Interoperable corpora: Why would we want it and how can we achieve it?

Vera Demberg & Merel Scholman
Universität des Saarlandes, Germany

TextLink Training School – Charles University

Introduction

- Discourse relations are semantic links between segments / arguments, e.g.:

  Hamsters turn into cannibals when **they are put on a diet.**

- Many discourse relations can be described in terms of logic
- In logic and semantics, P and Q are used to refer to statements

P & Q

Here's a short intro to how P and Q can work:

- **P & Q** = The situation described in P holds and the situation described in Q holds (additive/temporal)
  I visited the Prague Castle.$_{(P)}$ **I** <u>also</u> **went to the Charles Bridge.**$_{\mathbf{(Q)}}$
- **P** $\rightarrow$ **Q** = The situation in P leads to the situation in Q (causal/conditional)
  I am in Prague,$_{(P)}$ <u>so</u> **I tried Kulajda.**$_{\mathbf{(Q)}}$
- **P**$<$**X**[1] & **Q** $\rightarrow \neg$**X (**$\neg$ **X can be the same as Q)** = The situation described in P causes the expectation of X but it leads to the unexpected situation described in Q. (concession)
  <u>Although</u> **the cheese was rather strong,**$_{\mathbf{(P)}}$ *I liked it.*$_{(Q)}$

[1] A$<$B means A causes B

Introduction

- Discourse relation frameworks aim to describe these links between P and Q using labels
- These frameworks are then used to annotate different corpora
- Examples are the Penn Discourse Treebank, Rhetorical Structure Theory, GraphBank
- Each framework makes different distinctions regarding to which relations can hold between P and Q, e.g.:

Introduction

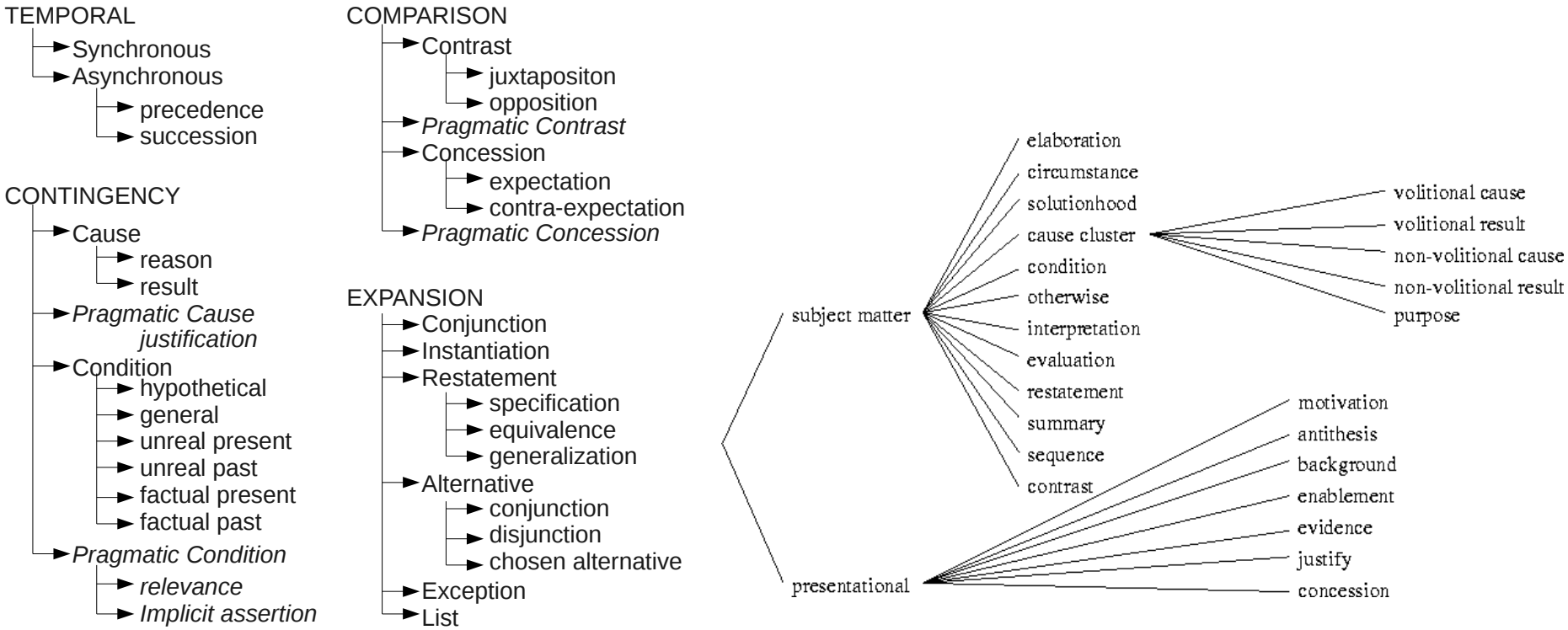


Table 1
Contentful conjunctions used to illustrate coherence relations.

Cause–effect	because; and so
Violated expectation	although; but; while
Condition	if . . . (then); as long as; while
Similarity	and; (and) similarly
Contrast	by contrast; but
Temporal sequence	(and) then; first, second, . . . ; before; after; while
Attribution	according to . . . ; . . . said; claim that . . . ; maintain that . . . ; stated that . . .
Example	for example; for instance
Elaboration	also; furthermore; in addition; note (furthermore) that; (for, in, on, against, with, . . .) which; who; (for, in, on, against, with, . . .) whom
Generalization	in general

Introduction

- It would be great if one could make use of all these corpora to investigate a specific research question
- However, the different distinctions made by frameworks makes comparison difficult
- In other words, the corpora are not interoperable

- Today, we will present a proposal to "translate" relation labels from one framework to another, so that researchers can make use of different corpora.
- Let's first look at two of the most well-known discourse annotated corpora to see why interoperability is an issue

Outline

1. Two large corpora and their frameworks
   - PDTB
   - RST

2. Use cases – What can we do with interoperable corpora?

Existing resources

- Different frameworks are based on different sets of relations, e.g.,
  - Grosz & Sidner (1986): 2 relations
  - PDTB (2008): 43 relations
  - RST-DT (2003): 78 relations
- Frameworks can also be different when adapted to different languages or modalities, e.g.,
  - RST Basque Treebank: different version of RST compared to RST-DT, includes other labels such as PREPARATION.
  - Prague Dependency Treebank: PDTB-style, but several changes have been made, e.g. the different conditional subtypes in PDTB have been merged into one type
  - Italian LUNA corpus: PDTB-style, but several labels have been introduced for spoken discourse, such as GOAL and speech-act labels

Existing resources – PDTB and RST-DT

- This part of the lecture: focus on two of the largest English discourse-annotated corpora – PDTB & RST
- Mapping discussed the rest of the day is illustrated using these two frameworks
- So first, we briefly discuss both frameworks to make sure everybody is on the same page

PDTB – The framework

- Penn Discourse Treebank (2008)
- Focus on low-level relations (within/between adjacent sentences), not on relations between relations
- Strong focus on discourse connectives
- Relations have two (and only two) arguments: Arg1 and Arg2
- Placement Arg2 depends on position of connective: '*Arg1* because **Arg2**', or 'Because **Arg2**, *Arg1*'

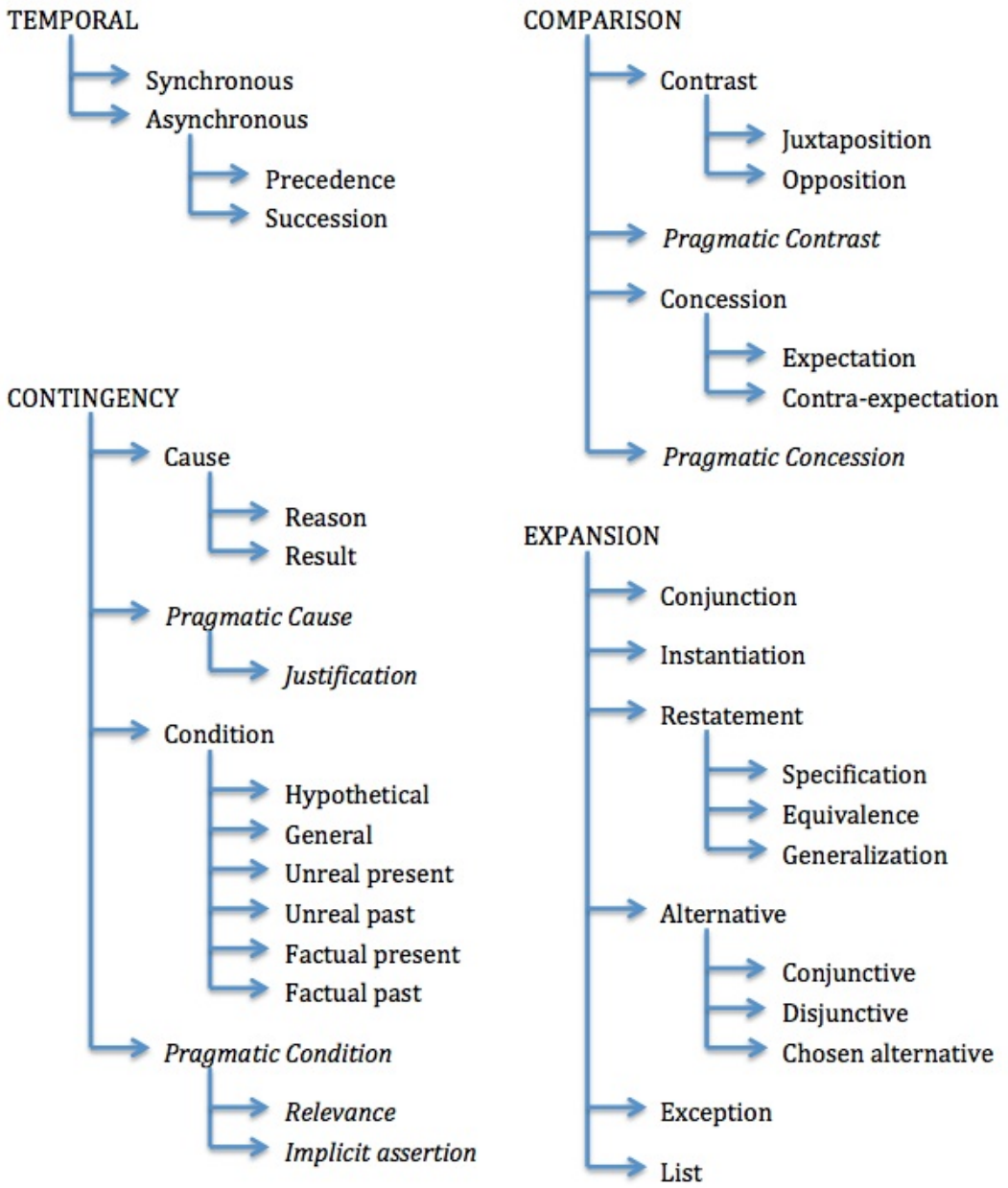
PDTB – Relation labels

- Hierarchical set of relation labels
- Three levels:

1. **Class** level: 4 major semantic classes
2. **Type** level: further refines the semantics of the class levels
3. **Subtype** level: defines semantic contribution of each argument

- When an annotator is uncertain of fine-grained sense (subtype), s/he can choose higher level (type) $\rightarrow$ good for inter-annotator agreement

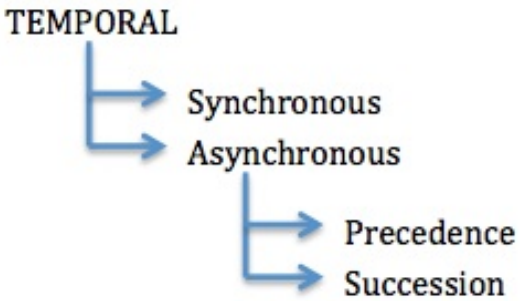
PDTB – Hierarchy



PDTB – Temporal

Temporal:
Arguments are temporally related (overlapping or ordered)

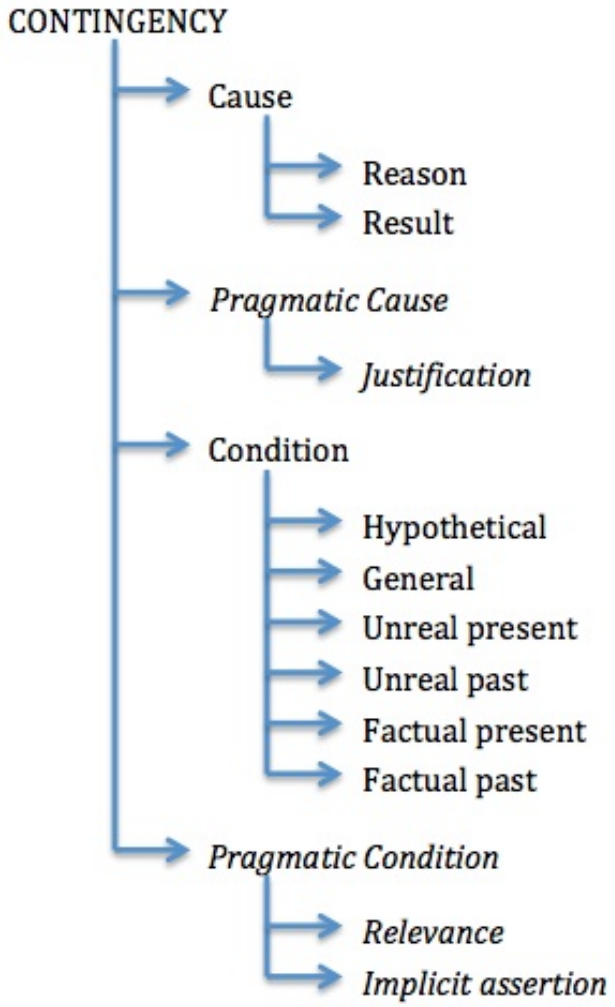
- *John was singing* while **he was washing his apple.**
  Synchronous
- *John washed his apple* and then **he ate it.**
  Asynchronous.Precedence
- *John ate his apple* after **he washed it.**
  Asynchronous.Succession



PDTB – Contingency

Contingency:
Event in one of the segments causally influences the other

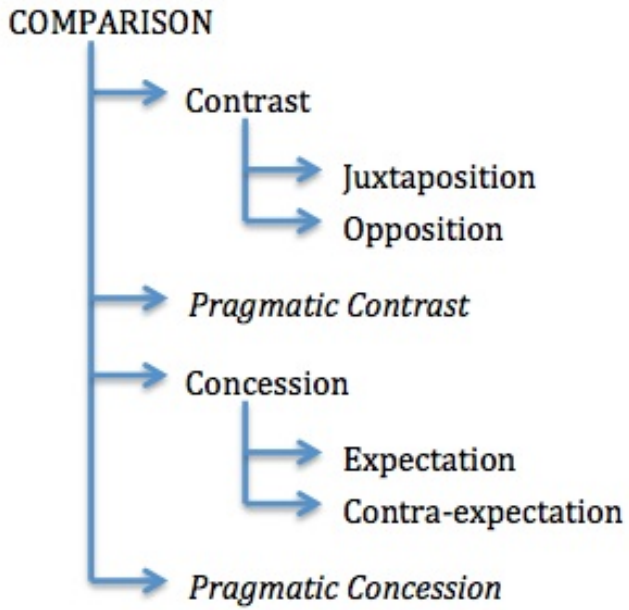
- *John was singing* so **his roommates left.**
  Cause.Result
- *John was singing* because **he wanted his roommates to leave.** Cause.Reason
- *John is manipulative* because **he sings in order to drive people away.**
  Pragmatic cause
- If **John likes singing,** *he should take lessons.*
  Condition



PDTB – Comparison

Comparison:
Discourse relation that highlights differences between the situations

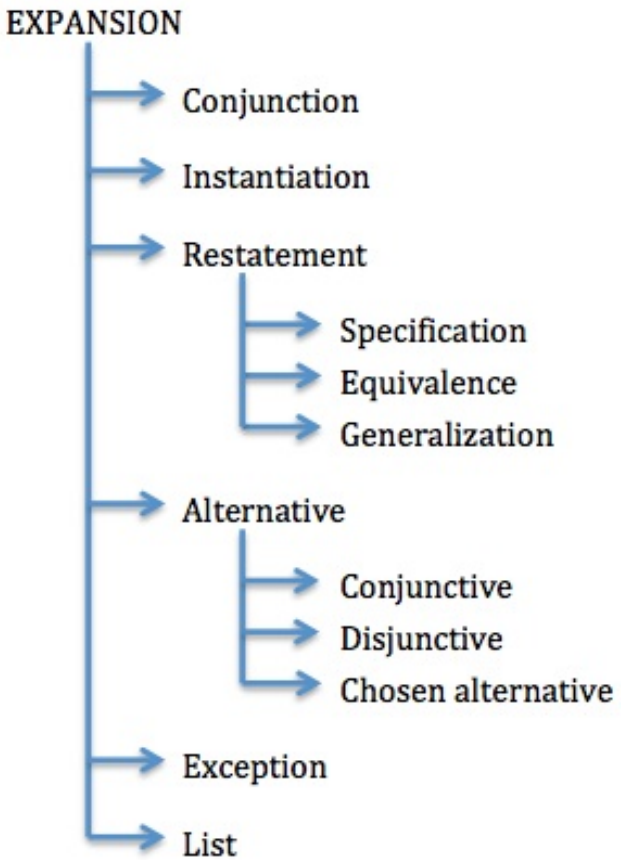
- *John likes apples* but **Mary likes pears.**
  Contrast
- Although **John likes fruit,** *he doesn't like pears.*
  Concession.Expectation
- *John likes fruit,* but **he doesn't like pears.**
  Concession.Contra-expectation



PDTB – Expansion

Expansion:
Events that "expand the discourse" (not temporal, causal, contrastive)

- *John likes apples* and **Mary does too.**
  Conjunction
- *John likes fruits.* For example, **he enjoys eating apples.**
  Instantiation
- *John likes fruits.* More specifically, **he likes apples.**
  Restatement.Specification
- *John doesn't eat vegetables.* Instead, **he eats a lot of fruit.**
  Alternative.Chosen alternative
- *John doesn't eat vegetables,* except **for when he's sick.**
  Exception
  PDTB manual:
  ¬ Arg1&Arg2 &¬Arg2→Arg1



PDTB – Exercise: annotate some relations

Use the subset of PDTB relations on the "mini manual" handout for this exercise. Write down the PDTB labels at the appropriate spot on the items handout.

1. *The student sometimes placed his jeans in the freezer overnight* because **ice-cold temperatures prevent dirty smells.**
2. *The beer was brewed with a chocolate extract.* **It** also **contains peppermint.**
3. *Experts say such long hours for flight attendants are dangerous.* For instance, **tired attendants might not react quickly enough during an emergency.**
4. *My mom ate bags of M&Ms while she was pregnant with me* so **chocolate is in my blood.**
5. Rather than keep the loss a secret from the outside world, *Michelle blabs about it to a sandwich man* while **ordering lunch over the phone.**
6. *They've been assured that the police doesn't have anything to do with the population count.* Still, **a lot of people are afraid of counteractions.**

PDTB – Corpora

Original corpus:

- English: Penn Discourse Treebank – Newspaper text, million words

Related corpora include:

- Chinese Discourse Treebank – Newspaper text, 70K words
- Czech: Prague Discourse Treebank – Newspaper text, 50K sentences
- English: Biomedical Discourse Relation Bank – Biomedical articles, 112K words
- Eng, Tur, Deu, Por, Pol, Rus: TED-MDB – TED talks, 6 texts
- Hindi Discourse Relation Bank – Newspaper text, 400K words
- Italian: Luna Corpus – Spoken dialog, 25K words
- Modern Standard Arabic: Leeds Arabic DTB – Newspaper text, 166K words
- Turkish: METU-TDB Corpus – Several written genres, 500K words

1. Two large corpora and their frameworks
   - PDTB
   - RST

2. Use cases – What can we do with interoperable corpora?

RST – The framework

- Rhetorical Structure Theory
- Original proposal: Mann and Thompson (1988)
- Developed for computer-based text generation

- Relations are formulated in terms of writer's intentions
- No strong focus on connectives like in PDTB

- Different versions available
- Version discussed here is developed by Carlson and Marcu (2003)

RST – Relation labels (C&M 2003)

- **Attribution**: attribution, attribution-negative
- **Background**: background, circumstance
- **Cause**: cause, result, consequence
- **Comparison**: comparison, preference, analogy, proportion
- **Condition**: condition, hypothetical, contingency, otherwise
- **Contrast**: contrast, concession, antithesis
- **Elaboration**: elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
- **Enablement**: purpose, enablement
- **Evaluation**: evaluation, interpretation, conclusion, comment
- **Explanation**: evidence, explanation-argumentative, reason
- **Joint**: list, disjunction
- **Manner-Means**: manner, means
- **Topic-Comment**: problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
- **Summary**: summary, restatement
- **Temporal**: temporal-before, temporal-after, temporal-same-time, sequence, inverted-sequence
- **Topic Change**: topic-shift, topic-drift

RST – Subset of labels: Temporal

Temporal labels in RST include the following:

- *John was singing* while **he was washing his apple.**

  TEMP.-SAME-TIME

- *John ate his apple* after **he washed it.**

  TEMP.-AFTER

- *John washed his apple* and then **he ate it.**

  TEMP.-BEFORE

- *John washed his apple.* **He recently started washing his apples before eating them.**

  BACKGROUND

RST – Subset of labels: Causal

Causal labels in RST include the following:

- *John was singing* so **his roommates left.** CAUSE
- *John's roommates left* when **he started singing.** RESULT
- *John and his roommates do not get along.* **They never spend time together.** EVIDENCE
- *John was singing* in order to **drive his roommates away.** PURPOSE

RST – Subset of labels: Contrastive

Constrastive labels in RST include the following:

- *John likes apples* but **Mary likes pears.** CONTRAST
- Although **John likes fruit,** *he doesn't like pears.* CONCESSION
- Although **he doesn't eat many pears,** *John enjoys eating apples.* ANTITHESIS

RST – Subset of labels: Additive

Additive labels in RST include the following:

- *John likes apples* and **John likes pears too.** ELAB.-ADDITIONAL
- *John likes fruits.* More specifically, **he likes apples.** ELAB.-GENERAL-SPECIFIC
- *John likes fruits.* For example, **he enjoys eating apples.** EXAMPLE

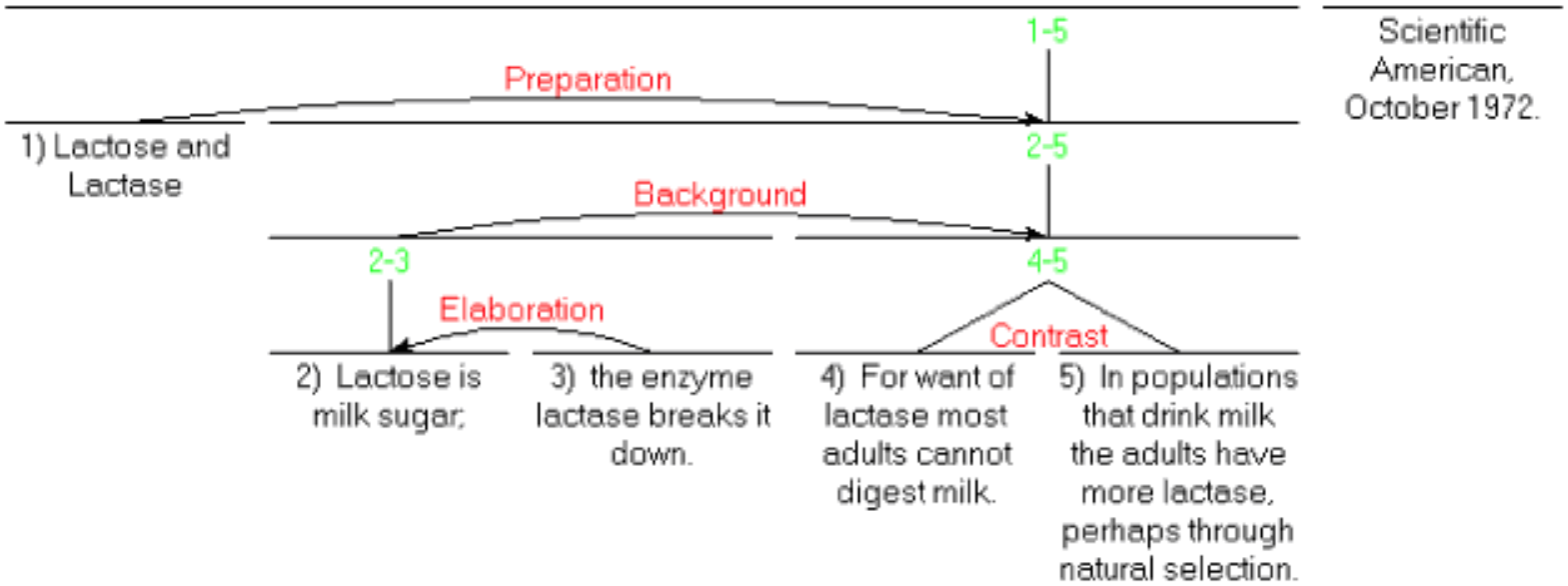
RST – Annotating a tree structure

RST creates tree structures of texts

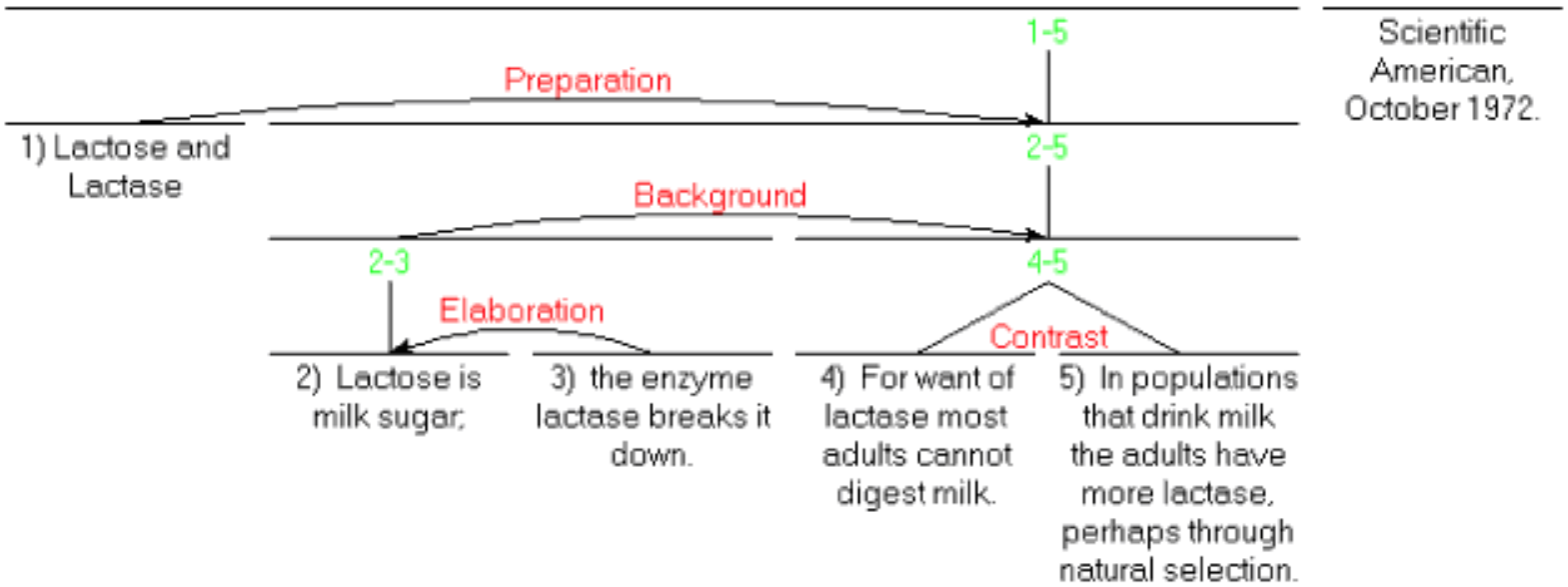
Procedure:

1. Divide the text into units
2. Examine each unit, and its neighbours. Is there a clear relation holding between them?
   - If yes, then mark that relation (e.g., Condition).
   - If not, the unit might be at the boundary of a higher-level relation. Look at relations holding between larger units (spans).
3. Continue until all the units in the text are accounted for.

RST – Tree structure



RST – Tree structure



Arrows point to the central part of the relation: the nucleus

RST – Nuclearity

- Arguments of RST relations are either nucleus or satellite
- Nucleus is central part of text, satellite is supportive of nucleus
  For example: Evidence relation (claim – argument):
  - Claim is more essential to the text than evidence
  - So claim is nucleus and evidence is satellite

RST – Nuclearity

- Arguments of RST relations are either nucleus or satellite
- Nucleus is central part of text, satellite is supportive of nucleus
  For example: Evidence relation (claim – argument):
  - Claim is more essential to the text than evidence
  - So claim is nucleus and evidence is satellite

- Writer's intentions are important: what does the writer want to achieve?
- Determining nuclearity can therefore rarely be done without taking the context of the relation into consideration

RST – Nuclearity

- Arguments of RST relations are either nucleus or satellite
- Nucleus is central part of text, satellite is supportive of nucleus
  For example: Evidence relation (claim – argument):
  - Claim is more essential to the text than evidence
  - So claim is nucleus and evidence is satellite

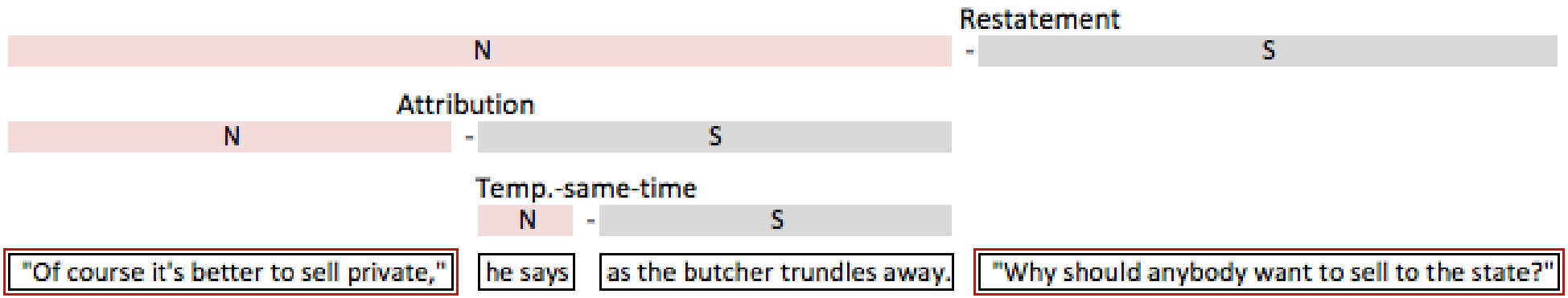
- Writer's intentions are important: what does the writer want to achieve?
- Determining nuclearity can therefore rarely be done without taking the context of the relation into consideration
- Connectives can change the nuclearity of very similar relations:
  1. *The earnings were fine and above expectations.*$_{\text{N}}$ Nevertheless, **Salomon's stock fell $1.125 yesterday.**$_{\text{S}}$
  2. Although **the earnings were fine and above expectations,**$_{\text{S}}$ *Salomon's stock fell $1.125 yesterday.*$_{\text{N}}$

RST – Nuclearity and trees

- Strong Nuclearity Principle:
  When a relation holds between two spans of text (higher up in the tree), it should also hold between the nuclei of these spans.

RST – Nuclearity and trees

- Strong Nuclearity Principle:
  When a relation holds between two spans of text (higher up in the tree), it should also hold between the nuclei of these spans.



→ RESTATEMENT actually holds between the nucleus of the nucleus and the satellite of RESTATEMENT

RST – Exercise: annotate some examples

Use the subset of RST relations on the handout for this exercise.

1. *The student sometimes placed his jeans in the freezer overnight* because **ice-cold temperatures prevent dirty smells.**
2. *The beer was brewed with a chocolate extract.* **It** also **contains peppermint.**
3. *Experts say such long hours for flight attendants are dangerous.* For instance, **tired attendants might not react quickly enough during an emergency.**
4. *My mom ate bags of M&Ms while she was pregnant with me* so **chocolate is in my blood.**
5. Rather than keep the loss a secret from the outside world, *Michelle blabs about it to a sandwich man* while **ordering lunch over the phone.**
6. *They've been assured that the police doesn't have anything to do with the population count.* Still, **a lot of people are afraid of counteractions.**

RST – Corpora

Original corpus:

- English: RST Discourse Treebank – Newspaper text, 176K words

Related corpora include:

- Basque: RST Basque Treebank – Abstracts, 15.5K words
- Chinese/Spanish Treebank – Several written genres, parallel corpus, 100 texts
- Dutch RUG Corpus – Several written genres, approx. 6K words
- German: Potsdam Commentary Corpus – Newspaper text, 44K words
- Portuguese: BP RST Corpus – Abstracts

PDTB vs. RST

Certain differences between these frameworks make it hard to compare between them:

- Difference in granularity (RST distinguishes many more labels than PDTB)

PDTB vs. RST

Certain differences between these frameworks make it hard to compare between them:

- Difference in granularity (RST distinguishes many more labels than PDTB)
- Difference in label names obscures similarities (PDTB's JUSTIFICATION vs. RST's EVIDENCE)

PDTB vs. RST

Certain differences between these frameworks make it hard to compare between them:

- Difference in granularity (RST distinguishes many more labels than PDTB)
- Difference in label names obscures similarities (PDTB's JUSTIFICATION vs. RST's EVIDENCE)
- Similarities in label names obscures differences (PDTB's CONTRAST vs. RST's COMPARISON)
  1. PDTB CONTRAST: *Most bond prices fell...* **Junk bond prices moved higher,** however.
  2. RST COMPARISON: *Instead of proposing a complete elimination of farm subsidies,* as **the earlier U.S. proposal did,** ...

  → RST manual: in COMPARISON relations, arguments are not in contrast.

PDTB vs. RST

Certain differences between these frameworks make it hard to compare between them:

- Difference in granularity (RST distinguishes many more labels than PDTB)
- Difference in label names obscures similarities (PDTB's JUSTIFICATION vs. RST's EVIDENCE)
- Similarities in label names obscures differences (PDTB's CONTRAST vs. RST's COMPARISON)
  1. PDTB CONTRAST: *Most bond prices fell...* **Junk bond prices moved higher,** however.
  2. RST COMPARISON: *Instead of proposing a complete elimination of farm subsidies,* as **the earlier U.S. proposal did,** ...

  → RST manual: in COMPARISON relations, arguments are not in contrast.

Interoperability of these frameworks could actually benefit the community greatly...

1. Two large corpora and their frameworks
   - PDTB
   - RST

2. Use cases – What can we do with interoperable corpora?

Use cases – What can we do with interoperable corpora?

A few examples:

- Query for a specific relation in multiple corpora = more data

Task: query for chosen_alternative in German TED talks

Not many instances of this relation in the corpus. We want to find more examples.

Look at German RST-style corpus PCC: annotated as PREFERENCE in RST
Rather than go there by air, I'd take the slowest train.

Use cases – What can we do with interoperable corpora?

A few examples:

- Query for a specific relation in multiple corpora = more data
- Compare how discourse relations are marked in different modalities/genres (e.g., written vs. spoken corpus)

Task: query for *so* in written/spoken corpora

so is used to mark RESULT relations in PDTB (written). We want to find out which relations it marks in spoken discourse.

in Crible et al.'s unified taxonomy: possible labels include CONSEQUENCE, CONCLUSION, TOPIC-SHIFTING

I've already had a meeting uhm an update meeting so the place hasn't burnt down or anything.

Use cases – What can we do with interoperable corpora?

A few examples:

- Query for a specific relation in multiple corpora = more data
- Compare how discourse relations are marked in different modalities/genres (e.g., written vs. spoken corpus)
- Check how discourse relation is marked in another language

Task: How are causals marked in Dutch?

Find different markers that occur in PDTB's CAUSE relations.

Look at the Dutch CCR-style corpus DiscAn: POSITIVE, CAUSAL relations
She went home early because she promised her husband she would.
"Ze kwam vroeg thuis omdat ze haar man beloofd had dat ze dat zou doen."

She arrived home late because I was already asleep.
"Ze kwam laat thuis want ik sliep al."

Use cases – What can we do with interoperable corpora?

A few examples:

- Query for a specific relation in multiple corpora = more data
- Compare how discourse relations are marked in different modalities/genres (e.g., written vs. spoken corpus)
- Check how discourse relation is marked in another language
- On a larger scale, compare how discourse relations are marked or distributed in one language vs. another

Task: Looking at contrastive relations in English/French

How are contrastive and non-contrastive relations distributed in English/French?

in PDTB: look at COMPARISON class vs. other classes
in Annodis: look at CONTRAST and ALTERNATION labels vs. other labels

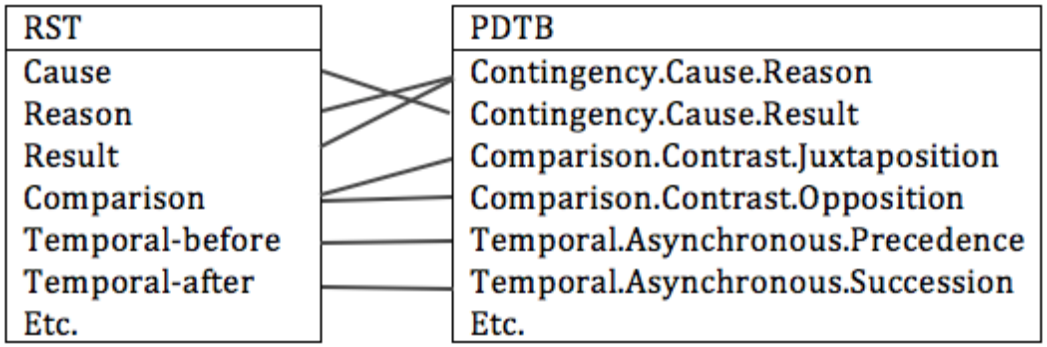
How can we make corpora interoperable?

- Given that there are so many differences between the frameworks, you have to know/study all the frameworks to identify the labels that are relevant for your work.
- Or is there an easier way to make these corpora interoperable?

- Different ways to create a mapping between frameworks:
  - One-to-one mapping
  - All-to-smallest common
  - All-to-decomposing features
- Let's look at these in more detail

Interoperable corpora: One-to-one mapping

- Construct one-to-one mappings for each combination of frameworks:
  - For every label in a framework, find the best matching corresponding label in another framework.

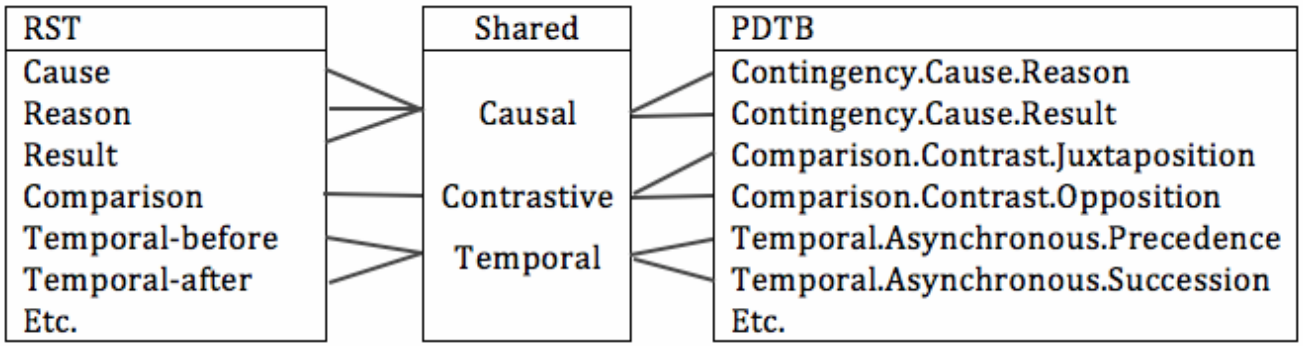


- Previous efforts:
  - Benamara & Taboada (2015): RST – SDRT
  - Chiarcos (2014): PDTB – RST

- Drawback: many mappings necessary to map to all frameworks, e.g.
  - 3 mapping for 3 frameworks (F1-F2, F2-F3, F1-F3)
  - 6 mappings for 4 frameworks (+ F1-F4, F2-F4, F3-F4), etc...

Interoperable corpora: All-to-smallest common mapping

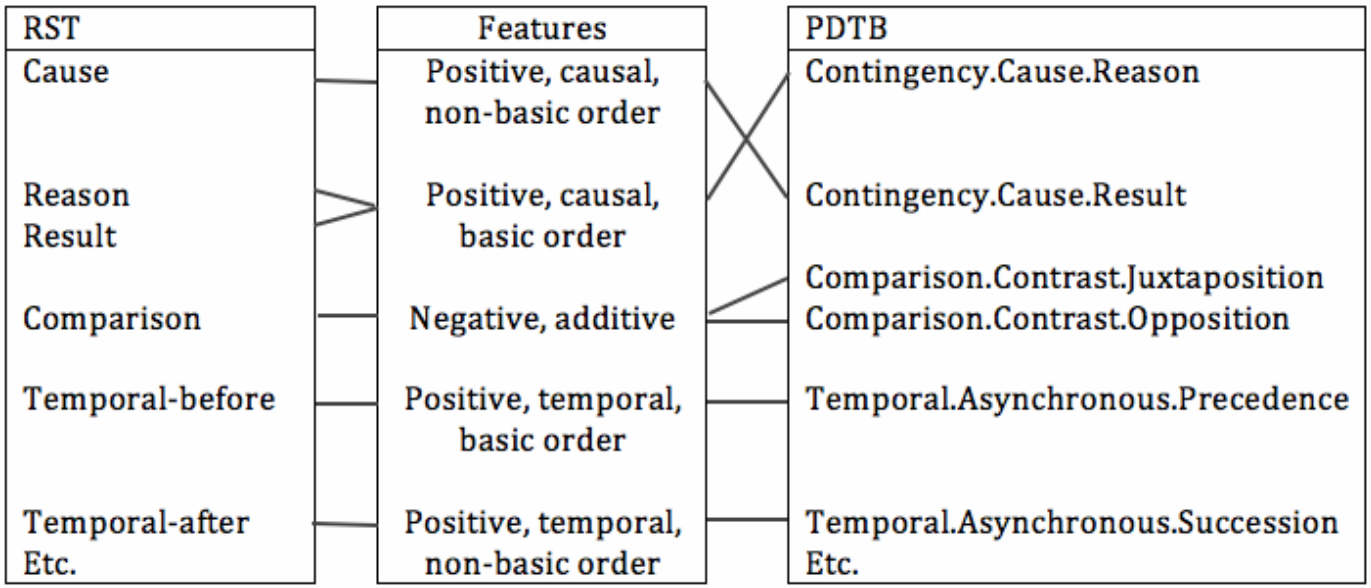
- Find set of common aspects between frameworks, map all relations to this set:



- Drawback: "smallest common" is probably very very small (2 distinctions: Y/N relation?)
- So we'd likely lose information

Interoperable corpora: All-to-decomposed features mapping

- Find common features of relation inventories, map all relations to their values for these features:



- Possible to easily add new frameworks by analysing the labels according to these features
- Labels can be underspecified for smaller inventories, so information will not be lost for bigger inventories.

Interoperable corpora: All-to-decomposed features mapping

- In favour of decomposed features, because it preserves the most amount of information
- In the next lecture, we will discuss how to go about these dimensions