The Penn Discourse Treebank

Eleni Miltsakaki*, Rashmi Prasad*, Aravind Joshi*, Bonnie Webber †

* University of Pennsylvania
3401 Walnut St., Philadelphia 19104, U.S.A.
{elenimi, rjprasad, joshi,}@linc.cis.upenn.edu

† University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, Scotland
bonnie@inf.ed.ac.uk

Abstract

This paper describes a new discourse-level annotation project – the Penn Discourse Treebank (PDTB) – that aims to produce a large-scale corpus in which discourse connectives are annotated, along with their arguments, thus exposing a clearly defined level of discourse structure. The PDTB is being built directly on top of the Penn Treebank and Propbank, thus supporting the extraction of useful syntactic and semantic features and providing a richer substrate for the development and evaluation of practical algorithms. We present a preliminary analysis of inter-annotator agreement – both the level of agreement and the types of inter-annotator variation.

1. Introduction

Large scale annotated corpora, such as the Penn Treebank (PTB) (Marcus et al., 1993), have played a critical role in natural language processing, complementing the equally critical role played by linguistic theory. With the demand for more powerful NLP applications comes a need for greater richness in annotation. At the sentence-level, Penn Propbank is adding predicate-argument annotation to sentences in PTB (Kingsbury and Palmer, 2002). At the discourse-level are efforts to produce corpora annotated with RST rhetorical relations (Carlson et al., 2003), coreference (Mitkov et al., 2000), (Muller et al., 2002) and temporal markers and relations (Gaizauskas et al., 2003).

This paper describes a new discourse-level annotation project – the Penn Discourse Treebank (PDTB) – that aims to produce a large-scale corpus in which discourse connectives are annotated, along with their arguments, thus exposing a clearly defined level of discourse structure. In this respect, the PDTB differs from the RST annotated corpus, in which the *basis* for an an RST relational assignment is lacking. The PDTB is being built on top of the sentence-level syntactic annotation of the PTB and its more recent semantic annotation (Penn Propbank). We believe that having interlinked annotations will support the extraction of useful syntactic and semantic features, thus providing a richer substrate for the development and evaluation of practical algorithms.

PTB and Propbank provide a sort of shallow semantic representation (predicate-argument structure, frames, and role sets), which can permit a level of inference in various NLP tasks, such as IE, QA, summarization, and MT tasks. PDTB can be seen as providing a next level of inferences due to discourse connectives, their arguments with their semantic roles. This level is deeper than that provided by PTB and Propbank, yet it is shallow in the sense that it peels off just those inferences licensed by the connectives.

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. The PDTB is expected to be released by November 2005. Intermediate versions of the annotated corpus will be made available before then. The PDTB corpus will include annotations of four types of connectives: subordinating conjunctions, coordinating conjunctions, adverbial connectives and implicit connectives. We describe each in more detail below. The final number of annotations in the corpus will amount to approximately 30,000: 10,000 implicit connectives, and 20,000 annotations of the 250 explicit connectives identified in the corpus. The final version of the corpus will also contain characterizations of the semantic roles associated with the arguments of each type of connective.

In this paper we present the results of annotating 10 explicit connectives, amounting to a total of 2,717 annotations, as well as the results of annotating 386 instances of implicit connectives in the Penn TreeBank. The list of 10 connectives includes the adverbial connectives 'therefore', 'as a result', 'instead', 'otherwise', 'nevertheless', and the subordinate conjunctions 'because', 'although', 'even though', 'when', and 'so that'. Currently, annotation is performed by four annotators. Individual annotation proceeds one connective at a time. *WordFreak*, the annotation tool being used by the annotators, identifies all instances of a given connective in the corpus, which are then annotated independently by four annotators.[1] This way, the annotators quickly gain experience with that connective and develop a better understanding of its predicate-argument characteristics. Similarly, for the annotation of implicit connectives, all instances (as specified in the guidelines below) are identified one text at a time. For this task, the annotators are required to read the entire text so that they can make well-

[1] *WordFreak* has been developed by Tom Morton at the University of Pennsylvania and has been substantially modified for our project by Jeremy Lacivita.

informed and reliable decisions about the implicit connectives and their arguments. In addition, after the arguments of each implicit connective have been identified, the annotators provide, if possible, an explicit connective that best expresses the inferred relation.

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

2.1. Discourse connectives

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

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

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

Coordinating conjunctions contain connectives such as 'and', 'but', and 'or'.

Adverbial connectives are sentence-modifying adverbs which express a discourse relation between two events or states, e.g., 'however', 'therefore', 'then', etc. In this class, we have also included prepositional phrases which express similar binary relations, such as 'as a result', 'in addition', 'in fact', etc. Example (2) shows the annotation of an instance of the adverbial connective 'as a result'.

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

Implicit connectives are identified between adjacent sentences that are not related by an explicit connective.[2] The annotation of implicit connectives is intended to capture discourse relations that are implicitly expressed between adjacent sentences. Annotators are asked to provide an explicit connective that best describes the inferred relation. For example, the explicit connective provided in (3) was 'in contrast'.

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

[2]There may, of course, be other implicitly expressed relations that we are not taking into account.

2.2. Annotation guidelines

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 conjunctions, (2) all coordinating conjunctions, (3) certain adverbials, and (4) implicit connectives between adjacent sentences. The adverbials include only those which convey a relation between two *abstract objects* such as events or states (Asher, 1993). For example, in (4) 'as a result' conveys a cause-effect relation between a limiting event and an operating event. In contrast, the semantic interpretation of 'strangely' in (5) only requires a single event/state which it classifies in the set of *strange* events/states.

(4) [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.

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

Implicit connectives are identified between adjacent sentences which are not related via any explicit connectives. Currently, we are not annotating implicit connectives intra-sententially (such as between a main clause and a free adjunct). We plan to do this at a later stage of the project.

What counts as a legal argument? Because we take discourse relations to hold between abstract objects, we require that an argument contains at least one predicate along with its arguments. Therefore, a legal argument can be a single clause, a single sentence, a sequence of clauses and/or sentences, or combinations of both. There are two exceptions to the requirement that an argument include predicative units – these are nominal phrases that express an event or a state, and discourse deictics that denote an event or state.

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 an optional tag **SUP1**, **SUP2** (*supplementary*) for cases when the annotator wants to mark textual spans s/he considers to be useful but *supplementary* information for the interpretation of an argument. Example (6) demonstrates its use. The spans providing supplementary information are shown in parentheses.

(6) *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).

3. Data analysis

To test the reliability of the annotation, we assessed inter-annotator agreement in terms of agree-

ment/disagreement on span identity for each token as a percentage of the pairs of spans that actually matched versus those that should have.[3] To use a most conservative measure, we used the *exact match* criterion. We present here agreement results on (a) 2717 tokens of 10 explicit connectives (mentioned in Section 2.) by 2 annotators, and (b) 386 tokens of implicit connectives, also by 2 annotators.[4]

For the 2717 explicit connectives, we computed percentage agreement for ARG1 and ARG2 annotations, treating them as independent tokens.[5] The total number of tokens is therefore twice the number of connectives, i.e, 5434. Using binary values for the *exact match* criterion, agreement for any ARG1 or ARG2 token was recorded as 1 when both annotators made identical textual selections, and 0 when the annotators made non-identical selections.

We achieved 90.2% agreement (4900/5434 tokens) on the ARG1 and ARG2 annotations of explicit connectives. Further distribution of the agreements by connective is given in Table 1. The second column gives the number of agreeing tokens for each connective and the third column gives the total number of (ARG1+ARG2) tokens available for that connective. The last column gives the percent agreement for the connective in that row, i.e., as a percentage of tokens for which agreement was 1 (column 2) versus the total number of tokens for that connective (column 3). The table shows that we achieved high agreement on argument annotations of subordinating conjunctions (92.4%). Average agreement on the adverbials was lower (71.8%). This difference between the two types is not surprising, since adverbial connectives are anaphoric (Webber et al., 2003) and locating the (sometimes non-adjacent) ARG1 argument of these connectives is believed to be a harder task.

We classified the 534 disagreements into 4 major types, given in Table 2. The third column gives the disagreement for each category as a percentage of the total disagreements. The majority of disagreements (79%) were due to "Partial Overlap", which subsumes the categories *higher verb*, *dependent clause*, *parenthetical*, *sentence*, and *other*. "Partial Overlap" means that there was some common span of text between the selections of the two annotators. *Higher verb* includes tokens where one of the annotators included the governing predicate for the clause marked by both annotators. An example of this is given in 7) and (8), where the higher clause "he knew" has been included in ARG1 by one annotator and not the other. *Dependent Clause* includes tokens where one of the annotators included extra clausal material that is syntactically dependent on the clause that was selected by both. *Sentence* means that one of the annotators included one or more additional sentences as part

[3]We did not use the kappa statistic (Siegel and Castellan, 1988) for computing inter-annotator agreement because the statistic requires the data tokens to be classified into discrete categories. The PDTB annotation constitutes selection of a span of text which can be of indeterminate length.

[4]SUP1 and SUP2 annotations were not considered in this test. Additional annotations by another 2 annotators are currently underway. The 2 annotators of the explicit connectives are different from the 2 annotators of the implicit connectives.

[5]In (Miltsakaki et al., 2004), we report on additional diagnostics using different classes of tokens.

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

Table 1: Agreement Distribution across Explicit Connectives, with ARG1 and ARG2 Annotations Counted Independently

of the annotation. For *parenthetical*, one of the annotators included a medial parenthetical, while the other did not – cf. Examples (9) and (10). *Other* included tokens with partial overlap between annotations, but in addition included a combination of more than one type, such as *higher verb+dependent clause*.

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

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

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

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

DISAGREEMENT TYPE	No.	%
Missing Annotations	72	13.5%
No Overlap	30	5.6%
Partial Overlap		
parenthetical	53	9.9%
higher verb	171	32.0%
dependent clause	182	34.1%
sentence	10	1.9%
other	6	1.1%
Unresolved	10	1.9%
TOTAL	534	100%

Table 2: Disagreement Classification for Explicit Connective ARG Annotations

Note that disagreements that contain a partial overlap could be counted as agreeing tokens if we relaxed the more conservative *exact match* measure to a *partial match* measure. Our subjective view was that in several cases, the 'extra' textual material, especially those fitting the *dependent*

clause and *parenthetical* category did not make any significant semantic contribution in terms of their inclusion or exclusion in the argument. With the *partial match* measure, excluding these cases reduces the disagreements to half the given number, giving us 94.5% agreement overall.

The *No Overlap* tokens were cases of true disagreement in that there was no overlap in the annotations selected by the annotators. These tokens constituted 5.6% of the disagreements. *Missing Annotations* (13.5%) was used for tokens where the annotation was missing for one annotator due to technical tool errors. *Unresolved* includes tokens which have introduced new issues for the annotation guidelines and cannot be resolved at this time.

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

Table 3: Disagreement Classification for Implicit Connective ARG Annotations

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

As a preliminary step in analyzing agreement on the type of explicit connective provided by the annotators in place of an implicit connective, we considered 5 groups of connectives conveying : a) additional information (e.g., 'furthermore', 'in addition') b) cause-effect relations (e.g., 'because', 'as a result'), c) temporal relations (e.g., 'then', 'simultaneously'), d) contrastive relations (e.g., 'however', 'although'), and e) restatement or summarization (e.g., 'in other words', 'in sum').[6] Agreement was then computed on these basic groups of connectives.[7] From the total of 386 tokens of implicit connectives, 9 were excluded from the analysis due to technical error (missing annotation). For the remaining 307 tokens, we achieved 72% agreement on the type of explicit connective that best conveyed the interpretation of the implicit connective.

For the argument annotations of the implicit connectives, we used the same diagnostic as for the explicit connectives above. On the 772 ARG1 and ARG2 tokens for the implicit connectives, we achieved 85.1% (657/772) agreement between 2 annotators. The analysis of the 115 disagreements is given in Table 3. Note that here again, the number of disagreements reduces to half using the *partial match* measure for the *parenthetical* and *dependent clause* classes, giving us 92.6% agreement overall.

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

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

4. Summary

In this paper we presented a new large scale annotation project, the Penn Discourse Treebank, which includes annotations of discourse connectives and their arguments. We reported preliminary results from inter-annotator agreement on completed annotations of 10 explicit connectives (2,717 annotations) and 386 tokens of implicit connectives. The high inter-annotator agreement that we achieved indicates that discourse connectives and their arguments expose a well-defined level of discourse structure that can be reliably annotated.

Acknowledgments

We would like to thank the reviewers for their useful comments. This work was partially supported by NSF Grant EIA 02-24417.

5. References

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

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

Gaizauskas, Rob, Patrick Hanks, James Pustejovsky, Roser Sauri, Andrew See, Andrea Setzer, Lisa Ferro, and Beth Sundheim., 2003. The timebank corpus. In *Corpus Linguistics 2003.*. Lancaster, U.K.

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

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

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

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, and Bonnie Webber, 2004. Annotating discourse connectives and their arguments. In *Proceedings of the Workshop on Frontiers in Corpus Annotation*. Boston, Massachussetts.

Mitkov, Ruslan, R. Evans, C. Orasan, C. Barbu, L. Jones, and V. Sotirova, 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000), Lancaster, U.K.*.

Muller, Christoph, Stefan Rapp, and Michael Strube, 2002. Applying co-training to reference resolution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Philadelpha PA*.

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

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