DATA AND ANNOTATION
THE IMPACT OF BIG DATA ON DISCOURSE ANNOTATION

NANCY IDE

DEPARTMENT OF COMPUTER SCIENCE

VASSAR COLLEGE

POUGHKEEPSIE, NEW YORK USA

Data in Discourse Analysis • Technische Universität Darmstadt • February 18-20, 2020

OUTLINE

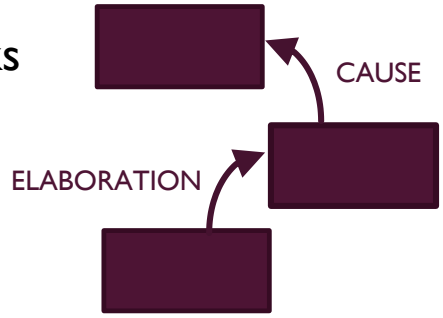
- Overview
  - Discourse analysis in Computational Linguistics (CL) /Natural Language Processing (NLP) over the years
- Theoretical choices
- Foundational theories
- Discourse analysis applied to (bigger) data
  - Different projects/practices
  - Evolution due to "data influence"

COMPUTATIONAL LINGUISTICS AND DISCOURSE ANNOTATION

- The field of Computational Linguistics (CL) / Natural Language Processing (NLP) has been transformed in recent years by the availability of **big data**
  - Annotated data used to train language models via **machine learning**

COMPUTATIONAL LINGUISTICS AND DISCOURSE ANNOTATION

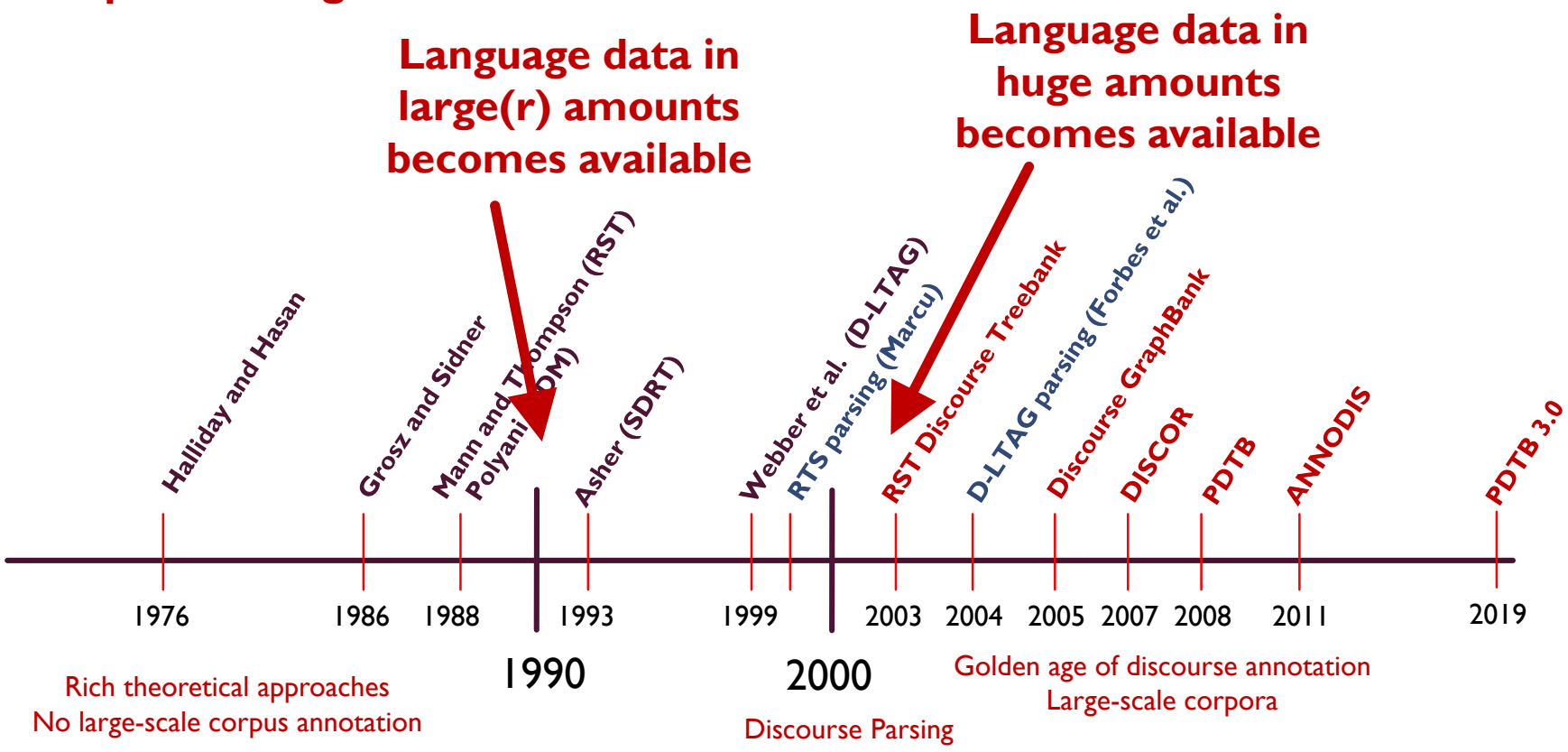
- Discourse annotation in the field of Computational Linguistics has focused on **discourse structure**
  - Dividing the text/document into relevant "units"
  - Identifying relations between/among the units
  - Providing descriptive labels for the relational links



- **Different theories of discourse structure provide different choices for how to do this**

The Big Picture

Theories
Discourse parsing
Corpus building



DISCOURSE ANALYSIS IN CL/NLP OVER THE YEARS

- Pre-1990
  - Theoretical development
    - Halliday and Hasan, Grosz and Sidner, Mann and Thompson (RST) , Polyani et al. (LDM)…
- 1990s
  - Rich theoretical approaches to discourse/text analysis not applied on a large scale
  - Annotation of discourse structure applied primarily to
    - identifying topical segments (Hearst, 1997)
    - inter-sentential relations (Nomoto and Matsumoto, 1999, Ts'ou *et al.* 2000)
    - hierarchical analyses of small corpora (Moser and Moore, 1995; Marcu *et al.* 1999)
  - Extraction of discourse structure from texts found applications in NLP
    - text summarization, information retrieval, machine translation, question answering
  - Late 90s: Discourse parsing

DISCOURSE ANALYSIS IN CL/NLP OVER THE YEARS

- 2000-present
  - **Golden age of discourse annotation**
  - Starts with trying to adapt existing theories to large-scale annotation
  - Later: theory-neutral
  - Annotation schemes affected by annotation needs
    - Try to find a balance between granularity of tagging and ability to identify discourse segments, relations, etc. consistently on a large scale
  - Data-driven approach
    - Nature of the data affecting annotation scheme design

DISCOURSE ANNOTATION

THEORETICAL CHOICES

SYNTAX OR SEMANTICS?

- Where do we introduce discourse structure?
  - Is it an extension of a syntactic parse of a text's constituent sentences?
  - Is it an extension of the semantic component?
- Most work on discourse structure takes the latter position
  - A discourse structure is a **semantic object**
    - a **graph** involving some sort of **semantic entities** as vertices and a **relational structure** over those entities

WHY DO WE CARE?

- **This choice has an effect on the design of an annotation scheme**
  - Which features are to be exploited to determine the nature of the discourse structure?
    - Syntactic (subject-verb inversion, sentence mood, modality…)
    - Semantic (antonyms for CONTRAST, hypernyms, etc.; verb or lexical classes such as anaphors)
    - Entities
    - Lexical (discourse markers, verbs *concede* and *cause* for CONCESSION and CAUSE…)
    - Morphological (tense for temporal relations, some non-finite verbs…)
    - Presentational (e.g. lists and headings)
      - Hovy and Arens (1991), Dale (1991), Bateman et al. (2001)

HOW ARE DISCOURSE STRUCTURES TO BE DEFINED?

- Some theories on the market
  - Rhetorical Structure Theory (RST) (Mann & Thompson, 1987)
  - Segmented Discourse Representation Theory (SDRT) (Asher, 1993)
  - Linguistic Discourse Model (LDM) (Polanyi *et al.*, 1988, 2004)
  - GraphBank model (Wolf & Gibson, 2005)
  - Penn Discourse Treebank model (PDTB) (Prasad *et al.*, 2008)
- Most define **hierarchical structures** by constructing complex discourse units (CDUs) from elementary discourse units (EDUs), i.e., "bottom-up", in recursive fashion

IMPLICATIONS

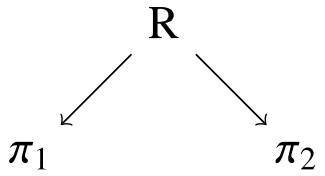
- **Annotation scheme designers have to weigh what theoretical work says with respect to what sort of annotation they want to do**
  - Some choices proposed by some theories may be suitable for some annotation tasks and not for others

WHAT ARE THE ELEMENTARY DISCOURSE UNITS?

- The first step in characterizing the discourse structure of a text is to determine the **elementary discourse units** (EDUs)
  - Minimal building blocks of a discourse tree
- Competing Hypotheses
  - Clauses (Grimes, 1975; Givon, 1983; Longacre, 1983; RST; DLTAG; SDRT)
  - Prosodic units (Hirschberg and Litman, 1993)
  - Sentences (Polanyi, 1988)
  - Intentionally defined discourse segments (Grosz and Sidner, 1986)
- Regardless of their theoretical stance, (almost) all agree that elementary discourse units are **non-overlapping spans of text**

ATTACHMENT DECISIONS

- Two approaches:
  - Discourse structures are **trees** (DLTAG, LDM, RST)
  - Discourse structures are some sort of **non-tree-like graph** (SDRT, Graphbank)
- Depends on how you answer:
  - Should the discourse annotations/structures make the semantic scope explicit for discourse relations?



  - I.e., does relation *R* have as its left argument the constituent π_1 and as its right argument the constituent π_2?

ATTACHMENT ISSUES

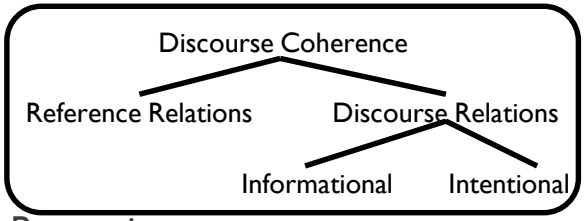
Theories supporting tree structures have a problem with long distance attachments

Some solutions:

1. **Add another layer of annotation** in which some nodes are labelled *nucleus* and others labeled *subordinate (satellite)*
   - Theories supporting tree structures have to make one of the two relations dominate the other
   - Use additional layer to compute the actual semantic scopes of discourse relations
2. **Adjust the conception of the discourse structure** to retain the scoping information

DISCOURSE RELATIONS

- The meaning and coherence of a discourse results partly from how its constituents relate to each other
  - Reference relations
  - Discourse relations



- **Informational**
  - Understanding Linguistic Structure is sufficient for Discourse Processing
  - Independent of how humans process discourse
- **Intentional**
  - Understanding Speaker Intentions is required for Discourse Processing

You'll want to book your reservation before the end of the day. Proposition 143 goes into effect tomorrow.

- **Intentional structure**: convince the caller to book a reservation before the end of the day
- **Informational structure**: explanation relation between two sentences

Most annotation schemes focus on **informational or semantic relations** (e.g, CONTRAST, CAUSE, CONDITIONAL, TEMPORAL, etc.) between abstract entities of appropriate sorts (e.g., facts, beliefs, eventualities, etc.), commonly called **Abstract Objects (AOs)** [Asher, 1993]

DISCOURSE RELATIONS

- **Theories and annotation schemes differ on what types of informational discourse relations there are, and how many**
  - Source of greatest difference among theories
  - Some (e.g. RST) have a large (50-80) number of relations, while others have few or none (e.g., LDM)
- Most annotation models include relations that allow for various kinds of
  - Expansion or elaboration of a given discourse unit
  - Explanatory links (why an event described in one discourse unit occurred)
  - Narrative and forward causal sequences
  - Structural relations like Parallel and Contrast
- **BUT no unique set of relations that is:**
  - **Suitable to accurately describe all attachments**
  - **Of a size and granularity appropriate for a substantial annotation task**
- **Devising such a set remains a controversial and difficult task**

DO WE NEED DISCOURSE RELATIONS?

- Some researchers have questioned the wisdom of identifying a specific set of relations
  - Grosz and Sidner, 1986
    - Trying to identify the "correct" set is a doomed enterprise, because there is no closed set
    - Do not disagree with the idea that relationships between adjacent clauses and blocks of clauses provide meaning and enforce coherence
    - But object to the notion that some small set of inter-clausal relations can describe English discourse adequately

DO WE NEED DISCOURSE RELATIONS?

- Others argue:
  - Discourse relations provide a level of description that is **capable of supporting a level of inference** potentially relevant to many NLP applications
    - Evidence from attempts to construct working systems that inter-clausal relations required to guide inference and planning processes
    - Without relations cannot e.g. plan an adequate multi-sentence paragraph by computer

SPECIFYING DISCOURSE RELATIONS

Broadly, there are two ways of specifying discourse relations:

- **Abstract specification**
  - Relations between two given Abstract Objects are always **inferred**, and **declared by choosing from a pre-defined set of abstract categories (relations)**
    - **Lexical elements** can serve as partial, ambiguous evidence for inference
- **Lexically grounded**
  - Relations grounded in **lexical elements**
  - Where lexical elements are absent, relations may be **inferred**

TRIGGERS

Similarly, there are two types of triggers for discourse relations considered by researchers:

- **Structure**
  - Discourse relations hold primarily between (adjacent) components with respect to some notion of structure

- **Lexical Elements and Structure**
  - Lexically-triggered discourse relations can relate the Abstract Object interpretations of **non-adjacent** as well as adjacent components
  - Discourse relations can be triggered by structure underlying adjacency, i.e., between adjacent components unrelated by lexical elements

EXAMPLES

Lexical Elements

- Cohesion in Discourse (Halliday & Hasan)

Structure

- Rhetorical Structure Theory (Mann & Thompson)
- Linguistic Discourse Model (Polanyi et al.)
- Discourse GraphBank (Wolf & Gibson)

Lexical Elements and Structure

- Discourse Lexicalized TAG (Webber, Joshi, Stone, Knott)

Different triggers encourage different annotation schemes

FROM THEORY TO ANNOTATION

A WHIRLWIND TOUR

HALLIDAY AND HASAN (1976)

Associate discourse relations with **conjunctive elements**

- Coordinating and subordinating conjunctions
- Conjunctive adjuncts (aka *discourse adjuncts*), including
  - Adverbs such as *but*, *so*, *next*, *accordingly*, *actually*, *instead*, etc.
  - Prepositional phrases (PPs) such as *as a result*, *in addition*, etc.
  - PPs with *that* or other referential item such as *in addition to that*, *in spite of that*, *in that* case, etc.
- Each element conveys a cohesive relation between
  - its **matrix sentence** and
  - **a presupposed predication** from the surrounding discourse

HALLIDAY AND HASAN (1976)

Explicitly reject any notion of structure in discourse

> Whatever relation there is among the parts of a text – the sentences, the paragraphs, or turns in a dialogue – it is not the same as structure in the usual sense, the relation which links the parts of a sentence or a clause. [pg. 6]

> Between sentences, there are no structural relations. [pg. 27]

H&H ANNOTATION SCHEME

- Each cohesive item in a sentence is labeled with:
  - **The type of cohesion**, e.g., for conjunctive elements:
    - **C** – Conjunction
    - **C.3** – Causal conjunction
    - **C.3.1** – Conditional causal conjunction
    - **C.3.1.1** – Emphatic conditional causal conjunction (e.g., *in that case, in such an event*)
  - **The discourse element it presupposes**
  - **The distance and direction to that item**
    - **Immediate** (same or adjacent sentence): o
    - **Non-immediate**
      - Mediated (# of intervening sentences): M[n]
      - Remote Non-mediated (# of intervening sentences): N[n]
      - Cataphoric: K

EXAMPLE

(6) **Then** we moved into the country, to a lovely little village called Warley. (7) It is about three miles from Halifax. (8) There are quite a few about. (9) There is a Warley in Worcester and one in Essex. (10) **But** the one not far out of Halifax had had a maypole, and a fountain. (11) By this time the maypole has gone, but the pub is still there called the Maypole.

[from *Meeting Wilfred Pickles*, by Frank Haley]

Sentence #	Cohesive item	Type	Distance	Presupposed item
6	**Then**	C.4.1.1	N.26	<preceding text>

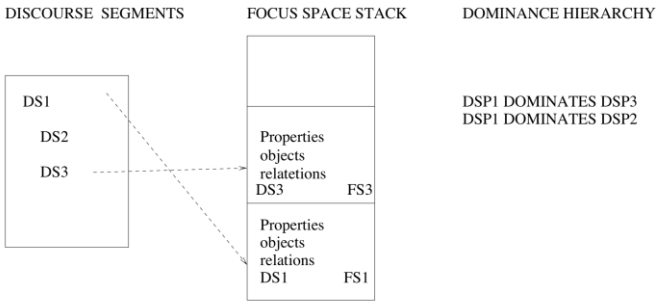
C.4 – Temporal conjunction
C.4.1 – Sequential temporal conjunction
C.4.1.1 – Simple sequential temporal conjunction (*then, next*)

Sentence #	Cohesive item	Type	Distance	Presupposed item
10	**But**	C.2.3.1	0	(S.9)

C.2 – Adversative conjunction
C.2.3 – Contrastive adversative conjunction
C.2.3.1 – Simple contrastive adversative conjunction (*but, and*)

GROSZ AND SIDNER (1986)

- Sidestep the issue of the structure of discourse imposed by semantics and define two very basic relations, DOMINANCE and SATISFACTION-PRECEDENCE
  - Carry purely intentional (that is, goal-oriented, plan-based) import
- Structure defined by a stack of **focus spaces**



- **Assumption: Two inter-clausal relations suffice to represent discourse structure**

Moore and Pollack later qualify this position, say both informational and intentional are needed

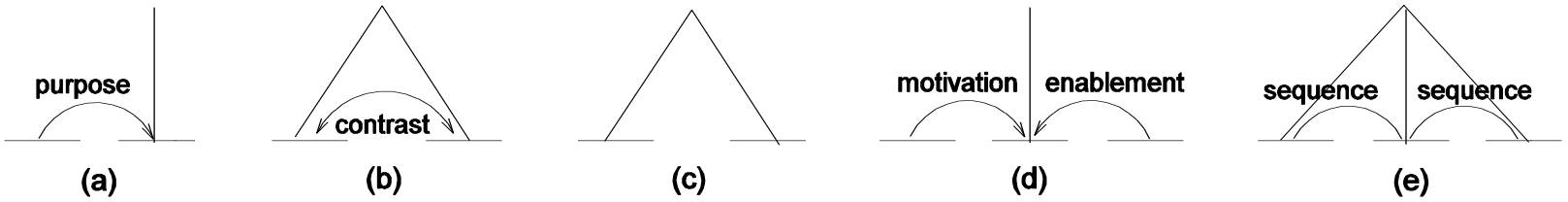
RHETORICAL STRUCTURE THEORY (RST)

- RST [Mann & Thompson, 1988] associate discourse relations **only** with **discourse structure**
  - Discourse structure reflects context-free rules called **schemas**
  - Applied to a text, schemas define a **tree** structure in which:
    - Each leaf is an **elementary discourse unit** (a **continuous text span**)
    - Each non-terminal covers a **contiguous, non-overlapping text span**
    - The root projects to a **complete, non-overlapping cover of the text**
    - Discourse relations (aka **rhetorical relations**) hold **only between children of the same non-terminal node**
    - **Clauses** should be minimal units of discourse, excluding subject and object clauses
      - mostly adverbial clauses that have a function at the discourse level
      - leave the door open for other definitions

RST SCHEMAS

RST schemas differ with respect to:

- what rhetorical relation, if any, hold between right-hand side (RHS) sisters;
- whether or not the RHS has a head (called a *nucleus*);
- whether or not the schema has binary, ternary, or arbitrary branching.



RST schema types in RST format

LINGUISTIC DISCOURSE MODEL (LDM)

- Polanyi 1988; Polanyi & van den Berg 1996; Polanyi et al. 2004
- *T*he LDM **resembles RST** in associating discourse relations only with **discourse structure**, in the form of a **tree** that projects to a complete, non-overlapping cover of the text

- The LDM **differs from RST** in distinguishing discourse structure from discourse interpretation
  - Discourse relations belong to discourse interpretation
- Discourse structure comes from three context-free rules, each with its own rule for **semantic composition (SC)**

LDM DISCOURSE STRUCTURE RULES

1. An N-ary branching rule for **discourse coordination** (lists and narratives)

   SC rule: The parent is interpreted as the information common to its children

2. A binary branching rule for **discourse subordination**, in which the **subordinate** child elaborates what is described by the **dominant** child

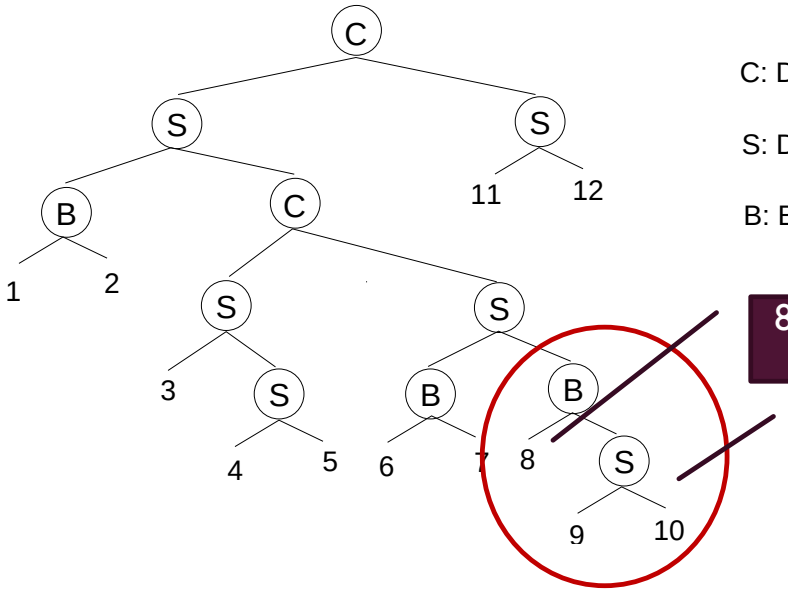
   SC rule: The parent receives the interpretation of its dominant child

3. An N-ary branching rule in which a **logical or rhetorical relation, or genre-based or interactional convention**, holds of the RHS elements

   SC rule: The parent is interpreted as the interpretation of its children and the relationship between them

EXAMPLE LDM ANNOTATION

[$_1$ Whatever advances we may have seen in knowledge management,]
[$_2$ knowledge sharing remains a major issue.] [$_3$ A key problem is] [$_4$ that documents only assume value] [$_5$ when we reflect upon their content.]
[$_6$ Ultimately,] [$_7$ the solution to this problem will probably reside in the documents themselves.] **[$_8$ In other words,] [$_9$ the real solution to the problem of knowledge sharing involves authoring,] [$_{10}$ rather than document management.]** [$_{11}$ This paper is a discussion of several new approaches to authoring and opportunities for new technologies] [$_{12}$ to support those approaches.]



C: Discourse coordination

S: Discourse subordination

B: Binary construction

8 is a rhetorical relation to the RHS

10 elaborates 9

DISCOURSE LEXICALIZED TAG (D-LTAG)

- Webber (2004)
- D-LTAG considers discourse relations triggered by **lexical elements**, focusing on
  a) the source of arguments to such relations
  b) the additional content that the relations contribute
- D-LTAG also considers discourse relations that may hold between **unmarked adjacent clauses**

MOTIVATION BEHIND D-LTAG

- D-LTAG holds that the sources of discourse meaning resemble the sources of sentence meaning, i.e,
  - **structure:** e.g., verbs, subjects and objects conveying pred-arg relations
  - **adjacency:** e.g., noun-noun modifiers conveying relations implicitly
  - **anaphora:** e.g., modifiers like *other* and *next*, conveying relations anaphorically
- D-LTAG is a **lexicalized grammar** for discourse, associating a lexical entry with the set of trees that represent its local **discourse** configurations

D-LTAG

What lexical entries head local discourse structures?

Discourse connectives:

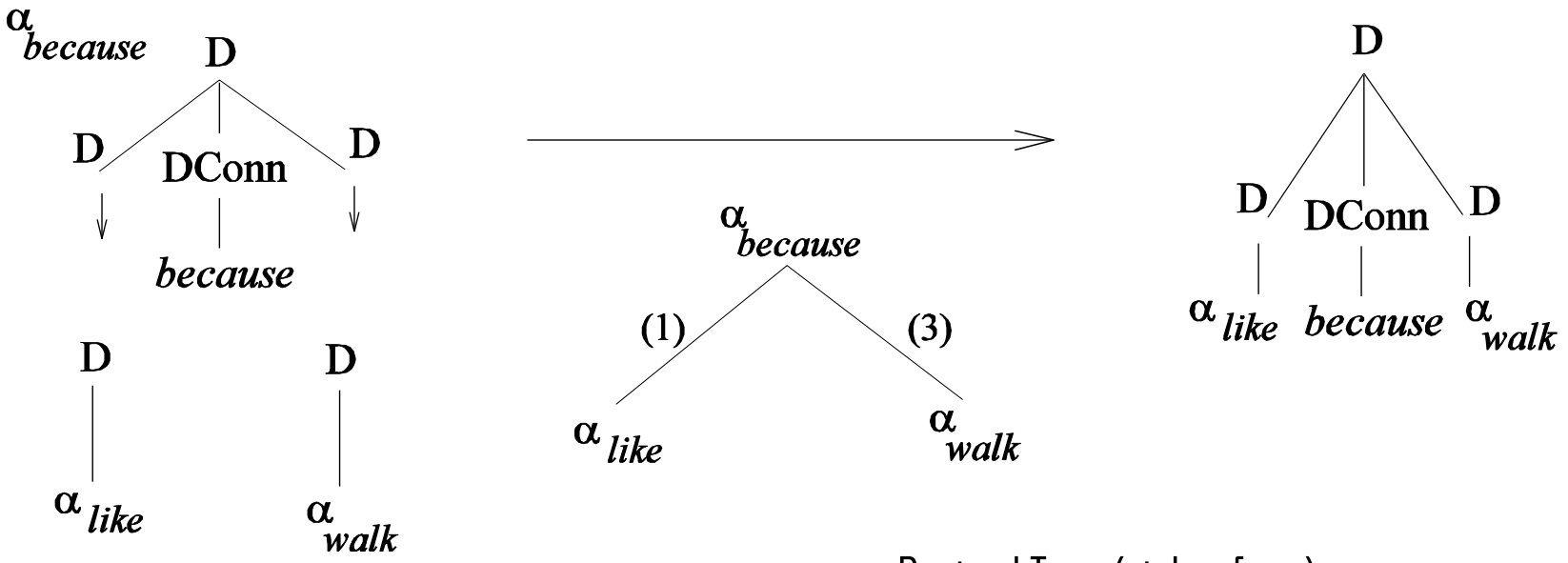
- coordinating conjunctions
- subordinating conjunctions and subordinators
- paired (parallel) constructions
- discourse adverbials

N.B. While these all have two arguments, D-LTAG does **not** take one to be **dominant** *(*ie, a *nucleus*) and the other **subordinate** (ie, a *satellite*).

EXAMPLE

Structural Arguments to Conjunctions

John likes Mary because **she walks Fido.**



Derived Tree (right of →)

Derivation Tree (below →)

GOLDEN AGE BEGINS…

RST DISCOURSE TREEBANK

- Carlson et al., 2003
- Main goal: create a **reference corpus for community-wide use**
- Two essential considerations
  - the corpus needed to be **consistently annotated**
  - must be made **publicly available**
- Two principle goals
  - grounded in **a particular theoretical approach**
  - **sufficiently large** to offer potential for wide-scale use, including
    - linguistic analysis
    - **training of statistical models of discourse**
    - other computational linguistic applications

First attempt to apply a theory of discourse to annotation on a large scale

RST DISCOURSE TREEBANK

- Use RST for three reasons:
  1. Yields **rich annotations** that uniformly **capture intentional, semantic, and textual features** that are specific to a given text
  2. Previous research (Marcu *et al.* 1999) showed that texts can be RST-annotated by multiple judges at relatively high levels of agreement ?
     - **Aimed to produce annotation protocols that would yield even higher agreement figures**
  3. Previous research showed RST trees can
     - Play a crucial role in building
       - Natural language generation systems (Hovy, 1993; Moore and Paris, 1993; Moore, 1995)
       - Text summarization systems (Marcu, 2000; Ide and Cristea 2000)
     - Be used to increase the naturalness of machine translation outputs (Marcu *et al.* 2000)
     - Be used to build essay-scoring systems that provide students with discourse-based feedback (Burstein *et al* . 2001)

RST DISCOURSE TREEBANK

- Adjacent spans are linked together via **rhetorical relations**
- Create a hierarchical structure
- **Mononuclear relations**
  - hold between **two spans** and reflect the situation in which one span, the **nucleus**, is more salient to the discourse structure, while the other span, the **satellite**, represents supporting information
- **Multinuclear relations**
  - hold among **two or more spans**, each of which has equal weight in the discourse structure
- A total of **53 mononuclear and 25 multinuclear relations** were used for the tagging of the RST Corpus
- In addition, **three relations used to impose structure on the tree**
  - textual-organization, span, same-unit (used to link parts of units separated by an embedded unit or span)
- **The final inventory of rhetorical relations is data driven, based on extensive analysis of the corpus**

RST DISCOURSE TREEBANK

- The annotated RST Discourse Treebank illustrates a tension between
  - Mann and Thompson's sole focus on discourse relations associated with structure underlying adjacency
  - Carlson et al.'s recognition *based on examination of the data* that **rhetorical relations can hold between elements other than adjacent clauses**

RST DISCOURSE TREEBANK

EDUS

- RST holds that there is a relation between clauses "whether or not they are grammatically or lexically signaled"
- **Applying this intuitive notion to the task of producing a large, consistently annotated corpus proved to be extremely difficult**
  - Boundary between discourse and syntax can be blurry
- **Goal: find a balance between granularity of tagging and ability to identify units consistently on a large scale**
  - Chose the **clause** as the elementary unit of discourse
  - Used **lexical and syntactic clues** to help determine clause boundaries

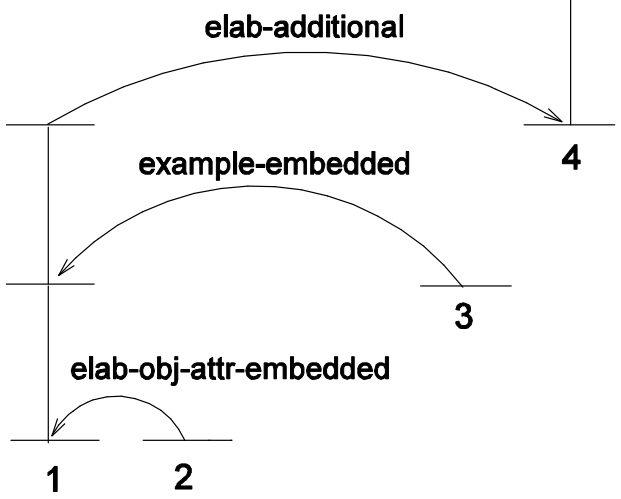
RST DISCOURSE TREEBANK

EMBEDDED CLAUSES

- Extend RST to cover appositive, complement and relative clauses, in order to capture more rhetorical relations
  - To do this, add **embedded versions of RST schemas**

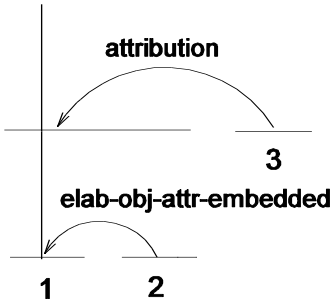
[In addition to the practical purpose[1]] [they serve,[2]] [to permit or prohibit passage for example[3]], [gates also signify a variety of other things.[4]]

Fine-grained segmentation



RST DISCOURSE TREEBANK

ADDITIONAL RELATIONS



(1) This is in part because of the effect
(2) of having the number of shares outstanding,
(3) she said.

from [Carlson et al, 2001]

- Add an **ATTRIBUTION** relation to relate a reporting clause and its complement clause, for speech act and cognitive verbs

N.B. Mann and Thompson reject **ATTRIBUTION** (aka **QUOTE**) as a rhetorical relation

"A reporting clause functions as evidence for the attributed material and thus belongs with it"

RST DISCOURSE TREEBANK

ANNOTATION PROCEDURE

Step 1: Segment the text into EDUs

Step 2: Connect pairs of units and label their status as *nucleus (N)* or *satellite (S)*

(N.B. Similar content may be expressed with different nuclearity)

He tried[N] hard, but[N] he failed.

Although[S] he tried hard, he[N] failed.

He tried[S] hard, yet[N] he failed.

Step 3: Assess which of 53 mono-nuclear and 25 multi-nuclear relations holds in each case

- Steps (2) and (3) can be interleaved, with (2) always preceding (3)
- The **result must be a singly-rooted hierarchical cover of each text**

THE DISCOURSE GRAPHBANK

- Wolf & Gibson 2005
- 135 texts from Associated Press and *Wall Street Journal* newswire data
- DG associates all discourse relations with **discourse structure**, but
  - Does not take that structure to be a tree
  - Same discourse unit can be an argument to many discourse relations
  - Admits two bases for structure:
    - Adjacent clauses can be **grouped** by common attribution or topic
    - **Any two adjacent or non-adjacent segments or groupings** can be linked by a discourse relation

  The first can yield hierarchical structure, while the second cannot

DISCOURSE GRAPHBANK

ANNOTATION PROCEDURE

Step 1: Create **EDUs** by inserting a segment boundary at every

- sentence boundary
- semicolon, colon or comma that marks a clause boundary
- quotation mark
- conjunction (coordinating, subordinating or adverbial)

[The economy,] [according to some analysts,] [is expected to improve by early next year.]

Step 2: Create **groupings** of adjacent segments that are either

- enclosed by pairs of quotation marks
- attributed to the same source
- part of the same sentence
- topically centered on the same entities or events

[[The securities-turnover tax has been long criticized by the West German financial community][because it tends to drive securities trading and other banking activities out of Frankfurt into rival financial centers,][especially London,][where trading transactions isn't taxed.]]

DISCOURSE GRAPHBANK

ANNOTATION PROCEDURE

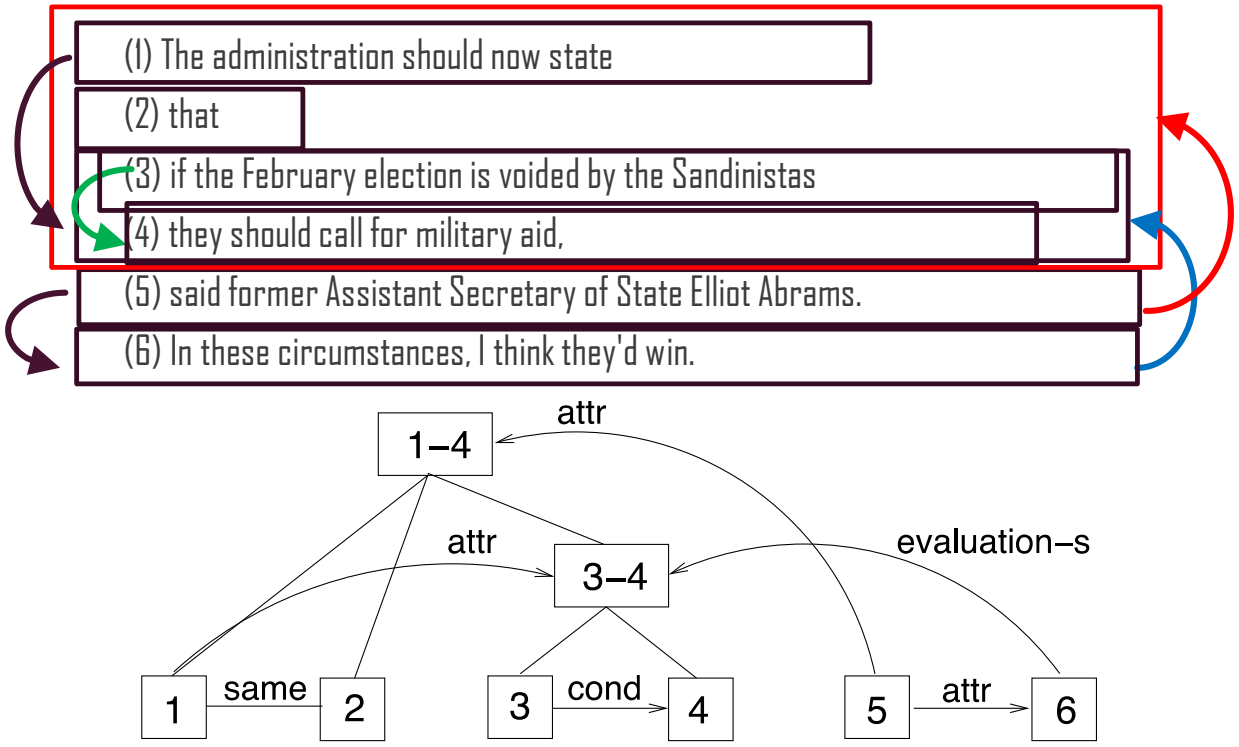
Step 3:

- Proceeding left-to-right, assess the possibility of a **discourse relation** holding between the current segment or grouping and each discourse segment or grouping to its left
- If one holds, create a new **non-terminal node labeled with the selected discourse relation**, whose children are the two selected segments or groupings

☞ **This produces a relatively flat discourse structure, in which arcs can cross and nodes can have multiple parents**

DISCOURSE GRAPHBANK

EXAMPLE ANALYSIS



While this is a much more complex structure than a tree, debate continues as to how to interpret W&G's results

ANNODIS CORPUS

- Based on Segmented Discourse Representation Theory (SDRT) (Asher, 1993, Asher and Lascarides, 2003)
  - Compute the logical form of a discourse
    - Uses compositional semantics and non-linguistic information such as real world knowledge as clues
  - Supports default reasoning
- Grew out of an earlier attempt DISCOR (Baldridge et al., 2007)
- Modified SDRT to **accommodate annotation task** as well as expand the theory

ANNODIS

- Investigated top-down and bottom-up approaches:
  - Top-down: start by finding the representation of a text's macro-organization, focus on "multi-level" text spans and signals of global text organization
  - Bottom-up: define hierarchical structures by constructing complex discourse units (CDUs) from elementary discourse units (EDUs), i.e., "bottom-up", in recursive fashion
  - Can give equivalent results, but typically emphasize different parts of discourse structure
- Developed two annotation models with some common characteristics in order to bring the two closer and permit annotation comparison

CORPUS CONTENTS

- Wanted a diversified corpus, with a variety of genre, length and type of discursive organization
  - Other major corpora include mainly newswire (*Wall Street Journal*)

- ANNODIS divided in two parts
  - Bottom-up approach : short texts (a few hundred words each)
  - Top-down approach : longer (several thousands words each), complete and more complex documents

BOTTOM-UP APPROACH

- Focused on providing a complete structure of a text, starting from the segmentation into EDUs
  - mostly clauses, appositions, some adverbials
- Modified SDRT to accommodate the annotation task
  - **Merged certain relations of earlier-developed DISCOR/SDRT relation set that proved difficult for experts to detect reliably**
  - Introduced new relations
    - Entity-elaboration, to account for appositions
    - Also used a "Frame" relation, which relates a framing adverbial and EDUs within its scope
    - Remaining relations are more or less common to all the theories of discourse or correspond to well-defined subgroups in fine-grained theories
- **Intermediate level of granularity was chosen as a compromise between informativeness and reliability of the annotation process**
  - Corresponds to the level chosen in the PDTB and a coarse-grained RST

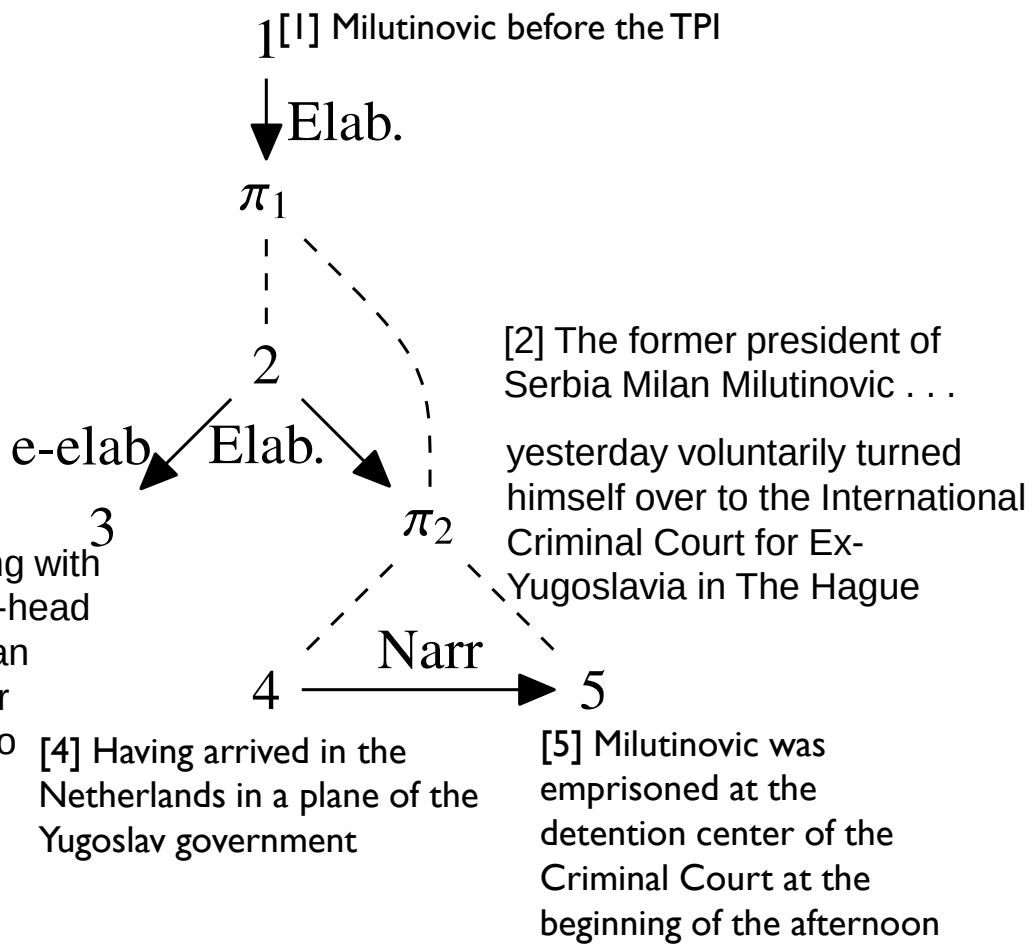
TOP-DOWN APPROACH

- Concerned with strategies regarding textual continuity and discontinuity
- **To translate this into a realistic annotation program**, devised an annotation model focusing on the detection of two discourse structures highlighting the continuity/discontinuity dichotomy:
  - Topical chains
    - segments made up of sentences containing topical co-referential expressions
  - Enumerative structures
    - segments (in effect CDUs) consisting of three sub-segments:
      - (optional) trigger announcing the enumeration
      - items composing the enumeration
      - (optional) closure that summarizes/closes the enumeration

EXAMPLE OF DISCOURSE GRAPH

- Nodes correspond to discourse units
- EDUs represented by their numbering
- CDUs start with π
- Dotted edges represent inclusion in a CDU
- Edges with arrows represent rhetorical relations
  - Elab. = Elaboration
  - e-elab = Entity Elaboration
  - Narr. = Narration

[Milutinovic before the TPI.]_1[The former president of Serbia Milan Milutinovic, [accused along with the Yugoslav ex-head of State Slobodan Milosevic for war crimes in Kosovo,]_3 yesterday voluntarily turned himself over to the International Criminal Court for Ex-Yugoslavia in The Hague]_2 [Having arrived in the Netherlands in a plane of the Yugoslav government,]_4 Milutinovic was emprisoned at the detention center of the Criminal Court at the beginning of the afternoon]_5



EDUS

- SDRT originally mute on the subject of EDU segmentation
  - In general, followed common practice of segmenting into sentences and/or tensed clauses
- Examination of the semantic behavior of appositives, non-restrictive relative clauses and other parenthetical material showed that such syntactic structures also contribute EDUs
  - provide semantic contents that do not fall within the scope of discourse relations or operators between the constituents in which they occur
- Developed **guidelines for the segmentation of text into EDUs**
  - had not been done before
  - allow discourse segments to be embedded in one another

PENN DISCOURSE TREEBANK (PDTB)

- Prasad et al., 2008
- Provides annotations of discourse relations, their arguments, senses, and attributions
- Corpus is the PTB-II portion of the *Wall Street Journal* corpus
  - ~1 million words

PDTB

- Work on discourse relations prior to PDTB focused on **discourse graphs** and **discourse trees** that describe discourse structure over an entire text by linking individual relations
- Annotating dependencies across relations presumes an understanding of the nature of representation for high-level discourse structure
- Currently little agreement on a theory
- **PDTB has taken an approach that avoids biasing the annotation towards one or the other theory**
- Chose to specify discourse relations at a low-level that is clearly defined and well-understood
  - Each discourse relation annotated independently of other relations--dependencies across relations are not marked

KEY IDEAS OF PDTB

- Discourse relations described at the **informational** (vs. intentional) level of meaning
- Discourse relations with explicit cues in the text annotated by marking the **lexical items** that express them
- When cues are implicit, annotators **insert a connective** that best expresses the inferred relation, which can then itself be annotated
  - Lexical grounding of the relations intended to **boost annotator confidence** in reasoning about the relations and **increase annotation reliability**
- **PDTB Annotation scheme developed in an iterative manner, based on feedback from annotators and lessons from earlier annotation experiments**

PDTB

- Takes a **theory-neutral** approach to annotating discourse relations
  - No commitments made about the nature of high-level discourse structure representation
  - **No dependencies between different relations marked** after annotating individual relations and their arguments
- Goals
  - Allow the corpus to be useful for researchers working within different frameworks
  - **Provide a resource for research towards a "data-driven, emergent theory of discourse structure"**
    - To address different proposals about the representational nature of discourse structure
      - Trees (Mann and Thompson, 1988; Polanyi, 1987)
      - Graphs (Wolf and Gibson, 2005)
      - DAGs (Asher and Lascarides, 2003; Webber et al, 2003; Lee et al, 2008)

RELATION TYPES

- **PDTB annotates both explicit and implicit relations**
- Two types of explicit relations
  1. Signaled by explicit connectives
     - Include subordinating conjunctions (e.g., *because*, *when*, *since*, *although*), coordinating conjunctions (e.g., *and*, *or*, *nor*), or adverbs and prepositional phrases (e.g., *however*, *otherwise*, *then*, *as a result*, *for example*)
  2. Signaled by "alternative lexicalizations" (AltLex)
     - belong to syntactic classes other than those admitted for connectives
     - only annotated between adjacent sentences to conform to practice for implicit connectives

RELATION TYPES

- Cases where annotators cannot not supply an implicit connective annotated as one of the following :
  - AltLex
  - EntRel
    - Cases where only an *entity-based coherence* relation can be perceived between the sentences
      - Ex: **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
    - Cases where no discourse relation or entity-based relation can be perceived between the sentences

EXAMPLE

Discourse **relations** (e.g., causal, contrastive, temporal) triggered by explicit words or phrases (**underlined**) or by adjacency

Arguments are two abstract objects (AO) such as events, states, and propositions, labeled Arg1 (***italics***) and Arg2 (**bold**).

Sense tags provided for explicit, AltLex, and implicit relations (**in parentheses**)

1. *Big buyers like P&G say there are other spots on the globe, and in India, where the seed could be grown. . . .* But **no one as made a serious effort to transplant the crop**. (Comparison:Concession:Contra-expectation)

2. *Some have raised their cash positions to record levels.* Implicit=because **High cash positions help buffer a fund when the market falls**. (Contingency:Cause:Reason)

3. *But a strong level of investor withdrawal is much more unlikely this time around,* [fund managers said] . **A major reason is that investors already have sharply scaled back their purchases of stock funds since Black Monday**. (Contingency:Cause:Reason)

Explicit realizations can occur via grammatically defined **connectives** or grammatically non-conjunctive expressions called **Alternative lexicalizations** (AltLex)

For adjacent sentences not related by an explicit connective or AltLex, an **implicit discourse relation** can be inferred. Annotator has to **insert a connective** to express the inferred relation

Discourse **relations** (e.g., causal, contrastive, temporal) triggered by explicit words or phrases (**underlined**) or by adjacency

Arguments are two abstract objects (AO) such as events, states, and propositions, labeled Arg1 (***italics***) and Arg2 (**bold**).

Sense tags provided for explicit, AltLex, and implicit relations (**in parentheses**)

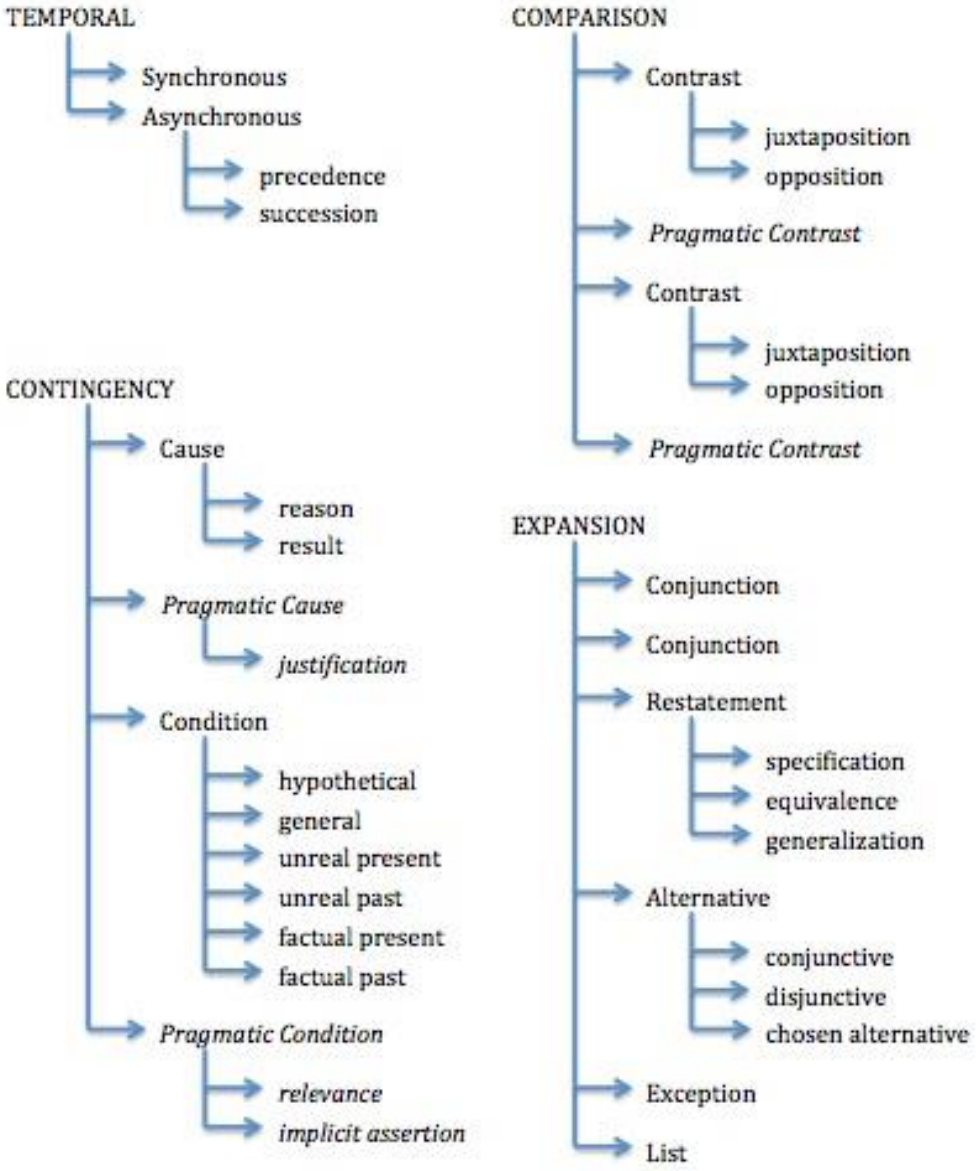
Adjacent sentences might not be related by a discourse relation when the sentences are linked by an entity-based coherence relation (**EntRel**) or not related at all via adjacency (**NoRel**)

4. *Pierre Vinken, . . ., will join the board as a nonexecutive director Nov. 29.* EntRel **Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group**.

5. *Jacobs is an international engineering and construction concern.* NoRel **Total capital investment at the site could be as much as $400 million,** . . ..

SENSE TAGS

- **Sense tags** in the PDTB are provided for the explicit, implicit and AltLex relations
  - Discourse connectives can have more than one meaning
    - E.g., *since* has three different senses, one purely 'Temporal', another purely 'Causal', and a third both 'Causal' and 'Temporal'
- Hierarchical organization of sense tags
  - **Intended to address issues of inter-annotator reliability**
    - Allows annotators to select a tag from a level that is comfortable to them


TEMPORAL
Synchronous
Asynchronous
precedence
succession
CONTINGENCY
Cause
reason
result
Pragmatic Cause
justification
Condition
hypothetical
general
unreal present
unreal past
factual present
factual past
Pragmatic Condition
relevance
implicit assertion
COMPARISON
Contrast
juxtaposition
opposition
Pragmatic Contrast
Contrast
juxtaposition
opposition
Pragmatic Contrast
EXPANSION
Conjunction
Conjunction
Restatement
specification
equivalence
generalization
Alternative
conjunctive
disjunctive
chosen alternative
Exception
List

EDUS

- Arguments of discourse relations **not constrained to be single clauses**
  - Can include multiple clauses or multiple sentences
- Non-clausal arguments allowed when clearly associated with an eventive reading
  - E.g., *nominalizations*, *discourse deictics* (e.g., *this*, *that*, *so*) that refer to abstract objects, verb phrases that appear to be analyzable as clausal coordination with subject ellipsis, and particles that function as reponses to questions, such as *yes*, *no*.
- **Minimality principle**
  - An argument must contain the minimal amount of information needed to complete the interpretation of the relation
  - Any other span of text perceived to be relevant (but not necessary) to the interpretation of arguments is optionally annotated as **supplementary information**
- **Arguments of explicit connectives can be located anywhere in the text**

ATTRIBUTION ANNOTATION

- Attribution not considered a discourse relation in PDTB
- But they are annotated for discourse relations and their arguments because of highly frequent use in the *Wall Street Journal* texts that constitute the corpus

CONTINUING STUDY

PREDICTING ARGUMENTS OF DISCOURSE CONNECTIVES

- Prior to the PDTB, discourse parsing focused on building a single tree structure that covers a text
  - proved to be extremely difficult
- Low-level annotation of discourse relations in the PDTB has stimulated research on the somewhat easier task of **discourse chunking** (Webber et al, 2012)
  - Still has benefits for applications

INVESTIGATING DISCOURSE RELATION LEXICALIZATION

- (Prasad et al, 2010b) show
  - discourse relations can be signaled by a wider variety of syntactic types than previously assumed
  - the set of **discourse relation markers** is open-ended
- The task of identifying discourse relations is much more challenging for discourse parsing research than previously believed

DAS AND TABOADA STUDY (2013)

Relation	Agreement	Disagreement
Antithesis	3	-
Attribution	19	1
Background	1	3
Cause-result	-	1
Circumstance	1	1
Condition	2	-
Contrast	3	-
Elaboration	47	17
Example	-	2
Explanation	-	4
Hypothetical	1	-
List	5	-
Manner	-	2
Problem-solution	2	1
Purpose	5	-
Same-unit	6	-
Summary	-	1
Temporal	2	-
Total	97	33

Table 3. Agreement and disagreement per relation

Additional annotation of RST Discourse Bank

Added layer of "signal" types

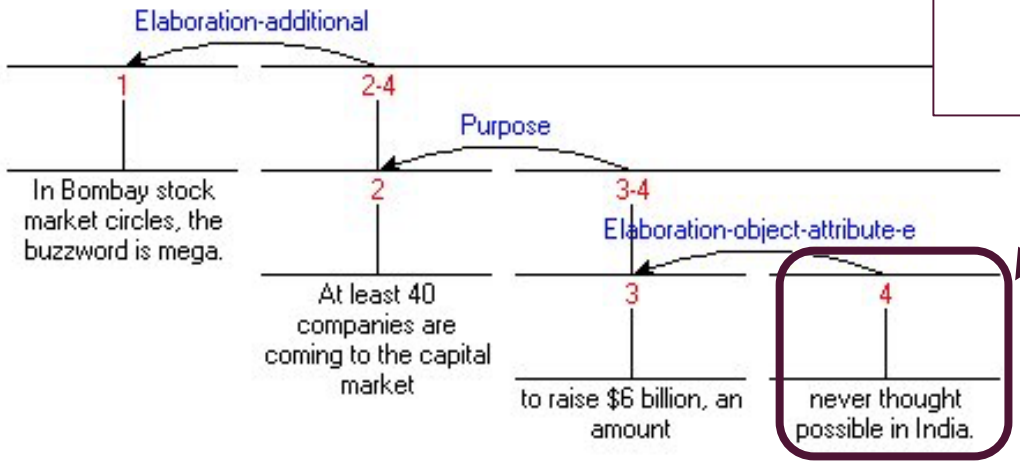
Kappa value 0.68 for annotations (moderate agreement)

THEORETICALLY-BASED PROBLEMS CONTINUE…

Das and Taboada study had problems due to

- Disagreements concerning relations
  - RST Discourse Treebank uses a very large set of 78 relations, including a high number of subtypes of Elaboration
  - Annotators had to keep all these distinctions in mind as they annotated

- Disagreements with EDU segmentation
  - disagree with the notion that noun and relative clauses stand in any kind of discourse relation to the words that they modify
    - should unit 4 be considered a span, and instead included as a unit with the noun that it modifies (amount)



SUMMARY

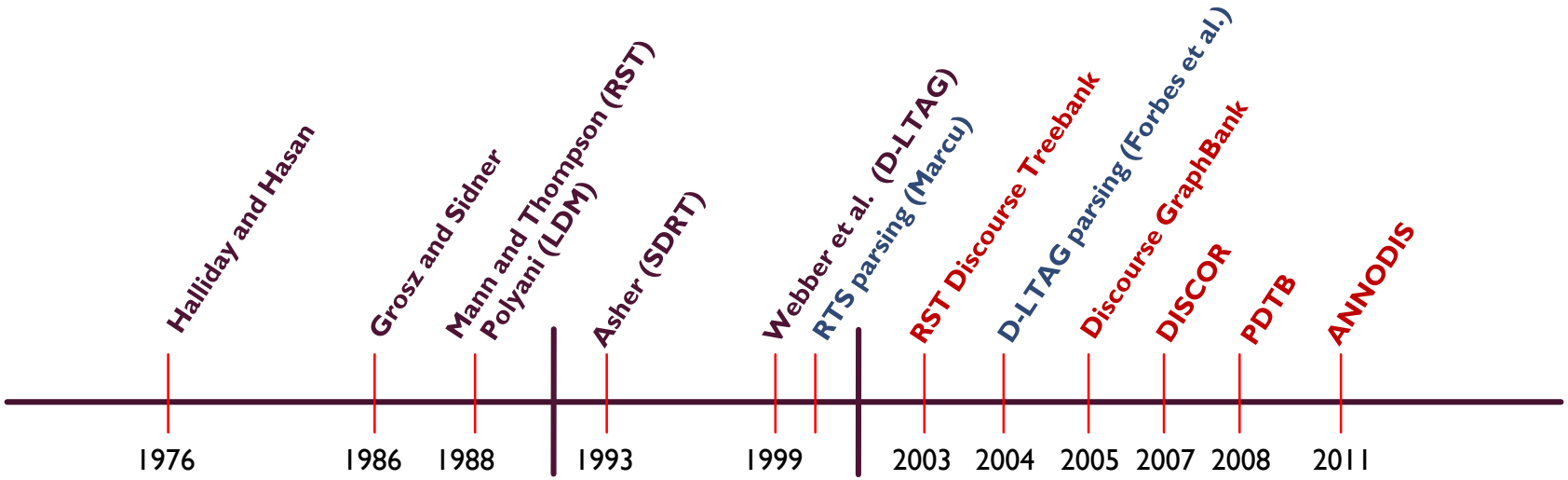
- Discourse annotation is highly subjective
  - No clear answer to many questions
  - No obvious universally acceptable theory

SOME THOUGHTS…

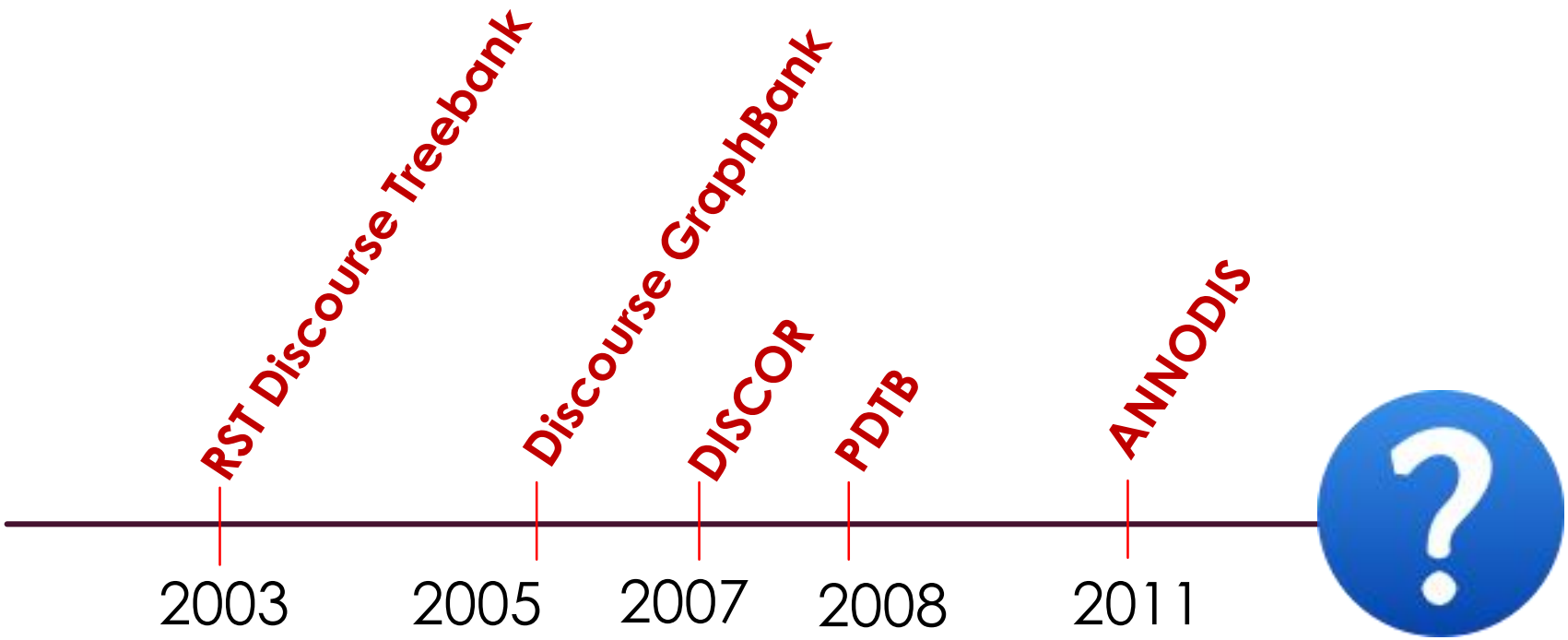
- …after listening to talks in this conference
- Computational linguists do not care/think (much) about some of the concerns outlined by Lou Burnard this morning
  - E.g., interpretive implications of "markup"
- Concerns for CL/NLP are practical, functional
  - With discourse annotation, progressed from primarily theoretical (humanistic?) analyses to increasing concern for
    - What can be identified reliably by annotators
    - What works for machine learning
    - What helps my application
- At the same time, discourse annotation in CL/NLP is still defined by concerns born of the subjectivity that informs analysis in many humanities disciplines
  - So far the answer to this situation seems to be "let the data drive the theory"

Theories
Discourse parsing
Corpus building

The Big Picture



Golden age of discourse annotation



Adapt existing theories → Theory-neutral

Data-driven approach to determining relations, etc. →

What next?

THANK YOU

Acknowledgements

Some material based on the following:

Joshi, A., Prasad, R., Webber, B. (2006) Discourse Annotation: Discourse Connectives and Discourse Relations. COLING/ACL 2006 Tutorial (ppt).

Asher, N. et al. (2017) ANNODIS and Related Projects: Case Studies on the Annotation of Discourse Structure. In Ide, N., Pustejovsky, J. (eds.) *Handbook of Linguistic Annotation.* Springer, pp. 1241-64.

Prasad, R., Webber, B., Joshi, A. (2017) The Penn Discourse Treebank: An Annotated Corpus of Discourse Relations. In Ide, N., Pustejovsky, J. (eds.) *Handbook of Linguistic Annotation.* Springer, pp. 1197-1217.