# The Penn Discourse TreeBank 1.0 Annotation Manual

**The PDTB Research Group**

March 29, 2006

**Contributors:**

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi
Institute for Research in Cognitive Science,
University of Pennnsylvania
*rjprasad,elenimi,nikhild,aleewk,joshi@linc.cis.upenn.edu*

and

Bonnie Webber
Division of Informatics,
University of Edinburgh
*bonnie@inf.ed.ac.uk*

# Contents

# 1 Introduction

## 1.1 Background and overview

An important aspect of discourse understanding and generation involves the recognition and processing of *discourse relations*. Following the views towards discourse structure in Webber and Joshi (1998); Webber *et al.* (2003), where discourse connectives as treated as discourse-level predicates that take two *abstract objects* such as events, states, and propositions (Asher, 1993) as their arguments, the Penn Discourse TreeBank (PDTB) aims to annotate the *argument structure*, *semantics* and *attribution* of discourse connectives and their arguments.[1] This document presents a report of the annotation guidelines for the first release of the Penn Discourse TreeBank corpus, PDTB-1.0, distributed free of charge under terms of the GNU General Public License (GPL). The corpus and associated software is available for download from the PDTB webpage, `http://www.seas.upenn.edu/~pdtb`.

Discourse connectives in the PDTB are distinguished primarily into `Explicit` discourse connectives, that include a set of lexical items drawn from well-defined syntactic classes, and `Implicit` discourse connectives, which are inserted between paragraph-internal adjacent sentence-pairs not related explicitly by any of the syntactically-defined set of `Explicit` connectives. In the latter case, the reader must attempt to infer a discourse relation between the adjacent sentences, and "annotation" consists of *inserting* a connective expression that *best* conveys the inferred relation. Connectives *inserted* in this way to express inferred relations are called `Implicit` connectives. Multiple discourse relations (Webber *et al.*, 1999) can also be inferred, and are annotated by inserting multiple `Implicit` connectives.

Adjacent sentence-pairs between which `Implicit` connectives cannot be inserted are further distinguished and annotated as three types: (a) `AltLex`, for when a discourse relation is inferred, but insertion of an `Implicit` connective leads to a *redundancy* in the expression of the relation due to the relation being *alternatively lexicalized* by some "non-connective" expression; (b) `EntRel`, for when no discourse relation can be inferred and where the second sentence only serves to provide some further description of an entity in the first sentence (a relation akin to *entity-based coherence* (Knott *et al.*, 2001)); and (c) `NoRel`, for when no discourse relation or entity-based coherence relation can be inferred between the adjacent sentences. Annotations of `Implicit` connectives, `ALtLex`, `EntRel`, and `NoRel` are collectively referred to as "implicit relation annotations" in this report.

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

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.[2] Supplements to `Arg1` and `Arg2`, called `Sup1` for material supplementary to `Arg1`, and `Sup2`, for material supplementary to `Arg2`, are annotated to mark material that is relevant but not "mininally necessary" for the interpretation of the relation.

For this release, semantic classification of connectives is done only for implicit relation annotations, and apply only to `Implicit` connectives and `AltLex` relations since discourse relations are inferred for just these two types. Semantic classes are classified broadly into seven types, partly motivated by the feature-based classification in Knott (1996). A more fine-grained classification will be done for the second release, including semantic classification for `Explicit` connectives.[3] Semantic classes are annotated as features on relations.

Attribution, which is a relation of "ownership" between abstract objects and individuals, is annotated for both (explicit and implicit) connectives, and their arguments. Attribution is annotated in terms of three features, the primary feature type encoding the *source* of attribution, and the other two encoding *factuality* and *polarity*.

The annotation guidelines described in this document draw and expand on earlier reports presented in annotation tutorials and papers, notably Miltsakaki *et al.* (2004a,b); Prasad *et al.* (2004); Dinesh *et al.* (2005); Prasad *et al.* (2005); Webber *et al.* (2005). The reader is assumed to be familiar with the details of the annotation framework and background as presented in these papers. The rest of this section discusses the source corpus and annotation style of the PDTB, and presents an overview of the annotations contained in the first release of the corpus. Section 2 presents the annotation guidelines for the argument structure of `Explicit` connectives. Annotation guidelines for implicit relations, their argument structure, and semantic classification are presented in Section 3. Finally, Section 4 discusses the guidelines for attribution annotation.

[2]The assumption of the arity constraint on a connective's 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.

[3]Some preliminary experiments on semantic classification of `Explicit` connectives have already been conducted (Miltsakaki *et al.*, 2005).

## 1.2 Source corpus and annotation styles

The PDTB annotations are done on the Wall Street Journal (WSJ) articles in the Penn TreeBank (PTB) II corpus (Marcus *et al.*, 1993). Annotation of connectives and their arguments consists of recording the text spans that anchor them in the WSJ RAW files, but the final annotation representation follows the "stand-off" annotation technique, such that the text spans are represented in terms of their character offsets in the WSJ RAW files.

Connectives and their arguments are also linked to the "parsed" PTB files in a similar stand-off style, with the reference to the PTB structural description of a PDTB connective or argument annotation being represented as a set of tree node *Gorn* address.[4] Other aspects of the annotation, such as semantic classes and attribution, are simply represented as features. A complete description of the representation format of the PDTB annotations is provided in a separate document downloadable from the corpus distribution webpage.[5]

Because of the stand-off annotation style of the PDTB, the corpus can be effectively used only in conjunction with the primary source data, the WSJ RAW and PTB parsed files, which must be obtained independently from the Linguistic Data Consortium (LDC).[6]

## 1.3 Summary of annotations

PDTB, Release 1.0 contains 100 distinct types of `Explicit` connectives (such as *because*, *when*, *as a result*, *and*, etc.), with a total of 18505 tokens. `Explicit` connectives have been annotated across the entire corpus (25 sections). Appendix A gives the distribution of `Explicit` connective types and tokens. Modified connectives such as *only because*, *just when*, etc. are treated as belonging to the same type as that of the head. The complete list of modified connectives per type is given in Appendix B. Implicit relations have been annotated in three PTB sections (Sections 08, 09, and 10) for this release, totalling 2003 tokens (1496 tokens of `Implicit` connectives, 19 tokens of `AltLex`, 435 tokens of `EntRel`, and 53 tokens of `NoRel`). The complete distribution of the types of implicit relations, along with semantic classes, is provided in Appendix C.

Since the PDTB also provides links to the PTB parsed texts, only 2304 texts from the PTB distribution were chosen for PDTB annotations, since other files could not be converted to stand-off. Of these 2304 texts, only 1808 texts are contained in this first distribution of the

[4] The links to the PTB parsed texts were generated programmatically. We have used these links in our experiments, but all have not been examined by a human.

[5] `http://www.seas.upenn.edu/~pdtb/pdtb-corpus-1.0/pdtb-fileformats.pdf`

[6] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T7`

PDTB: the remaining 496 texts either did not have any occurrences of Explicit connectives annotated for this release (see Appendix A), or have not yet been annotated for Implicit connectives (see above).[7]

## 1.4 Notation conventions

In what follows, the 4-digit number in parentheses after an example gives the WSJ RAW file number containing the example. In all examples, annotated Explicit connectives are underlined, and annotated Implicit connectives are shown in small caps, along with their associated semantic classes (in parentheses). For clarity, Implicit connectives are further indicated by the marker, "Implicit =". For the arguments of connectives, the text whose interpretation is the basis for Arg1 appears in italics, while that of Arg2 appears in bold. For example, in (1), the subordinating conjunction *because* is an Explicit discourse connective that establishes a causal relation between "the campaign board refusing to pay Mr. Dinkins" (Arg1) and "Mr. Dinkin's campaign records being incomplete" (Arg2). In Example (2), the Implicit connective *so* is inserted to express the inferred CONSEQUENCE relation between the second and third sentences, i.e., between "Motorola no longer delivering junk mail" (Arg1) and "the mail going into the trash" (Arg2).

(1) *The city's Campaign Finance Board has refused to pay Mr. Dinkins $95,142 in matching funds* because **his campaign records are incomplete**. (0041)

(2) Motorola is fighting back against junk mail. *So much of the stuff poured into its Austin, Texas, offices that its mail rooms there simply stopped delivering it.* Implicit= SO (CONSEQUENCE) **Now, thousands of mailers, catalogs and sales pitches go straight into the trash**. (0989)

AltLex, EntRel, and NoRel annotations are also indicated by underlined markers, namely as "AltLex =", "EntRel", and "NoRel". The semantic class, recorded only for AltLex among these three types, is shown in small caps and parentheses. Also, for AltLex, the elsewhere lexicalizing expression of the relation is shown in square brackets.

Supplementary annotations are shown in parentheses, with Sup1 and Sup2 marked as subscripts, as seen for Sup1 in Example (3).

(3) ($_{Sup1}$ Workers described "clouds of blue dust") *that hung over parts of the factory,* even though **exhaust fans ventilated the area**. (0003)

[7]Subsequent releases of the PDTB will have implicit relations annotated across the entire corpus, and possibly additional Explicit connectives that were missed for this release.

# 2 Explicit Connectives and their Arguments

## 2.1 Identifying `Explicit` connectives

`Explicit` connectives in the PDTB are drawn from the following grammatical classes:

- *Subordinating conjunctions* (e.g., *because*, *when*, *since*, *although*):

  (4) Since **McDonald's menu prices rose this year**, *the actual decline may have been more.* (1280)

  (5) *The federal government suspended sales of U.S. savings bonds* because **Congress hasn't lifted the ceiling on government debt**. (0008)

- *Coordinating conjunctions* (e.g., *and*, *or*, *nor*):[8]

  (6) *The House has voted to raise the ceiling to $3.1 trillion*, but **the Senate isn't expected to act until next week at the earliest**. (0008)

  (7) *The subject will be written into the plots of prime-time shows*, and **viewers will be given a 900 number to call**. (2100)

- *(ADVP and PP) adverbials* (e.g., *however*, *otherwise*, *then*, *as a result*, *for example*).[9]

  (8) *Working Woman, with circulation near one million, and Working Mother, with 625,000 circulation, are legitimate magazine success stories.* **The magazine Success**, however, **was for years lackluster and unfocused**. (1903)

  (9) *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**. (0629)

[8] Coordinating conjunctions appearing in only clausal conjunctions have been annotated. Coordinating conjunctions appearing in VP coordinations are not annotated, such as the conjunction *and* in (1):

(1) More common chrysotile fibers are curly *and* are more easily rejected by the body, Dr. Mossman explained. (0003)

[9] The adverbial *in fact* was annotated as a discourse connective, although we now think (on theoretical grounds) that it is probably not one (Forbes-Riley *et al.*, 2006). We will be examining the annotation of *in fact* to see whether there is empirical evidence to back up this theoretically motivated decision.

However, not all expressions satisfying these roles are annotated as discourse connectives, since they do not denote relations between two abstract objects (AOs). For example, discourse markers within the class of adverbials, such as *well*, *anyway*, *now*, etc. (Hirschberg and Litman, 1987), are not annotated since they signal the organizational or focus structure of the discourse, rather than relating AOs. And clausal adverbials such as *strangely*, *probably*, *frankly*, *in all likelihood* etc. are also not annotated as discourse connectives since they take a single AO as argument, rather than two (Forbes-Riley *et al.*, 2006).

Of the remaining types of `Explicit` connectives that were identified (see Appendix A), there are of course several instances in the corpus where the expression is homonymous with a discourse connective, but in fact serves another function, such as to relate non-AO entities (e.g., the use of *and* to conjoin noun phrases in Example (10), and the use of *for example* to modify a noun phrase in Example (11)), to relativize extracted adjuncts (e.g., the use of *when* to relativize the time NP in Example (12)), and so on. Such expressions are not annotated as discourse connectives.

(10) Dr. Talcott led a team of researchers from the National Cancer Institute *and* the medical schools of Harvard University *and* Boston University. (0003)

(11) These mainly involved such areas as materials – advanced soldering machines, *for example* – and medical developments derived from experimentation in space, such as artificial blood vessels. (0405)

(12) Equitable of Iowa Cos., Des Moines, had been seeking a buyer for the 36-store Younkers chain since June, *when* it announced its intention to free up capital to expand its insurance business. (0156)

## 2.2 Modified connectives

Several connectives can be further modified by adverbs such as *only*, *even*, *at least*, and so on. Some examples of modified connectives are provided in Examples (13-15). (Modifiers in these examples are shown in parentheses for clarity.) Such modification is extremely productive, and rather than listing each bare and modified occurrence of a connective as a separate type, modified forms are treated as the same type as that of the head - the bare form.[10] Appendix B lists all the annotated modified forms for each connective type.

[10]In the annotation, the head of a connective is given as a feature. For bare forms, the connective itself is given as the head. See the description of PDTB file formats (Footnote 5.

(13) *That power can sometimes be abused*, (particularly) since **jurists in smaller jurisdictions operate without many of the restraints that serve as corrective measures in urban areas**. (0267)

(14) *You can do all this* (even) if **you're not a reporter or a researcher or a scholar or a member of Congress**. (0108)

(15) *We're seeing it* (partly) because **older vintages are growing more scarce**. (0071)

While we have annotated modified connectives such as described above, certain types of post-modified connectives have not been annotated for this release, in particular those post-modified by prepositions, for example *because (of)...*, *as a result (of)...*, *instead (of)...*, and *rather (than)....* While in many cases such expressions relate noun phrases lacking an AO interpretation, (Example 16), there are also several cases (Example 17) where they appear as discourse connectives. We plan to handle such cases in future releases of the corpus.

(16) The products already available are cross-connect systems, used *instead of* mazes of wiring to interconnect other telecommunications equipment. (1064)

(17) *instead of* featuring a major East Coast team against a West Coast team, it pitted the Los Angeles Dodgers against the losing Oakland A's. (0443)

## 2.3 Parallel connectives

In addition to modified forms of connectives, we have also annotated a small set of "parallel" connectives, that is, pairs of connectives where one part presupposes the presence of the other, and where both together take the same two arguments (Examples (18-20). Such connectives are listed as distinct types and are annotated discontinuously. In Appendix A, parallel connectives are shown with two dots between the two parts of the pair (e.g., *on the one hand..on the other hand*, *if..then*, *either..or*).

(18) On the one hand, Mr. Front says, *it would be misguided to sell into "a classic panic."* On the other hand, **it's not necessarily a good time to jump in and buy**. (2415)

(19) If *the answers to these questions are affirmative*, then **institutional investors are likely to be favorably disposed toward a specific poison pill**. (0275)

(20) Either *sign new long-term commitments to buy future episodes* or **risk losing "Cosby" to a competitor**. (0060)

## 2.4 Conjoined connectives

Conjoined connectives like *when and if* and *if and when* are treated as complex connectives and listed as distinct types. Examples are shown in (21-22).

(21) When and if **the trust runs out of cash** – which seems increasingly likely – *it will need to convert its Manville stock to cash.* (1328)

(22) Hoylake dropped its initial $13.35 billion ($20.71 billion) takeover bid after it received the extension, but said *it would launch a new bid* if and when **the proposed sale of Farmers to Axa receives regulatory approval**. (2403)

## 2.5 Linear order of connectives and arguments

Connectives and their arguments can appear in any relative order. 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. 23), `Arg2-Arg1` (Ex. 24), or `Arg2` may appear between discontinuous parts of `Arg1` (Ex. 25), depending on the relative position of the subordinate clause with respect to its matrix clause.

(23) *The federal government suspended sales of U.S. savings bonds* because **Congress hasn't lifted the ceiling on government debt**. (0008)

(24) Because **it operates on a fiscal year**, *Bear Stearns's yearly filings are available much earlier than those of other firms.* (1948)

(25) *Most oil companies*, when **they set exploration and production budgets for this year**, *forecast revenue of $15 for each barrel of crude produced.* (0725)

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 (26) shows, arguments of discourse adverbials can appear between their other discontinuously annotated argument. In this example, the text span associated with `Arg1` appears between the discontinuous spans associated with `Arg2`.

(26) 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**. (0155)

The position of connectives in the `Arg2` clause they modify is restricted to initial position for subordinating and coordinating conjunctions, but adverbials are free to appear anywhere in their `Arg2` clause, as shown below:

(27) *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**. (0072)

(28) *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** ... (0761)

(29) *Polyvinyl chloride capacity "has overtaken demand* and **we are experiencing reduced profit margins** as a result", ... (2083)

## 2.6 Location of Arguments

There is no restriction on how far an argument can be from its corresponding connective. So arguments can be found in the same sentence as the connective (Examples 30-32), in the sentence immediately preceding that of the connective (Examples 33-35), or in some non-adjacent sentence (Example 36).

(30) *The federal government suspended sales of U.S. savings bonds* because **Congress hasn't lifted the ceiling on government debt**. (0008)

(31) *Most balloonists seldom go higher than 2,000 feet* and **most average a leisurely 5-10 miles an hour**. (0239)

(32) *In an invention that drives Verdi purists bananas, Violetta lies dying in bed during the prelude*, rising deliriously **when** then **she remembers the great parties she used to throw**. (1154)

(33) *Why do local real-estate markets overreact to regional economic cycles*? Because **real-estate purchases and leases are such major long-term commitments that most companies and individuals make these decisions only when confident of future economic stability and growth**. (2444)

(34) *Metropolitan Houston's population has held steady over the past six years.* And **personal income, after slumping in the mid-1980s, has returned to its 1982 level in real dollar terms**. (2444)

(35) *Such problems will require considerable skill to resolve.* However, **neither Mr. Baum nor Mr. Harper has much international experience**. (0109)

(36) Mr. Robinson of Delta & Pine, the seed producer in Scott, Miss., 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.* (Sup1 That's because pollination, while easy in corn because the carrier is wind, is more complex and involves insects as carriers in crops such as cotton). "It's one thing to say you can sterilize, and another to then successfully pollinate the plant," he said. Nevertheless, he said, **he is negotiating with Plant Genetic to acquire the technology to try breeding hybrid cotton**. (0209)

## 2.7 Types and extent of arguments

### 2.7.1 Simple clauses

With a few exceptions to be discussed below (Section 2.7.2), the simplest syntactic realization of an abstract object as a connective's argument is taken to be a clause, tensed or non-tensed. Further, the clause can be a matrix clause, a complement clause, or a subordinate clause. Some examples of single clausal realizations are shown in Examples (37-42). For clause types such as non-finite clauses and relative clauses, the argument selection assumes the presence of implicit subjects and traces of extracted complements available in the syntactic structure of the clause in the PTB, so that the complete interpretation of the argument is assumed to be derivable from the selection.

(37) A Chemical spokeswoman said *the second-quarter charge was "not material"* **and that no personnel changes were made** as a result. (0304)

(38) In Washington, House aides said Mr. Phelan told congressmen that the collar, *which banned program trades through the Big Board's computer* when **the Dow Jones Industrial Average moved 50 points**, didn't work well. (0088)

(39) *Knowing a tasty – and free – meal* when **they eat one**, the executives gave the chefs a standing ovation. (0010)

(40) Alan Smith, president of Marks & Spencer North America and Far East, says that Brooks Brothers' focus is to boost sales *by broadening its merchandise assortment* while **keeping its "traditional emphasis."** (0530)

(41) Radio Shack says it has a policy *against selling products* if **a salesperson suspects they will be used illegally**. (1058)

(42) "We have been a great market *for inventing risks* **which other people** then **take, copy and cut rates**." (1302)

### 2.7.2 Non-clausal arguments

In some exceptional cases, non-clausal elements are treated as realizations of abstract objects.

**2.7.2.1 VP coordinations** Verb phrases in VP coordinations are taken to denote an abstract object and can be annotated as arguments. However, the subject of the VP coordinates is included in the argument selection only for the first VP coordinate (`Arg1` of *then* in Example 43). Subjects for non-initial coordinates are not included in the selection (`Arg2` of *then* in Example 43 and `Arg1` of *because* in Example 44), and will have to be retrieved via independent heuristics to arrive at the complete interpretation of the argument.[11]

(43) *It acquired Thomas Edison's microphone patent* **and** then **immediately sued the Bell Co.** (`Sup2` claiming that the microphone invented by my grandfather, Emile Berliner, which had been sold to Bell for a princely $50,000, infringed upon Western Union's Edison patent). (0091)

(44) She became an abortionist accidentally, *and continued* because **it enabled her to buy jam, cocoa and other war-rationed goodies**. (0039)

**2.7.2.2 Nominalizations** Nominalizations are annotated as arguments of connectives in two strictly restricted contexts. The first context is when they allow for an "existential" interpretation, as in Example (45), where the `Arg1` selection can be interpreted existentially as "that there will be major new liberalizations":

(45) Economic analysts call his trail-blazing liberalization of the Indian economy incomplete, and many are hoping *for major new liberalizations* if **he is returned firmly to power**. (2041)

The second context is when they involve a clearly observable case of a "derived nominalization", as in Example (46), where the `Arg1` selection can be assumed to be transformationally derived from "such laws to be resurrected":

(46) But in 1976, the court permitted *resurrection of such laws*, if **they meet certain procedural requirements**. (0426)

[11] Note that coordinating conjunctions in VP coordinations are not annotated for this release (Fn. 8).

**2.7.2.3 Anaphoric expressions denoting abstract objects** Anaphoric expressions, when they themselves denote an abstract object, such as demonstratives like *this* and *that*, and VP anaphora like *so*, can be annotated as arguments of connectives. Such annotation assumes that an anaphora resolution mechanism will yield the interpretation of the argument.[12]

(47) "It's important to share the risk *and even more so* <u>when</u> **the market has already peaked**." (0782)

(48) Investors who bought stock with borrowed money – that is, "on margin" – may be more worried than most following Friday's market drop. *That's* <u>because</u> **their brokers can require them to sell some shares or put up more cash to enhance the collateral backing their loans**. (2393)

(49) Evaluations suggest that good ones are – *especially so* <u>if</u> **the effects on participants are counted**. (2412)

**2.7.2.4 Responses to Questions** In some contexts such as question-answer sequences, where the response to a question only includes response particles like "yes" and "no", the response particles are themselves annotated as arguments, with the preceding question annotated as "Sup" to indicate the question-answer relation.

(50) Underclass youth are a special concern. (`Sup1` Are such expenditures worthwhile, then)? *Yes*, <u>if</u> **targeted**. (2412)

(51) (`Sup1` Is he a victim of Gramm-Rudman cuts)? *No*, <u>but</u> **he's endangered all the same**: His new sitcom on ABC needs a following to stay on the air. (0528)

### 2.7.3 Multiple clauses/sentences, and the Minimality Principle

In addition to single clauses, abstract object arguments of connectives can also be realized as multiple clauses and multiple sentences. Example (52) shows multiple sentences selected for the `Arg1` argument of *still*. Multiple clause and multiple sentence arguments can also be annotated discontinuously if they so appear in the text.

[12]We have attempted to mark the (AO) antecedent of the anaphor as "Sup", but this has not been done consistently for this release.

(52) *Here in this new center for Japanese assembly plants just across the border from San Diego, turnover is dizzying, infrastructure shoddy, bureaucracy intense. Even after-hours drag; "karaoke" bars, where Japanese revelers sing over recorded music, are prohibited by Mexico's powerful musicians union.* Still, **20 Japanese companies, including giants such as Sanyo Industries Corp., Matsushita Electronics Components Corp. and Sony Corp. have set up shop in the state of Northern Baja California**. (0300)

There are no restrictions on how many or what types of clauses can be included in these complex selections, except for the *Minimality Principle*, according to which only as many clauses and/or sentences should be included in an argument selection as are *minimally required* and *sufficient* for 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` and `Sup2`, for `Arg1` and `Arg2` respectively.

## 2.8 Conventions

This section describes certain conventions that we have followed in the annotation. For such cases, we do not make any claims about whether and how they contribute to the interpretation of the relations. They were mostly adopted for convenience of annotation.

### 2.8.1 Inclusion of clausal complements and non-clausal adjuncts

For all clauses that are selected as arguments of connectives, all complements of the main clausal predicate and all non-clausal adjuncts (e.g., "a speciality chemicals concern" in `Arg2` of Example 53), adverbs (e.g., "for example" in `Arg1` of Example 54), complementizers (e.g., "that" in `Arg1` and `Arg2` of Example 55), conjunctions (e.g., "But" in `Arg1` of Example 56), and relative pronouns (e.g., "whom" in `Arg1` of Example 57) modifying the clause are obligatorily included in the argument (except for the connective that is itself being annotated), even if these elements are not necessary for the "minimal" interpretation of the relation (see Section 2.7.3).

(53) Although **Georgia Gulf hasn't been eager to negotiate with Mr. Simmons and NL, a specialty chemicals concern**, *the group apparently believes the company's management is interested in some kind of transaction.* (0080)

(54) players must abide by strict rules of conduct even in their personal lives – *players for the Tokyo Giants, for example, must always wear ties* when **on the road**. (0037)

(55) There seems to be a presumption in some sectors of (Mexico's) government *that there is a lot of Japanese money waiting behind the gate*, and **that by slightly opening the gate, that money will enter Mexico**. (0300)

(56) *But the Reagan administration thought otherwise*, and **so may the Bush administration**. (0601)

(57) That impressed Robert B. Pamplin, Georgia-Pacific's chief executive at the time, *whom Mr. Hahn had met* while **fundraising for the institute**. (0100)

Inclusion of non-clausal elements is obligatory even when it warrants discontinuous annotation (Examples 58-61).

(58) They found students in an advanced class a year earlier who said she gave them similar help, *although* because **the case wasn't tried in court**, *this evidence was never presented publicly.* (0044)

(59) He says *that* when **Dan Dorfman, a financial columnist with USA Today, hasn't returned his phone calls**, *he leaves messages with Mr. Dorfman's office saying that he has an important story on Donald Trump, Meshulam Riklis or Marvin Davis.* (1376)

(60) Under two new features, participants will be able to transfer money from the new funds to other investment funds *or*, if **their jobs are terminated**, *receive cash from the funds.* (0204)

(61) *Last week*, when **her appeal was argued before the Missouri Court of Appeals**, *her lawyer also relied on the preamble.* (1423)

Non-clausal attributing phrases are also included obligatorily in the clausal argument they modify, such as "according to..." phrases in the `Arg1` of both the following examples:

(62) *No foreign companies bid on the Hiroshima project, according to the bureau.* But **the Japanese practice of deep discounting often is cited by Americans as a classic barrier to entry in Japan's market**. (0501)

(63) *Even so, according to Mr. Salmore, the ad was "devastating"* because **it raised questions about Mr. Courter's credibility**. (0041)

Note that verbs of attribution along with their subject are in general excluded from an argument when the attribution does not itself play a role in the interpretation of the relation.

But while the constraint against excluding non-verbal attribution phrases stands as an exception to the general guideline for attribution annotation, there is no loss of information in examples like those above, since attribution is currently recorded as features (Section 4), and is, thus, still reflected in the annotations. However, we may reconsider the guidelines when we annotate "attribution spans" for the second release.

### 2.8.2 Punctuation

For practical reasons in the annotation process, all punctuation at the boundaries of connective and argument selections was excluded. However, in the annotation links to the PTB parsed files, some heuristics are used to extend the annotation spans to include certain boundary punctuations. So while the text annotation does not include punctuations occurring at the edges of arguments, they can be obtained from the linked annotations in some cases. Punctuation heuristics and extensions are described in detail in the PDTB file format documentation (Footnote 5).

# 3 Implicit Connectives and their Arguments

## 3.1 Introduction

The goal of annotating `Implicit` connectives in the PDTB is to capture relations between abstract objects that are not realized explicitly in the text (by one of a set of the syntactically-defined `Explicit` connectives - see Section 2.1) and are left to be inferred by the reader. For example, in (64), a CAUSAL relation is inferred between "raising cash positions to record levels" and "high cash positions helping to buffer a fund", even though no `Explicit` connective appears in the text to express this relation. Similarly, in (65), a CONSEQUENCE relation is inferred between "the increase in the number of rooms" and "the increase in the number of jobs", though there is no `Explicit` connective visible to express this relation.

(64) Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* `Implicit` = BECAUSE (CAUSAL) **High cash positions help buffer a fund when the market falls.** (0983)

(65) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* `Implicit` = SO (CONSEQUENCE) **By a**

**rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**. (0994)

In the PDTB, such inferred relations between adjacent sentences in the text are marked as `Implicit` connectives, by the insertion of a connective expression that best expresses the inferred relation. So in Examples (64) and (65), the `Implicit` connectives *because* and *so* are inserted to capture the perceived CAUSAL and CONSEQUENCE relations respectively.

Multiple discourse relations between adjacent sentences are also allowed to be inferred, and are annotated as multiple `Implicit` connectives. In Example (66), two `Implicit` connectives, *when* and *for example*, are inserted to express how `Arg2` presents one instance of the circumstances under which "Mr. Morishita comes across as an outspoken man of the world". Similarly, in Example (67), the two `Implicit` connectives *since* and *for example* are provided to express how `Arg2` presents one instance of the reasons for the claim that "the third principal did have garden experience".

(66) *The small, wiry Mr. Morishita comes across as an outspoken man of the world.* `Implicit` = WHEN FOR EXAMPLE (TEMPORAL, ADD.INFO) **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.**[13] (0800)

(67) *The third principal in the S. Gardens adventure did have garden experience.* `Implicit` = SINCE FOR EXAMPLE (CAUSAL, ADD.INFO) **The firm of Bruce Kelly/David Varnell Landscape Architects had created Central Park's Strawberry Fields and Shakespeare Garden**. (0984)

The decision to "lexically encode" inferred relations in this way was made with the aim of achieving high reliability among annotators while avoiding the difficult task of training them to reason about pre-defined abstract relations. The annotation of inferred relations was thus done intuitively, and involved reading adjacent sentences (and in some cases, the preceding text as well - see Section 3.5.2), making a decision about whether or not a relation could be inferred between them, and providing an appropriate `Implicit` connective to express the inferred relation, if any. Three distinct pre-defined labels, `AltLex`, `EntRel` and `NoRel` (see Section 3.6), were used for cases where an `Implicit` connective could not be provided: `AltLex` for cases where the insertion of an `Implicit` connective to express an inferred relation

[13] "ADD.INFO" is short for "ADDITIONAL-INFO"

led to a *redundancy* in the expression of the relation; `EntRel` for cases where only an *entity-based coherence* relation could be perceived between the sentences; and `NoRel` for cases where no discourse relation or entity-based coherence relation could be perceived between the sentences. As mentioned earlier, the four aforementioned annotation types are referred to as "implicit relation annotations" in this report.

`Implicit` connectives are annotated between all successive pairs of sentences within paragraphs (see Section 3.4), where sentence delimiters are taken to be the period ("."), the semi-colon (";"), and the colon (":"). Example (68) shows a CAUSAL relation inferred between sentences separated by a semi-colon, while Example (69) shows an ADD.INFO relation inferred between sentences separated by a colon.[14]

(68) In a typical leverage strategy, a fund tries to capture the spread between what it costs to borrow and the higher return on the bonds it buys with the borrowed money. *If the market surges, holders can make that much more profit*; `Implicit` = BECAUSE (CAUSAL) **the leverage effectively acts as an interest-free margin account for investors.** (0983)

(69) *Many small investors are facing a double whammy this year*: `Implicit` = IN PARTICULAR (ADD.INFO) **They got hurt by investing in the highly risky junk bond market**, and the pain is worse because they did it with borrowed money. (0983)

For this release of the PDTB, implicit relations are annotated for three sections of the PTB - Sections 08, 09, and 10, totalling 2004 tokens. There are 1496 `Implicit` connective tokens, 19 `AltLex` tokens, 435 `EntRel` tokens, and `53` NoRel tokens.[15] The complete distribution of the types of implicit relations in the PDTB, along with associated semantic classes, where applicable, is given in Appendix C.

In the PDTB annotation files, the left character offset of the second sentence is represented as the "placeholder" for an implicit relation. `Implicit` connectives are represented as features on the annotation. Further details on the representation of implicit relation annotations in the corpus can be found in the documentation of PDTB file formats (see Footnote 5).

[14] Sometimes the colon is used as part of the expression of attribution. Such cases are not annotated since they do not indicate potential sentence boundaries:

(1) Commenting on the budget mess this week, President Bush said: "The perception out there is that it's the fault of Congress. And you can look to the leadership and ask them why that is the perception of the American people."

[15] The second release of the PDTB will contain implicit relation annotations across the entire corpus.

## 3.2 Semantic Classification

For the first release of the PDTB, inferred discourse relations, which are annotated either as `Implicit` connectives or as `AltLex` (Section 3.6.1), have been classified broadly into seven semantic types, and are associated with the annotations as features. A more fine-grained semantic classification will be followed for the second release.[16] The currently used semantic classification is given below.

- Additional-info[17] (includes continuation, elaboration, exemplification, similarity)
- Causal
- Temporal
- Contrast (includes opposition, concession, denial of expectation)
- Condition
- Consequence
- Restatement/summarization[18]

## 3.3 On getting at intentions

Of course, in such annotation there is a danger in saying that the inferred relations are the ones the writer in fact "intended". To a large extent, annotators rely on their own world knowledge as well as reasoning on this knowledge to derive these inferences. But it's quite possible that these may be rejected upon consultation with the author. In short, texts can be open to more than one interpretation, both globally and locally. And annotation of implicit relational inferences is thus inherently biased.

However, to some extent, we have been able to overcome the problem of such conflicts. By having multiple annotators (for implicit relations, we have three, and for hard cases, at least two more opinions are solicited), we increase the likelihood of deriving the intended

[16] We will also consider making distinctions for inferred relations in terms of levels of interpretation, thus distinguishing between, for example, *semantic*, *epistemic*, or *speech act* interpretations of the relation (Sweetser, 1990).

[17] This is represented as add.info in the annotations

[18] This is represented as restate/sum in the annotations

inferences. That is, our estimate of overall reliability is based on at least three judgements per token.

## 3.4 Unannotated implicit relations

For this first release, implicit relations have not been annotated in some circumstances.

### 3.4.1 Implicit relations across paragraphs

We have not annotated for implicit relations between adjacent sentences separated by a paragraph boundary, as in many cases there isn't one. But because a relation may exist, we may carry out such annotation in subsequent releases of the corpus. For example, in (292) a CAUSAL relation can be inferred between the last sentence of the first paragraph and the first sentence of the second paragraph, in that the latter provides one reason why the 1% charge is in fact the best bargain available.

(70) The Sept. 25 "Tracking Travel" column advises readers to "Charge With Caution When Traveling Abroad" because credit-card companies charge 1% to convert foreign-currency expenditures into dollars. In fact, this is the best bargain available to someone traveling abroad. In contrast to the 1% conversion fee charged by Visa,

foreign-currency dealers routinely charge 7% or more to convert U.S. dollars into foreign currency. On top of this, the traveler who converts his dollars into foreign currency before the trip starts will lose interest from the day of conversion. At the end of the trip, any unspent foreign exchange will have to be converted back into dollars, with another commission due. (0980)

### 3.4.2 Intra-sentential relations

Implicit relations between adjacent clauses in the same sentence have also been delayed for a later phase. For example, we have not annotated intra-sentential relations between a main clause and any *free adjunct*. As between adjacent sentences, different relationships can hold between these clauses (Webber and Di Eugenio, 1990):

(71) The market for export financing was liberalized in the mid-1980s, forcing the bank to face competition. (0616)

(72) Second, they channel monthly mortgage payments into semiannual payments, reducing the administrative burden on investors. (0029)

(73) Mr. Cathcart says he has had "a lot of fun" at Kidder, adding the crack about his being a "tool-and-die man" never bothered him. (0604)

So in Examples (71) and (72), the event expressed in the free adjunct can be considered a CONSEQUENCE of that expressed in the main clause (which might be annotated with an `Implicit` *so* or *thereby*), while in Example (73) the event expressed in the free adjunct merely follows (as CONTINUATION) that expressed in the main clause (which might be annotated with an `Implicit` *then*).

### 3.4.3 Implicit relations in addition to explicitly expressed relations

We have only annotated implicit relations between adjacent sentences with no `Explicit` connective between them, even though the presence of a an `Explicit` connective, in particular a discourse adverbial, in a sentence does not preclude the presence of either another `Explicit` connective relating with the previous text (Example 74) or an `Implicit` connective (Example 75). In both examples shown, the sentences are related via a CAUSAL as well as a CONDITIONAL relation, with the difference being that the CAUSAL relation is expressed with an `Explicit` *because* in Example (74), while the same relation is inferred in Example (75).

(74) If the light is red, stop **because** **otherwise** you'll get a ticket.

(75) If the light is red, stop. **Otherwise** you'll get a ticket.

However, for the current release, we have only annotated multiple `Explicit` relations, as in Example (74), and multiple `Implicit` relations, as in Examples (66-67), but not multiple relations when one of them is explicitly expressed in the text, as in Example (75).

### 3.4.4 Implicit relations between non-adjacent sentences

Finally, this release of the PDTB does not annotate implicit relations between non-adjacent sentences, even if such a relationship clearly holds. For example, even if the discourse adverbial *then* were removed from Example (76), the event expressed by clause (76d) would still be understood as holding after that expressed by clause (76b). Nevertheless, we neither require nor allow the annotators to annotate the one or more `Implicit` connectives that express the connection holding between clauses (76b) and (76d). This is a practical decision rather than one that has any deep significance.

(76) a. John loves Barolo.
b. So he ordered three cases of the '97.
c. But he had to cancel the order
d. because he (then) discovered he was broke.

## 3.5 Extent of Arguments

### 3.5.1 Sub-sentential arguments

While implicit relations are annotated between adjacent sentences, this does not mean that the arguments of an inferred relation need span complete sentences. As with the `Explicit` connectives, annotators were asked to select only as much of the adjacent sentences as was minimally necessary for the interpretation of the inferred relation. Furthermore, as for `Explicit` connectives, parts of the text that were seen as "relevant " (but not "necessary") to the interpretation of the relation could be marked as *supplementary* information. For instance, in Example (77), for the inferred ADD.INFO/EXEMPLIFICATION relation, the matrix clause is excluded from `Arg1`, and is marked as `Sup1` (shown in parentheses) - its "relevance" being due to its containment of the referent of the relative pronoun *when* in `Arg1`.

(77) (Average maturity was as short as 29 days at the start of this year), *when short-term interest rates were moving steadily upward.* `Implicit` = FOR EXAMPLE (ADD.INFO) **The average seven-day compound yield of the funds reached 9.62% in late April**. (0982)

Parts of the sentence may also be left out without being labeled as Sup, when they are not considered relevant to the interpretation of the relation, as seen for `Arg2` in Example (78).

(78) *Meanwhile, the average yield on taxable funds dropped nearly a tenth of a percentage point, the largest drop since midsummer.* `implicit` = IN PARTICULAR (ADD.INFO) **The average seven-day compound yield**, which assumes that dividends are reinvested and that current rates continue for a year, **fell to 8.47%, its lowest since late last year, from 8.55% the week before, according to Donoghue's**. (0982)

Attribution is also a cause for selection of sub-sentential spans, as seen for Arg1 in Example (79), and for both `Arg1` and `Arg2` in Example (80).

(79) *"Lower yields are just reflecting lower short-term interest rates,"* said Brenda Malizia Negus, editor of Money Fund Report. Implicit = SINCE (CAUSAL) **Money funds invest in such things as short-term Treasury securities, commercial paper and certificates of deposit, all of which have been posting lower interest rates since last spring**. (0982)

(80) Ms. Terry did say *the fund's recent performance "illustrates what happens in a leveraged product" when the market doesn't cooperate.* Implicit = STILL (CONTRAST) "**When the market turns around**," she says, "**it will give a nice picture" of how leverage can help performance**. (0983)

### 3.5.2 Multiple sentence arguments

In addition to selecting sub-sentential clauses, arguments (`Arg1` as well as `Arg2`) are also allowed to span over multiple sentences (discontinuously, if necessary) if such an extension is "minimally required" for the interpretation of the relation.[19] For instance, for the inferred ADD.INFO/EXEMPLIFICATION relation in Example (81), the example of "legal controversies always assuming a symbolic significance far beyond the particular case" is given not just by the sentence following it, but rather by a combination of the three following sentences.

(81) Legal controversies in America have a way of assuming a symbolic significance far exceeding what is involved in the particular case. They speak volumes about the state of our society at a given moment.
*It has always been so.* Implicit = FOR EXAMPLE (ADD.INFO) **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. The result was a world-famous trial exposing profound cultural conflicts in American life between the "smart set," whose spokesman was H.L. Mencken, and the religious fundamentalists, whom Mencken derided as benighted primitives. Few now recall the actual outcome: Scopes was convicted and fined $100, and his conviction was reversed on appeal because the fine was excessive under Tennessee law**. (0946)

Similar scenarios obtain for `Explicit` connectives, except that for `Implicit` connective annotations, the extension to multiple sentences is subject to the strict constraint of *adjacency*.

[19] Extension of arguments to multiple sentences is restricted for `EntRel` and `NoRel` (see Section 3.6).

That is, at least some part of the spans selected for `Arg1` and `Arg2` must belong to the initially selected adjacent sentences.

Lists, when they comprise the `Arg1` argument, are also taken to be *minimal*. So `Arg1` is extended to include the complete list (Example 82).

(82) All the while, Ms. Bartlett had been busy at her assignment, serene in her sense of self-tilth. *As she put it in a 1987 lecture at the Harvard Graduate School of Design: "I have designed a garden, not knowing the difference between a rhododendron and a tulip." Moreover, she proclaimed that "landscape architects have been going wrong for the last 20 years" in the design of open space. And she further stunned her listeners by revealing her secret garden design method: Commissioning a friend to spend "five or six thousand dollars . . . on books that I ultimately cut up." After that, the layout had been easy. "I've always relied heavily on the grid and found it never to fail.*" `Implicit` = IN ADDITION (ADD.INFO) **Ms. Bartlett told her audience that she absolutely did not believe in compromise or in giving in to the client "because I don't think you can do watered-down versions of things**." (0984)

Example (83) shows an example where multiple sentences are selected for both `Arg1` and `Arg2`, as "minimally" required for `Arg2`, and as a "list" for `Arg1`.

(83) While the model was still on view, *Manhattan Community Board 1 passed a resolution against South Gardens. The Parks Council wrote the BPCA that this "too 'private' . . . exclusive," complex and expensive "enclosed garden . . . belongs in almost any location but the waterfront.*" `Implicit` = SIMILARLY (ADD.INFO) **Lynden B. Miller, the noted public garden designer who restored Central Park's Conservatory Garden, recalls her reaction to the South Gardens model in light of the public garden she was designing for 42nd Street's Bryant Park: "Bryant Park, as designed in 1933, failed as a public space, because it made people feel trapped. By removing the hedges and some walls, the Bryant Park Restoration is opening it up. It seems to me the BPCA plan has the potential of making South Gardens a horticultural jail for people and plants**." (0984)

Implicit relations are also annotated to record the relation of a parenthetical sentence with its preceding sentence. However, when annotating the relation between a parenthetical and its subsequent sentence, `Arg1` is (at least) extended to the sentence occurring before the

parenthetical. So given a three sentence text containing S1, (S2), and S3, where (S2) is the parenthetical, two relations are marked: one between [S1] as `Arg1` and [(S2)] as `Arg2`, and the other between [S1,(S2)] as `Arg1` and [S3] as `Arg2`.

## 3.6 Non-insertability of Implicit Connectives

In many cases, an `Implicit` connective cannot be inserted between adjacent sentences. These have been classified into 3 types: `AltLex`, `EntRel`, and `NoRel`.[20] We describe each of these types below.

### 3.6.1 `AltLex` (Alternative lexicalization)

These are cases where a discourse relation *is* inferred between the adjacent sentences but where providing an `Implicit` connective leads to *redundancy in the expression of the relation.* This is because the relation is *alternatively lexicalized* by some "non-connective expression". Such non-connective lexicalizations fall into one of two types. The first as an adjunct prepositional phrase, such as *on top of this* (Example 84), *after that* (Example 85), *after the study* (Example 86), etc., where the head of the prepositional phrase expresses the relation and the prepositional complement is an anaphoric expression referring to `Arg1` of the relation.[21] In examples below, the `AltLex` expression is shown in square brackets for clarity.[22]

(84) *In contrast to the 1% conversion fee charged by Visa, foreign-currency dealers routinely charge 7% or more to convert U.S. dollars into foreign currency.* AltLex = (ADD.INFO) **[On top of this], the traveler who converts his dollars into foreign currency before the trip starts will lose interest from the day of conversion.** (0980)

(85) And she further stunned her listeners by revealing her secret garden design method: *Commissioning a friend to spend "five or six thousand dollars . . . on books that*

[20] Note that in previous work (Prasad *et al.*, 2005), we used different labels for two of these categories, NOCONN-ENT for `EntRel`, and NOCONN for `AltLex`.

[21] We note that it is possible to analyze such PPs as discourse connectives (Forbes-Riley *et al.*, 2006). If we want to do this in the future, the current marking of the `AltLex` in the annotation will allow us to distinguish such cases and annotate them later on the basis of their internal structure (see Footnote 22).

[22] `AltLex` annotations in the corpus are anchored on the text span character offsets of the elsewhere lexicalizing expressions, similar to the manner in which `Explicit` connectives are represented. See the documentation of the PDTB file formats (Footnote 5).

*I ultimately cut up."* AltLex = (TEMPORAL) [**After that**], **the layout had been easy.** (0984)

(86) Mr. Breeden, in his first testimony to Congress since taking the SEC post, said *the agency is studying the Friday the 13th market plunge, including how current circuit breakers affected the market that day and the following Monday.* AltLex = (TEMPORAL) [**After the study**], **the SEC would be willing to consider adding new circuit breakers or fine-tuning the current ones**, he added. (0987)

In the second type, the relation is lexicalized in the core clausal structure of Arg2, namely the clausal predicate and its arguments, though unlike the PPs above, these cases may or may not involve the realization of `Arg1` as an anaphoric expression. In Examples (87) and (88), for instance, the CAUSAL and ADD.INFO/EXEMPLIFICATION relations respectively are incorporated into the subject of `Arg2`, with `Arg1` being realized anaphorically as the complement of the NP subject head. In Example (89), however, while the CAUSAL relation is incorporated into the subject, there is no anaphoric expression referring to `Arg1`.

(87) I read the exerpts of Wayne Angell's exchange with a Gosbank representative ("Put the Soviet Economy on Golden Rails," editorial page, Oct. 5) with great interest, since the gold standard is one of my areas of research. Mr. Angell is incorrect when he states that the Soviet Union's large gold reserves would give it "great power to establish credibility." *During the latter part of the 19th century, Russia was on a gold standard and had gold reserves representing more than 100% of its outstanding currency, but no one outside Russia used rubles.*
*The Bank of England, on the other hand, had gold reserves that averaged about 30% of its outstanding currency, and Bank of England notes were accepted throughout the world.* AltLex = (CAUSAL) [**The most likely reason for this disparity**] **is that the Bank of England was a private bank with substantial earning assets, and the common-law rights of creditors to collect claims against the bank were well established in Britain.** (0985)

(88) The Soviet Union should keep these lessons in mind as it seeks *to establish the ruble as an international currency.* AltLex = (ADD.INFO) [**One way to make the ruble into a major international currency**] **would be to leave reserves of gold and earning assets in a Swiss bank with distributions based on Swiss laws**. (0985)

(89) *As of March 1, its Flint office, with about 2,500 employees, stopped delivering bulk mail and non-subscription magazines. Employees were told that if they really wanted*

*the publications, they would have to have them sent home instead.* AltLex = (CAUSAL) **[The reason:] overload, especially of non-subscription magazines**. (0989)

Example (90) shows a case where the CONSEQUENCE relation is lexically incorporated into the verb, and the subject realizes Arg1 as an anaphoric expression.

(90) *Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject.* AltLex = (CONSEQUENCE) **[Mayhap this metaphorical connection made] the BPC Fine Arts Committee think she had a literal green thumb**. (0984)

Annotation of the arguments of AltLex relations follows the same guidelines as provided for Implicit connectives. That is, they are subject to the *adjacency constraint*, they can be discontinuous, and they must include all and only the amount of text "minimally" required for the interpretation of the relation.

### 3.6.2 EntRel (Entity-based coherence)

EntRel captures cases where the implicit relation between adjacent sentences is not between their AO interpretations, but is rather one resulting from a connection akin to *entity-based coherence* (Knott *et al.*, 2001). That is, the connection is due only to some entity being realized in both sentences, where realization can be direct (Examples 91-92) or indirect (Examples 93-94),[23]. Note that entity realization here also includes reification of an abstract object mentioned in the first sentence, such as with the demonstrative *this* in Example (95), and the definite description *the appointments* in Example (96).[24]

(91) *Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern.* EntRel **Mr. Milgrim succeeds David Berman, who resigned last month**. (0945)

(92) *The purchase price was disclosed in a preliminary prospectus issued in connection with MGM Grand's planned offering of six million common shares.* EntRel **The luxury**

[23]We use the term *indirect realization* as used in Centering Theory (Grosz *et al.*, 1995), to refer to *inferrables*, and more generally, the phenomenon of *bridging*.

[24]We note that currently, some EntRel annotations are a consequence of the constraint against annotating non-adjacent relations (Section 3.4.4). For the second phase of the PDTB, we plan to look again at EntRel annotations and consider distinguishing such cases.

**airline and casino company, 98.6%-owned by investor Kirk Kerkorian and his Tracinda Corp., earlier this month announced its agreements to acquire the properties, but didn't disclose the purchase price**. (0981)

(93) *Last year the public was afforded a preview of Ms. Bartlett's creation in a tablemodel version, at a BPC exhibition.* EntRel **The labels were breathy: "Within its sheltering walls is a microcosm of a thousand years in garden design ... At the core of it all is a love for plants**." (0984)

(94) K mart developed the centers, which range in size from about 150,000 square feet to just over 250,000 square feet. *Most are anchored by a K mart store.* EntRel **The retailer reportedly will lease its stores back from the developer, who plans to expand the small centers.** (0988)

(95) She has done little more than recycle her standard motifs – trees, water, landscape fragments, rudimentary square houses, circles, triangles, rectangles – and fit them into a grid, as if she were making one of her gridded two-dimensional works for a gallery wall. *But for South Gardens, the grid was to be a 3-D network of masonry or hedge walls with real plants inside them.* EntRel **In a letter to the BPCA, kelly/varnell called this "arbitrary and amateurish."** (0984)

(96) *Ronald J. Taylor, 48, was named chairman of this insurance firm's reinsurance brokerage group and its major unit, G.L. Hodson & Son Inc. Robert G. Hodson, 65, retired as chairman but will remain a consultant. Stephen A. Crane, 44, senior vice president and chief financial and planning officer of the parent, was named president and chief executive of the brokerage group and the unit, succeeding Mr. Taylor.* EntRel **The appointments are effective Nov. 1**. (0948)

`EntRel` annotations are not associated with any semantic class, their labels being self-evident of their semantic type. Argument selection for `EntRel` is subject to the *adjacency constraint*, though the selection can be discontinuous. The "minimality" constraint here is somewhat restricted, in that the selection should be minimal *upto* the level of the sentence. In particular, for `EntRel` we only identify the minimal set of (complete) sentences that mention the entities reified in the `Arg2` sentence. Thus, unlike `Explicit`, `Implicit` and `AltLex` annotations, arguments of the `EntRel` relation cannot contain sub-sentential spans, including those obtained by excluding attribution. In Example (97), for instance, the entire sentences are selected as `Arg1` and `Arg2`, even though the "remodeling" and "refurbishing" event entities in `Arg1` that are reified and predicated of in `Arg2` are embedded as conjoined arguments in the sentential complement, and even though the reification and predication of the same entities in `Arg2` should strictly exclude two levels of attribution (see Section 4).

(97) *Proceeds from the offering are expected to be used for remodeling the company's Desert Inn resort in Las Vegas, refurbishing certain aircraft of the MGM Grand Air unit, and to acquire the property for the new resort.* EntRel **The company said it estimates the Desert Inn remodeling will cost about $32 million, and the refurbishment of the three DC-8-62 aircraft, made by McDonnell Douglas Corp., will cost around $24.5 million**. (0981)

Example (98) illustrates an annotation of `EntRel` where multiple sentence arguments are required. The last sentence only provides an additional predication about the two mentioned ads, but since the antecedent of the referring expression, *both ads*, is "split" across the previous two sentences, both sentences are selected as `Arg1` of the `EntRel` relation.

(98) HOLIDAY ADS: Seagram will run two interactive ads in December magazines promoting its Chivas Regal and Crown Royal brands. *The Chivas ad illustrates – via a series of pullouts – the wild reactions from the pool man, gardener and others if not given Chivas for Christmas. The three-page Crown Royal ad features a black-and-white shot of a boring holiday party – and a set of colorful stickers with which readers can dress it up.* EntRel **Both ads were designed by Omnicom's DDB Needham agency**. (0989)

*Supplementary* annotations are disallowed for arguments of `EntRel`. We also do not do any further annotation within the arguments to identify the entity or entities realized across the arguments: annotation of anaphoric relations not associated directly with discourse relations is outside the scope of this project.

### 3.6.3 `NoRel` (No relation)

These are cases where no discourse relation or entity-based coherence relation can be inferred between the sentences in the selected pair. Examples (99-101) show cases where the NOREL label was used.

(99) The products already available are cross-connect systems, used instead of mazes of wiring to interconnect other telecommunications equipment. *This cuts down greatly on labor, Mr. Buchner said.* NoRel **To be introduced later are a multiplexer, which will allow several signals to travel along a single optical line; a light-wave system, which carries voice channels; and a network controller, which directs data flow through cross-connect systems**. (1064)

(100) Jacobs Engineering Group Inc. 's Jacobs International unit was selected to design and build a microcomputer-systems manufacturing plant in County Kildare, Ireland, for Intel Corp. *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million, according to Intel**. (1081)

(101) While the model was still on view, Manhattan Community Board 1 passed a resolution against South Gardens. The Parks Council wrote the BPCA that this "too 'private' . . . exclusive," complex and expensive "enclosed garden . . . belongs in almost any location but the waterfront." Lynden B. Miller, the noted public garden designer who restored Central Park's Conservatory Garden, recalls her reaction to the South Gardens model in light of the public garden she was designing for 42nd Street's Bryant Park: "Bryant Park, as designed in 1933, failed as a public space, because it made people feel trapped. *By removing the hedges and some walls, the Bryant Park Restoration is opening it up.* NoRel **It seems to me the BPCA plan has the potential of making South Gardens a horticultural jail for people and plants**." (0984)

As also noted for the `EntRel` annotations (see Footnote 24), we note that some of the `NoRel` labels are a consequence of the constraint against annotating non-adjacent relations (Section 3.4.4). That is, if no relation is perceived between two adjacent sentences, a relation cannot be marked with some non-adjacent sentence(s), even if such a relation exists. In Example (101), for instance, even while no relation is inferred due to adjacency between the last two sentences, the annotation of possible non-adjacent relational inferences is not allowed, such as the ADD.INFO/SIMILARITY relation between the claim in the last sentence and the claim in the third last sentence. As the example shows, the `NoRel` label has to be marked between the two initially selected adjacent sentences.[25]

For `Norel` annotations, only the two initially selected sentences can be annotated as the arguments. *Supplementary* annotations are disallowed. And obviously, because of the absence of a relation, no semantic class annotation is recorded.

[25]We plan to distinguish such cases for the second release of the PDTB.

# 4 Attribution

## 4.1 Introduction

The relation of "attribution" is a relation of "ownership" between abstract objects and individuals or agents. That is, attribution has to do with ascribing beliefs and assertions expressed in text to the agent(s) holding or making them ((Riloff and Wiebe, 2003; Wiebe *et al.*, 2004, 2005)). Since we take discourse connectives to convey semantic predicate-argument relations between abstract objects, one can distinguish a variety of cases depending on the attribution of the discourse relation or its arguments. For example, a discourse relation may hold either between the attributions (and the agents of attributions) themselves or just between the abstract object arguments of the attribution, as shown below:[26]

(102) 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.* (0267)

(103) *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**. (0098)

In Example (102), the TEMPORAL discourse relation denoted by *when* 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. In Example (103), on the other hand, the CONTRAST relation denoted by *while* holds between the agent arguments of the attribution relation, which means that the attribution relation is part of the contrast as well.

Abstract object arguments of attributions can be discourse relations as well, as seen in Example (104), where the TEMPORAL relation between the two arguments is also being quoted and thus attributed to an individual other than the writer of the text.

[26] In this report, the text span associated with attributions is shown boxed for illustrative purposes only. Attribution text spans have not been annotated in this version of the PDTB, and the guidelines for this have not been specified yet. We note that while some attribution spans can be identified clearly as the *reporting frames* of Huddleston and Pullum (2002), others are less clearly categorized this way, sometimes appearing as, for example, adverbial phrases, and sometimes not appearing at all (when they have to be inferred anaphorically from the prior context).

(104) "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. (2300)

In addition to `Explicit` connectives, attribution in the PDTB is also marked for `Implicit` connectives and their arguments. `Implicit` connectives express relations that are inferred by the reader. In such cases, the writer intends for the reader to infer a discourse relation. As with `Explicit` connectives, implicit relations intended by the writer of the article are distinguished from those intended by some other agent or speaker introduced by the writer of the text. For example, while the implicit relation in Example (105) is attributed to the writer, in Example (106), both `Arg1` and `Arg2` have been expressed by another speaker whose speech is being quoted: in this case, the implicit relation is attributed to the other speaker.

(105) *The gruff financier recently started socializing in upper-class circles.* `Implicit` = FOR EXAMPLE (ADD.INFO) Although he says he wasn't keen on going, **last year he attended a New York gala where his daughter made her debut**. (0800)

(106) "*We've been opposed to" index arbitrage "for a long time*," said Stephen B. Timbers, chief investment officer at Kemper, which manages $56 billion, including $8 billion of stocks. `Implicit` = BECAUSE (CAUSE) "**Index arbitrage doesn't work, and it scares natural buyers" of stock**. (1000)

Attribution is also annotated for `AltLex` relations, but not for `EntRel` and `NoRel`.

Attribution annotation is done with feature labels, and involves a three-way categorization in terms of (1) the *source* of attribution, (2) the *factuality* of attribution, and (3) the *polarity* of attribution.[27] We discuss each of these in turn. (In what follows, attribution features and values assigned to examples are shown below the examples; REL stands for 'Relation' - explicit or implicit.)

## 4.2 Source

The *source* feature distinguishes primarily between two *sources* of attribution, the Writer of the text ("Wr"), or some other Speaker (or Agent) mentioned by the Writer ("Ot"). Writer attribution is the default, and is also used for cases indicating an ambiguity between "Wr" and "Ot".

[27] The annotation of *factuality* and *polarity* will most likely undergo revision for the second release.

When the *source* feature value of an argument is the same as that of the connective, the argument's source is marked as "Inh", to indicate that it "inherits" the attribution value of the connective. With respect to the attribution source associated with discourse connectives and their arguments, there are broadly two possibilities:

**Case 1** A connective and both its arguments are attributed to the same source, either the Writer, as for Example (107), or the Speaker (Bill Bierdermann) in Example (108):

(107) *At that price, CBS was the only player at the table* when **negotiations with the International Olympic Committee started in Toronto Aug. 23**. (1057)
*Source:* REL=Wr; `Arg1`=Inh; `Arg2`=Inh

(108) "*The public is buying the market* when **in reality there is plenty of grain to be shipped**," said Bill Biedermann, Allendale Inc. research director. (0192)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh

**Case 2** One or both arguments have a different attribution value from that of the connective. In Example (109), the connective and `Arg1` are attributed to the Writer, whereas `Arg2` is attributed to another Speaker (here, the purchasing agents).

(109) *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**. (0178)
*Source:* REL=Wr; `Arg1`=Inh; `Arg2`=Ot

In some cases, in particular when the source is "Ot", attributions are implicit or unrealized, and have to be inferred from the prior context. Such inferred attributions are also annotated.

(110) "There are certain cult wines that can command these higher prices," says Larry Shapiro of Marty's, one of the largest wine shops in Dallas. "What's different is that it is happening with young wines just coming out. *We're seeing it* partly because **older vintages are growing more scarce**." (0071)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh

As the previous examples illustrate, if an explicit attribution associated with a connective or argument does not play a role in the interpretation of the discourse relation, the phrase corresponding to the agent and verb of attribution is not included in the annotation. The exception to this is the obligatory inclusion of non-clausal attributions, such as "*according*

*to*" phrases (Examples (111-112), because of the general guideline for not excluding non-clausal adjuncts (Section 2.8.1). However, attributions are still reflected in the annotations for such cases as well, because of the use of features.[28]

(111) *No foreign companies bid on the Hiroshima project,* *according to the bureau*. But **the Japanese practice of deep discounting often is cited by Americans as a classic barrier to entry in Japan's market**. (0501)
*Source:* REL=Wr; Arg1=Ot; Arg2=Inh

(112) *Even so,* *according to Mr. Salmore*, *the ad was "devastating"* because **it raised questions about Mr. Courter's credibility**. (0041)
*Source:* REL=Ot; Arg1=Inh; Arg2=Inh

## 4.3 Factuality

The *factuality* feature distinguishes primarily between assertions and beliefs attributed to agents, by making a distinction between reporting verbs/phrases (marked "Fact"), such as *say*, *announce*, etc., and verbs/phrases of propositional attitude (marked "NonFact"), such as *believe*, *think* etc. In general, *factuality* is meant to capture the observable *degree of commitment* of an individual towards a proposition. All the non-Writer attributions in the examples above illustrate "Fact" attributions. Examples (113-115) illustrate "NonFact" attributions.

(113) Lawyers worry *that* if **they provide information about clients**, *that data could quickly end up in the hands of prosecutors.* (0049)
*Source:* REL=Ot; Arg1=Inh; Arg2=Inh
*Factuality:* REL=NonFact; Arg1=Null; Arg2=Null

(114) Jay Goldinger, with Capital Insight Inc., reasons *that* while **the mark has posted significant gains against the yen as well** – the mark climbed to 77.70 yen from 77.56 yen late Tuesday in New York – *the strength of the U.S. bond market compared to its foreign counterparts has helped lure investors to dollar-denominated bonds, rather than mark bonds.* (0059)
*Source:* REL=Ot; Arg1=Inh; Arg2=Inh
*Factuality:* REL=NonFact; Arg1=Null; Arg2=Null

[28]We note that non-clausal attributions will be further identified when the text spans associated with attributions are annotated for the second release.

(115) A spokesman for Temple estimated *that Sea Containers' plan* – if **all the asset sales materialize** – *would result in shareholders receiving only $36 to $45 a share in cash.* (0063)
*Source:* REL=Ot; Arg1=Inh; Arg2=Inh
*Factuality:* REL=NonFact; Arg1=Null; Arg2=Null

The "Null" feature value for *factuality* is used for arguments, and it means that the *factuality* of a relation's argument needs to be derived by independent (here, undefined) considerations under the scope of the relation. However, when an argument appears under an attribution distinct from that of the relation, its *factuality* is indicated explicitly, based on its own local attribution. So while the *factuality* of all the arguments in Examples (113-115) above are undefined with the "Null" value, Examples (116-119) below have *factuality* defined separately for one or more of their arguments because they appear under an attribution distinct from that of the relation.

(116) 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.* (0267)
*Source:* REL=Wr; Arg1=Ot; *Arg2*=Inh
*Factuality:* REL=Fact; Arg1=Fact; Arg2=Null

(117) *Factory orders and construction outlays were largely flat in December* while purchasing agents said **manufacturing shrank further in October**. (0178)
*Source:* REL=Wr; Arg1=Inh; Arg2=Ot
*Factuality:* REL=Fact; Arg1=Null; Arg2=Fact

(118) *No foreign companies bid on the Hiroshima project,* *according to the bureau*. But **the Japanese practice of deep discounting often is cited by Americans as a classic barrier to entry in Japan's market**. (0501)
*Source:* REL=Wr; Arg1=Ot; Arg2=Inh
*Factuality:* REL=Fact; Arg1=Fact; Arg2=Null

(119) "*Having the dividend increases is a supportive element in the market outlook*, but I don't think **it's a main consideration**," he says. (0090)
*Source:* REL=Ot; Arg1=Inh; Arg2=Inh
*Factuality:* REL=Fact; Arg1=Null; Arg2=NonFact

The default for *factuality* is "Fact" for relations, and "Null" for arguments.

A single connective or argument can appear with multiple types of (stacked) attributions, as in the following examples. In such cases, the attribution syntactically closest to the connective or argument is selected for annotation.

(120) Mary Elizabeth Ariail, another social-studies teacher, says [she believed] *Mrs. Yeargin wanted to keep her standing high* so **she could get a new job that wouldn't demand good hearing**. (0044)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=NonFact; `Arg1`=Null; `Arg2`=Null

(121) But, says the general manager of a network affiliate in the Midwest, "[I think] if **I tell them I need more time**, *they'll take 'Cosby' across the street* (0060)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=NonFact; `Arg1`=Null; `Arg2`=Null

(122) One station manager says [he believes] *Viacom's move is a "pre-emptive strike"* because **the company is worried that "Cosby" ratings will continue to drop in syndication over the next few years**. (0060)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=NonFact; `Arg1`=Null; `Arg2`=Null

(123) "*Having the dividend increases is a supportive element in the market outlook*, but [I don't think] **it's a main consideration**," he says. (0090)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=Fact; `Arg1`=Null; `Arg2`=NonFact

## 4.4 Polarity

The feature values for *polarity* are "Pos" and "Neg", with "Pos" as the default value.

The *polarity* feature is annotated (on all explicit and implicit relations and their arguments) to primarily distinguish two cases when verbs of attribution are negated on the surface - syntactically (e.g., *didn't say*, *don't think*) or lexically (e.g., *denied*). The distinction is based on (a) whether surface negation is interpreted in a "lower" position, i.e., as taking scope over the complement of the attribution (Horn, 1978), or (b) whether surface negation is interpreted *in situ*.

Example (124) illustrates the first case. The 'but' clause entails an interpretation such as "I think it's not a main consideration", for which the negation must be interpreted lower than

its surface position. In particular, the interpretation of the CONTRAST relation denoted by *but* requires that `Arg2` should be interpreted directly under the scope of negation.

(124) "*Having the dividend increases is a supportive element in the market outlook*, but I don't think **it's a main consideration**," he says. (0090)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=Fact; `Arg1`=Null; `Arg2`=NonFact
*Polarity:* REL=Pos; `Arg1`=Pos; `Arg2`=Neg

To capture such entailments with surface negated attribution verbs, an argument of a connective is marked "Neg" for *polarity* when the interpretation of the connective requires the surface negation to take semantic scope over the lower argument. In general, "Neg" is marked wherever negation is interpreted. Thus, in Example (124), *polarity* is marked as "Neg" for `Arg2`.

In the absence of explicit negative attributions, the polarity is marked as the default "Pos", such as for Writer attributions, or Other positive attributions (such as for the relation and `Arg1` of Example 124).

Examples (125-127) illustrate the second case, i.e., the *in situ* interpretation of negation. Such interpretations imply that the assertion, question, or belief cannot be attributed to the individual in question. Since negation in these cases is not interpreted lower, the arguments of the connectives are all marked as the default "Pos". The *in situ* interpretation is indicated by marking "Neg" for the *polarity* of the relation.

(125) In an interview with the trade journal Automotive News,
Mr. Iacocca declined to say *which plants will close* or **when Chrysler will make the moves**. (0435)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=Fact; `Arg1`=Null; `Arg2`=Null
*Polarity:* REL=Neg; `Arg1`=Pos `Arg2`=Pos

(126) Researchers couldn't estimate *the cost of the drug* when it reaches the market (1934)
(read `Arg1` as a nominalization, i.e., as *what the drug would cost* or as *what the cost of the drug would be* - see Section 2.7.2)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=NonFact; `Arg1`=Null; `Arg2`=Null
*Polarity:* REL=Neg; `Arg1`=Pos `Arg2`=Pos

(127) Mr. Driscoll didn't elaborate *about who the potential partners were* or **when the talks were held**. (0067)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=Fact; `Arg1`=Null; `Arg2`=Null
*Polarity:* REL=Neg; `Arg1`=Pos; `Arg2`=Pos

We note that there is a third possibility for how surface negation can be interpreted, although we have not observed this in the corpus annotated so far. The constructed example in (128a) illustrates such a case. In addition to entailing (128b) - in which case it would be annotated parallel to Example (124) above - (128a) can also entail (128c), such that the negation is intrepreted lower, as taking scope over the "relation" (Lasnik, 1975), rather than one of the arguments. As the *polarity* annotations for (128c) show, such cases, if they do arise, are currently indistinguishable from the polarity annotation of Examples (125-127).

(128) a. John doesn't think *Mary will get cured* because **she took the medication**.
b. $\models$ John thinks *that* because **Mary took the medication**, *she will not get cured.*
c. $\models$ John thinks *that Mary will get cured* not because **she took the medication** (but because she has started practising yoga.)
*Source:* REL=Ot; `Arg1`=Inh; `Arg2`=Inh
*Factuality:* REL=NonFact; `Arg1`=Null; `Arg2`=Null
*Polarity:* REL=Neg; `Arg1`=Pos; `Arg2`=Pos

# Appendix A

This Appendix provides a distribution of the type of `Explicit` connectives in the PDTB, Release 1.0. Tables 1 and 2 are split tables, their "Total" adding to 18505.

| `Explicit` Connective | No. |
|---|---|
| accordingly | 6 |
| additionally | 7 |
| after | 577 |
| afterward | 11 |
| also | 1748 |
| alternatively | 6 |
| although | 328 |
| and | 3003 |
| as | 743 |
| as a result | 78 |
| as an alternative | 2 |
| as if | 16 |
| as long as | 24 |
| as soon as | 20 |
| as though | 6 |
| as well | 6 |
| because | 860 |
| before | 326 |
| before and after | 1 |
| besides | 19 |
| but | 3312 |
| by comparison | 11 |
| by contrast | 27 |
| by then | 8 |
| consequently | 10 |

| `Explicit` Connective | No. |
|---|---|
| conversely | 2 |
| earlier | 15 |
| either..or | 4 |
| else | 1 |
| except | 10 |
| finally | 32 |
| for | 3 |
| for example | 197 |
| for instance | 98 |
| further | 9 |
| furthermore | 11 |
| hence | 4 |
| however | 487 |
| if | 1234 |
| if and when | 3 |
| if..then | 38 |
| in addition | 165 |
| in contrast | 12 |
| in fact | 83 |
| in other words | 17 |
| in particular | 15 |
| in short | 4 |
| in sum | 2 |
| in the end | 10 |
| in turn | 30 |

Table 1: Distribution of `Explicit` connective types

| Explicit Connextive | No. |
|---|---|
| indeed | 104 |
| insofar as | 2 |
| instead | 112 |
| later | 92 |
| lest | 2 |
| likewise | 8 |
| meantime | 15 |
| meanwhile | 193 |
| moreover | 101 |
| much as | 9 |
| neither..nor | 3 |
| nevertheless | 44 |
| next | 7 |
| nonetheless | 27 |
| nor | 31 |
| now that | 22 |
| on the contrary | 4 |
| on the one hand..on the other hand | 1 |
| on the other hand | 37 |
| once | 85 |
| or | 98 |
| otherwise | 24 |
| overall | 12 |
| plus | 1 |
| previously | 49 |

| Explicit Connective | No. |
|---|---|
| rather | 17 |
| regardless | 2 |
| separately | 74 |
| similarly | 18 |
| simultaneously | 6 |
| since | 184 |
| so | 263 |
| so that | 31 |
| specifically | 10 |
| still | 197 |
| then | 340 |
| thereafter | 11 |
| thereby | 12 |
| therefore | 26 |
| though | 320 |
| thus | 113 |
| till | 3 |
| ultimately | 18 |
| unless | 95 |
| until | 162 |
| when | 990 |
| when and if | 1 |
| whereas | 5 |
| while | 782 |
| yet | 101 |

Table 2: Distribution of Explicit connective types (Cont.)

# Appendix B

This Appendix lists modified forms and variants of `Explicit` connectives in the PDTB, 1.0.

| `Explicit` Connective | Modified forms and variants |
|---|---|
| accordingly | - |
| additionally | - |
| after | after, six years after, two weeks after, only after, even after, shortly after, a day after, nearly two months after, one day after, about a week after, soon after, a week after, just after, only two weeks after, less than a month after, long after, only three years after, almost immediately after, nearly a year and a half after, some time after, months after, within a year after, five years after, sometimes after, just a day after, reportedly after, four days after, immediately after, just five months after, two days after, three months after, years after, 18 months after, minutes after, just 15 days after, in the first 25 minutes after, seven years after, 29 years and 11 months to the day after, a year after, about three weeks after, just minutes after, a few weeks after, right after, more than a year after, within minutes after, a few months after, a month after, 25 years after, eight months after, just a month after, especially after, particularly after, a few hours after |
| afterward | shortly afterwards, shortly afterward, afterward, afterwards |
| also | - |
| alternatively | - |
| although | - |
| and | - |
| as | as, just as, especially as, even as, particularly as |
| as a result | as a result, largely as a result |
| as an alternative | - |
| as if | - |
| as long as | as long as, only as long as |
| as soon as | as soon as, just as soon as |
| as though | - |
| as well | - |

Table 3: Modified forms and variants of `Explicit` connectives

| Explicit Connective | Modified forms and variants |
|---|---|
| because | because, only because, in part because, partly because, primarily because, merely because, largely because, simply because, not only because, just because, particularly because, in large part because, mainly because, at least partly because, apparently because, not because, perhaps because, presumably because, especially because |
| before | before, even before, almost before, just before, fully eight months before, in the 3 1/2 years before, two years before, a day or two before, just days before, long before, a decade before, two days before, several months before, years before, shortly before, a week before, five minutes before, since before, an average of six months before, about six months before, a full five minutes before, two months before, just eight days before |
| before and after | - |
| besides | - |
| but | - |
| by comparison | - |
| by contrast | - |
| by then | - |
| consequently | - |
| conversely | - |
| earlier | - |
| either..or | - |
| else | - |
| except | - |
| finally | - |
| for | - |
| for example | - |
| for instance | - |
| further | - |
| furthermore | - |
| hence | - |
| however | - |

Table 4: Modified forms and variants of Explicit connectives (Cont.)

| `Explicit` **Connective** | **Modified forms and variants** |
|---|---|
| if | if, even if, only if, if only, typically if, especially if, particularly if |
| if and when | - |
| if..then | - |
| in addition | - |
| in contrast | - |
| in fact | - |
| in other words | - |
| in particular | - |
| in short | - |
| in sum | - |
| in the end | - |
| in turn | - |
| indeed | - |
| insofar as | - |
| instead | - |
| later | later, later on |
| lest | - |
| likewise | - |
| meantime | in the meantime, meantime |
| meanwhile | meanwhile, in the meanwhile |
| moreover | - |
| much as | as much as, so much as, much as |
| neither..nor | - |
| nevertheless | - |
| next | - |
| nonetheless | - |
| nor | - |
| now that | - |
| on the contrary | - |
| on the one hand..on the other hand | - |
| on the other hand | - |
| once | - |

Table 5: Modified forms and variants of `Explicit` connectives (Cont.)

| `Explicit` Connective | Modified forms and variants |
|---|---|
| or | - |
| otherwise | - |
| overall | - |
| plus | - |
| previously | - |
| rather | - |
| regardless | - |
| separately | - |
| similarly | - |
| simultaneously | almost simultaneously, simultaneously |
| since | since, ever since, particularly since, especially since |
| so | - |
| so that | - |
| specifically | - |
| still | still, even still |
| then | then, even then |
| thereafter | thereafter, shortly thereafter |
| thereby | - |
| therefore | - |
| though | even though, though |
| thus | - |
| till | - |
| ultimately | - |
| unless | - |
| until | until, at least until, only until, just until |
| when | when, only when, usually when, at least not when, especially when, just when, except when, even when, back when, at least when, particularly when |
| when and if | - |
| whereas | - |
| while | while, even while |
| yet | - |

Table 6: Modified forms and variants of `Explicit` connectives (Cont.)

# Appendix C

This Appendix gives the distribution of the types of implicit relations annotated in the PDTB, 1.0., along with the semantic classes for `Implicit` connectives and `AltLex` relations. (Multiple `Implicit` connectives are shown together as a distinct type and are separated by "_". "Restate/Sum" is a single semantic class, short for "Restatement/Summarization"; "Add.Info" is short for "Additional-information".) Tables 10, 11, and 12 are split, their "Total" adding to 1496.

| **Implicit relation type** | **No.** |
|---|---|
| `Implicit` | 1496 |
| `AltLex` | 19 |
| `EntRel` | 435 |
| `NoRel` | 53 |
| Total | 2003 |

Table 7: Distribution of types of implicit relations

| **Semantic class** | **No.** |
|---|---|
| Add.Info | 793 |
| Temporal | 58 |
| Contrast | 281 |
| Cause | 192 |
| Consequence | 138 |
| Restate/Sum | 23 |
| Add.Info_Add.Info | 3 |
| Temporal_Add.Info | 1 |
| Add.Info_Temporal | 8 |
| Contrast_Add.Info | 2 |
| Add.Info_Cause | 1 |
| Cause_Add.Info | 11 |
| Cause_Temporal | 1 |
| Contrast_Contrast | 1 |
| Add.Info_Consequence | 2 |
| TOTAL | 1515 |

Table 8: Distribution of semantic classes for `Implicit` connectives and `AltLex`

| AltLex semantic class | No. |
|---|---|
| Add.Info | 2 |
| Cause | 3 |
| Consequence | 9 |
| Restate/Sum | 1 |
| Temporal | 4 |
| Total | 19 |

Table 9: Distribution of semantic classes for AltLex

| `Implicit` connective | Semantic class | No. |
|---|---|---|
| accordingly | Consequence | 2 |
| afterwards | Temporal | 1 |
| after | Temporal | 10 |
| also | Add.Info | 38 |
| although | Contrast | 25 |
| and_as a result | Add.Info_Consequence | 1 |
| and_so | Add.Info_Consequence | 1 |
| and_then | Add.Info_Temporal | 8 |
| and | Add.Info | 184 |
| as a result | Consequence | 26 |
| as | Cause | 21 |
| as | Temporal | 1 |
| at the same time | Add.Info | 1 |
| because_for example | Cause_Add.Info | 10 |
| because_for instance | Cause_Add.Info | 1 |
| because_previously | Cause_Temporal | 1 |
| because | Cause | 151 |
| besides | Add.Info | 2 |
| but_also | Contrast_Add.Info | 1 |
| but_in addition | Contrast_Add.Info | 1 |
| but_nevertheless | Contrast_Contrast | 1 |
| but | Contrast | 86 |
| by comparison | Add.Info | 3 |
| by comparison | Contrast | 11 |
| by contrast | Contrast | 12 |
| consequently | Consequence | 10 |
| earlier | Temporal | 5 |
| even though | Contrast | 5 |
| finally | Add.Info | 1 |
| finally | Temporal | 1 |

Table 10: Distribution of `Implicit` connectives and their semantic classes

| Implicit connective | Semantic class | No. |
|---|---|---|
| first | Add.Info | 3 |
| for example | Add.Info | 107 |
| for instance | Add.Info | 21 |
| for one thing | Add.Info | 1 |
| for one thing | Cause | 1 |
| furthermore | Add.Info | 46 |
| further | Add.Info | 4 |
| however | Contrast | 61 |
| in addition | Add.Info | 13 |
| in comparison | Add.Info | 1 |
| in contrast | Contrast | 4 |
| in fact_for example | Add.Info_Add.Info | 1 |
| in fact | Add.Info | 57 |
| in fact | Contrast | 2 |
| in other words | Restate/Sum | 15 |
| in particular_because | Add.Info_Cause | 1 |
| in particular_for example | Add.Info_Add.Info | 2 |
| in particular | Add.Info | 87 |
| in short | Restate/Sum | 1 |
| in sum | Restate/Sum | 4 |
| in the end | Consequence | 1 |
| in turn | Add.Info | 3 |
| indeed | Add.Info | 38 |
| indeed | Restate/Sum | 1 |
| instead | Contrast | 13 |
| later | Temporal | 1 |

Table 11: Distribution of Implicit connectives and their semantic classes (Cont.)

| `Implicit` connective | Semantic class | No. |
|---|---|---|
| meanwhile | Add.Info | 4 |
| meanwhile | Contrast | 2 |
| meanwhile | Temporal | 1 |
| moreover | Add.Info | 16 |
| nevertheless | Contrast | 8 |
| on the contrary | Contrast | 1 |
| on the other hand | Contrast | 13 |
| or | Add.Info | 2 |
| particularly | Add.Info | 2 |
| previously | Temporal | 2 |
| rather | Contrast | 11 |
| similarly | Add.Info | 7 |
| simultaneously | Temporal | 1 |
| since | Cause | 16 |
| so that | Consequence | 1 |
| so | Consequence | 74 |
| specifically | Add.Info | 111 |
| still | Contrast | 5 |
| subsequently | Temporal | 4 |
| that is | Restate/Sum | 1 |
| then | Temporal | 17 |
| therefore | Consequence | 9 |
| though | Contrast | 1 |
| thus | Consequence | 6 |
| what's more | Add.Info | 1 |
| when_for example | Temporal_Add.Info | 1 |
| when | Temporal | 6 |
| whereas | Contrast | 11 |
| while | Add.Info | 38 |
| while | Contrast | 10 |
| while | Temporal | 4 |

Table 12: Distribution of `Implicit` connectives and their semantic classes (Cont.)

# References


Asher, N. (1993). *Reference to Abstract Objects*. Kluwer, Dordrecht.

Dinesh, N., Lee, A., Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2005). Attribution and the (non)-alignment of syntactic and discourse arguments of connectives. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor, Michigan.

Forbes-Riley, K., Webber, B., and Joshi, A. (2006). Computing discourse semantics: The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, **23**, 55–106.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203–225.

Hirschberg, J. and Litman, D. J. (1987). Now let's talk about NOW: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 163–171.

Horn, L. (1978). Remarks on neg-raising. In P. Cole, editor, *Syntax and Semantics 9: Pragmatics*. Academic Press, New York.

Huddleston, R. and Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge Univ. Press, Cambridge, UK.

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

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

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: the interaction of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 181–196. Benjamins, Amsterdam.

Lasnik, H. (1975). On the semantics of negation. In *Contemporary Research in Philosophical Logic and Linguistic Semantics*, pages 279–313. Dordrecht: D. Reidel.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004a). Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, Boston, MA.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004b). The Penn Discourse Treebank. In *Proceedings of the 4th International Conference on Language Rescourses and Evaluation (LREC 2004)*, Lisbon, Portugal.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A., and Webber, B. (2005). Experiments on sense annotation and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005)*, Barcelona, Spain.

Prasad, R., Miltsakaki, E., Joshi, A., and Webber, B. (2004). Annotation and data mining of the Penn Discourse Treebank. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 88–95, Barcelona, Spain.

Prasad, R., Joshi, A., Dinesh, N., Lee, A., Miltsakaki, E., and Webber, B. (2005). The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for NLG*.

Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP03)*, pages 105–112, Sapporo, Japan.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge University Press, Cambridge.

Webber, B. and Di Eugenio, B. (1990). Free adjuncts in natural language instructions. In *Proceedings of COLING90*, pages 395–400.

Webber, B. and Joshi, A. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In M. Stede, L. Wanner, and E. Hovy, editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 86–92. Association for Computational Linguistics, Somerset, New Jersey.

Webber, B., Knott, A., and Joshi, A. (1999). Multiple discourse connectives in a lexicalized grammar for discourse. In *Proceedings of the Third International Workshop on Computational Semantics, Tilberg, The Netherlands.*, pages 309–325.

Webber, B., Joshi, A., Stone, M., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, **29**(4), 545–587.

Webber, B., Joshi, A., Miltsakaki, E., Prasad, R., Dinesh, N., Lee, A., and Forbes, K. (2005). A short introduction to the PDTB. In *Copenhagen Working Papers in Language and Speech Processing*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language. *Computational Linguistics*, **30**(3), 277–308.

Wiebe, J., Wilson, T., , and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, **1**(2).