Discourse Annotation
Penn Discourse Treebank (PDTB)

Aravind K Joshi
University of Pennsylvania
Philadelphia PA 19104

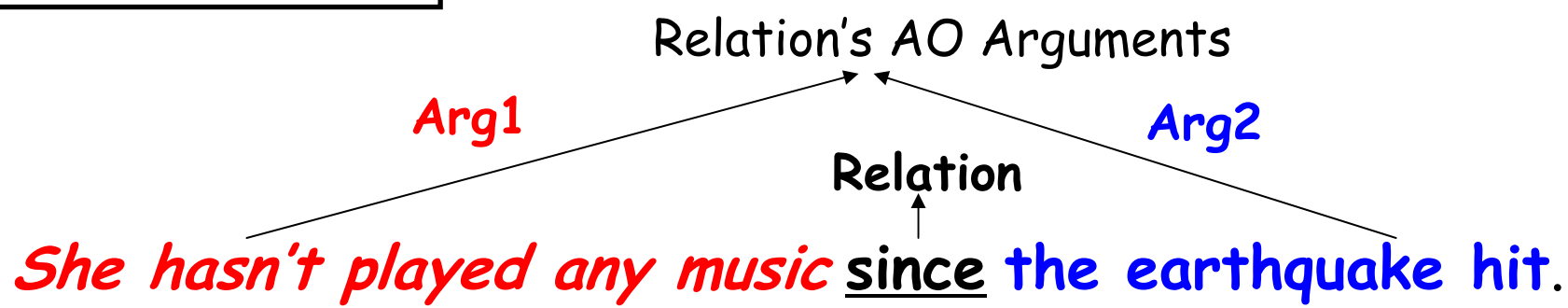
PDTB Group: Current members: Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, and Gerry Campion

Discourse Relations

Relations between the **abstract objects** (AOs) we talk about in discourse

e.g., Causal, Temporal
Contrast, Elaboration

Events, states,
Propositions



The meaning and coherence of a discourse results partly from how its constituents relate to each other.

Recognizing such relations has long-standing benefits for NLP and resulting applications (QA, QG, summarization, IE, MT)

A Paragraph Text from the WSJ Corpus

- But Mr. Noriega was convinced the Reagan White House wouldn't act against him, recalls his close ally Jose Blandon, because he had an insurance policy: his involvement with the Contra rebels in Nicaragua. Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. Mr. Noriega also conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua. And on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.

☞Most NLP processing focuses on sentence-level information.
But a great deal of useful and desirable information can be extract from discourse relations. Discourse NLP has focused on coreference.

Discourse Relations in a WSJ Paragraph

- **But** Mr. Noriega was convinced the Reagan White House wouldn't act against him, recalls his close ally Jose Blandon, **because** he had an insurance policy: his involvement with the Contra rebels in Nicaragua. **IMPLICIT (specifically)** Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. Mr. Noriega **also** conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua. **And** on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.

☞ Useful and desirable information from discourse relations -- identifiable via explicit phrases and implicit relations between adjacent sentences

Inferring from Discourse Relations

Other argument is first sentence of previous paragraph: ***Although working for U.S. Intelligence, Mr. Noriega was hardly helping the U.S. exclusively***

- **But Mr. Noriega was convinced the Reagan White House wouldn't act against him**, recalls his close ally Jose Blandon, because he had an insurance policy: his involvement with the Contra rebels in Nicaragua. Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. Mr. Noriega also conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua. And on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.

☞ Relation prevents the false inference from use of Arg1 as fact - that the Reagan White House would act against Noriega

☞ Challenges: Sense ambiguity (Concession, not Contrast);
long-distance Arg1.

Inferring from Discourse Relations

- ***But Mr. Noriega was convinced the Reagan White House wouldn't act against him***, recalls his close ally Jose Blandon, **because** **he had an insurance policy: his involvement with the Contra rebels in Nicaragua**. Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. Mr. Noriega also conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua. And on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.

- Causal relations answer why-questions

- Challenges: Arg1 identification (3 candidates in same sentence);
  Sense ambiguity (Cause, not Justification)

Inferring from Discourse Relations

- But Mr. Noriega was convinced the Reagan White House wouldn't act against him, recalls his close ally Jose Blandon, because he had an insurance policy: his involvement with the Contra rebels in Nicaragua. Mr. Blandon says ***the general allowed the Contras to set up a secret training center in Panama***. **Mr. Noriega <u>also</u> conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua**. And on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.

☞ (Unambiguous) Conjunction relation suggests grouping of facts towards a topic

☞ Challenges: Identify Arg1

Inferring from Discourse Relations

- But Mr. Noriega was convinced the Reagan White House wouldn't act against him, recalls his close ally Jose Blandon, because he had an insurance policy: his involvement with the Contra rebels in Nicaragua. Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. ***Mr. Noriega also conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua***. **And** **on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua**.

☞ More Conjunction

Inferring from Discourse Relations

- *But Mr. Noriega was convinced the Reagan White House wouldn't act against him, recalls his close ally Jose Blandon, because he had an insurance policy: his involvement with the Contra rebels in Nicaragua*. **Implicit (Specifically)** **Mr. Blandon says the general allowed the Contras to set up a secret training center in Panama. Mr. Noriega also conveyed intelligence from his spy operation inside the Nicaraguan capital of Managua. And on at least one occasion, in the spring of 1985, he helped arrange a sabotage attack on a Sandinista arsenal in Nicaragua.**

☞ Identify summary and elaboration sentences.
N.B. Adjacent conjunction relations allow grouping to form argument of a relation

☞ Challenges: Implicit relation sense detection between adjacent sentences

Creating a Corpus of Discourse Relations

Our Goal

Annotate a large-scale corpus of discourse relations to extend the scope of discourse-level NLP research and resulting applications

Penn Discourse Treebank (PDTB)

Assumptions and Methodology

- Direct marking of high-level discourse structure is difficult
- Little agreement on high-level discourse representation structures
- Instead, keep the annotation **low-level** and **theory-neutral**:
  - Mark individual relations without further composition
  - This allows corpus to be usable with different frameworks
  - Also allows for "emergent" high-level discourse structure
- **Lexically-grounded approach** ⟨ leads to high reliability
- **Stand-off** annotations ⟨ Can easily merge with other annotation layers

The Source Data

Penn Treebank II (PTB-II) portion of the Wall Street Journal (WSJ) Corpus

- 2159 texts
- Approx. 1 million words
- Approx 50K sentences

- Richly annotated (in part or whole) at other layers. E.g.,

  - POS and Syntactic constituency (Penn Treebank)
  - Semantic roles (Propbank)
  - Coreference (Ontonotes)
  - Events (Timebank)
  - Opinions (MPQA)

Discourse Relation Triggers in PDTB

Two kinds of triggers:

- **Lexical:** Discourse Relations can be **grounded in lexical items**.

Abstract Objects related by lexically anchored discourse relations can be adjacent or non-adjacent in the text

- **John went to the store** because **he had to buy glue**.
- **John went to the store**. **Then** **he went home**.
- **John went to the store**. He had to buy glue. **Then** **he went home**.

- **Structural, through Adjacency:** Discourse Relations can be triggered by **structure underlying adjacency**. Such relations are **implicit** and have to be **inferred** (but may be partly supported by text).

- **John went to the store**. **[Implicit=because (causal)]** **He had to buy glue**.

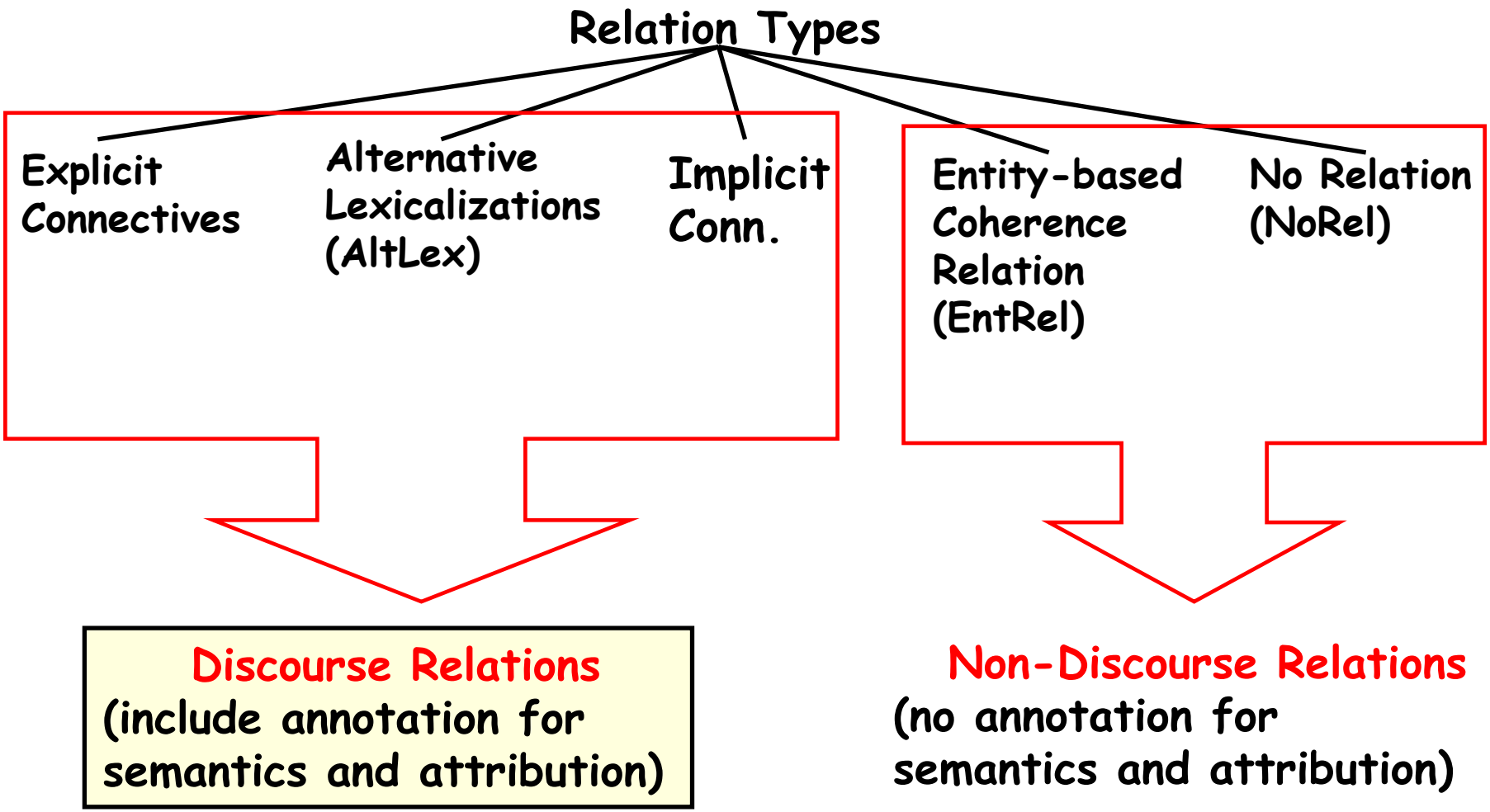
Arguments of Discourse Relations

- Arguments of discourse relations are abstract objects (AOs) - events, actions, states, propositions
- Arity of any relation is 2
- Arguments are named **Arg1** and **Arg2**, based purely on syntactic conventions:
  - For **explicit relations**, **Arg2** is the argument syntactically associated with the explicit phrase. **Arg1** is other argument.
  - For **implicit relations**, **Arg2** is the second sentence in the adjacent sentence pair. **Arg1** is the other argument

What is Annotated in PDTB

- **Discourse relations**, as their anchoring text span offsets
  - Explicitly realized relations
  - Implicit relations (text span offset linked to Arg2 span)
- **Arguments** of relations, as their anchoring text span offsets
- **Senses (semantics)** of relations, as features
- **Attribution** of relations and their arguments, as the text span offsets anchoring attribution phrases (when explicit), and features capturing the attribution semantics.

PDTB Annotation Overview



Explicit Connectives

Explicit connectives are drawn from well-defined syntactic classes:

- Subordinating conjunctions (e.g., *when*, *because*, *although*, etc.)
  - ***The federal government suspended sales of U.S. savings bonds*** **<u>because</u> Congress hasn't lifted the ceiling on government debt.**
- Coordinating conjunctions (e.g., *and*, *or*, *so*, *nor*, etc.)
  - ***The subject will be written into the plots of prime-time shows***, **<u>and</u> viewers will be given a 900 number to call.**
- Discourse adverbials (e.g., *then*, *however*, *as a result*, etc.)
  - ***In the past, the socialist policies of the government strictly limited the size of … industrial concerns to conserve resources and restrict the profits businessmen could make***. **<u>As a result</u>, industry operated out of small, expensive, highly inefficient industrial units.**

Ambiguity in Connective Identification

Lexical items used as explicit connectives can have non-discourse functions as well.

Filtering criterion: Arguments must denote Abstract Objects (AOs).

The following are rejected because the AO criterion is not met:

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

- 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.

- 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.

Modified Connectives

Connectives can be **modified** by adverbs and focus particles:

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

Initially identified connective (**since**, **if**) is extended to include modifiers.

☞ These modifications make the task of connective identification challenging, since one can't simply have a list of connectives!

Parallel Connectives

Paired connectives take the same arguments:

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

- Treated as complex connectives – annotated discontinuously

Complex Connectives

Multiple relations can sometimes be expressed as a conjunction of connectives:

- **When and if** the trust runs out of cash -- which seems increasingly likely -- *it will need to convert its Manville stock to cash*.
- 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.

- Treated as complex connectives

Linear Order of Arguments

- No constraints on relative order. Discontinuous annotation is allowed.

  - **Linear:**
    - *The federal government suspended sales of U.S. savings bonds* **because** **Congress hasn't lifted the ceiling on government debt**.

  - **Interposed:**
    - *Most oil companies*, **when** **they set exploration and production budgets for this year**, *forecast revenue of $15 for each barrel of crude produced*.
    - *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 even more tightly than restrictions aimed at individuals**.

Location of Arg1

- **Same sentence as Conn and Arg2:**
  - *The federal government suspended sales of U.S. savings bonds* **because** **Congress hasn't lifted the ceiling on government debt**.

- **Sentence immediately previous to Conn and Arg2:**
  - *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**.

- **Previous sentence non-contiguous to Conn and Arg2:**
  - Mr. Robinson … said *Plant Genetic's success in creating genetically engineered male steriles doesn't automatically mean it would be simple to create hybrids in all crops*. 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**.

Syntactic Realization of Arguments

- Simplest syntactic realization of an Abstract Object argument is:
  - A **clause**, tensed or non-tensed, or ellipsed.

  The clause can be a matrix, complement, coordinate, or subordinate clause.

- A Chemical spokeswoman said ***the second-quarter charge was "not material"*** **and that no personnel changes were made** **as a result**.
- 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.
- ***Knowing a tasty -- and free -- meal*** **when** **they eat one**, the executives gave the chefs a standing ovation.
- ***Players for the Tokyo Giants, for example, must always wear ties*** **when** **on the road**.

☞ **Syntactically implicit elements for non-finite and extracted clauses are assumed to be available.**

Exceptions to Non-Clausal Arguments

- **VP conjuncts:**
  - *It acquired Thomas Edison's microphone patent* and **then** immediately sued the Bell Co.
  - She became an abortionist accidentally, *and continued* **because** it enabled her to buy jam, cocoa and other war-rationed goodies.

- **Nominalizations: allowed only when clausal transformation OK**
  - 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.
  - But in 1976, the court permitted *resurrection of such laws*, **if** they meet certain procedural requirements.

Exceptional Non-Clausal Arguments

- **Anaphoric expressions denoting Abstract Objects:**
  - "It's important to share the risk ***and even more so*** **when the market has already peaked**."
  - Investors who bought stock with borrowed money -- that is, "on margin" -- may be more worried than most following Friday's market drop. ***That's*** **because their brokers can require them to sell some shares or put up more cash to enhance the collateral backing their loans**.

- **Responses to questions:**
  - Are such expenditures worthwhile, then? ***Yes***, **if targeted**.
  - Is he a victim of Gramm-Rudman cuts? ***No***, **but he's endangered all the same**.

N.B. Referent is annotated as Supplementary material (next slide)

Multiple Clauses: Minimality Principle

- Any number of clauses can be selected as arguments:
  - ***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**.

But, the selection is constrained by a **Minimality Principle**:

- Only as many clauses and/or sentences should be included as are minimally required for interpreting the relation. Any other span of text that is perceived to be relevant (but not necessary) should be annotated as **supplementary information**:
  - Sup1 for material supplementary to ***Arg1***
  - Sup2 for material supplementary to ***Arg2***

Supplements to Arguments

Example of **Sup1**:

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***. **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**.

Implicit Connectives

When there is no Explicit connective present to relate adjacent sentences, it may be possible to **infer** a discourse relation between them **due to adjacency**.

- ***Some have raised their cash positions to record levels***.
  Implicit=because **High cash positions help buffer a fund when the market falls.**

- ***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 **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.**

Such discourse relations are annotated by **inserting an "Implicit connective" that "best" captures the relation**.

Non-insertability of Implicit Connectives

Three types of cases where **Implicit connectives cannot be inserted** between adjacent sentences.

- **AltLex**: A discourse relation is inferred, but insertion of an Implicit connective leads to redundancy because the relation is **alternatively lexicalized** by some non-connective expression:
  - *A few years ago, the company offered two round-trip tickets on Trans World Airlines to buyers of its Riviera luxury car*. **The promotion helped Riviera sales exceed the division's forecast by more than 10%,** Buick said at the time.

Non-insertability of Implicit Connectives

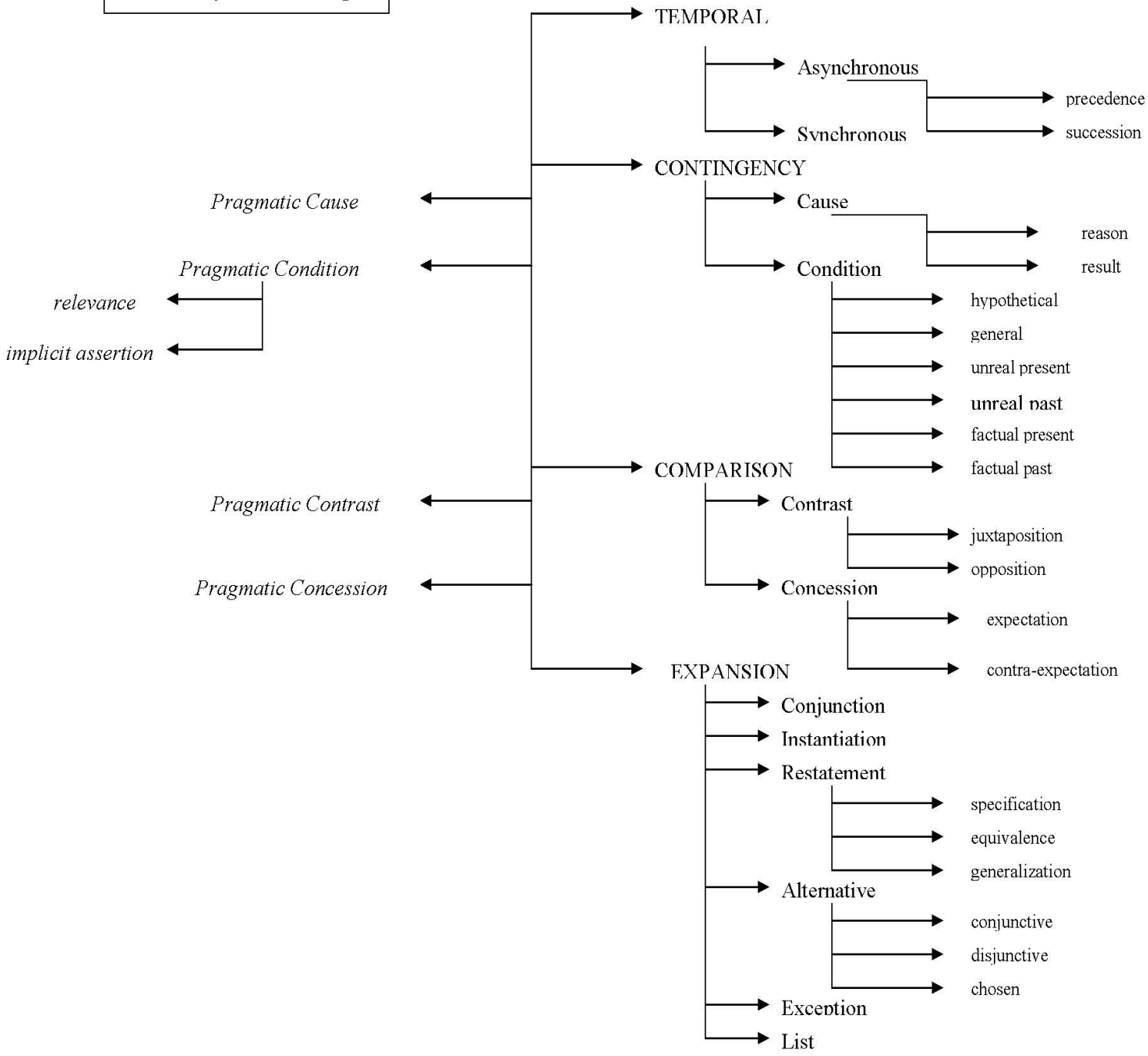
- **EntRel**: the coherence is due to an entity-based description continuation relation.
  - *Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern*. **EntRel** **Mr. Milgrim succeeds David Berman, who resigned last month**.
- **NoRel**: Neither discourse nor entity-based relation is inferred.
  - This conforms to the 'soft-landing' scenario," said Elliott Platt. *"I don't see any signs that inventories are excessive."* **A soft landing is an economic slowdown that eases inflation without leading to a recession**.

NoRel typically occurs because of the possibility of non-adjacent implicit relations which are not annotated.

Annotations of Senses in PDTB

- Sense annotations provided for Explicit, Implicit and Altlex tokens
- A hierarchical sense classification scheme
  - 3 tiers
  - Lower tiers specify more refined meanings
  - Some pragmatic dimensions included
  - Annotators have freedom to specify meaning at any depth, depending on their confidence of interpretation

Hierarchy of sense tags



First and Second level

- TEMPORAL
  - Asynchronous
  - Synchronous
- CONTINGENCY
  - Cause
  - Condition
- COMPARISON
  - Contrast
  - Concession
- EXPANSION
  - Conjunction
  - Instantiation
  - Restatement
  - Alternative
  - Exception
  - List

Third level

- TEMPORAL: Asynchronous
  - Precedence
  - Succession
- TEMPORAL: Synchronous

  No subtypes
- CONTINGENCY: Cause
  - Reason
  - Result

CONTINGENCY: Condition
- hypothetical
- general
- factual present
- factual past
- unreal present
- unreal past

Third level: subtype

- COMPARISON: Contrast
  - Juxtaposition
  - Opposition
- COMPARISON: Concession
  - Expectation
  - Contra-expectation
- EXPANSION: Restatement
  - Specification
  - Equivalence
  - Generalization
- EXPANSION: Alternative
  - Conjunctive
  - Disjunctive
  - Chosen alternative

Attribution

- **What is Attribution?**
  - Relation of "ownership" between Agents and Abstract Objects
  - Abstract objects:
    - *Assertions* (agent's commitment towards truth of proposition)
    - *Beliefs* (same as above, but different in degree)
    - *Facts* (agent's knowledge or evaluation of presupposed proposition)
    - *Eventualities* (agent's intention/attitude toward considered eventuality)

  ☛ But Attribution is not a discourse relation!
  (Not a relation between two AOs)

Why Annotate Attribution?

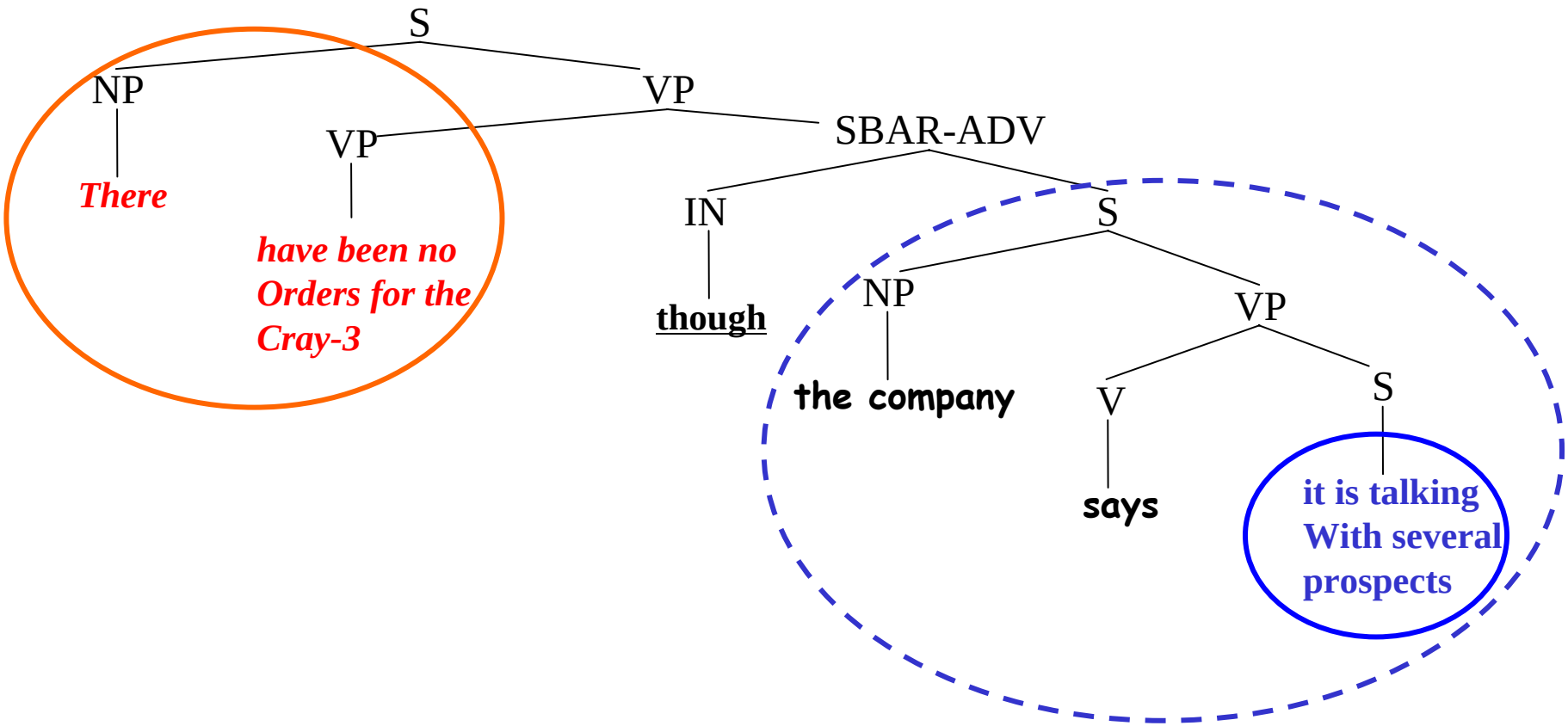
- Discourse-level annotation revealed attribution as the source for conflicts between (assumed) syntactic and semantic dependencies (Dinesh et al., 2005)

  - **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*.

  - *There have been no orders for the Cray-3 so far*, **though** the company says it is talking with several prospects.

- Worth annotating, from both a linguistic and NLP perspective!

Syntax-Discourse Mismatches: Attribution



– – – – – **Sentence semantics:** concession relation between "there being no orders for the Cray-3" and "the company saying something".

_____ **Discourse Semantics:** concession relation between "there being no orders for the Cray-3" and "there being a possibility of some prospects".

Syntax-Discourse Mismatches: Attribution

- Mismatches occur with other relations as well, such as causal relations:

  Credit analysts said *investors are nervous about the issue* <u>because</u> they say the company's ability to meet debt payments is dependent on too many variables, including the sale of assets and the need to mortgage property to retire some existing debt.

- ✓ **Discourse semantics:** causal relation between "investors being nervous" and "problems with the company's ability to meet debt payments"

- ✓ **Sentence semantics:** causal relation between "investors being nervous" and "credit analysts saying something"!

NLP Problem: Attribution

- Attribution cannot always be excluded by default

➢ *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**.

Attribution Features

Attribution **text spans** are annotated on relations and arguments, with **4 features**

Source: encodes the different agents to whom proposition is attributed

- **Wr: Writer agent**
- **Ot: Other non-writer agent**
- **Arb: Generic/Atbitrary non-writer agent**
- **Inh: Used only for arguments; attribution inherited from relation**

Type: encodes different types of Abstract Objects

- **Comm: Verbs of communication (assertions)**
- **PAtt: Verbs of propositional attitude (beliefs)**
- **Ftv: Factive verbs (facts)**
- **Ctrl: Control verbs (considered eventualities)**
- **Null: Used only for arguments with no explicit attribution**

Attribution Features

Polarity: when a surface negated attribution is interpreted lower

Neg: Lowering negation
Null: No Lowering of negation

Determinacy: indicates that the annotated TYPE of the attribution relation cannot be taken to hold in context

Indet: when the context cancels the entailment of attribution
Null: when no such embedding contexts are present

Attribution Features

Polarity: How surface negated attributions can take *narrow semantic scope* over the attributed content – over the relation or over one of the arguments:

- "*Having the dividend increases is a supportive element in the market outlook*, **but [I don't think]** **it's a main consideration**," **[he says]**.

Arg2 for the Contrast relation: **it's not a main consideration**
Neg on Arg2

Determinacy: How negation and modality associated with attributions can *cancel the attribution:*

- **[John didn't say]** ***that he left*** **because** **he was tired**
  Indet on Rel

PDTB Annotation Summary

Total No. of Tokens:	40600
Explicit Connectives:	18505
Implicit Relations:	16224
AltLex:	624
EntRel:	5210
NoRel:	254

PDTB Resources

- **PDTB** is available from the LDC
- PDTB website:
  - **http://www.seas.upenn.edu/~pdtb**
- **Tools** are available to browse and query the PDTB annotations, together with the alignments with PTB:
  - **http://www.seas.upenn.edu/~pdtb/PDTBAPI/**
    (linked from PDTB website; PTB-II distribution required to use the tools)
- The PDTB **annotation manual** (PDTB-Group, 2008) provides:
  - The guidelines followed for the annotation
  - Full Corpus distributions for annotations
- **Papers** on PDTB posted on PDTB website.
  Overview paper: "The Penn Discourse Treebank 2.0." - Prasad et al. (LREC, 2008)

A Note on "Stand-off" Annotation

- Text span annotations are represented in terms of "character offsets" in the raw text files
- Text span annotations are aligned with the Penn TreeBank (PTB), and represented as their "tree node address" in the PTB parsed files.
- Additional layers can be easily aligned as long as they are themselves stand-off

☞ Because of the stand-off representation of annotations, PDTB must be used with the PTB-II distribution, which contains the WSJ raw and PTB parsed files.

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T7

Experimental Research with PDTB

- **Identifying arguments of connectives**
(Dinesh etal 2005, Wellner 2007; Elwell and Baldridge 2008, Wellner 2009, Prasad etal 2010, Lin et al., 2010)

- **Identifying senses of explicit and implicit relations**
(Miltsakaki etal 2005, Pitler et al 2008, Pitler et al. 2009, Pitler and Nenkova 2009, Lin et al. 2009, Louis et al. 2010, Zhou et al 2010)

- **Applications:**
  - Predicting Readability (Pitler and Nenkova 2008)
  - Summarization (Louis et al. 2010)
  - Question Generation (Mannem et al 2010)

- **Experimental Linguistic Research:**
  - Local Coherence features (Louis and Nenkova 2010)
  - Alternative Lexicalizations of Relations (Prasad et al 2010)
  - Genre and Discourse Relations (Webber 2009)