted above that *although* and *even though* pattern oppositely to one another. This suggests (1) that making the feature extraction procedure specific to particular connectives, as in the PDTB, will reveal distributional patterns that are lost when more abstract relations are the focus of the annotation, and (2) that a larger set of annotated tokens can show more reliable distributional patterns.

In sum, data mining of PDTB with respect to subordinating conjunctions has shown radically different distribution patterns regarding the relative position of the arguments. Some of these have confirmed and strengthened previous theoretical claims and some have suggested new and promising research directions. Further work in this area will also be extremely relevant for NLG sentence planning components employing discourse relations (Walker et al. (2003), Stent et al. (2004), among others), where the sentence planner needs to make decisions regarding cue placement. Finally, while our approach is "syntactic", with the distribution of the connectives and their arguments being explored in terms of whether they are subordinating conjunctions, coordinating conjunctions, or adverbial connectives, one can also explore the patterning of connectives in terms of semantic categories, once their semantic role annotation is complete (cf. Section 2.2). The latter could be especially interesting to cross-linguistic studies of discourse, as well as to applications such as multilingual generation and MT are envisaged.[11]

4 Summary

In this paper we have presented the Penn Discourse TreeBank (PDTB), a large-scale discourse-level annotated corpus that is being developed towards the creation of a multi-layered annotated corpus, integrating the Penn TreeBank, PropBank and

[11] We thank an anonymous reviewer for pointing this out.

the PDTB. The PDTB encodes low-level discourse structure information, marking discourse connectives as indicators of discourse relations, and their arguments. We have reported high inter-annotator agreement for the PDTB annotation. Our data mining experience and preliminary results show that the multi-layered corpora is a rich source of information that can be exploited towards the development of powerful and efficient natural language understanding and generation systems as well as towards large-scale corpus-based research.

Acknowledgments

We are very grateful to Tom Morton and Jeremy Lacivita for the development and modification of the *WordFreak* annotation tool. Special thanks to Jeremy for providing continuous technical support. Thanks are also due to our annotators, Cassandre Creswell, Driya Amandita, John Laury, Emily Pawley, Alan Lee, Alex Derenzy and Steve Pettington.

References

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

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

Cassandre Creswell, Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2002. The Discourse Anaphoric Properties of Connectives. In *Proceedings of DAARC2002*. Edições Colibri.

Barbara di Eugenio, Johanna D. Moore, and Massimo Paolucci. 1997. Learning Features that Predict Cue Usage. In *Proceedings of ACL/EACL 97*.

Kate Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. 2003. D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information*, 12(3):261–279.

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

Dan Hardt and Maribel Romero. 2002. Ellipsis and the Structure of Discourse. In *Proceedings of Sinn und Bedeutung VI*.

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

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of LREC-02*.

Ivana Kruijff-Korbayová and Bonnie Webber. 2001. Information Structure and the Semantics of ‘otherwise’. In *Proceedings of ESSLLI 2001: Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

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*.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating Discourse Connectives and their Arguments. In *Proceedings of the NAACL/HLT Workshop: Frontiers in Corpus Annotation*.

Marc Moens and Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14(2):15–28.

Megan G. Moser and Johanna Moore. 1995. Investigating Cue Selection and Placement in Tutorial Discourse. In *Proceedings of ACL95*.

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

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of ACL-2004*.

Marilyn Walker, Rashmi Prasad, and Amanda Stent. 2003. A Trainable Generator for Recommendations in Multimodal Dialogue. In *Eurospeech, 2003*.

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*.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999a. Discourse Relations: A Structural and Presuppositional Account Using Lexicalized TAG. In *Proceedings of ACL-99*.

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)*.

Bonnie Webber, Alistair Knott, and Aravind Joshi. 2000. Multiple Discourse Connectives in a Lexicalized Grammar for Discourse. In *Proceedings of IWCS-00*.

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

Sandra Williams and Ehud Reiter. 2003. A Corpus Analysis of Discourse Relations for Natural Language Generation. In *Proceedings of Corpus Linguistics*.